EPFL

Thèse n° 9780

Distributional Regression and Autoregression via Optimal Transport

Présentée le 19 septembre 2023

Faculté des sciences de base
Chaire de statistique mathématique
Programme doctoral en mathématiques

pour l'obtention du grade de Docteur ès Sciences

par

Laya GHODRATI

Acceptée sur proposition du jury

Prof. M. Colombo, présidente du jury
Prof. V. Panaretos, directeur de thèse
Prof. E. Cazelles, rapporteuse
Prof. A. Petersen, rapporteur
Dr Y. Zemel, rapporteur

École
polytechnique
fédérale
de Lausanne

2023

Abstract

We present a framework for performing regression when both covariate and response are probability distributions on a compact and convex subset of $\mathbb{R}^d$. Our regression model is based on the theory of optimal transport and links the conditional Fréchet mean of the response to the covariate via an optimal transport map. We define a Fréchet-least-squares estimator of this regression map, and establish its consistency and rate of convergence to the true map under full observation of the regression pairs.

For the specific case when $d = 1$, we obtain additional results: we establish the minimax rate of estimation of such a regression function, by deriving a lower bound that matches the convergence rate attained by the Fréchet least squares estimator. Additionally, we find an upper-bound for the convergence rate of an estimator when observing only samples from the covariate and response distributions. Also in this case, the computation of the estimator is shown to reduce to a standard convex optimisation problem, and thus our regression model can be implemented with ease. We illustrate our methodology using real and simulated data.

We explore the problem of defining and fitting models of autoregressive time series of probability distributions on a compact interval of $\mathbb{R}$. In this context, an order-1 autoregressive model is a Markov chain that specifies a certain structure (regression) for the one-step conditional Fréchet mean with respect to a natural probability metric. We construct and investigate different models based on iterated random function systems of optimal transport maps. While the properties and interpretation of these models depend on how they relate to the iterated transport system, they can all be analyzed theoretically in a unified way. We present such a theoretical analysis, including convergence rates, and illustrate our methodology using real and simulated data. Our models generalise or extend certain existing models of transportation-based regression and autoregression, and in doing so also provides some new insights on those previous models.

Keywords: Distributional Regression, Distributional Time Series, Optimal Transport, Wasserstein Metric

Résumé

Nous présentons un cadre pour réaliser une régression lorsque la covariable et la réponse sont des distributions de probabilité sur un sous-ensemble compact et convexe de $\mathbb{R}^d$. Notre modèle de régression est fondé sur la théorie du transport optimal et lie la moyenne conditionnelle de Fréchet de la réponse à la covariable à travers une application de transport optimal. Nous définissons un estimateur des moindres carrés de Fréchet pour cette application, et nous établissons sa consistance et son taux de convergence vers la véritable application, sous l'observation complète des paires de régression.

Dans le cas particulier où $d = 1$, nous obtenons des résultats supplémentaires : nous établissons le taux d'estimation minimax d'une telle fonction de régression, en dérivant une borne inférieure qui correspond au taux de convergence atteint par l'estimateur des moindres carrés de Fréchet. De plus, nous trouvons une borne supérieure pour le taux de convergence d'un estimateur lorsque nous observons uniquement des échantillons issus des distributions de la covariable et de la réponse. Dans ce cas également, le calcul de l'estimateur se réduit à un problème d'optimisation convexe standard, permettant ainsi une mise en œuvre aisée de notre modèle de régression. Nous illustrons notre méthodologie à l'aide de données réelles et simulées.

Nous explorons le problème de la définition et de l'ajustement de modèles de séries temporelles autorégressives de distributions de probabilités sur un intervalle compact de $\mathbb{R}$. Dans ce contexte, un modèle autorégressif d'ordre 1 est une chaîne de Markov qui spécifie une certaine structure (régression) pour la moyenne conditionnelle de Fréchet à une étape, par rapport à une métrique de probabilité naturelle. Nous élaborons et étudions différents modèles basés sur des systèmes de fonctions aléatoires itérées d'applications de transport optimal. Bien que les propriétés et l'interprétation de ces modèles dépendent de la manière dont ils se rapportent au système de transport itéré, ils peuvent tous être analysés théoriquement de manière unifiée. Nous présentons une telle analyse théorique, y compris les taux de convergence, et illustrons notre méthodologie à l'aide de données réelles et simulées. Nos modèles généralisent ou étendent certains modèles existants de régression et d'autorégression basés sur le transport, et apportent par conséquent de nouvelles perspectives sur ces modèles antérieurs.

Mots-clés : Régression distributionnelle, Séries temporelles distributionnelles, Transport optimal, Métrique de Wasserstein

Acknowledgements

I want to express my gratitude to those whose scientific contributions directly influenced my thesis:

I am grateful to my supervisor, Victor Panaretos, for not only accepting me to work with him but also granting me the freedom to choose my topic and pace. His continuous encouragement, guidance and support have been invaluable throughout my thesis journey. I also extend my appreciation to the members of the jury for their time and consideration.

My thanks go to Yoav Zemel for his meticulous attention to detail and for his insightful feedback that refined my work. His patience and willingness to answer my questions has been greatly appreciated. I am grateful to Neda Mohammadi for her collaboration in my initial PhD year and for being a supportive friend throughout the highs and lows of my research. I would like to thank Kartik Waghmare for being available on numerous occasions to discuss my research progress. I'm really thankful to Arya Akhavan, a great friend whose kindness and hard work have been a source of motivation. The few weeks Arya was visiting Lausanne were probably the most productive for my research. Finally, I wish to thank Ivo Maceira for his unmatched coolness, which, although not entirely contagious, has taught me a thing or two. I also appreciate his help with coding and writing.

This thesis is dedicated to my parents, Monireh and Zia, and my sister, Mahdis.

Contents

List of Symbols

$\|.\|$	the Euclidean norm on $\mathbb{R}^d$
$\|.\|_p$	usual norm on the Lebesgue space $L^p(\Omega)$
$\|f\|_{L^p(\mu)}$	L^p norm of a function $f : [0,1] \to \mathbb{R}$ with respect to a measure μ
$\|f\|_{C^\beta}$	β-Hölder norm of a function $f : [0,1] \to \mathbb{R}$
$Leb(A)$	Lebesgue measure of a subset $A \subseteq \mathbb{R}^d$
δ_x	Dirac measure at a point x
$\mathcal{P}(\mathcal{X})$	set of Borel probability measures on a space $\mathcal{X}$
$\mathcal{W}_{2,\mathrm{ac}}(\mathcal{X})$	set of absolutely continuous Borel probability measures on $\mathcal{X}$
$\mathcal{W}_2(\mathcal{X})$	set of probability measures $\left\{\mu \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} \|x\|^2 \,\mathrm{d}\mu(x) < \infty\right\}$
$d_{\mathcal{W}}(\mu, \nu)$	Wasserstein distance
id	identity map of $\mathbb{R}^d$
$T\#\mu$	the push-forward measure defined as $[T\#\mu](A) = \mu(T^{-1}(A))$
$T_{\mu\to\nu}$	transport map that pushes forward μ to ν
$\mathrm{supp}(\mu)$	support of a measure μ
F_ν and F_ν^{-1}	cumulative distribution function and quantile function of a probability measure $\nu \in \mathcal{P}(\mathbb{R})$
$O_{\mathbb{P}}(1)$	a sequence that is bounded in probability.
$o_{\mathbb{P}}(1)$	a sequence that converges to zero in probability
$a_n \lesssim b_n$	$a_n \le cb_n$ where c is a constant independent of n
I	Identity matrix
∇f	The gradient of a function $f : \mathbb{R}^d \to \mathbb{R}$
$\nabla^2 f$	The Hessian matrix of a function $f : \mathbb{R}^d \to \mathbb{R}$

Chapter 1

Introduction

As datasets continue to grow in size and complexity, traditional statistical methods that focus on summarizing or aggregating data through scalar or vector values are becoming inadequate, as such summarization or aggregation processes inevitably lead to the loss of essential information. While some preprocessing of data is typically required, it is essential to preserve the inherent characteristics of the data's original structure and create methods specifically tailored to this structure.

Many complex data types, including images, histograms, and point clouds, can be represented as probability distributions. With the growing prevalence of these data types across various fields, it has become crucial to develop models capable of handling data where every single datum is depicted as a probability distribution. This may involve observing either the entire distribution or samples drawn from the distributions. The primary focus of this thesis is the development of models tailored for regression and time series analysis of probability distributions.

Analyzing distributions is a challenging task since they do not reside in a linear space and are infinite-dimensional. As a result, when distributions constitute the data itself, standard statistical methods developed either for linear or finite-dimensional data are not applicable directly. For instance, functional data analysis [34] has been developed for infinite-dimensional data, but it is primarily limited to data within linear spaces. Some of the previous methods, apply certain transformations to the data to map them into a space with linear structure [38, 22, 58, 39]. However, to tackle this non-linearity, our strategy is to examine the data in its natural metric space and integrate this geometry into our models.

Optimal transport theory addresses the question of what constitutes a canonical metric space for distributions by introducing the concept of Wasserstein space for distributions [55]. This theory provides a foundation for developing new methods that can effectively model and analyze complex data represented as probability distributions while preserving their inherent non-linear structure.

In previous methodologies, the emphasis was on the tangent structure within Wasserstein space [15, 79]. In contrast, the models presented in this thesis adopt a shape constraint approach and concentrate directly at the level of probability distributions. This method offers a clear and straightforward interpretation by relating response and covariate distributions through an optimal transport map and incorporating specific deformations as additional noise. The specific deformation noise leads to interpreting the model as specifying the conditional Fréchet mean of the response given the covariate. Additionally, the regression operator can be understood pointwise at the level of the original distributions. Particularly in the one-dimensional case, the model's effect in mass transportation is equivalent to quantile re-arrangement. By providing a more direct and interpretable way of modeling the relationship between probability distributions, the shape constraint method serves as a valuable alternative to previous methods that rely on the tangent structure of Wasserstein space.

Moving forward, the study of distributional autoregression models is a logical next step. This addresses scenarios where there's a dependency between probability distributions. By regarding the sequence of probability distributions $\{\mu_n\}_{n=1}^N$ as a Markov chain in Wasserstein space, autoregressive relationships can be modeled through the establishment of a connection between the conditional Fréchet mean at time $n + 1$ and the chain's state at time n. We generalize previous methods of transportation-based autoregressive models [80] to obtain functional dynamics and interpretable classes of distributional autoregressions. This is achieved by extending the functional structure of the regression operator we've developed and adopting previous methods that use a contractive parameter to ensure the stationarity of the dynamics.

1.1 Structure of the Thesis

Chapter 2 This chapter provides a brief overview of the main concepts and definitions of optimal transport theory that is used in this thesis, it continues with a discussion on some of the potential applications of distribution-on-distribution regression and ends by reviewing some of the previous methods.

Chapter 3 This chapter is based on the published articles [24] and [26]. We introduce a model for performing regression when both covariate and response are probability distributions on a closed interval of $\mathbb{R}$. After specifying the model and establishing the identifiability of the regression map, we define a Fréchet-least-squares estimator of this regression map and establish its consistency and rate of convergence to the true map under full observation of the regression pairs. Additionally, we show that our estimator is minimax optimal when the distributions are fully observed. We illustrate the performance of the model and the estimator through a simulation and by real data analysis for mortality and quantum dot data.

Chapter 4 This chapter is based on the preprint [27]. It considers the generalization of the distribution-on-distribution regression model of Chapter 3 to higher dimensions, $d > 1$. Some more restrictive structural assumption needs to be imposed on the component of the regression model than when $d = 1$. These assumptions lead to the identifiability of the model and also enable us to use empirical process theory to derive the rate of convergence of our estimator.

Chapter 5 This chapter is based on the preprint [25]. It develops transportation-based autoregressive models, examining three distinct notions of autoregression that capture various distributional time series characteristics. The models are compared to existing approaches, with stationarity conditions established and identifiability, consistency, and convergence rates demonstrated. Finally, the chapter highlights the finite sample performance of the proposed methodology using both simulated and real data.

Chapter 6 The final chapter outlines some potential directions for research in each of the last 3 chapters. Additionally, it introduces an alternative model for distribution-on-distribution regression and provides some preliminary results.

Chapter 2

Overview

This chapter provides an introduction to optimal transport theory and the problem of distribution-on-distribution regression, along with a discussion of possible applications and an overview of previous methods.

2.1 Overview of Optimal Transport

In this section, we give a brief overview of optimal transport theory including some of the main results and the notation which is used throughout this thesis. We will go through some definitions and intuitions without going into too much depth or formalities. For more background see, e.g. Villani [73], Villani et al. [74], Santambrogio [64] and Ambrosio et al. [3].

The Monge Problem In 1781, Monge introduced the optimal transport problem, which addresses a practical question in civil engineering: how to effectively move a pile of sand to fill a hole of the same total volume. To tackle this problem mathematically, Monge represented the sand pile and hole as probability measures, μ, and ν, both belonging to the set of Borel probability measures $\mathcal{P}(\mathcal{X})$ on a space $\mathcal{X}$ which in this original example we take as $\mathbb{R}^2$. The successful transportation of sand requires finding a map $T : \mathcal{X} \to \mathcal{X}$ that ensures the amount of sand arriving at each target location $B \subseteq \mathcal{X}$ (measured by $\nu(B)$) equals the amount of sand sent to it, i.e., $\nu(B) = \mu(T^{-1}(B))$. This map T is referred to as a transport map, pushing forward μ to ν, which is denoted as $\nu = T\#\mu$. The objective is to find the transport map T that minimizes total displacements:

$$\min_{T\#\mu=\nu} \int \|x - T(x)\| \,\mathrm{d}\mu(x).$$

The Monge problem can be generalized in different ways, the most natural one being the extension of the ground space $\mathcal{X}$ to $\mathbb{R}^d$ for any $d \geq 1$. Furthermore, rather than minimizing the total displacements, the objective function can be any general cost function of the form $\int c(x, T(x)) \,\mathrm{d}\mu(x)$, where $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In this thesis, we focus on the case where the ground space is $\mathbb{R}^d$ and where the cost function is given by $c(x, y) = \|x - y\|^2$.

The Monge formulation of optimal transport is a difficult non-linear optimization problem, and even more troublesome is the fact that this formulation can be ill-posed since the set of optimal transport maps pushing forward μ to ν might be empty: For instance, in the case where μ is a Dirac at a point x_0 and ν is any probability measure but Dirac, there is no map T such that $\nu = T\#\mu$ since $T\#\delta_{x_0} = \delta_{T(x_0)}$ for any T, so no T can exist.

The Kantorovich Problem Kantorovich's reformulation of the Monge problem in 1942 removed the problems that the latter suffered and guaranteed the existence of a solution. Kantorovich relaxed the original formulation by allowing mass at a given point x to be split and sent to an infinite number of locations, rather than being sent to a single location $T(x)$. Mathematically speaking, this split of mass is represented by a transport plan γ, which is a distribution over $\mathbb{R}^d \times \mathbb{R}^d$ with marginals ν and μ. A transport plan γ tells us that the amount of mass that should be sent from $A \subset \mathbb{R}^d$ to $B \subset \mathbb{R}^d$ is $\gamma(A \times B)$. The problem formulated by Kantorovich is as follows: Given two probability measures μ and ν on $\mathbb{R}^d$, the task is to find an optimal transport plan that minimizes a ground-cost function (assuming the quadratic case for now). That is, we want to find:

$$d^2_{\mathcal{W}}(\nu, \mu) := \inf_{\gamma \in \Gamma(\nu,\mu)} \int_{\Omega} \|x - y\|^2 \, \mathrm{d}\gamma(x, y), \tag{2.1}$$

where $\Gamma(\nu, \mu)$ is the set of transport plans of μ and ν. In contrast to the Monge problem, the Kantorovich problem always has a minimizer (see Theorem 4.1 Villani et al. [74] for a more general result). Moreover, it takes the form of an infinite dimensional linear programming.

Wasserstein Distance The optimal transport cost between two distributions as defined by (2.1), defines a distance between the two distributions (see chapter 6 of Villani et al. [74]), which is called the Wasserstein distance. Wasserstein space of distributions over $\mathbb{R}^d$, denoted by $\mathcal{W}_2(\mathbb{R}^d)$, is a set of probability measures on $\mathbb{R}^d$ that have finite second moments and is equipped with the $d^2_{\mathcal{W}}$ distance[1]. This distance makes $\mathcal{W}_2(\mathbb{R}^d)$ a complete separable metric space.

Wasserstein distance has become a canonical choice for comparing distributions due to its several desirable properties. For example, it incorporates the geometry of the underlying space. One way to see this, is that $d^2_{\mathcal{W}}(\delta_x, \delta_y) = \|x - y\|^2$ and therefore the mapping $x \longrightarrow \delta_x$ is an isometric embedding of $\mathbb{R}^d$ in $\mathcal{W}_2(\mathbb{R}^d)$.

Moreover, the Wasserstein distance can be used to compare discrete (e.g., empirical) and continuous distributions.

[1] The finiteness of second moments guarantees that $d^2_{\mathcal{W}}(\mu, \nu) < \infty$ for all $\mu, \nu \in \mathcal{W}_2(\mathbb{R}^d)$

Brenier's Theorem One of the most fundamental results in optimal transport is Brenier's Theorem [9] which gives a condition under which the Kantorovich and Monge problems are equivalent. Specifically, for two measures μ and ν in $\mathcal{W}_2(\mathbb{R}^d)$, if μ is absolutely continuous with respect to the Lebesgue measure, there exists a unique and equal solution to both problems. This solution can be characterized as a deterministic transport map $T : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, which is μ-almost surely the gradient of a convex function $\varphi : \mathbb{R}^d \longrightarrow \mathbb{R}$. In other words, the unique optimal way to transport one distribution to another is by pushing the source distribution forward with the gradient of a convex function.

An optimal transport map between μ and ν is sometimes denoted as $T_{\mu\to\nu}$. When the map T can be represented as the gradient of a convex function (i.e., $T = \nabla\varphi$), the convex function φ is known as the Kantorovich potential of μ and ν.

Univariate Case In the special case when $\mathcal{X} = \mathbb{R}$, an optimal transport map has an explicit characterization. Let $F_\mu(x) = \mu((-\infty, x))$ be the distribution function of a measure μ. The quantile function of μ is defined as:

$$F_\mu^{-1}(t) = \sup\{t \in \mathbb{R} : F_\mu(x) \leq t\}, \quad t \in [0,1].$$

Let $\mu, \nu \in \mathcal{W}_2(\mathbb{R})$, then:

$$d_{\mathcal{W}}^2(\mu,\nu) = \int_0^1 \left|F_\mu^{-1}(p) - F_\nu^{-1}(p)\right|^2 \mathrm{d}p. \tag{2.2}$$

If the source measure μ is absolutely continuous (i.e. F_μ is continuous), then the optimal map T between μ and ν is a non-decreasing map with the explicit expression

$$T = F_\nu^{-1} \circ F_\mu. \tag{2.3}$$

Monotonicity of Optimal Maps One of the implications of Brenier's theorem is that optimal maps are monotone operators, representing a generalization of the fact that optimal maps are non-decreasing functions in the univariate case. To see this, note that the gradient of a convex differentiable function φ is a monotone operator:

$$\forall (x, x') \in \mathbb{R}^d \times \mathbb{R}^d, \quad \langle \nabla\varphi(x) - \nabla\varphi(x'), x - x' \rangle \geq 0.$$

However, unlike the one-dimensional case, not all monotone operators in higher dimensions are gradients of convex functions. For example, a rotation is a monotone operator that cannot be an optimal map[2].

[2] Let A be a matrix, then the operator Ax defined by A is related to a quadratic form $\varphi(x) = \langle Bx, x\rangle/2$, resulting in $A = \nabla\varphi = (B + B^\top)/2$. This means A should be symmetric.

Couplings It is sometimes useful to interpret the optimal transport problem in probabilistic terms as the search for an optimal coupling between two distributions. A coupling of two distributions μ and ν is a pair of random variables X and Y with distributions μ and ν respectively. The Wasserstein distance can then be expressed as:

$$d^2_{\mathcal{W}}(\mu, \nu) = \inf \mathbb{E} \|X - Y\|^2 ,$$

where the pair (X, Y) spans all possible couplings of μ and ν.

A coupling (X, Y) is called deterministic if there exists a measurable map T such that $Y = T(X)$. For instance, if $T\#\mu = \nu$ and X is a random variable with distribution μ, and $Y = T(X)$, then Y is distributed according to ν and therefore (X, Y) is a deterministic coupling and vice versa. Consequently, we can interpret the Monge problem as finding an optimal deterministic coupling.

It is important to note that, given the definitions of the Wasserstein distance, any coupling (X, Y) of μ and ν with joint distribution γ provides an upper bound for the Wasserstein distance

$$d^2_{\mathcal{W}}(\mu, \nu) \leq \mathbb{E}_{(X,Y)\sim\gamma} \|X - Y\|^2 .$$

In this thesis, we sometimes use the fact that for $\mu, \nu, b \in \mathcal{W}_2(\mathbb{R}^d)$, and for maps T_1 and T_2 such that $T_1\#b = \mu$ and $T_2\#b = \nu$, the following inequality is valid:

$$d_{\mathcal{W}}(\mu, \nu) \leq \|T_1 - T_2\|_{L^2(b)} . \tag{2.4}$$

To see this, suppose Z is a random variable with distribution b, and $X = T_1(Z)$ and $Y = T_2(Z)$, then (X, Y) is a coupling for (μ, ν), therefore the inequality holds. The inequality becomes an equality if and only if, the maps T_1 and T_2 are optimal, i.e. $T_1 = T_{b\to\mu}$ and $T_2 = T_{b\to\nu}$, and the measures μ, ν, b are compatible, meaning that $T_{\mu\to\nu} \circ T_{b\to\mu} = T_{b\to\nu}$ (see Section 2.3.2 of Panaretos and Zemel [55]).

In the special case where $d = 1$, the equality $T_{\mu\to\nu} \circ T_{b\to\mu} = T_{b\to\nu}$ always holds. This is because the optimal maps are characterized by nondecreasing functions, and the composition of two nondecreasing functions is also nondecreasing (and thus optimal). Therefore, we have:

$$d_{\mathcal{W}}(\mu, \nu) = \left\|T_{b\to\mu} - T_{b\to\nu}\right\|_{L^2(b)} . \tag{2.5}$$

2.1.1 Statistical Inference in Wasserstein Space

Optimal transport and Wasserstein distance play significant roles in statistics and machine learning. For an in-depth review of Wasserstein distance applications in statistical theory and methodology, consult Panaretos and Zemel [54]. A key question for statisticians is how to accurately estimate optimal transport-related objects, like

Wasserstein distance, optimal map, or barycenter, using available data.

In this section we briefly overview three of the problems in statistical optimal transport that will be important to this thesis, in the following contexts:

(I) Since we typically observe distributions through samples, it is crucial to have efficient methods for estimating the underlying densities.

(II) The estimation of our regression operator is intricately linked to the estimation of optimal transport maps. Under similar regularity assumptions that aid in estimating optimal maps, we can estimate our regression parameters.

(III) We will utilize the Fréchet functional to construct a sum-of-squares functional within the context of distribution-on-distribution regression. Additionally, we will use assumptions from previous studies on the Fréchet mean to determine the rate of convergence.

(I) Density Estimation in Wasserstein Distance An essential problem involves estimating the density of a distribution based on independent samples from it while measuring error using the Wasserstein distance. Given n i.i.d. samples from a distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$, the statistical literature often employs the plug-in approach, focusing on the empirical distribution μ_n. However, it is known that

$$d^2_{\mathcal{W}}(\mu, \mu_n) \sim n^{-2/d},$$

(see Niles-Weed and Berthet [50]). This indicates that achieving a specific precision requires a sample size exponential in dimension.

One solution is to leverage smoothness assumptions to address the curse of dimensionality in this setting [50]. In particular, Niles-Weed and Berthet [50] achieves a faster rate of convergence for a wavelet density estimator over measures with densities in Besov classes and demonstrates its minimax optimality.

(II) Estimation of Optimal Transport Map Another crucial question in statistical optimal transport is determining how to estimate the optimal transport map, T, between two unknown distributions μ and ν, based on independent and identically distributed samples drawn from each distribution.

Several studies [21, 31, 35, 46, 48] have proposed estimators for the map T by imposing smoothness assumptions on the distributions μ and ν or the underlying optimal map T. In particular, Hütter and Rigollet [35] established a minimax lower bound for the problem when the map T is in an α-Hölder ball and showed that their estimator achieves minimax optimality up to a polylogarithmic factor. Manole et al. [46] proposed an estimator that achieves the optimal convergence rate when the sampling domain is a d-dimensional torus. However, research into computationally efficient statistical estimators for optimal maps is less developed in the literature [48].

The proposed estimator by Hütter and Rigollet [35] cannot be feasibly computed because it requires projection onto the space of smooth and strongly convex functions, which is NP-hard [48]. The estimator by Manole et al. [46] also requires computation time growing exponentially in d. In contrast, Muzellec et al. [48] explored an alternative estimator that can be computed in polynomial time if the underlying map is sufficiently smooth, but their estimator does not achieve the minimax optimal rate.

(III) Wasserstein Fréchet Mean Estimating the mean of a random object is a fundamental task in statistics. However, since the Wasserstein space doesn't have a linear structure, the usual definition of mean in linear spaces cannot be extended to this space. Agueh and Carlier [1] introduced the notion of Fréchet mean (or barycenter) in the Wasserstein space, which relies on the metric structure of the space: For probability measures $\mu_1, \cdots, \mu_n \in \mathcal{W}_2(\mathbb{R}^d)$ with weights $\lambda_1, \cdots, \lambda_n$, a Fréchet mean is a minimizer of the functional

$$b \longrightarrow \sum_{i=1}^{n} \lambda_i d_{\mathcal{W}}^2(b, \mu_i).$$

The existence of such a minimizer is guaranteed, and it is unique if at least one μ_i is absolutely continuous with respect to the Lebesgue measure.

The notion of Fréchet mean can also be defined at the population level. Let Λ be a random measure in $\mathcal{W}_2(\mathbb{R}^d)$, and denote the distribution of Λ as P. A Fréchet mean of Λ is a minimizer of the Fréchet functional

$$F(b) = \frac{1}{2}\mathbb{E} d_{\mathcal{W}}^2(b, \Lambda) = \frac{1}{2}\int_{\mathcal{W}_2(\Omega)} d_{\mathcal{W}}^2(b, \lambda)\, \mathrm{d}P(\lambda) \quad b \in \mathcal{W}_2(\mathbb{R}^d). \tag{2.6}$$

The Fréchet functional always admits a minimizer and if Λ has a finite Fréchet functional and is absolutely continuous with positive probability, the Fréchet mean is unique.

Le Gouic and Loubes [41] proved that if a random measure has a finite Fréchet functional and a unique Fréchet mean, then the empirical Fréchet mean will almost surely converge to this unique Fréchet mean. For one-dimensional cases, the empirical Fréchet mean converges to the population counterpart at a parametric rate[3], as shown by Panaretos and Zemel [53] and Bigot et al. [8]. Ahidar-Coutrix et al. [2] and Le Gouic et al. [42] investigated the convergence rate of the empirical Fréchet mean in certain general geodesic spaces, particularly in the Wasserstein space. A key component of their study was establishing quadratic growth of the Fréchet functional

[3] a rate of the form c/n based on n independent samples

around its minimizer, which on the Wasserstein space takes the form

$$F(b) - F(b^*) \geq Cd^2_{\mathcal{W}}(b, b^*),$$

where b^* is the minimizer of F and C is a constant. Ahidar-Coutrix et al. [2] illustrated that specific regularity conditions on the Kantorovich potentials corresponding to optimal maps between the Fréchet mean and the distributions in the support of P guarantee the quadratic growth. Using these conditions, Ahidar-Coutrix et al. [2] derived an upper bound for the convergence rate of the Fréchet mean. Subsequently, Le Gouic et al. [42] demonstrated that if Kantorovich potentials are α-strongly convex and β-smooth [4], with $\beta - \alpha < 1$, a parametric convergence rate for the Fréchet mean can be obtained.

Zemel and Panaretos [78] proposed a gradient descent algorithm to compute the Fréchet mean, proving its convergence under certain assumptions. Chewi et al. [18], established the rate of convergence of the gradient descent algorithm when P is supported on Gaussian probability measures.

2.2 Distributional Data Analysis

In this thesis, our main focus is on developing regression methods to estimate the relationships between two distributions. Additionally, we also investigate different autoregressive models for distributional time series.

Throughout the upcoming chapters, we will provide precise definitions of these problems. As for the rest of this chapter, we will discuss scenarios where the objective is to infer relationships between variables in the form of distributions, to motivate the problem at hand.

2.2.1 Possible Applications of Distribution-on-Distribution Regression

In the following section, we provide several examples, although not an exhaustive list, of scenarios in which a distribution-on-distribution regression framework could be a valuable tool for modeling data.

Setting-dependent observations Multi-dimensional measurements can provide valuable information about both the distributions of individual variables and the relationships between them. However, in certain situations, obtaining multi-dimensional measurements isn't feasible, and only the marginal distributions of each variable can be observed.

[4] A twice differentiable convex function is α-strongly convex if its Hessian matrix's eigenvalues are always at least α, and β-smooth if these eigenvalues are always no more than β, for all points $x \in \mathbb{R}^d$.

Additionally, we can imagine a situation where the marginal distributions themselves are not fixed and depend on a particular setting in data collection, but the relationship between those marginals is somewhat fixed.

More formally, suppose there are k distinct settings, and within each setting, we have an input variable X^k and output variable Y^k and we observe independent samples from them. Here, the distributions of X^k and Y^k could be influenced by the specific setting k and connected through a regression operator at the distribution level.

• **Light-emitting quantum dots:** One example is when the relationship between input and output measurements of a physical system is not observable. For instance, measurements of physical quantities may be aggregates of its many constituents, and the individual-to-individual correspondence is lost: In particular, shining light on a single "quantum dot" of nanometric size and measuring its size-dependent spectral response might be unfeasible, while it is feasible to do so on a collection of such dots, therefore only the marginal distributions of sizes and emitted radiation is observed [63].

Inferring the relationship between two parameters from their marginal densities without extra assumptions is ill-posed if the density function of each parameter is fixed. However, it is increasingly the case where the distributions themselves are not fixed and they are dependent on the experimental setting. For example, different experimental settings of quantum dot production result in different size distributions (see Section 3.5). Correct identification of the probabilistic relationship between the parameters might be possible by combining information from different settings.

• **Censored Data:** In many cases, it is crucial to establish the connection between two independent datasets, which is made harder when the relationship between them is concealed due to privacy concerns. To address such situations, various methods have been developed. One example is unmatched regression [12, 62, 66] which has been specifically designed for cases where we have access to independent samples of input and output variables, each with fixed distributions. However, this approach may not be suitable when considering multiple settings with varying underlying distributions, necessitating the development of alternative methods to handle such scenarios.

For example, political scientists often aim to estimate the impact of demographic factors on voting behavior, even though census data and vote counts are collected separately. Consider a scenario where voting results and population demographics are available across different regions of a country. The objective would be to explore the relationship between these variables, acknowledging that the distribution of various demographic factors varies from region to region.

Another example involves insurance companies seeking to predict the average number of doctor visits for specific age groups within a population. Berzel et al. [7] considered the problem of estimating the conditional distribution of the number of visits to the doctor in a determined population, based on some demographic factors, including age. In this study, the data consisted of the age and gender of a group of people as well as their number of doctor visits during some years. However, we can consider the situation in which individual clinics may possess data on patient ages and visit counts but cannot share this information due to privacy concerns. Instead, they may publish distributions of the number of visits and age distributions of visitors. It can be assumed that clinics in different locations will have distinct age distributions for their patients. By combining information from multiple clinics, the goal would be to establish a regression relationship between the distribution of the number of doctor visits and age distribution.

In such situations, a distribution-on-distribution regression method may offer a better data model and yield more accurate results.

Atmospheric Flow We might be interested in understanding how clouds move and estimating physical variables such as wind and pressure using pictures of cloud formations at different times. We can represent clouds as continuous distributions of cloud particles and analyze their shapes and intensities between consecutive images by applying optimal transport theory. The works of Cullen [20], and Alessio Figalli provide valuable insights into how particles move optimally in this scenario, and by solving the optimal transport problem, we can gain insights into these physical variables.

However, given a single pair of before and after pictures of clouds, we cannot infer anything about areas that were not covered by clouds at some point. To overcome this limitation, we can use multiple pairs of cloud images which show particles in various locations, and assuming that the weather conditions were stationary, we can use distribution-on-distribution regression to gain insights beyond what we could learn from a single pair of images. By combining information from multiple pairs of images, we can leverage the power of optimal transport theory to obtain a robust estimate of wind direction and enhance our knowledge of atmospheric dynamics. This might allow us to make more accurate predictions about physical variables such as wind and pressure, ultimately leading to a more comprehensive understanding of wind patterns and atmospheric conditions. See Figure 4.1 for a related toy model demonstration.

Income inequality and life expectancy Studies investigating the effects of income inequality on health have produced varied results across different countries [44]. While there is substantial evidence linking higher individual income to reduced mortality rates [44], the debate continues on how rising inequality impacts mortality distribution beyond the individual-level relationship [70].

It might be worthwhile to explore the possibility that average life expectancy depends, not only on an individual's income level but also on the overall income distribution in their population. Distribution-on-distribution regression methods could potentially offer a more accurate explanation of this relationship.

However, the specific approach developed in this thesis might not be the most suitable for this problem. For example, a population with a bimodal income distribution could result in a skewed age-at-death distribution towards younger ages compared to a society with a unimodal income distribution and similar average or median income (see Figure 2.1). Our model, which assumes that the relationship between the covariate and response distributions is approximately explained via an optimal transport map, might not capture this scenario effectively.

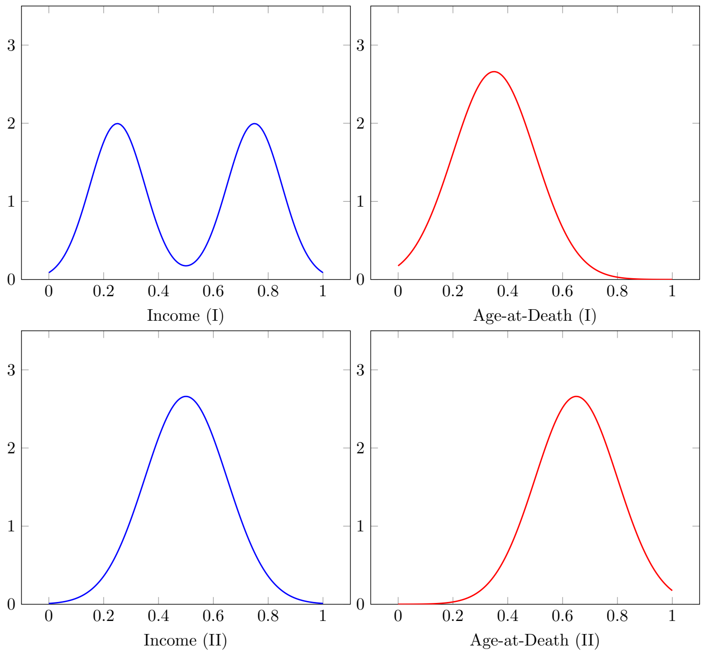


Figure 2.1: Two pairs of distributions (I and II) illustrating a possible shortcoming of optimal transport regression, as the optimal maps associated to each pair are significantly different in shape. Units are normalized.

2.2.2 Previous Methods

Various statistical methods for distributional data have been developed (see Petersen et al. [59] for an excellent review). Specially, various studies develop regression methods where the predictor is a distribution and the response is a scalar, for example, see Chen et al. [15], Póczos et al. [60], Tang et al. [69], Oliva et al. [52], Szabó et al. [68], Bachoc et al. [6].

Here we provide a summary of past methodologies for regression when both the predictor and response variables are distributions (hence distribution-on-distribution regression).

FDA approach Functional data analysis (FDA) is concerned with the analysis of data that can be represented as objects residing in function spaces of infinite dimensions. Probability density functions can also be treated as functional data, subject to the constraints that they are nonnegative and have an integral sum of one. Because of these constraints, directly applying FDA methods to probability distributions may lead to issues. One solution is to first map probability densities to a Hilbert space of functions, and then apply FDA methods, such as function-to-function regression, to the transformed distribution.

It is essential for the mapping between distributions and Hilbert space to be invertible. This is because we often want to interpret the response of the regression in the original space of distributions. An invertible method called the log quantile density (LQD) transformation was introduced by Petersen et al. [58]. They utilized functional regression models with LQD-transformed functions as predictors and either LQD-transformed functions or scalars as responses.

However, the LQD transformation has its limitations. It does not consider the geometry of the probability distribution space, and the resulting transformation map is not isometric. This leads to deformations that alter the distances between pairs of objects and can create problems with interpretability. Moreover, the model is only applicable when the covariate and response distributions are supported on a closed interval of $\mathbb{R}$.

Kernel Smoothing Oliva et al. [51] applied a Nadaraya-Watson type kernel regression for distribution-on-distribution regression. Specifically, they considered a situation with a set of distribution pairs (P_i, Q_i), where the P_i are supported on a cube in $\mathbb{R}^k$, the Q_i are supported on a cube in $\mathbb{R}^l$, and $Q_i = f(P_i)$, with f being a transformation. The distributions P_i, Q_i are only observed through random finite samples drawn from each (which is the only source of uncertainty in the observations).

For a new distribution P_0, the response is estimated as a locally weighted average, using kernels as weights, and is given by:

$$\hat{f}(P_0) = \sum Q_i W(P_i, P_0), \quad \text{where} \quad W(P_i, P_0) = \frac{K(\frac{d(P_i, P_0)}{h})}{\sum K(\frac{d(P_j, P_0)}{h})}.$$

Here, d represents a distance measure between distributions, K is a nonnegative real-valued kernel function, and h is the bandwidth for kernel regression.

The advantage of this regression method is that the predictor and response distributions are not restricted to $\mathbb{R}$, and the method is non-parametric. A drawback of this method is its unsuitability for extrapolation, as the response distribution for any new predictor distribution is a convex combination of response distributions in the training set. For example, this limitation comes into play when the new input is concentrated spatially far from the previous samples (see Appendix 2 of Chen et al. [16]). Additionally, Nadaraya-Watson type estimators rely on tuning parameters and generally suffer from the curse of dimensionality.

Wasserstein Regression The ideas of modeling distributions within the Wasserstein space for regression and autoregression were studied by Chen et al. [15] and Kokoszka et al. [39]. They used the structure of the tangent space to create a regression operation. This was achieved by using the log transform, which moved the regressor and response to appropriate tangent spaces. As a result, a (linear) regression model is established in a familiar framework akin to Hilbert space. This approach allowed the authors to take advantage of well-known techniques of functional regression and build a suitable asymptotic theory.

Although this method has the potential to be used when distributions are supported on compact subsets of $\mathbb{R}^d$ for $d \geq 1$, only the case where $d = 1$ was explored. In section 3.2.3, we delve into this method more deeply and draw a comparison with the regression model we propose.

Chapter 3

Distribution-on-Distribution Regression in One Dimension

The work in this chapter was done in collaboration with my supervisor Victor Panaretos and it is published in two articles [24, 26]. This chapter integrates the content of these published works, adhering closely to the original text with only slight modifications to avoid redundancies. Moreover, an appendix focused on the analysis of quantum dot data is included.

3.1 Introduction

Functional data analysis [34] considers statistical inference problems whose sample and parameter spaces constitute function spaces. This framework encompasses data that are best viewed as realisations of random processes, and presents challenges arising from the infinite dimensionality of the function spaces, typically taken to be separable Hilbert spaces. On the other hand, non-Euclidean statistics [56] treats inference problems whose sample and parameter spaces are finite dimensional manifolds. Such problems present with a different set of challenges, linked with the non-linearity of the corresponding spaces, which often arises due to non-linear constraints satisfied by the data/parameters.

When the data/parameters of interest are, in fact, probability distributions, one has a problem that is simultaneously functional and non-Euclidean: on the one hand the data can be seen as random processes, and on the other they satisfy non-linear constraints, such as positivity and integral constraints. Thus, the functional data analysis of probability distributions features interesting challenges stemming from this dual nature of the ambient space, for example the finite measurement of intrinsically infinite dimensional objects, and the lack of a linear structure which is crucial to basic statistical operations, such as averaging or, more generally, regression toward a mean. See Petersen et al. [59] for an excellent overview.

One approach to dealing with the non-linear nature of probability distributions is to apply a suitable transformation and map the problem back to a space with a linear structure [38, 22, 58, 39]. A seemingly more natural approach is to embrace the intrinsic non-linearity, and to analyse the data in their native space, equipped with a canonical metric structure. In the case of probability distributions, the Wasserstein metric [54, 55] has been exhibited as a canonical choice [53], primarily because it captures deformations, which are typically the main form of variation for probability distributions.

The case of inferring the Fréchet mean of a collection of random elements in the Wasserstein space is by now well understood [53, 8, 78, 42]. The deep links to convexity and the tangent space structure of the Wasserstein space play an important role in motivating and deriving the analysis of this case. The next step is to understand the notion of regression of one probability distribution on another. The first to do so were Chen et al. [15], and, independently, Zhang et al. [79], the latter paper focussing on autoregression. They used the tangent space structure to define a regression operation: using the log transform, the regressor and response are lifted to suitable tangent spaces, where a (linear) regression model is defined in a more familiar Hilbertian setting [47, 32]. This allows the authors to use the well-developed toolbox of functional regression, and derive appropriate asymptotic theory.

In this chapter, we propose an alternative notion of distribution-on-distribution regression, following a different path. Rather than taking a geometrical approach, via the tangent bundle structure, we follow a shape-constraint approach, namely exploiting convexity. Our model is defined directly at the level of the probability distributions, and stipulates that the response distributions are related to the covariate distributions by means of an optimal transport map, and further deformational noise. A key advantage of this approach is its clean and transparent interpretation, since the regression operator can be interpreted *pointwise* at the level of the original distributions, and its effect consists in mass transportation, or equivalently, quantile re-arrangement. Further to this, the approach requires minimal regularity conditions, and does not suffer from ill-posedness issues as inverse problems do. We show that our estimator is minimax optimal and that computational implementation of the estimator reduces to a standard convex optimisation problem. The usefulness of the approach is exhibited when revisiting the analysis of the mortality data of Chen et al. [15], where the approach is seen to lead to similar (if more expansive) qualitative conclusions, but with the advantage of an arguably improved interpretability. Additionally, we apply our method to the analysis of quantum dot data.

3.2 Distribution-on-Distribution Regression

3.2.1 Fréchet Functionals and Regression Operators

Let $\Omega, \Omega' \subseteq \mathbb{R}$ and (μ, ν) be a pair of random elements in $\mathcal{W}_2(\Omega) \times \mathcal{W}_2(\Omega')$ with joint distribution P. Then, similar to a standard nonparametric regression model, we can define a regression operator $\Gamma : \mathcal{W}_2(\Omega) \longrightarrow \mathcal{W}_2(\Omega')$ as the minimizer of the conditional Fréchet functional, viewed as a function of μ,

$$\operatorname*{argmin}_{b} \int_{\mathcal{W}_2(\Omega)} d^2_{\mathcal{W}}(b, \nu)\, \mathrm{d}P(\nu \mid \mu) = \Gamma(\mu)$$

assuming that for any μ, the Fréchet mean of the conditional law $P(\cdot \mid \mu)$ of ν given μ is unique , which can be enforced by means of regularity assumptions on the pair (μ, ν).

The difference between the above formulation and the standard regression formulation is that we have replaced the notion of expectation with a Wasserstein-Fréchet mean, an approach termed as “Fréchet Regression” by Petersen and Müller [57]. Postulating a specific form on the regression operator Γ amounts to defining a certain type of regression model. If Γ is left unconstrained, except for possessing some degree of regularity, then we would speak of a nonparametric regression model. However, assumptions on Γ are needed to ensure its identifiability, and simply assuming it is regular will not suffice in this more general context.

For instance, the approach of Chen et al. [15] and Zhang et al. [79] consists in constraining Γ to be in a certain sense linear, in that it can be represented as a linear operator at the level of the tangent bundle. Identifiability, and indeed fitting and asymptotic theory, can then be derived by appealing to the inclusion of the tangent spaces in Hilbert spaces.

Here we impose a different constraint on Γ, and consequently define a different notion of regression. Namely we impose a *shape constraint*, by assuming that $\Gamma(\mu) = T\#\mu$, where T is an increasing map. This is developed in the next section which postulates a regression model on the pair (μ, ν) that guarantees the uniqueness of the conditional Fréchet mean $\Gamma(\mu)$ of ν given μ, and imposes mild conditions ensuring the identifiability of Γ.

3.2.2 The Regression Model and the Fréchet-Least-Squares Estimator

Henceforth, we will take the domain Ω to be a compact interval of $\mathbb{R}$. Let $\{(\mu_i, \nu_i)\}_{i=1}^N$ be an independent collection of regressor/response pairs in $\mathcal{W}_2(\Omega) \times \mathcal{W}_2(\mathbb{R})$. Motivated by the discussion in the previous paragraph, we define the regression model

$$\nu_i = T_{\epsilon_i}\#(T_0\#\mu_i), \quad \{\mu_i, \nu_i\}_{i=1}^N, \tag{3.1}$$

where $T_0 : \Omega \longrightarrow \Omega$ is an unknown optimal map and $\{T_{\epsilon_i}\}_{i=1}^N$ is a collection of independent and identically distributed random optimal maps satisfying $\mathbb{E}\{T_{\epsilon_i}(x)\} = x$ almost everywhere on Ω. These represent the “noise” in our model. The regression task will be to estimate the unknown T_0 from the observations $\{\mu_i, \nu_i\}_{i=1}^N$. To be able to do so, we need to ensure that T_0 is identifiable, and for this we now introduce some conditions.

In the spirit of Section 3.2.1, let P be the probability law induced on $\mathcal{W}_2(\Omega) \times \mathcal{W}_2(\mathbb{R})$ by model (3.1). We denote by P_M and P_N the marginal distributions induced on the typical regressor μ and the typical response ν, respectively.

Denote by Q the measure that is linear average of P_M, i.e.

$$Q(A) = \int_{\mathcal{W}_2(\Omega)} \mu(A)\, \mathrm{d}P_M(\mu).$$

Define the parameter set of optimal transport maps $\mathcal{T}$ as:

$$\mathcal{T} := \{T : \Omega \to \Omega : 0 \leq T'(x) < \infty \text{ for } Q\text{-almost every } x \in \Omega\}.$$

Implicit in the definition of $\mathcal{T}$ is that its elements are assumed differentiable Q-a.e. We will also assume:

Assumption 3.2.1. *The model* (3.1) *is induced by by map T_0 that is a detereministic element of class $\mathcal{T}$.*

Assumption 3.2.2. *The error maps $T_\epsilon : \Omega \to \mathbb{R}$ are i.i.d. non-decreasing random maps satisfying $\mathbb{E}(T_{\epsilon_i}(x)) = x$ and $\mathbb{E}(T_\epsilon^2(x)) < \infty$ for almost every x on Ω.*

With these assumptions in place, we can now establish identifiability:

Theorem 3.2.3. *Assume that the law P induced by model (3.1) satisfies Assumptions 3.2.2 and 3.2.1. Then, the regressor operator $\Gamma(\mu) = T_0\#\mu$ in model* (3.1) *is identifiable over the parameter class $\mathcal{T}$ in the $L^2(Q)$ topology. Specifically, for any $T \in \mathcal{T}$ such that $\|T - T_0\|_{L^2(Q)} > 0$, it holds that*

$$M(T) > M(T_0),$$

where

$$M(T) := \frac{1}{2} \int_{\mathcal{W}_2(\Omega)\times\mathcal{W}_2(\Omega)} d_{\mathcal{W}}^2(T\#\mu, \nu)\, \mathrm{d}P(\mu, \nu). \tag{3.2}$$

Remark 3.2.4. *[Identifiability Q-almost everywhere] Theorem 3.2.3 establishes the identifiability of T_0 up to Q-null sets. Consequently, if the random covariate measure μ is almost surely supported on a strict subset $\Omega_0 \subset \Omega$, we can identify T_0 on Ω_0 (which coincides with the support of Q) but not on $\Omega \setminus \Omega_0$. Of course, if the measure Q is equivalent to Lebesgue measure, in the sense of mutual absolute continuity, identifiability will also hold Lebesgue almost everywhere on Ω. Additional conditions on the law of the random covariate measure μ can yield this equivalence. Suppose the covariate measures have density with respect to the Lebesgue measure. Then a a simple extra condition is to require $\int_{\mathcal{W}_2(\Omega)} \inf_{x\in\Omega} f_\mu(x)\, \mathrm{d}P_M(\mu) > 0$, yielding that $f_Q(x) > 0$, where*

f_μ and f_Q are the Lebesgue densities of the measures μ and Q. However this condition implies that supp(μ) $= \Omega$ with positive probability, which can be restrictive as we would like our model to encompass situations where none of the covariate measures are fully supported on Ω. A considerably weaker condition that guarantees the equivalence of Q to Lebesgue measure is to require the existence of a cover $\{E_m\}_{m\geq 1}$ of Ω such that $P_M\{E_m \subseteq \mathrm{supp}(f_\mu)\} > 0$ for all m – intuitively, this enables different covariate measures to give information on T_0 on different subsets of Ω, but requires that they collectively provide information on all of Ω. As an example let $\Omega = [0,1]$ and let μ be defined as the normalised Lebesgue measure on $S = [U, U+1/3] \mod 1$, where U is a uniform random variable on $[0,1]$. In this case none of the realisations of μ are supported on Ω, but the "cover condition" is satisfied.

Further to identifiability, the theorem gives a way to estimate T_0 by means of M-estimation. We can define an estimator $\hat{T}_N$ as the minimizer of the sample counterpart of M,

$$M_N(T) := \frac{1}{2N}\sum_{i=1}^{N} d_{\mathcal{W}}^2(T\#\mu_i, \nu_i), \qquad \hat{T}_N := \arg\min_{T\in\mathcal{T}} M_N(T), \tag{3.3}$$

where (μ_i, ν_i) are independent samples from P for $i = 1, \dots, N$. In effect this a "Fréchet least square" estimator. The existence and uniqueness of a minimizer is not a priori obvious, but we establish both in the next section 3.2.4.

Remark 3.2.5 (Pure Intercept Model). *When all the input measures are equal, $\mu_1 = \ldots = \mu_N$, our regression model reduces to a "pure intercept model", which is equivalent to the problem of estimating a Fréchet mean. To see this, let μ_0 a fixed measure. From the assumption that $\mathbb{E}\{T_{\epsilon_i}(x)\} = x$ a.e., one can deduce that the conditional Fréchet mean of the measure ν, given the measure μ_0 is equal to $\nu_0 = T_0\#\mu_0$. Estimation of T_0 is then equivalent to estimation of the Fréchet mean ν_0 of the output measures, since $T_0 = T_{\mu_0\to\nu_0} = F_{\nu_0}^{-1} \circ F_{\mu_0}$.*

3.2.3 Interpretation and Comparison

It was argued in the introduction that the proposed regression model has the advantage of being easily interpretable, and now we elaborate on this point. The fact that the regressor operator $\Gamma(\mu)$ takes the form

$$\Gamma(\mu) = T_0\#\mu, \tag{3.4}$$

where $T_0 : \Omega \longrightarrow \Omega$ is a monotone map, has a simple interpretation in terms of mass transport: the effect of the Fréchet mean in this regression is to transport the probability mass assigned by μ on a subinterval $(a,b) \subset \Omega$ onto the transformed subinterval $(T_0(a), T_0(b))$. Therefore, the model can be directly interpreted at the level of the quantity that the input/output measures are modelling. In particular, the model can be interpreted at the level of quantiles. Since

$$F^{-1}_{T_0\#\mu}(\alpha) = (T_0 \circ F^{-1}_{\mu})(\alpha) = T_0\{F^{-1}_{\mu}(\alpha)\}, \qquad \alpha \in (0,1),$$

we can see that the mean effect of the regression is to move the α-quantile of μ, say q_α, to the new location $T_0(q_\alpha)$. Each response distribution ν_i will then further deviate from its conditional Fréchet mean $T_0\#\mu_i$ by means of a random monotone "error" map $T_\epsilon : \Omega \to \mathbb{R}$ whose expectation is the identity map,

$$F^{-1}_{\nu_i}(\alpha) = T_{\epsilon_i}\left[T_0\{F^{-1}_{\mu}(\alpha)\}\right], \qquad \alpha \in (0,1).$$

This highlights the analogy with a classical regression setup, except that the addition operation is replaced by the composition operation at the level of quantiles, or equivalently, by the push-forward operation at the level of distributions. In particular, the assumption that $\mathbb{E}\{T_{\epsilon_i}(x)\} = x$ is directly analogous to the classical assumption that the errors have zero mean: one can directly see that $\mathbb{E}\{T_{\epsilon_i}(x)\} = x$ for almost all $x \in \Omega$ implies that

$$\mathbb{E}\{F^{-1}_{\nu_i}(\alpha)\} = \mathbb{E}\Big(T_{\epsilon_i}\left[T_0\{F^{-1}_{\mu}(\alpha)\}\right]\Big) = T_0\{F^{-1}_{\mu}(\alpha)\}, \qquad \alpha \in (0,1).$$

Assuming that we have obtained an estimator $\hat{T}_N$ of the regression map T based on N regressor/response pairs, we can then define the *fitted distributions*,

$$\hat{\nu}_i = \hat{T}_N\#\mu_i.$$

We can also define the ith *residual map* $T_{e_i}(x) : \Omega \to \Omega$ as the optimal transport map $T_{e_i} = T_{\hat{\nu}_i \to \nu_i}$ that pushes forward the fitted value $\hat{\nu}_i$ to the observed response ν_i. The residual maps can be plotted in a "residual plot" and contrasted to the identity map, by analogy to the classical regression case. This can help identify outlying observations, and also to appreciate in what manner the fitted values differ from the observe values. In particular, it can reveal in which regions of the support of the measures the model provides a good fit, and where less so. It can also serve to identify clusters of observations whose residuals are similar, suggesting the potential presence of a latent indicator variable, i.e. that separate regressions ought to be fit to different groups of observations. Finally, the residual plot can serve as a diagnostic tool for the validity of the model. Since the residual map T_{e_i} can be seen as a proxy for the latent error map T_{ϵ_i}, deviations of the average of the residual maps from the identity can serve as a means to diagnose departures from the assumed model. Note that, contrary to classical regression, where the residuals sum to zero by construction, the residual maps T_{e_i} are *not* constrained to have mean equal to the identity.

By comparison, Chen et al. [15] introduce (linear) regression in Wasserstein space by means of a geometric approach, that is in a sense a linear model between tangent spaces. Namely, for $\bar{\mu}$ and $\bar{\nu}$, the Fréchet means of the regressor and response measures, they postulate a regressor operator of the form

$$\Gamma(\mu) = \left\{\mathcal{B}(T_{\bar{\mu}\to\mu} - I) + I\right\}\#\bar{\nu}, \tag{3.5}$$

where $I(x) = x$ is the identity map on Ω, and $\mathcal{B} : L^2(\bar{\mu}) \to L^2(\bar{\nu})$ is a bounded linear operator with some assumptions, so that the terms involved be well-defined. Again, linearity guarantees identifiability. The expression appears convoluted, but the geometrical interpretation is simple: $T_{\bar{\mu}\to\mu} - I$ represents the image of μ under the log map at $\bar{\mu}$ (see Section 2.3 of Panaretos and Zemel [55]). Equivalently, $T_{\bar{\mu}\to\mu} - I$ is the lifting of μ to the tangent space $\mathrm{Tan}_{\bar{\mu}}\{\mathcal{W}_2(\Omega)\} \subset L^2(\bar{\mu})$ at $\bar{\mu}$. Once the regressor μ is lifted onto $\mathrm{Tan}_{\bar{\mu}}\{\mathcal{W}_2(\Omega)\} \subset L^2(\bar{\mu})$, the action of the regression operator is to map it to its image in $L^2(\bar{\nu})$ via the bounded linear operator $\mathcal{B} : L^2(\bar{\mu}) \to L^2(\bar{\nu})$, as in a standard functional linear model. The final step is to push forward $\bar{\nu}$ by this image plus the identity, i.e. $\mathcal{B}(T_{\bar{\mu}\to\mu} - I) + I$, which retracts back onto $\mathcal{W}_2(\Omega)$ and yields a measure (if $\mathcal{B}(T_{\bar{\mu}\to\mu} - I) + I$ is a monotone map, then this is equivalent to exponentiation, see Section 2.3 of Panaretos and Zemel [55]). The model is most easily interpretable on the tangent space, where it states that the expected lifting of the response ν_i at $\bar{\nu}$ is related to the lifting of the regressor μ_i at $\bar{\mu}$ by means of the linear operator $\mathcal{B}$. Similarly, fitted values are defined on the tangent space, and then can be retracted by the same push-forward operation.

The two approaches do not directly compare, and neither captures the other as a special case. Similarly, there is no reason to a priori expect that one model would typically outperform the other in terms of fit, and one can expect this to depend on the specific data set at hand. Thus, our method should be seen as an alternative rather than an attempt at an improved or more general version of regression. An apparent advantage of the regressor function (3.4), however, is an arguably easier and more direct interpretation of the regression effect, directly at the level regressor/response, through a monotone re-arrangement of probability mass, as discussed above. Indeed this allows a direct point-wise interpretation of the regression effect. The regressor (3.5) on the other hand allows for a traditional (functional) regression interpretation via the linear operator $\mathcal{B}$, albeit acting on the logarithms of regressor/response, which makes it harder to interpret the regression effect at the level of the original measures, since there are two transformations involved, one non-linear and one linear. Similar points can be made with regards to the residuals and residual plots. Another potential advantage is at the level of regularity conditions imposed on Γ for the purposes of theory. Equation (3.5) leads to an inverse problem on the tangent space, as is standard with functional linear models, and thus requires more delicate technical assumptions on the problem, in addition to regularisation. By contrast, the shape-constrained approach (3.4) only requires monotonicity on the regressor T_0. It also avoids the instabilities of an inverse problem.

The utility of our model illustrated in Section 3.4, which considers an example where the age-at-death distribution ν_i for country i in 2013 serves as a response distribution, and the age-at-death distribution μ_i of the same country in 1983 serves as the regressor. Interestingly, it leads to similar fits and qualitative conclusions as the analysis of the same data by Chen et al. [15], while exhibiting a clean and more expansive interpretation. Indeed, our definition of residual maps help identify effects related to changes in infant mortality not easily detectable when looking only at the fitted distributions, and to identify an interesting clustering of observations. See Section 3.4 for more details.

3.2.4 Existence and Uniqueness of the Estimator

In this section, we establish the existence and uniqueness of the estimator $\hat{T}_N$. To show the existence, we use a variant of the Weierstrass theorem, namely Kurdila and Zabarankin [40, Thm 7.3.6], stated for convenience as Theorem 3.6.1 in the Appendix. This requires establishing the convexity and Gateaux differentiability of the functional M_N, and this we do in the next lemma:

Lemma 3.2.6 (Strict Convexity and Differentiability). *Let $\mathcal{T}$ be the parameter set and suppose we have N independent observations (μ_i, ν_i) that are realizations of P. Then both the empirical functional $M_N(T)$ and the population functional $M(T)$ are strictly convex with respect to $T \in \mathcal{T}$. Moreover the functionals M and M_N are Gateaux-differentiable on the set of optimal maps in $\mathcal{T}$ with respect to the $L^2(Q)$ and $L^2(Q_N)$ distances, respectively. The corresponding derivatives of M in the direction $\eta \in L^2(Q)$ is:*

$$D_\eta M(T) = \int \int_\Omega \eta(x)\{T(x) - T_{\mu,\nu}(x)\}\, \mathrm{d}\mu(x)\, \mathrm{d}P(\mu,\nu), \tag{3.6}$$

and the derivative of M_N in the direction $\eta \in L^2(Q_N)$ is

$$D_\eta M_N(T) = \frac{1}{N}\sum_{i=1}^{N} \int_\Omega \eta(x)\{T(x) - T_{\mu_i,\nu_i}(x)\}\, \mathrm{d}\mu_i(x), \tag{3.7}$$

where $T_{\mu,\nu}$ is the optimal map from μ to ν.

Since $\mathcal{T}$ is a convex, closed, and bounded subset of $L^2(Q)$ functions, we may now apply the Weierstrass theorem cited above to conclude:

Proposition 3.2.7 (Existence and Uniqueness of the Estimator)**.** *There exists a unique solution $\hat{T}_N \in \mathcal{T}$ to the Fréchet sum-of-squares minimization problem* (3.3)*, with uniqueness being in the $L^2(Q_N)$ sense.*

3.2.5 Consistency and Rate of Convergence

In this section, we establish the asymptotic properties of the proposed estimators both in the case of the fully observed set of measures $\{\mu_i, \nu_i\}$ and the case where one only indirectly observes input/output distributions through i.i.d. samples from each. A natural risk function to measure the quality of the estimator is the Fréchet mean squared error:

$$R(T) := \mathbb{E}_{\mu \sim P_M} d^2_{\mathcal{W}}(T_0 \# \mu, T \# \mu) = \int_{\mathcal{W}_2(\Omega)} d^2_{\mathcal{W}}(T_0 \# \mu, T \# \mu)\, \mathrm{d}P_M(\mu).$$

Using the equation (2.5) we can rewrite the above risk as follows:

$$\begin{aligned}\int d^2_{\mathcal{W}}(T_0 \# \mu, T \# \mu)\, \mathrm{d}P_M(\mu) &= \int \|T_0 - T\|^2_{L^2(\mu)}\, \mathrm{d}P_M(\mu) \\ &= \int \int_\Omega \left|T_0(x) - T(x)\right|^2 \mathrm{d}\mu(x)\, \mathrm{d}P_M(\mu) \\ &= \|T_0 - T\|^2_{L^2(Q)}\end{aligned}$$

Thus, we can obtain consistency and convergence rates in Fréchet mean squared error using the criterion $\|T_0 - \hat{T}_N\|_{L^2(Q)}$, in particular:

Theorem 3.2.8. *In the context of model* (3.1)*, suppose that Assumptions 3.2.1 and 3.2.2 hold true. Then, the estimator $\hat{T}_N$ defined in* (3.3) *is a consistent estimator for T_0 satisfying*

$$N^{1/3} \left\|\hat{T}_N - T\right\|_{L^2(Q)} = O_{\mathbb{P}}(1). \tag{3.8}$$

In many practical applications, one does not have not access to the measures (μ_i, ν_i). Instead, one has to make do with observing random samples from each μ_i and ν_i. In this case, a standard approach is to use smoothed proxies in lieu of the unobservable measures, usually assuming some more regularity. Therefore in this case we have to assume the input distributions have density with respect to the Lebesgue measure.

Assumption 3.2.9. *Let μ be a measure in the support of P_M. Then μ is absolutely continuous with the respect to the Lebesgue measure on Ω.*

We also denote by Q_N the empirical counterpart of Q, namely $Q_N(A) = \frac{1}{N}\sum_{i=1}^N \mu_i(A)$, where $\{\mu_i\}$ are independent random measures with law P_M.

Also note that in the presence of Assumption 3.2.9, the Q-a.e. existence of T' is automatically guaranteed, since Lebesgue's theorem on the differentiation of monotone functions states that a monotone function automatically has a derivative Lebesgue almost everywhere in the interior of Ω, and Assumption 3.2.9 implies that Q is dominated by Lebesgue measure.

Let μ_i^n and ν_i^n be consistent estimators of μ_i and ν_i obtained from smoothing a random sample of size n from each respective measure. Given such estimators, define a new estimator of T_0 as

$$\hat{T}_{n,N} := \arg\min_{T \in \mathcal{T}_B} \frac{1}{2N} \sum_{i=1}^{N} d_{\mathcal{W}}^2(T\#\mu_i^n, \nu_i^n), \tag{3.9}$$

where

$$\mathcal{T}_B := \{T : \Omega \to \Omega : 0 \leq T'(x) < B \text{ for } Q\text{-almost every } x \in \Omega\}.$$

Note that here one can use any estimators of μ_i and ν_i which are consistent in Wasserstein distance, provided μ_i^n is absolutely continuous.

Then, the rate of convergence of $\hat{T}_{n,N}$ will depend on the rate of convergence of μ_i^n and ν_i^n to μ_i and ν_i, respectively in the Wasserstein distance:

Theorem 3.2.10. *In the context of model* (3.1)*, suppose that Assumptions 3.2.2 and 3.2.9 holds true, and furthermore that there exists a $B < \infty$ such that $T_0 \in \mathcal{T}_B$, and $T_\epsilon \in \mathcal{T}_B$ almost surely. Then, the estimator $\hat{T}_{n,N}$ defined in* (3.9) *satisfies*

$$\left\|\hat{T}_{n,N} - T_0\right\|_{L^2(Q)} = O_{\mathbb{P}}(N^{-1/3}) + O_{\mathbb{P}}({r_n}^{-1/2}), \tag{3.10}$$

where r_n^{-1} is the rate of convergence in the Wasserstein distance of μ_i^n to μ_i and ν_i^n to ν_i.

Precise values of r_n can be obtained by choosing specific estimators and imposing additional regularity on the underlying regressor/response measures. For instance, one can follow the estimation approach of [50] and obtain the minimax rate of convergence over measures with densities in Besov classes.

Remark 3.2.11. *Note that $B \in (0, \infty)$ can be any finite constant, however large. Its precise value does not influence the rate* (3.10) *itself, but only the constants. It is therefore not to be interpreted as a regularisation parameter.*

3.2.6 Minimax Lower Bound

We establish the minimax optimality of the estimator $\hat{T}_N$. Since there is no ill-conditioning inherent in the setup of model (3.1), one might hope for a rate of $O(N^{-1/2})$

when the measures are completely observed (as opposed to being sampled from or observed with error). Our purpose is to show that, in the most general case, $O(N^{-1/3})$ is the "right rate", by establishing a link between model (3.1) and classical isotonic regression. We also discuss additional conditions that could lead to improved rates.

Theorem 3.2.12. *In the context of model* (3.1) *and under Assumptions 3.2.1 and 3.2.2, it holds that:*

$$\inf_{\hat{T}} \sup_{T_0 \in \mathcal{T}} \mathbb{E}\left\{ \left\| \hat{T}_N - T_0 \right\|^2_{L^2(Q)} \right\} \geq N^{-2/3},$$

where the infimum is taken over all measurable functions $\hat{T}$ *of the data* $\{(\mu_i, \nu_i)\}_{i=1}^N$ *ranging in* $\mathcal{T}$.

Here we are only concerned with the lower bound (with respect to N) when one observes the covariate/response measures completely, as an indicator of the minimax estimation rate intrinsic[1] to model (3.1).

Corollary 3.2.13. *Under the same conditions, the Fréchet-least-squares estimator* $\hat{T}_N$ *attains the lower bound given by Theorem 3.2.12 and the upper bound given by Theorem 3.2.8 and 3.2.12, and consequently is minimax optimal.*

The minimax rate obtained here is not directly comparable to rates such as those obtained in the classic case of functional *linear* regression ([32] ; see also Cuevas [19], and Goia and Vieu [29], Aneiros et al. [5] for broader reviews). This is not only because the nature of the model is intrinsically *non-linear*, but also because the relationship posited by the model does not lead to an ill-posed inverse problem. They are also distinct from rates pertaining to fully non-parametric functional regression (Chagny and Roche [14], Brunel et al. [10]; see also Ling and Vieu [43]). The reason is that the regression operator, though non-parametrically specified, is constrained to be *monotone*. In this sense, the model can be seen as a "shape-constrained non-parametric functional regression model". This distinguishes the form of the estimation procedure (e.g. kernel averaging is unsuitable) and affects the minimax rate itself. In particular, the fact that both response and covariate are distributions play a distinct role in the analysis – e.g. the use of Dirac deltas to connect to classical isotonic regression.

3.2.7 Computation

Since the domain Ω is one-dimensional, we have that

$$d^2_{\mathcal{W}}(\nu, \mu) = \int_0^1 \left| F_\mu^{-1}(p) - F_\nu^{-1}(p) \right|^2 \mathrm{d}p.$$

[1] If the covariate/response measures are observed indirectly, additional regularity is asserted on the covariate/response measures in order to be able to recover them. But such assumptions are extrinsic to the structure of the model (3.1) itself.

Furthermore, since the regressors μ_i are assumed absolutely continuous (Assumption 3.2.9), we can always write $\nu_i = T_{\mu_i \to \nu_i} \# \mu_i$ for an optimal map $T_{\mu_i \to \nu_i}$. We can therefore manipulate the Fréchet sum-of-squares and use a Riemann approximation to write

$$\begin{aligned}
\sum_{i=1}^{N} d_{\mathcal{W}}^2(T \# \mu_i, \nu_i) &= \sum_{i=1}^{N} \left\| T \circ F_{\mu_i}^{-1} - F_{\nu_i}^{-1} \right\|_{L^2}^2 \\
&= \sum_{i=1}^{N} \int_0^1 \left| T \circ F_{\mu_i}^{-1}(p) - F_{\nu_i}^{-1}(p) \right|^2 \mathrm{d}p \\
&= \sum_{i=1}^{N} \int_0^1 \left| T \circ F_{\mu_i}^{-1}(p) - T_{\mu_i \to \nu_i} \circ F_{\mu_i}^{-1}(p) \right|^2 \mathrm{d}p \\
&= \sum_{i=1}^{N} \int_{\Omega} \left| T(x) - T_{\mu_i \to \nu_i}(x) \right|^2 \mathrm{d}\mu_i(x) \qquad (3.11) \\
&\approx \sum_{i=1}^{N} \sum_{j=1}^{m} \left| T(x_j) - T_{\mu_i \to \nu_i}(x_j) \right|^2 \mu_i(I_j), \qquad (3.12)
\end{aligned}$$

for m user-defined nodes $\{x_j\}_{j=1}^m$ in an interval partition $\{I_j\}_{j=1}^m$ of Ω. Writing $y_{ij} = T_{\mu_i \to \nu_i}(x_j)$, $w_{ij} = \mu_i(I_j)$ and $z_j = T(x_j)$, we reduce the above approximate minimization of the Fréchet sum-of-squares to the solution of the following convex optimization problem:

$$\begin{aligned}
&\text{minimise } f(z) = \sum_{i=1}^{N} \sum_{j=1}^{m} w_{ij} h_i(y_{ij}, z_j), \\
&\text{subject to } z_1 \leq z_2 \leq \cdots \leq z_m,
\end{aligned} \qquad (3.13)$$

where $h_i(y_{ij}, z_j) = |y_{ij} - z_j|^2$. The above problem resembles an isotonic regression problem with repeated measurements, and can be solved via the Pool-Adjacent-Violater-Algorithm (PAVA) [45], which has a linear time complexity.

Remark 3.2.14. *To be strictly faithful to the assumptions of Theorem 3.2.10, the computation could incorporate additional constraints of the form* $(z_{i+1} - z_i) \leq B(x_{i+1} - x_i)$*, as a discretization of* $T' \leq B$*. From a practical point of view, though, we always have* $(z_{i+1} - z_i) \leq \left(|\Omega| / \min_{1 \leq j \leq m} |I_j|\right)(x_{i+1} - x_i)$*, since* $T : \Omega \to \Omega$ *is monotone. So maintaining the original formulation of Section 3.2.7 implicitly corresponds to some* $B > |\Omega| / \min_{1 \leq j \leq m} |I_j|$ *in Theorem 3.2.10.*

3.3 Simulated Examples

In this section we illustrate the estimation framework and finite sample performance of the method by means of some simulations. First we generate random predictors $\{\mu_i\}_{i=1}^N$. We consider random distributions that are mixtures of three independent Beta components. We choose the parameters of the Beta distributions to be uniformly distributed random variables on $[1, 10]$, with densities

$$f_{\mu_i}(x) = \sum_{j=1}^{3} \pi_j b_{\alpha_{i,j},\beta_{i,j}}(x), \quad \alpha_{i,j} \sim \text{Uniform}[1, 10], \quad \beta_{i,j} \sim \text{Uniform}[1, 10].$$

The $\{\pi_j\}_{j=1}^3$ are arbitrary fixed mixture weights in $[0, 1]$, such that $\sum_{j=1}^3 \pi_j = 1$. As for the noise maps T_{ϵ_i}, we use the class of random optimal maps introduced in Panaretos and Zemel [53]. Let k be an integer and define $\zeta_k : [0, 1] \to [0, 1]$ by

$$\zeta_0(x) = x, \quad \zeta_k(x) = x - \frac{\sin(\pi k x)}{|k|\pi}, \qquad k \in Z \setminus \{0\}.$$

These are strictly increasing smooth functions satisfying $\zeta_k(0) = 0$ and $\zeta_k(1) = 1$ for any k. These maps can be made random by replacing k by an integer-valued random variable K. If the distribution of K is symmetric around zero, then it is straightforward to see that $E[\zeta_K(x)] = x$, for all $x \in [0, 1]$, as required in the definition of model (3.1). We generate a discrete family of random maps by the following procedure, which is slightly different from the mixture family of maps introduced in [53]: for $J > 1$ let $\{K_j\}_{j=1}^J$ be i.i.d. integer-valued symmetric random variables, and $\{U_{(j)}\}_{j=1}^{J-1}$ be the order statistics of $J - 1$ i.i.d. uniform random variables on $[0, 1]$, independent of $\{K_j\}_{j=1}^J$. The random maps are then defined as

$$T_\epsilon(x) = \sum_{j=1}^{J-1} I(U_{(j)} \le x \le U_{(j+1)}) \left[\zeta_{K_j}\left(\frac{x - U_{(j)}}{U_{(j+1)} - U_{(j)}} \right) \left(U_{(j+1)} - U_{(j)} \right) + U_{(j)} \right].$$

As for the optimal map T_0 constituting the regression operator, we set $T_0 = \zeta_4$. After having generated the random μ_i and T_{ϵ_i}, we generate the response distributions according to model (3.1), i.e. $\nu_i = T_{\epsilon_i} \# T \# \mu_i$. Figure 3.1 depicts representative sample pairs of predictor and response densities.

For estimation, we consider the case where we only observe n independent samples from each pair of distributions $(\mu_i, \nu_i)_{i=1}^N$. For simplicity, we use kernel density estimation, rather than the estimators in [50], to obtain the proxies μ_i^n and ν_i^n for the distributions μ_i and ν_i. Subsequently, for each i, we estimate Q_i^n, where Q_i^n is the optimal map such that $\nu_i^n = Q_i \# \mu_i^n$ and solve the convex optimisation problem described in Section 3.2.7 to obtain the estimator $\hat{T}_{n,N}$. Figure 3.2 contrasts the

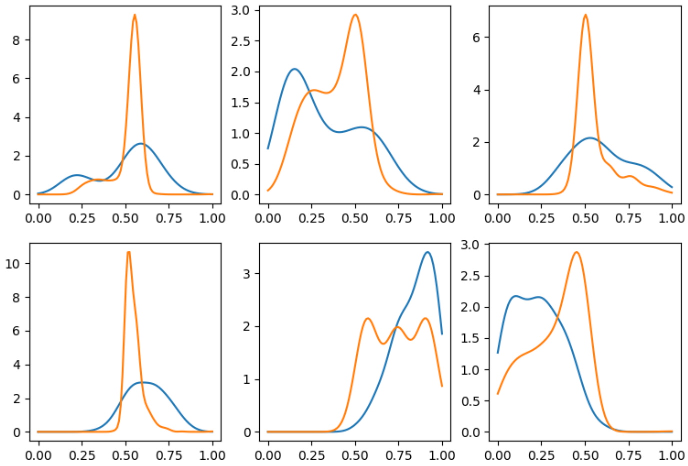


Figure 3.1: Examples of simulated predictor (blue) and corresponding response (orange) densities.

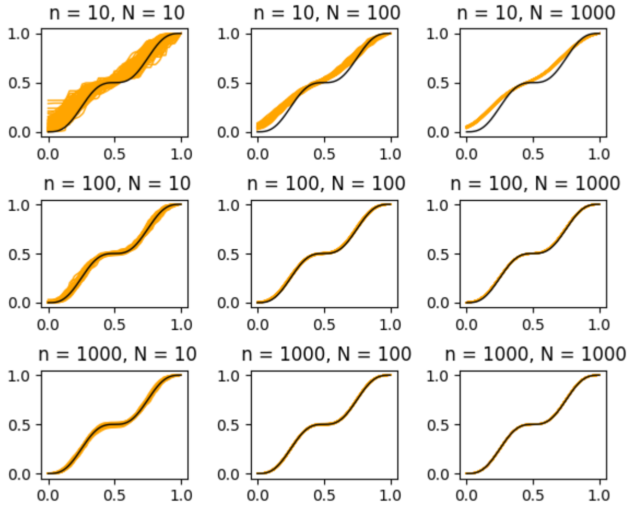


Figure 3.2: Estimated (yellow) versus true (black) regression map for each of 100 replications of the combinations of $N \in \{10, 100, 1000\}$ and $n \in \{10, 100, 1000\}$.

estimated and true regression maps in each replication, for all nine combinations $N \in \{10, 100, 1000\}$ and $n \in \{10, 100, 1000\}$. It is apparent that the dominant source of error is the bias due to partial observation, i.e. due to observing the measures through finite samples of size n. When n is moderately large (e.g. $n = 100$) we see that the agreement between estimated and true map is very good, even for small values of N. To quantitatively summarise the behaviour of the mean squared er-

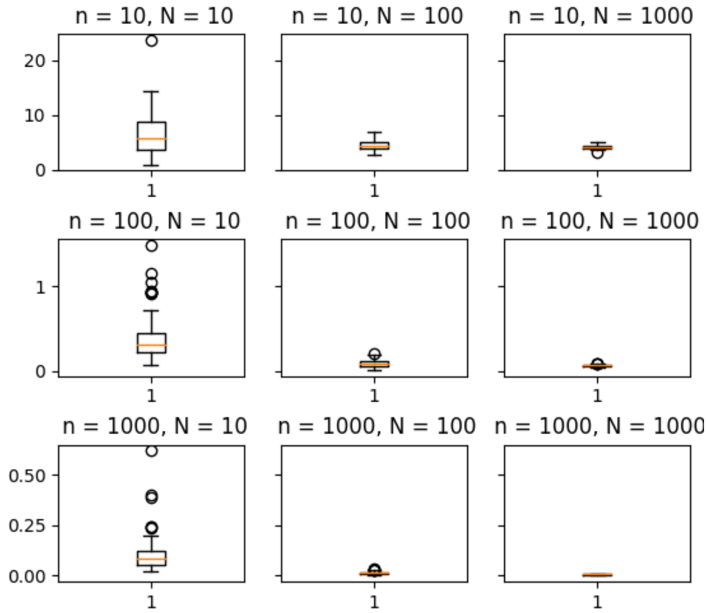


Figure 3.3: Boxplots for the squared L^2 deviation between the true regression map and the estimated regression maps based on 100 replications for the nine combinations of $N \in \{10, 100, 1000\}$ and $n \in \{10, 100, 1000\}$. The y-axis scale is common for different values of N.

ror in N, we construct boxplots for the error $\|\hat{T}_{n,N} - T_0\|_{L^2}$ in Figure 3.3, each based on 100 replications for the corresponding combination of $n \in \{10, 100, 1000\}$ and $N \in \{10, 100, 1000\}$. The scale used is the same for each value of n, in order to focus the behaviour with respect to N.

3.4 Analysis of Mortality Data

We consider the age-at-death distributions for $N = 37$ countries in the years 1983 and 2013, obtained from the Human Mortality Database of UC Berkeley and the Max Planck Institute for Demographic Research [2]. Death rates are provided by single years of age up to 109, with an open age interval for 110+. We use Gaussian kernel density smoothing, to obtain age-at-death densities from the count data. Denote by μ_i the age-at-death distribution for the ith country at year 1983 and ν_i the age-at-death distribution for the same country at year 2013. We use the distributions μ_i and ν_i as predictor and response distributions respectively. We chose these two years to allow comparison with Chen et al. [15], who illustrate their methodology on the same data set, and same pair of years.

[2] openly accessible at `www.mortality.org`

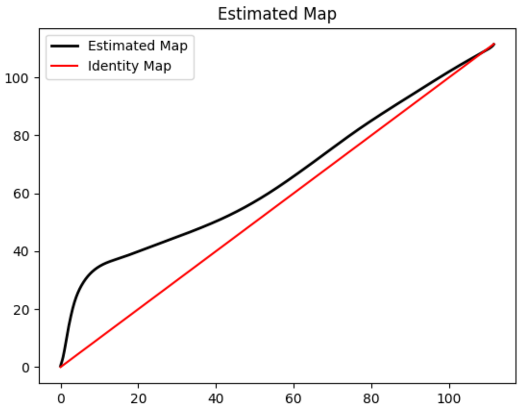


Figure 3.4: Estimated Regression Map for the age-at-death distributional regression (black) contrasted to the identity (red).

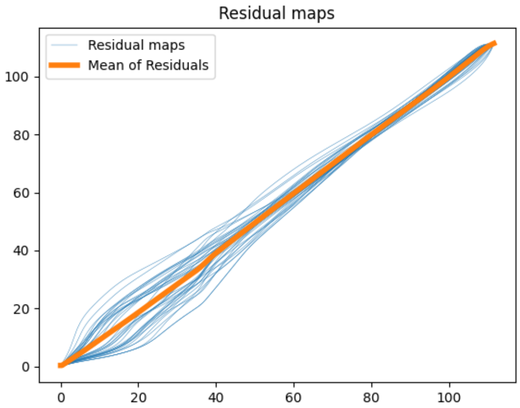


Figure 3.5: Residual maps of all the 37 countries (blue) and their average map (orange).

We fit the model (3.1) by means of the approach described in Section 3.2.7 to obtain the estimated regression map based on the $N = 37$ countries. This is depicted in Figure 3.4. The map dominates the identity map pointwise, indicating that the regression effect is to transport the mass of the age-at-death distribution to the right at visually all locations. Said differently, the map indicates an effect of net improvement in mortality across all ages. The most pronounced such effect is observed in young ages (between 0-10), where the regression map rises steeply: The proportion of the population dying at ages 0-10 in 1983 is redistributed approximately over the range 0-30 in 2013. The form of the map restricted to $[0, 10] \mapsto [0, 30]$ is approximately linear, indicating that this redistribution is achieved by conserving the actual shape

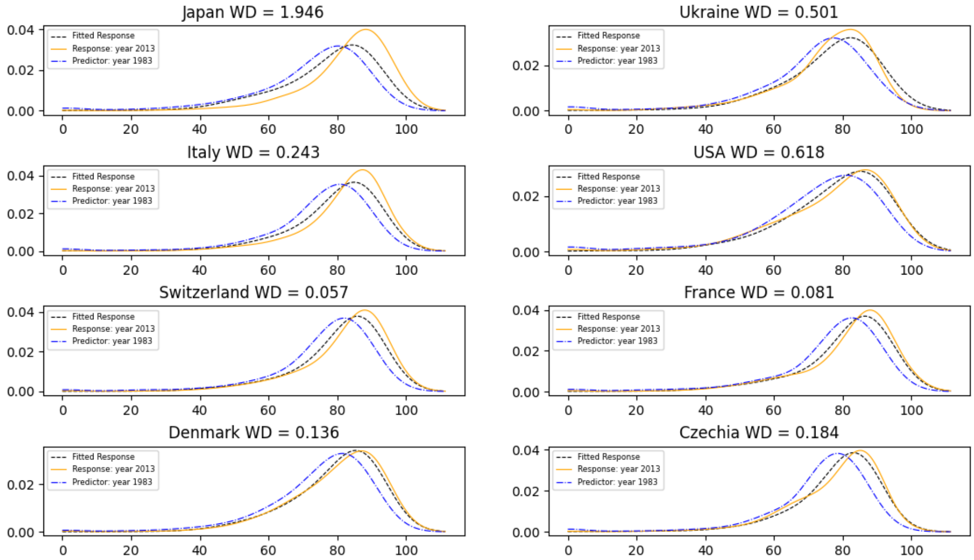


Figure 3.6: Distribution-on-distribution regression for the mortality distributions of Japan, Ukraine, Italy and USA in the year 2013 on those in 1983. Here WD stands for the Wasserstein distance between the observed and fitted densities at year 2013, indicating goodness-of-fit.

of the distribution but scaling by a constant. The effect is still visible though less pronounced in the early adult to middle age range: The proportion of the population dying at ages between 20 and 60 in 1983 is approximately redistributed over ages 40-60 in 2013. The regression map is approximately parallel to the identity map on the range 60-80, shifted upwards by about 10 years indicating a translation of that interval by that amount of years between 1983 and 2013, i.e. the proportion of the population dying between 60-80 in 1983 has shifted to ages 70-90, but the shape of the distribution of that proportion over each of these two 20 year periods is approximately conserved. Overall, the regression map approximately resembles a piecewise linear map, allowing to interpret it locally by translations and dilations.

It is not easy to directly compare the effects expressed via this estimated regression map with the effects reflected by the estimated regression coefficient function $\hat{\beta}$, that is, the integral kernel of the operator $\mathcal{B}$ in Equation (3.5) obtained in Chen et al. [15, Figure 3], when fitting their model to the same data. This is largely due to fact that the $\hat{\beta}$ acts on tangent space elements, and thus is rather subtle to interpret. In interpreting their estimated regression operator, those authors remarked that the estimated $\hat{\beta}(s,t)$ was stratified according to the s argument so that, "*if the log-transformed predictor is non-negative or non-positive throughout its domain, then the fit for the log-transformed response is determined by the comparison of the abso-*

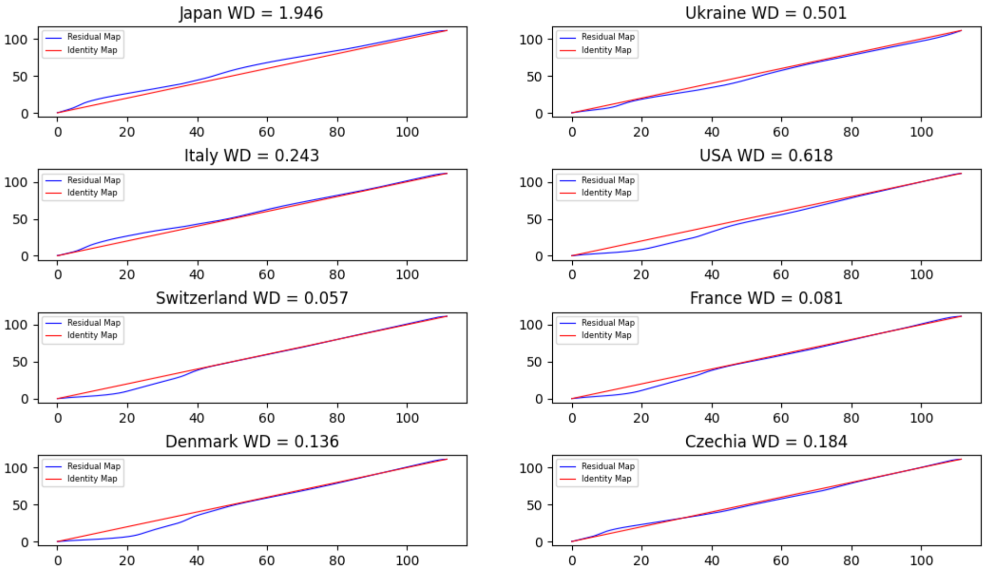


Figure 3.7: Residual maps $T_{\hat{\nu}_i \to \nu_i}$ (blue) vs Identity map (red) for the eight countries in Figure 3.6.

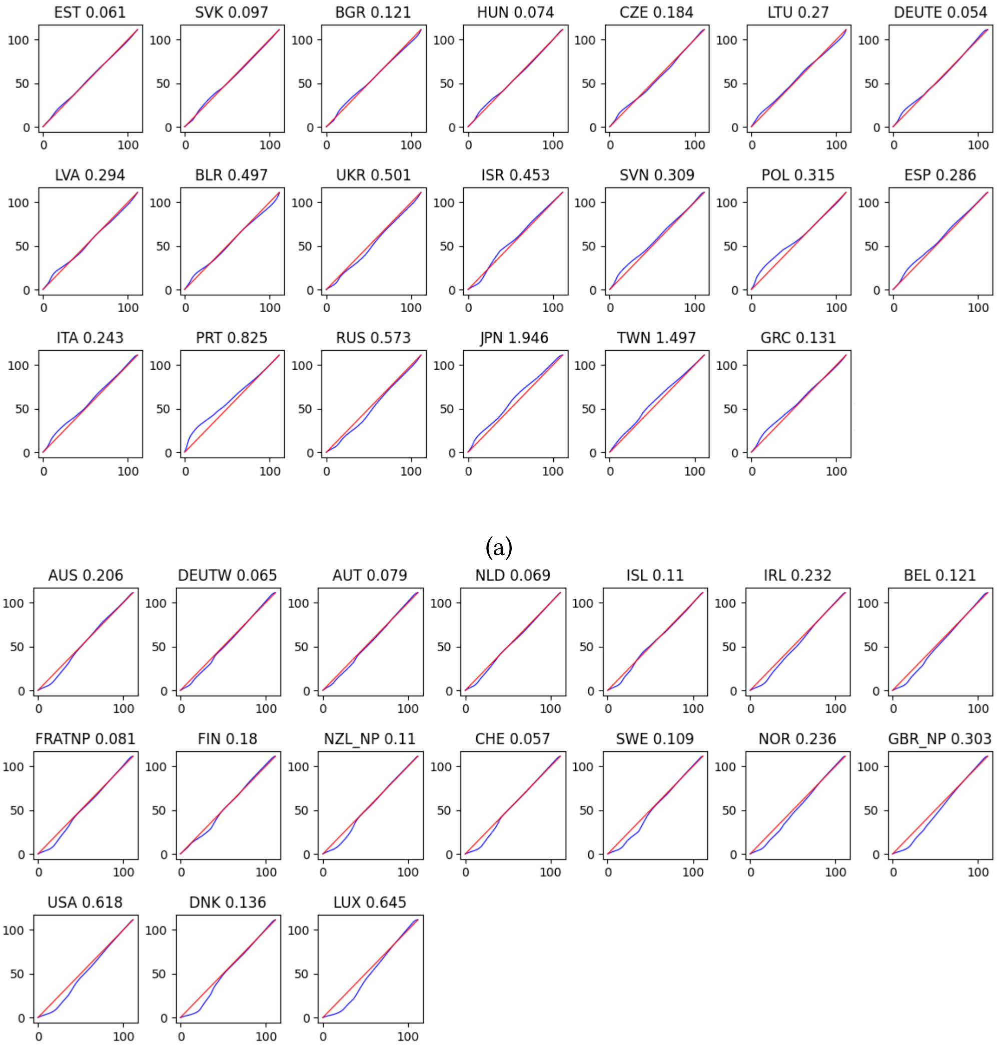
lute values of the log transformed predictor over the positive and negative strata of the estimated coefficient $\beta(\cdot, t)$".

Using the estimated map $\hat{T}_N$ we can then compute the fitted age-at-death distributions for the year 2013, namely $\hat{\nu}_i = \hat{T}_N \# \mu_i$. Figure (3.6) depicts the predictor and response densities as well as fitted response densities for a sample of 8 different countries. The first four of these countries (Japan, Ukraine, Italy and USA) were also selected as representative examples in Chen et al. [15]. All eight countries exhibit a negatively-skewed age-at-death distribution. Comparing the actual distributions for the years 1983 and 2013 we can observe the decreasing trend in infant death counts and peaks shifting to older ages, as dictated by the fitted regression map. Contrasting observed and fitted distributions for 2013 allows for better comparison with the model output in [15], than does comparing the estimated regression operators.

Indeed, the main observations made in [15] are also apparent from our fitted model. In the case of our model, besides looking at the shape of the predicted densities, we can also take advantage of the direct interpretability of the residual maps $T_{e_i} = T_{\hat{\nu}_i \to \nu_i}$, where $T_{\hat{\nu}_i \to \nu_i}$ is the optimal map between the fitted response $\hat{\nu}_i$ and actual response ν_i. The collection of residual maps is plotted in Figure (3.7). It is apparent that the pointwise variability declines for progressively older ages, illustrating that it is harder to fit mortality at younger ages. One can then focus on the residual maps of specific countries. For example, doing so in the case of Japan and Ukraine, we reproduce the observation in [15] that "*for Japan, the rightward mortality shift is seen to be more expressed than suggested by the fitted model, so that longevity extension is more than is anticipated, while the mortality distribution for Ukraine seems to shift to the right at a slower pace than the fitted model would suggest*". Similarly, we recover the same inference as in [15] regarding the US: "*while the evolution of the mortality distributions for Japan and Ukraine can be viewed as mainly a rightward shift over calendar years, this is not the case for USA, where compared with the fitted response, the actual rightward shift of the mortality distribution seems to be accelerated for those above age 75* [note: 65 in our case]*, and decelerated for those below age 70* [note: 65 in our case]". In terms of fit as measured by the Wasserstein distance between response and fit, both models have a harder time fitting Japan, ours doing slightly worse. On the flip side, our model fits Italy better, and the US and Ukraine considerably better (we only contrast countries explicitly mentioned in [15]).

Figure 3.5 features the overlay of all residual maps, in order to explore the goodness-of-fit of the model as well as the validity of the model assumptions. As the figure shows, the mean of residuals almost matches the identity map, which provides evidence in support of our model specification, in that the residual effects after correcting for the regression should have mean identity, reflected by the assumption that $E\{T_\epsilon(x)\} = x$. Note that, contrary to usual least squares where the residuals have empirical mean zero, the residual maps need not have mean identity exactly.

Finally, we can scrutinise the individual residual maps for each of the 37 countries which we plot separately in Figures (3.8a) and (3.8b). The separation into two figures is deliberate, and is based on an apparent clustering: in Figure (3.8a) one can observe more of a rightward shift of fitted mortalities compared to the observed moralities for the countries concerned. This contrasts to countries in Figure (3.8b) which feature less of a rightward shift than fitted by the model. In a sense, these are clusters of "underfitted" and "overfitted" observations. Interestingly countries in Figure (3.8a) belong mostly to Eastern Europe plus Portugal, Spain, Italy, Israel, Japan and Taiwan. Countries in figure (3.8b) belong to western/northern European countries plus USA, New Zealand and Australia. Thanks to the pointwise interpretability of the residual maps, one can notice a particular contrast between these two groups of countries in terms of their fitted/observed infant mortality rates. This may be related to the fact that countries in Figure (3.8a) experienced a more pronounced improvement in their health care systems over the period 1983-2013, compared countries in Figure (3.8b) where healthcare was of comparably high quality already in 1983. It is interesting to note that Japan and Taiwan feature residual maps that everywhere dominate the identity.



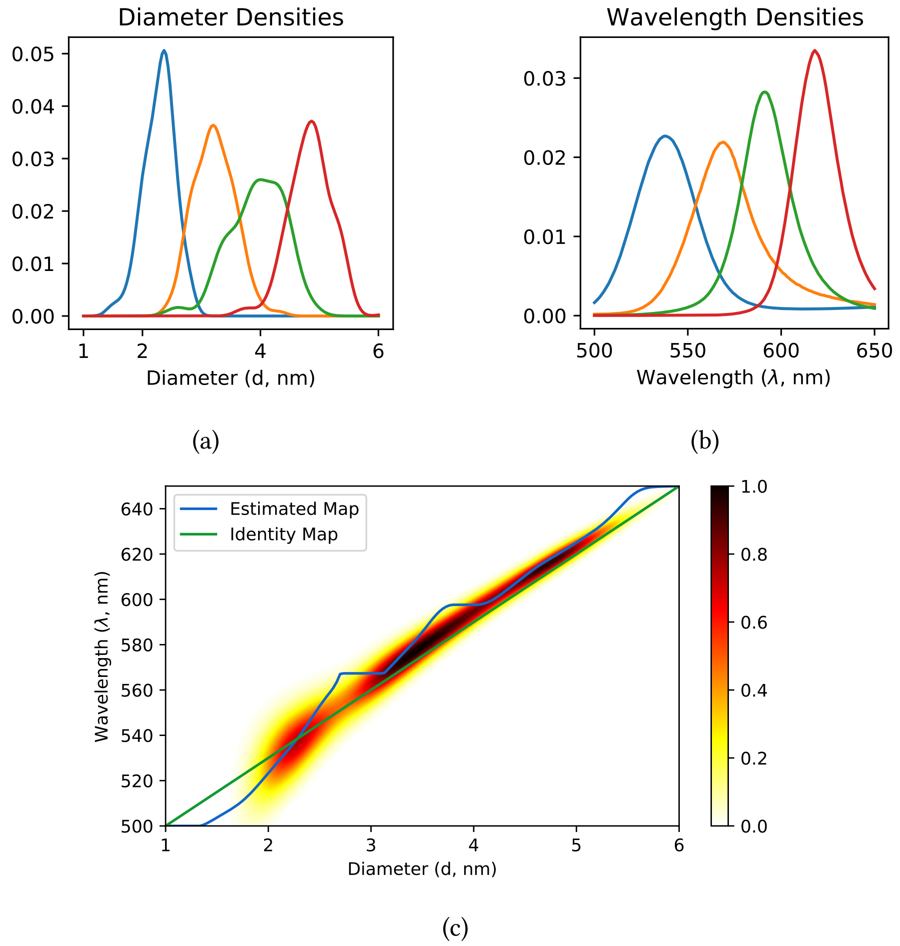
(a)

(b)

Figure 3.8: Residual maps (blue), the identity map (red) and the Wasserstein distance between the observed and fitted densities at year 2013 for each country. The countries are clustered in two groups (a) and (b). The list of abbreviations can be found in Table 3.1 in the Supplement.

3.5 Analysis of Quantum Dot Data

Röding et al. [63] present a method for resolving the relationship between parameters from the observation of their marginal distributions. In their first use case, given 4 pairs of diameter (input) and wavelength (output) distributions of “quantum dot” experimental setting, they estimate the joint distribution [63, Fig. 1] which we plot



(a)

(b)

(c)

Figure 3.9: (a) and (b) are the marginal distributions of diameter and wavelength respectively. (c) is the joint distribution as estimated by [63] and our estimated map.

as the colormap in Figure (3.9c). Quantum dots are nanometer-sized structures produced in large quantities which have useful light-emitting properties, for example, in the production of LEDs [77]. The same batch of quantum dots will have a natural variability in their shapes and sizes which influence their spectra of emission, that is, what wavelengths of light are emitted when some external source (electric current, laser) excites them, and with what intensity are those wavelengths emitted [77].

Given the 4 pairs of distributions shown in Figures (3.9a) and (3.9b), we apply our regression model to obtain the estimated map as shown in Figure (3.9c). Our map does not always follow the areas of maximum weight of the joint distribution but seems to be attracted to them. Note that Röding et al. [63] fitted lognormal distributions to the raw data, while we used a kernel density estimation with bandwidth determined by Scotts' rule to obtain the diameter distributions, and we normalized the raw wavelength intensity data so it is interpretable as a density.

The different treatment of the raw data could be a cause of the extra oscillations observed in the estimated maps. Regardless of the treatment, we think these oscillations should reduce if the number of distribution pairs N were to increase, as these oscillations are a result of overfitting for a small N.

Let us define the line of average λ given a fixed d on the joint distribution as $\bar{\lambda}(d)$. This line is approximately linear with an apparent change of slope around $d = 2\mathrm{nm}$. Even if the input diameter distribution were controlled so as to reach a Dirac $\delta(d - d_0)$ form, the output wavelength may still be distributed around $\bar{\lambda}(d_0)$ due to other uncontrolled parameters: For example, the exact shape of the quantum dot is a relevant factor [37], not only its diameter, among other factors. The joint distribution shows such a finite variation and will show it even for an increasing number of pair distributions N.

It would be interesting to check if the estimated map is related to the average line $\bar{\lambda}(d_0)$ and converges to it for large N. In this case, we would be correctly modelling the deviations from $\bar{\lambda}(d)$ as the gaussian noise of our regression model.

3.6 Proofs

Proof of Lemma 3.2.6. Using the closed form of the Wasserstein distance when $d = 1$ as given by equality (2.2), one can write:

$$M(T) = \frac{1}{2} \int \int_0^1 \left| T\{F_\mu^{-1}(p)\} - F_\nu^{-1}(p) \right|^2 \mathrm{d}p \, \mathrm{d}P(\mu, \nu).$$

The expression above shows that M is convex with respect to T since the map $x \longrightarrow x^2$ is convex and also integration preserves convexity. To show the strict convexity we should prove that for all $0 < \beta < 1$ and all T_1, T_2 such that $\|T_1 - T_2\|^2_{L^2(Q)} > 0$,

$$M\big\{\beta T_1 + (1-\beta)T_2\big\} < \beta M(T_1) + (1-\beta)M(T_2).$$

In fact by expanding the squares in the equality and doing some algebra one can conclude that the equality happens if and only if $\|T_1 - T_2\|^2_{L^2(Q)} = 0$. Thus M, and similarly M_N, are strictly convex.

Notice that the domain of definition of M can be extended to the space of $L^2(Q)$ functions. Therefore the Gateaux derivative of M in the direction of $\eta \in L^2(Q)$ can be defined as:

$$D_\eta M(T) = \lim_{\epsilon \to 0} \frac{M(T+\epsilon\eta) - M(T)}{\epsilon}.$$

Expanding the first term we have:

$$\begin{aligned} M(T+\epsilon\eta) &= M(T) + \epsilon \int \int_0^1 \Big[T\{F_\mu^{-1}(p)\} - F_\nu^{-1}(p)\Big]\eta\{F_\mu^{-1}(p)\}\,\mathrm{d}p\,\mathrm{d}P(\mu,\nu) \\ &\quad + \frac{\epsilon^2}{2}\int\int_0^1 \big|\eta\{F_\mu^{-1}(p)\}\big|\,\mathrm{d}x\,\mathrm{d}P(\mu,\nu) \\ &= M(T) + \epsilon\int \langle T - F_\nu^{-1}\circ F_\mu, \eta\rangle_{L^2(\mu)}\,\mathrm{d}P(\mu,\nu) + \frac{\epsilon^2}{2}\int \|\eta\|^2_{L^2(\mu)}\,\mathrm{d}P(\mu) \\ &= M(T) + \epsilon\int \langle T - F_\nu^{-1}\circ F_\mu, \eta\rangle_{L^2(\mu)}\,\mathrm{d}P(\mu,\nu) + \frac{\epsilon^2}{2}\|\eta\|^2_{L^2(Q)}\,. \end{aligned} \tag{3.14}$$

The last equality is true since

$$\begin{aligned} \int \|\eta\|^2_{L^2(\mu)}\,\mathrm{d}P(\mu) &= \int\int_\Omega |\eta(x)|^2\,\mathrm{d}\mu(x)\,\mathrm{d}P(\mu,\nu) \\ &= \int\int_\Omega |\eta(x)|^2\,\mathrm{d}Q(x) \\ &= \|\eta\|^2_{L^2(Q)}\,. \end{aligned}$$

Since $\|\eta\|^2_{L^2(Q)} < \infty$, we can conclude

$$D_\eta M(T) = \int \langle T - F_\nu^{-1}\circ F_\mu, \eta\rangle_{L^2(\mu)}\,\mathrm{d}P(\mu,\nu) = \int\int_\Omega \{T(x) - T_{\mu,\nu}(x)\}\eta(x)\,\mathrm{d}\mu(x)\,\mathrm{d}P(\mu,\nu),$$

where $T_{\mu,\nu}$ is the optimal map from μ to ν. One can use a similar argument to derive the derivative of M_N.

□

Proof of Theorem 3.2.3. We prove that T_0 is the unique minimizer of the population functional in $\mathcal{T}$. Suppose $\nu = T_\epsilon\#(T_0\#\mu_0)$ for some fixed measure μ_0, where by assumption $\mathbb{E}\{T_\epsilon(x)\} = x$ almost everywhere. Thus according to Proposition 3.2.11 of [55], $T_0\#\mu_0$ is the Fréchet mean of the conditional probability law of ν given μ_0 or equivalently, for any μ_0

$$\arg\inf_{b\in\mathcal{W}_2(\Omega)} \int_{\mathcal{W}_2(\Omega)} d^2_{\mathcal{W}}(b,\nu)\,\mathrm{d}P(\nu|\mu_0) = T_0\#\mu_0,$$

where P is the joint distribution of (μ,ν) induced by Model (3.1). Now T_0 is a minimizer of the above functional, since for any T:

$$\begin{aligned} M(T) &= \int d^2_{\mathcal{W}}(T\#\mu,\nu)\,\mathrm{d}P(\mu,\nu)\\ &= \int\int d^2_{\mathcal{W}}(T\#\mu_0,\nu)\,\mathrm{d}P(\nu|\mu_0)\,\mathrm{d}P(\mu_0)\\ &\geq \int\int d^2_{\mathcal{W}}(T_0\#\mu_0,\nu)\,\mathrm{d}P(\nu|\mu_0)\,\mathrm{d}P(\mu_0)\\ &= \int d^2_{\mathcal{W}}(T_0\#\mu,\nu)\,\mathrm{d}P(\mu,\nu). \end{aligned}$$

Also since $d^2_{\mathcal{W}}(T\#\mu,\nu)$ is strictly convex w.r.t. $T\in\mathcal{T}$, and integration preserves strict convexity, the functional M is strictly convex. So T_0 is, in fact, the *unique* minimizer. □

To establish Proposition 3.2.7, we will use the following theorem.

Theorem 3.6.1 (Kurdila and Zabarankin [40], Theorem 7.3.6)**.** *Let X be a reflexive Banach space and suppose that $f : M \subseteq X \longrightarrow \mathbb{R}$ is Gateaux-differentiable on the closed, convex and bounded subset M. If any of the following three conditions holds true,*

1. *f is convex over M,*
2. *Df is monotone over M,*
3. *$D^2 f$ is positive over M,*

then all three conditions hold, and there exists an $x_0 \in X$ such that

$$f(x_0) = \inf_{x\in M} f(x).$$

Proof of Proposition 3.2.7. The set of maps $\mathcal{T}$ is closed, convex and bounded in the Hilbert space of $L^2(Q)$ functions. Thus the existence follows immediately from (3.2.6) and Theorem 3.6.1. Uniqueness also follows from strict convexity of M. □

To establish the consistency and rate of convergence of our estimator, we will make use of the theory of M-estimation. To this aim, we restate some key theorems from Van Der Vaart and Wellner [72].

Theorem 3.6.2 (Van Der Vaart and Wellner [72], Theorem 3.2.3). *Let M_n be random functions for positive integer n, and let M be a fixed function of θ such that for any $\epsilon > 0$*

$$\inf_{d(\theta,\theta_0)\geq\epsilon} M(\theta) > M(\theta_0), \tag{3.15}$$

$$\sup_{\theta} |M_n(\theta) - M(\theta)| \to 0 \quad \textit{in probability.} \tag{3.16}$$

Then any sequence of estimators $\hat{\theta}_n$ with $M_n(\hat{\theta}_n) \leq M_n(\theta_0) + o_{\mathbb{P}}(1)$ converges in probability to θ_0.

Theorem 3.6.3 (Van Der Vaart and Wellner [72], Theorem 3.2.5). *Let M_N be a stochastic process indexed by a semi-metric space Θ with semi-metric ρ, and let M be a deterministic function, such that for every θ in a neighborhood of θ_0,*

$$M(\theta) - M(\theta_0) \gtrsim \rho^2(\theta, \theta_0).$$

Suppose that, for every N and sufficiently small δ,

$$\mathbb{E}^* \sup_{\rho^2(\theta,\theta_0)<\delta} \sqrt{N}\left|(M_N - M)(\theta) - (M_N - M)(\theta_0)\right| \lesssim \phi_N(\delta),$$

for functions ϕ_N such that $\delta \to \phi_N(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ (not depending on N). Let

$$r_N^2 \phi_N\left(\frac{1}{r_N}\right) \leq \sqrt{N}, \quad \textit{for every } N.$$

If the sequence $\hat{\theta}_N$ satisfies $M_N(\hat{\theta}_N) \leq M_N(\theta_0) + O_{\mathbb{P}}(r_N^{-2})$, and converges in outer probability to θ_0, then $r_N\rho(\hat{\theta}_N, \theta_0) = O_{\mathbb{P}}^*(1)$. If the displayed conditions are valid for every θ and δ, then the condition that $\hat{\theta}_N$ is consistent is unnecessary.

Theorem 3.6.4 (Van Der Vaart and Wellner [72], Theorem 2.7.5). *The class $\mathcal{F}$ of monotone functions $f : \mathbb{R} \to [0, 1]$ satisfies*

$$\log N_{[\,]}(\epsilon, \|.\|_{L^2(Q)}, \mathcal{F}) \leq K\left(\frac{1}{\epsilon}\right),$$

for every probability measure Q, every $p \geq 1$, and a constant K that depends only on p.

Theorem 3.6.5 (Van Der Vaart and Wellner [72], Theorem 3.4.2). *Let $\mathcal{F}$ be class of* *measurable functions such that* $Pf^2 < \delta^2$ *and* $\|f\|_\infty < M$ *for every* f *in* $\mathcal{F}$. *Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} |\sqrt{N}(\hat{P} - P)f| \leq \tilde{J}_{[\,]}(\delta, \|.\|_{L^2(P)}, \mathcal{F}) \left(1 + \frac{\tilde{J}_{[\,]}(\delta, \|.\|_{L^2(P)}, \mathcal{F})}{\delta^2 \sqrt{N}} M \right),$$

where $\tilde{J}_{[\,]}(\delta, \|.\|_{L^2(P)}, \mathcal{F}) = \int_0^\delta \sqrt{1 + \log N_{[\,]}(\epsilon, \|.\|_{L^2(P)}, \mathcal{F})}\, \mathrm{d}\epsilon.$

Proof of Theorem 3.2.8. Recall that, from Lemma 3.2.7, $\hat{T}_N$ is the minimizer of the following criterion within the function class $\mathcal{T}$:

$$M_N(T) := \frac{1}{2N} \sum_{i=1}^{N} d^2_{\mathcal{W}}(T \# \mu_i, \nu_i).$$

And the “true” optimal map T_0 is the minimizer of the following criterion function,

$$M(T) := \frac{1}{2} \int d^2_{\mathcal{W}}(T \# \mu, \nu)\, \mathrm{d}P(\mu, \nu).$$

First we obtain an adequate upper bound for the bracketing number of the class of functions indexed by T of the form:

$$\mathcal{F}_u := \{ f_T(\mu, \nu) = d^2_{\mathcal{W}}(T \# \mu, \nu) - d^2_{\mathcal{W}}(T_0 \# \mu, \nu), \text{ s.t. } T \in \mathcal{T} \text{ and } \|T - T_0\|_{L^2(Q)} \leq u \},$$

where the domain of each function $f_T \in \mathcal{F}_u$ is $\mathcal{W}_2(\Omega) \times \mathcal{W}_2(\Omega)$. Let

$$\log N_{[\,]}(\epsilon, \|.\|_{L^2(P)}, \mathcal{F}_u)$$

denote the bracketing entropy of the function class $\mathcal{F}_u$. One can directly control this bracketing entropy by the bracketing entropy of the class of optimal maps $\mathcal{T}$ (denote by $\log N_{[\,]}(\epsilon, \|.\|_{L^2(Q)}, \mathcal{T})$). First note that $\|T_\epsilon - \mathrm{Id}\|_{L^2(\mu)}$ is bounded for any μ in the support of P_M and therefore for any $T_1, T_2 \in \mathcal{F}_u$:

$$\begin{aligned} \left\| f_{T_1} - f_{T_2} \right\|^2_{L^2(P)} &= \int |d^2_{\mathcal{W}}(T_1 \# \mu, \nu) - d^2_{\mathcal{W}}(T_2 \# \mu, \nu)|^2 \, \mathrm{d}P(\mu, \nu) \\ &\leq C \int \|T_1 - T_2\|^2_{L^2(\mu)} \, \mathrm{d}P_M(\mu) \\ &\leq C \, \|T_1 - T_2\|^2_{L^2(Q)} . \end{aligned} \tag{3.17}$$

From Lemma 3.6.4, we infer that $\log N_{[\,]}(\epsilon, \|.\|_{L^2(Q)}, \mathcal{T}) \leq K \left(\frac{1}{\epsilon}\right)$, as optimal maps are monotone functions. Therefore, by inequality (3.17) we get:

$$\log N_{[\,]}(\epsilon, \|.\|_{L^2(P)}, \mathcal{F}_u) \leq \log N_{[\,]}(\epsilon, \|.\|_{L^2(Q)}, \mathcal{T}) \lesssim \left(\frac{1}{\epsilon} \right).$$

The inequality (3.17) also shows that

$$Pf_T^2 \leq P\,\|T - T_0\|^2_{L^2(\mu)} = \|T - T_0\|^2_{L^2(Q)} \leq u^2,$$

for all $f_T \in \mathcal{F}_u$.

To get the rate of convergence, we first show that $M(T)$ has quadratic growth around its minimizer. For any map T, we can write $T = T_0 + \eta$, where $\eta = T - T_0$. Thus the equation (3.14), with $\epsilon = 1$ and also the fact $D_\eta M(T_0) = 0$ yields

$$\begin{aligned} M(T) - M(T_0) &= \frac{1}{2}\,\|\eta\|^2_{L^2(Q)} \\ &= \frac{1}{2}\,\|T - T_0\|^2_{L^2(Q)}\,. \end{aligned}$$

Next, we find a function $\phi_N(\delta)$ such that

$$\begin{aligned} \mathbb{E} \sup_{\|T-T_0\|_{L^2(Q)} \leq \delta, T \in \mathcal{T}} \sqrt{N}\Big|(M_N - M)(T) - (M_N - M)(T_0)\Big| &= \mathbb{E} \sup_{f \in F_\delta} \sqrt{N}|(P_N - P)f| \\ &\leq \phi_N(\delta). \end{aligned}$$

Since the functions in $\mathcal{F}_\delta$ are uniformly bounded and $Pf^2 \leq \delta^2$ for all $f \in \mathcal{F}_\delta$, the conditions of Theorem 3.6.5 are satisfied and we can choose

$$\phi_N(\delta) = \tilde{J}_{[\,]}(\delta, \|.\|_{L^2(P)}, \mathcal{F}_\delta)\left(1 + \frac{\tilde{J}_{[\,]}(\delta, \|.\|_{L^2(P)}, \mathcal{F}_\delta)}{\delta^2\sqrt{N}}\bar{c}\right),$$

where the constant $\bar{c}$ is a uniform upper bound for the functions in class $\mathcal{F}_\delta$. Since we noted that $\log N_{[\,]}(\epsilon, \|.\|_{L^2(P)}, \mathcal{F}_u) \lesssim \epsilon^{-1}$ for any $u > 0$, we can show

$$\tilde{J}_{[\,]}(\delta, \|.\|_{L^2(P)}, \mathcal{F}) \leq \int_0^\delta 1 + \sqrt{\log N_{[\,]}(\epsilon, \|.\|_{L^2(P)}, \mathcal{F}_\delta)}\,\mathrm{d}\epsilon \lesssim \sqrt{\delta}.$$

The above inequality and the required condition $\phi_N(\delta) \leq \delta_N^2\sqrt{N}$ gives the bound $\delta_N = N^{-1/3}$.

□

To establish the rate of convergence under imperfect observation we will make use of the following Lemma.

Lemma 3.6.6. *Let μ_n be a sequence of measures converging in Wasserstein distance to a measure μ at a rate of convergence r_n^{-1} and let $T \in \mathcal{T}$. Then $d^2_{\mathcal{W}}(T\#\mu_n, T\#\mu) \lesssim r_n^{-2}$.*

Proof. For simplicity and without loss of generality assume that $d^2_{\mathcal{W}}(\mu_n, \mu) = r_n^{-2}$

exactly. If S_n is the optimal map from μ_n to μ, then

$$\int \left|S_n(x) - x\right|^2 d\mu_n \leq r_n^{-2}.$$

Since T is differentiable almost everywhere, and satisfies $|T'(x)| \leq B$ for almost all $x \in \Omega$, then T is Lipschitz continuous with Lipschitz constant at most B. Thus

$$\begin{aligned} d_{\mathcal{W}}^2(T\#\mu_n, T\#\mu) &\leq \int \left|T\{S_n(x)\} - T(x)\right|^2 d\mu_n \\ &\leq B^2 \int \left|S_n(x) - x\right|^2 d\mu_n \\ &\lesssim r_n^{-2} \end{aligned} \tag{3.18}$$

□

Proof of Theorem 3.2.10. Define $M_{n,N}(T) := \frac{1}{N}\sum_{i=1}^N d_{\mathcal{W}}^2(T\#\mu_i^n, \nu_i^n)$. For any map $T \in \mathcal{T}$,

$$\begin{aligned} \mathbb{E}|M_{n,N}(T) - M_N(T)| &= \mathbb{E}\left|\frac{1}{N}\sum_{i=1}^N d_{\mathcal{W}}^2(T\#\mu_i^n, \nu_i^n) - \frac{1}{N}\sum_{i=1}^N d_{\mathcal{W}}^2(T\#\mu_i, \nu_i)\right| \\ &\leq \mathbb{E}\left|d_{\mathcal{W}}^2(T\#\mu_i^n, \nu_i^n) - d_{\mathcal{W}}^2(T\#\mu_i, \nu_i)\right| \\ &\leq 2C\mathbb{E}\left|d_{\mathcal{W}}(T\#\mu_i^n, \nu_i^n) - d_{\mathcal{W}}(T\#\mu_i^n, \nu_i)\right| \\ &\quad + \mathbb{E}\left|d_{\mathcal{W}}(T\#\mu_i^n, \nu_i) - d_{\mathcal{W}}(T\#\mu_i, \nu_i)\right| \\ &\leq 2C\mathbb{E}d_{\mathcal{W}}(\nu_i^n, \nu_i) + \mathbb{E}d_{\mathcal{W}}(T\#\mu_i^n, T\#\mu_i) \\ &\lesssim r_n^{-1} \qquad \text{(by Lemma 3.6.6),} \end{aligned} \tag{3.19}$$

where $C = \sup_{\mu,\nu} d_{\mathcal{W}}(\mu, \nu)$, and r_n^{-1} is the rate of estimation of an absolutely continuous measure from n samples. Thus the above inequality shows the uniform convergence of $M_{n,N}$ to M_N (at a rate independent of N). Also, since $\hat{T}_N$ is the unique minimizer of M_N, according to Theorem 3.6.2, $\hat{T}_{n,N}$ is a consistent estimator for $\hat{T}_N$, when N is fixed.

Now assuming N is fixed, we again use Theorem 3.6.3 for functionals $M_{n,N}$ and M_N. Since both functionals are differentiable, the first condition of the Theorem (quadratic growth) is satisfied. For the second condition we need to find an upper bound for

$$\mathbb{E} \sup_{\|T - \hat{T}_N\|_{L^2(Q)} < \delta} \sqrt{n}\left|(M_{n,N} - M_N)(T) - (M_{n,N} - M_N)(\hat{T}_N)\right| = \phi_n(\delta). \tag{3.20}$$

According to (3.19), we have $\phi_n(\delta_n) \lesssim r_n^{-1}\sqrt{n}$. We also need $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$, thus

$\delta_n^2 \sim r_n^{-1}$. Therefore

$$\left\|\hat{T}_{n,N} - \hat{T}_N\right\|_{L^2(Q)} = \delta_n = r_n^{-1/2},$$

and

$$\left\|\hat{T}_{n,N} - T_0\right\|_{L^2(Q)} \leq \left\|\hat{T}_{n,N} - \hat{T}_N\right\|_{L^2(Q)} + \left\|\hat{T}_N - T_0\right\|_{L^2(Q)},$$

thus

$$\left\|\hat{T}_{n,N} - T_0\right\|_{L^2(Q)} \lesssim r_n^{-1/2} + N^{-1/3}.$$

□

Before proving Theorem 3.2.12, we restate Fano's method in the format that we use to prove the theorem 3.2.12, which is taken from [75].

Given a class of distributions $\mathcal{P}$, we let θ denote a functional on the space $\mathcal{P}$ that is a mapping from a distribution $\mathbb{P}$ to a parameter $\theta(\mathbb{P})$ taking values on some space Ω. Let $\rho : \Omega \times \Omega \longrightarrow [0, \infty)$ be a given metric. Also let $\Phi : [0, \infty] \longrightarrow [0, \infty)$ be an increasing function. Then we define the ρ-minimax risk for the estimation of θ as:

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) := \inf_{\hat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\Phi\left(\rho(\hat{\theta}, \theta(\mathbb{P}))\right)\right].$$

The following theorem (proposition 15.2 [75]) gives a lower bound on the minimax error.

Theorem 3.6.7. *(Generalized Fano's inequality) Let $\{\theta^1, \cdots, \theta^M\}$ be a 2δ-separated set in the ρ semi-metric on $\Theta(\mathcal{P})$, and suppose that J is uniformly distributed over the index set $\{1, \cdots, M\}$, and $(Z|J = j) \sim P_{\theta^j}$. Then for any increasing function $\Phi : [0, \infty] \longrightarrow [0, \infty)$, the minimax risk is lower bounded as*

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \Phi(\delta)\left\{1 - \frac{I(Z; J) + \log 2}{\log M}\right\}, \tag{3.21}$$

where $I(Z; J)$ is the mutual information between Z and J.

In order to find an upper-bound for the mutual information in the inequality 3.21, we use lemma 15.5 of [75] which we restate here:

Lemma 3.6.8. *(Yang-Barron method) Let $N_{KL}(\epsilon; \mathcal{P})$ denote the ϵ-covering number of $\mathcal{P}$ in the square-root KL divergence. Then the mutual information is upper bounded as*

$$I(Z; J) \leq \inf_{\epsilon > 0}\{\epsilon^2 + \log N_{KL}(\epsilon; \mathcal{P})\}. \tag{3.22}$$

Proof of Theorem 3.2.12. The idea will be to imbed the setting of isotonic regression within the setting of the current estimation problem. We will then use Fano's method

(Theorem 15.2. and Lemma 15.5 from [75]) as restated above for our purposes, following the usual path for establishing the isotonic rate.

First and without loss of generality, we assume that $\Omega = [0, 1]$. Suppose P_M is supported on the set of measures $S := \{\delta_x \text{ s.t. } x \in [0, 1]\}$, where δ_x is a point mass at $x \in [0, 1]$. Suppose also that $\mathrm{d}P_M(\delta_x) = \mathrm{d}p(x)$, where p is a distribution on $[0, 1]$ with bounded density. Note that in this setting, we can see that the distribution p is equal to Q (defined in the first section).

Further let $\sigma^2 > 0$ and suppose that given $x \in [0, 1]$ the marginal distribution of the real-valued random variable $T_\epsilon(x)$ is centered Gaussian with variance σ^2, i.e.

$$T_\epsilon(x) \sim N(x, \sigma^2), \qquad \forall x \in [0, 1].$$

To see that such family of random maps exists, take each random map to be $T_\epsilon(x) = I(x) + \sigma Z$ where $I(x) = x$ is the identity map and $Z \sim N(0, 1)$ is a standard Gaussian. By construction such maps are increasing and their marginal distribution at any fixed point is a Gaussian.

In the setting we have constructed, both predictor and response distributions are supported on a single point (Dirac measures). We can thus conveniently represent them by identifying them with their singleton support. More precisely we represent each pair of predictor/response distributions (μ_i, ν_i) via their support (X_i, Y_i). Therefore, we assume that we observe the collection $\{(X_i, Y_i)\}_{i=1}^N$, where $X_i \in [0, 1]$ and are i.i.d. samples from distribution p, and Y_i are i.i.d. samples from the distribution $N(T_0(X_i), \sigma^2)$, i.e. the marginal distribution of Y_i given $X_i = x$ is $N(T_0(x), \sigma^2)$. The estimation of the true map T_0 in this setting is now equivalent to the estimation of an isotonic regression map from the sample pairs $\{(X_i, Y_i)\}_{i=1}^N$.

Note that any map T, induces a probability distribution $\mathbb{P}_T(X, Y) \in \mathcal{P}(\mathbb{R}^2)$, and denote by $\mathbb{P}_T^N$ the distribution induced on $\{(X_i, Y_i)\}_{i=1}^N$. Let $\mathcal{P}_T$ denote the following family of distributions on $\mathbb{R}^{2N}$ of the form

$$\mathcal{P}_T = \{\mathbb{P}_T^N, s.t.\ T \in \mathcal{T}\}.$$

We want to find an upper-bound for the ϵ-covering number of $\mathcal{P}_T$ in the square root KL divergence [75], denoted by $N_{KL}(\epsilon; \mathcal{P}_T)$. Since for all $T \in \mathcal{T}$ we can write $\mathbb{P}(X, Y) = p(X)\mathbb{P}_T(Y|X)$, we only need to control the ϵ-covering number of the conditional distributions $\mathbb{P}(Y|X)$.

The idea is to show $N_{KL}(\epsilon; \mathcal{P}_T)$ can be upper-bounded using the bracketing entropy of the set $\mathcal{T}$, denoted by $\log N_{[]}(\epsilon, \|.\|_{L^2(Q)}, \mathcal{T})$. First note that according to [72, Thm 2.7.5], we have the following upper-bound for the bracketing entropy of the set $\mathcal{T}$:

$$\log N_{[]}(\epsilon, \|.\|_{L^2(Q)}, \mathcal{T}) \leq K\left(\frac{1}{\epsilon}\right).$$

Country List Figure (3.8a)	
Country Name	Country Code
Estonia	EST
Slovakia	SVK
Bulgaria	BGR
Hungary	HUN
Czechia	CZE
Lithuania	LTU
East Germany	DEUTE
Latvia	LVA
Belarus	BLR
Ukraine	UKR
Israel	ISR
Slovenia	SVN
Poland	POL
Spain	ESP
Italy	ITA
Portugal	PRT
Russia	RUS
Japan	JPN
Taiwan	TWN
Greece	GRC

Country List Figure (3.8b)	
Country Name	Country Code
Australia	AUS
West Germany	DEUTW
Austria	AUT
Netherlands	NLD
Iceland	ISL
Ireland	IRL
Belgium	BEL
France	FRATNP
Finland	FIN
New Zealand	NZL-NP
Switzerland	CHE
Sweden	SWE
Norway	NOR
U.K.	GBR-NP
U.S.A.	USA
Denmark	DNK
Luxemburg	LUX

Table 3.1: Country abbreviations used in Figures 3.8a and 3.8b

Since $\mathbb{P}(Y|X)$ is a Gaussian distribution, for any two maps T_1 and T_2, we can control

$$KL(\mathbb{P}^N_{T_1}||\mathbb{P}^N_{T_2}) \leq \frac{N}{2\sigma^2}\left\|T_1 - T_2\right\|^2_{L^2(p)} = \frac{N}{2\sigma^2}\left\|T_1 - T_2\right\|^2_{L^2(Q)},$$

so we conclude that $\log N_{KL}(\epsilon; \mathcal{P}_T)$ is no larger than $\log N_{[\,]}(\mathcal{T}, \frac{\sigma\sqrt{2}}{\sqrt{N}}\epsilon, \|.\|_{L^2(Q)}) \lesssim \frac{\sqrt{N}}{\sigma\epsilon}$.

Now we can take any δ-packing on the set $\mathcal{T}$. We know $\log M(\mathcal{T}, \delta, \|.\|_{L^2(Q)}) \asymp \frac{1}{\delta}$, where $M(\mathcal{T}, \delta, \|.\|_{L^2(Q)})$ is the δ-packing number of the set $\mathcal{T}$. Take $\Phi(\delta) = \delta^2$, then using Theorem 3.6.7 and Lemma 3.6.8 (Appendix) we can write

$$\mathfrak{M}(\theta(\mathcal{P}); \|.\|_{L^2(Q)}) \geq \frac{\delta}{2}\left(1 - \frac{\log N_{[\,]}(\mathcal{T}, \frac{\sigma\sqrt{2}}{\sqrt{N}}\epsilon, \|.\|_{L^2(Q)}) + \epsilon^2 + \log 2}{\log M(\mathcal{T}, \delta, \|.\|_{L^2(Q)})}\right).$$

Finally choosing $\epsilon_N^{-2} \asymp \delta_N \asymp N^{-1/3}$ yields the desired rate.

□

Chapter 4

Distribution-on-Distribution Regression in Higher Dimension

The work in this chapter was done in collaboration with my supervisor Victor Panaretos and is publicly available as a preprint [27]. This chapter follows the priprint with slight changes.

4.1 Introduction

So far, distributional regression in the Wasserstein space has been confined to measures on the real line due to the geometrical, computational, and statistical complexities associated with higher dimensions. These complexities include the lack of closed-form solutions for optimal transport maps, the positive curvature of the space, and the curse of dimensionality. Despite these challenges, at least in principle, both the Wasserstein regression framework developed by Chen et al. [15], which uses the tangent structure of the Wasserstein space to develop a tangential Hilbert-type linear model, and the shape-constraint approach introduced in Chapter 3, have the potential to be studied in higher dimensions.

In this Chapter, we consider the distributional regression problem in higher dimensions, focussing on the shape-contraint approach, in light of its leaner technical assumptions and greater statistical interpretability. We incorporate and adapt concepts and techniques from prior studies in higher dimensional statistical optimal transport. Specifically, we employ strategies that address the curse of dimensionality in estimating optimal transport maps by imposing regularity conditions borrowed from [31, 35]. And, we draw from previous work on imposing geometric and shape constraints to derive the rate of convergence of Fréchet mean [2, 18]. We thus establish identifiability of the monotone map model in higher dimensions, introduce a regularised Fréchet-least-squares estimator, and establish its consistency and rate of convergence.

4.2 The Model and its Identifiability

Let Ω be a subset of $\mathbb{R}^d$ that is compact, convex, and with a non-empty interior. Let (μ, ν) be a pair of random elements in $\mathcal{W}_2(\Omega) \times \mathcal{W}_2(\Omega)$ with a joint distribution denoted by P. Similar to Chapter 3, we define a regression operator, $\Gamma : \mathcal{W}_2(\Omega) \to \mathcal{W}_2(\Omega)$, characterized as the minimizer of the conditional Fréchet functional as a function of μ:

$$\operatorname*{argmin}_b \int_{\mathcal{W}_2(\Omega)} d^2_{\mathcal{W}}(b, \nu)\, \mathrm{d}P(\nu \mid \mu) = \Gamma(\mu).$$

It is implicitly assumed that the Fréchet mean of the conditional probability distribution $P(\cdot \mid \mu)$ of ν given μ is unique for any μ. This uniqueness can be ensured through suitable regularity assumptions on the pair (μ, ν). A regression model consists in positing a specific structure for $\Gamma(\mu)$. The identifiability of such a model will typically require additional assumptions on the pair (μ, ν).

We consider generalizing the regression model (3.1), which is defined for distributions on $\mathbb{R}$: We assume that $\Gamma(\mu) = T_0\#\mu$, where T_0 is an optimal map, and the response distribution ν further deviates from its conditional Fréchet mean $T_0\#\mu$ by means of a random optimal map perturbation (with Bochner mean identity).

More explicitly, we consider the regression model

$$\nu_i = T_{\epsilon_i}\#(T_0\#\mu_i), \quad \{\mu_i, \nu_i\}_{i=1}^N, \tag{4.1}$$

where $\{\mu_i, \nu_i\}_{i=1}^N$ are probability distributions on Ω, $T_0 : \Omega \to \Omega$ is an unknown transport map and the $\{T_{\epsilon_i}\}_{i=1}^N$ are independent and identically distributed random optimal transport maps with $\mathbb{E}(T_{\epsilon_i}) = \mathrm{id}$, representing the noise in our model. The regression task is to estimate the unknown map T_0 from the observations $\{\mu_i, \nu_i\}_{i=1}^N$.

Let P denote the joint distribution on $\mathcal{W}_2(\Omega) \times \mathcal{W}_2(\Omega)$ induced by model (4.1). We denote by P_M the induced marginal distribution of μ and will be assuming that it is supported on regular measures:

Assumption 4.2.1. *Let μ be a measure in the support of P_M. Then μ is absolutely continuous with respect to the Lebesgue measure.*

The linear average (Bochner mean) of P_M, will be denoted as

$$Q(A) = \int_{\mathcal{W}_2(\Omega)} \mu(A)\, \mathrm{d}P_M(\mu).$$

Recall that a twice differentiable function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex and L-smooth if at any point $x \in \mathbb{R}^d$, its Hessian matrix is positive definite with its smallest eigenvalue being no less than α and no greater than L:

$$\alpha I \preceq \nabla^2 \varphi(x) \preceq LI, \quad \forall x \in \mathbb{R}^d,$$

where $A \preceq B$ signifies that the difference $B - A$ between two matrices is positive semi-definite.

Consider the following sets of potential functions:

$$\Phi_\alpha := \{\varphi : \text{ such that } \alpha I \preceq \nabla^2 \varphi(x) \preceq LI\} \quad \text{for } L > \alpha \geq 0,$$

$$\Phi := \cup_{\alpha > 0} \Phi_\alpha.$$

Throughout the chapter, we suppose L is fixed and known. Next, we define the following set of optimal maps:

$$\mathcal{T} := \{T : \Omega \to \Omega : T = \nabla\varphi, \text{ for some } \varphi \in \Phi\}.$$

We will require the following regularity of the optimal maps involved in the model, in order to ensure idenfitiability:

Assumption 4.2.2. *The map T_0 belongs to the class $\mathcal{T}$.*

Assumption 4.2.3. *The maps T_{ϵ_i} are i.i.d random elements in $\mathcal{T}$, with $\mathbb{E}(T_{\epsilon_i}) = \mathrm{id}$.*

With these assumptions and definitions in place, we can now establish identifiability:

Theorem 4.2.4. *(Identifiability) Assume that the law P induced by model* (4.1) *satisfies Assumptions 4.2.1, 4.2.2, and 4.2.3. Then, the regressor operator $\Gamma(\mu) = T_0 \# \mu$ in the model* (4.1) *is identifiable over the class of maps $T \in \mathcal{T}$, up to Q-null sets. Specifically, for any map $T \in \mathcal{T}$ such that $\|T - T_0\|_{L^2(Q)} > 0$, it holds that*

$$M(T) > M(T_0),$$

where for any $T \in \mathcal{T}$,

$$M(T) := \frac{1}{2} \int_{\mathcal{W}_2(\Omega) \times \mathcal{W}_2(\Omega)} d^2_{\mathcal{W}}(T\#\mu, \nu) \, \mathrm{d}P(\mu, \nu). \tag{4.2}$$

Remark 4.2.5 (Identifiability Q-almost everywhere)**.** *Theorem 4.2.4 establishes that T_0 is identifiable, up to Q-null sets, and this holds true under minimal conditions on the input measures μ. The intuition behind this form of identifiability is similar to the one provided in Remark 3.2.4 for the case of one dimension: If the measure Q is supported on a subset $\Omega_0 \subset \Omega$, the map T_0 cannot be identified on $\Omega \setminus \Omega_0$. But, if the measure Q is mutually absolutely continuous with the Lebesgue measure, then identifiability is also true almost everywhere on Ω with respect to Lebesgue measure.*

This equivalence can be achieved by enforcing additional conditions on the law of random covariate measures μ. One straightforward condition is to require that the input measures, μ, have a bounded density from below with positive probability. However, this implies that the support of μ equals Ω with positive probability, which may be restrictive since we want our model to include scenarios where none of the covariate measures have the full support on Ω.

A weaker condition to ensure the equivalence of Q with the Lebesgue measure is to assume the existence of a cover $\{E_m\}_{m\geq 1}$ of Ω such that the probability $P_M\{E_m \subseteq \mathrm{supp}(f_\mu)\} > 0$ is greater than zero for all m. This condition suggests that different covariate measures can provide information about T_0 on different subsets of Ω, but collectively, they must provide information about all of Ω. Consider an example where Ω is the d-dimensional unit cube and μ is defined as the normalized Lebesgue measure on $S = \big([U_1, U_1 + 1/3] \mod 1\big) \times \cdots \times \big([U_d, U_d + 1/3] \mod 1\big)$. Here, $\{U_i\}_{i=1}^d$ are independent uniform random variables on $[0, 1]$. In this scenario, none of the μ realizations are supported on Ω, yet the "cover condition" is met. This remark directly extends the analogous observations in the one-dimensional case.

Figure 4.1 illustrates the output of Model (4.1) when $d = 2$. In the first plot of each column, blue dots represent samples from a covariate distribution μ. The black dots are sampled from the (conditional Fréchet mean) distribution $T_0\#\mu$. The flow curves depict the effect of $T_0 - \mathrm{id}$, with the colour indicating its magnitude. Then, we examine 4 different random maps T_ϵ. In the next 4 plots, we observe samples from the response distribution $\nu = T_\epsilon\#T_0\#\mu$, represented by red dots. In each plot, the flow curves represent $T_\epsilon - \mathrm{id}$.

4.3 Statistical Analysis

To obtain a consistent estimator and derive its rate of convergence, we use empirical process theory. This requires us to make additional regularity assumptions on the model. We consider the case where the true map T_0 satisfies a Hölder condition (defined below).

Definition 4.3.1. *(Hölder Space) For any vector $\boldsymbol{k} \in \mathbb{N}^d$ with coordinates $(k_1, \cdots, k_d)$ write $|\boldsymbol{k}| = \sum_{i=1}^d k_i$ and define the differential operator*

$$D^{\boldsymbol{k}} = \frac{\partial^{|\boldsymbol{k}|}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}.$$

For any real number $\beta > 0$, we define the Hölder norm of smoothness β of a $\lfloor \beta \rfloor$-

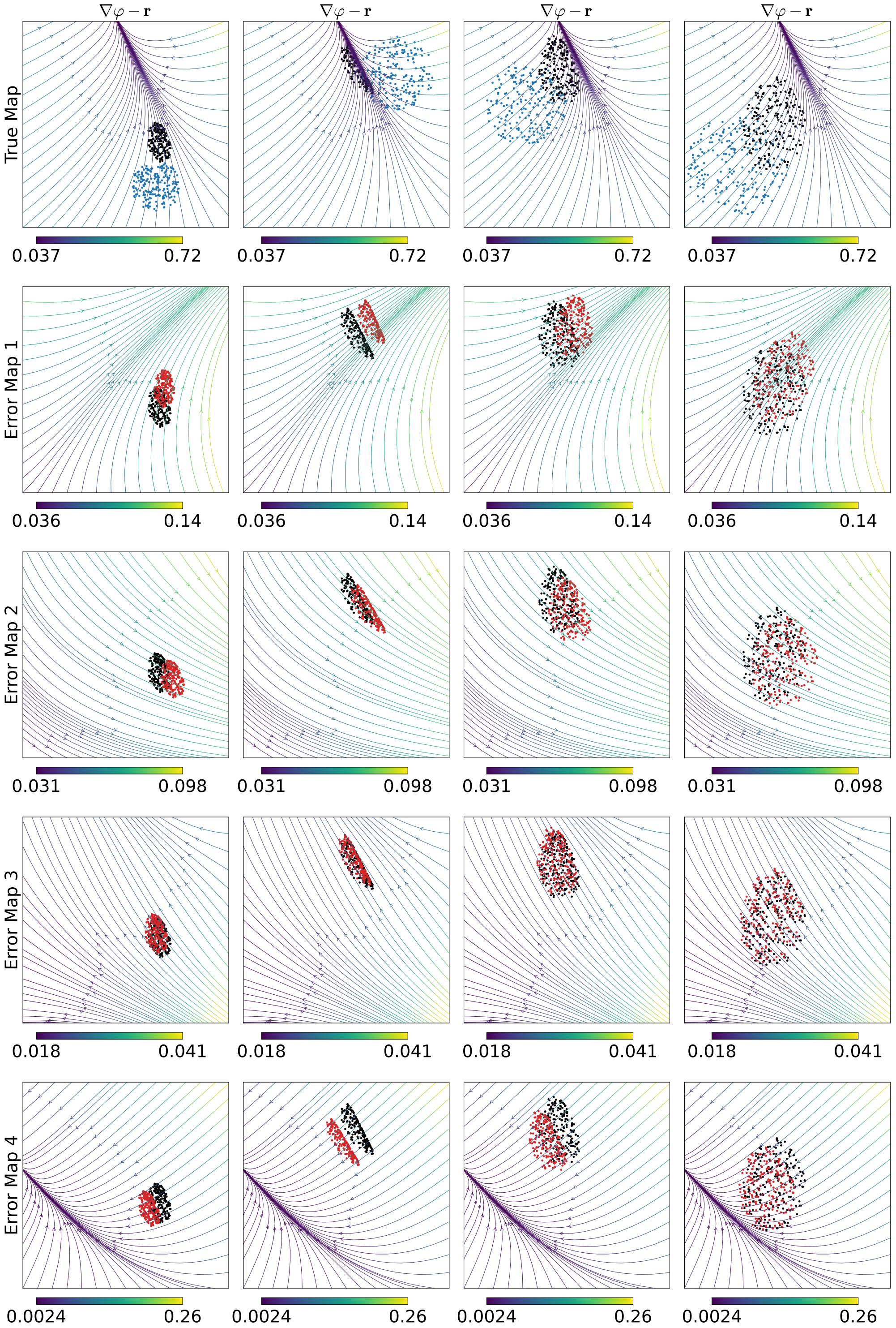


Figure 4.1: Illustration of Model (4.1) for $d = 2$, showing samples from μ (blue), $T_0\#\mu$ (black), and $\nu = T_\epsilon\#T_0\#\mu$ (red) for four different realisations of the error map, along with corresponding flow curves.

times differentiable function $f : \Omega \to \mathbb{R}$ *as*

$$\|f\|_{C^\beta} := \max_{|\boldsymbol{k}|\le\lfloor\beta\rfloor} \sup_x |D^{\boldsymbol{k}} f(x)| + \max_{|\boldsymbol{k}|=\lfloor\beta\rfloor} \sup_{x\neq y} \frac{|D^{\boldsymbol{k}} f(x) - D^{\boldsymbol{k}} f(y)|}{\|x-y\|^{\beta-\lfloor\beta\rfloor}}. \tag{4.3}$$

The Hölder ball of smoothness β *and radius* $L > 0$*, denoted by* $C^\beta_R(\Omega)$*, is then defined as the class of* $\lfloor\beta\rfloor$*-times continuously differentiable functions with Hölder norm bounded by the radius L:*

$$C^\beta_R(\Omega) := \{f \in C^{\lfloor\beta\rfloor}(\Omega) : \|f\|_{C^\beta} \le R\}.$$

We might occasionally drop the argument Ω when the underlying space can be understood from the context. Now consider the following sets of maps:

$$\mathcal{T}_{\beta,\gamma,R} := \{T : T = \nabla\varphi, \varphi \in \Phi \cap \overline{C^{\beta+\gamma}_R}^{\|\cdot\|_{C^\beta}}\}$$

$$\mathcal{T}_{\beta,R} := \{T : T = \nabla\varphi, \varphi \in \Phi_0 \cap C^\beta_R\}.$$

It holds that $\mathcal{T}_{\beta,\gamma,R} \subset \mathcal{T}_{\beta,3R}$ (by way of Lemma 4.5.4 in Section 4.5). Our assumption now is:

Assumption 4.3.2. *The map* T_0 *belongs to the set* $\mathcal{T}_{\beta,\gamma,R}$*, where* β*,* γ*, and* R *are positive constants and* $\lfloor\beta+\gamma\rfloor = \lfloor\beta\rfloor$*.*

Remark 4.3.3. *The* β*-Hölder function classes appear frequently in optimization and such regularity assumption is standard in non-parametric regression. In statistical optimal transport theory (see e.g. [31, 35]) similar assumptions are imposed to estimate optimal transport maps when observing samples from distributions.*

Our objective is to obtain an estimator, denoted by $\hat{T}_{N,(\beta,\gamma,R)}$, of the unknown map T_0. To do so, we define $\hat{T}_{N,(\beta,\gamma,R)}$ as the constrained minimizer of the sample version of the functional M,

$$\hat{T}_{N,(\beta,\gamma,R)} := \operatorname*{argmin}_{T\in\mathcal{T}_{\beta,\gamma,R}} M_N(T), \qquad M_N(T) := \frac{1}{2N}\sum_{i=1}^N d^2_{\mathcal{W}}(T\#\mu_i, \nu_i), \tag{4.4}$$

Here, (μ_i, ν_i) are independent samples drawn from P for $i = 1, \dots, N$. The constraint amounts to the requirement that $T \in \mathcal{T}_{\beta,\gamma,R}$. The minimizer might not be unique. We can take any of the minimizers of (4.4) as $\hat{T}_{N,(\beta,\gamma,R)}$. But it will exist in $\mathcal{T}_{\beta,3R}$, provided $\beta > 2$:

Theorem 4.3.4. *The minimization problem* (4.4) *has a solution in* $\mathcal{T}_{\beta,3R}$ *when* $\beta > 2$*.*

Remark 4.3.5. *The functional* $d^2_{\mathcal{W}}(T\#\mu, \nu)$ *is not necessarily convex with respect to* T*, meaning that we cannot expect* $d^2_{\mathcal{W}}([aT_1 + (1-a)T_2]\#\mu, \nu) \le a d^2_{\mathcal{W}}(T_1\#\mu, \nu) + (1-$

$a)d^2_{\mathcal{W}}(T_2\#\mu, \nu)$. The reason is that the Wasserstein distance is not necessarily convex with respect to geodesics. And, if we set T_2 as the identity function, $[aT_1 + (1-a)Id]\#\mu$ represent the geodesic between μ and $T_1\#\mu$. Under this scenario, selecting T_1, μ and ν as per example 9.1.5 of [3] results in the violation of the inequality.

Nevertheless, when ν is close to μ we can show convexity. Denote by $T_{\mu\to\nu}$ the optimal transport from μ to ν. Suppose $T_{\mu\to\nu}$ is gradient of a λ-strongly convex function. Then based on Theorem 6 of Manole et al. [46] and inequality (2.4), *we have:*

$$\begin{aligned} d^2_{\mathcal{W}}([aT_1 + (1-a)T_2]\#\mu, \nu) &\leq \left\| aT_1 + (1-a)T_2 - T_{\mu\to\nu} \right\|^2_{L^2(\mu)} \\ &\leq \lambda^2 d^2_{\mathcal{W}}([aT_1 + (1-a)T_2]\#\mu, \nu), \end{aligned} \tag{4.5}$$

and similarly

$$\begin{aligned} ad^2_{\mathcal{W}}(T_1\#\mu, \nu) + (1-a)d^2_{\mathcal{W}}(T_2\#\mu, \nu) &\leq a \left\| T_1 - T_{\mu\to\nu} \right\|^2_{L^2(\mu)} + (1-a) \left\| T_2 - T_{\mu\to\nu} \right\|^2_{L^2(\mu)} \\ &\leq a\lambda^2 d^2_{\mathcal{W}}(T_1\#\mu, \nu) + (1-a)\lambda^2 d^2_{\mathcal{W}}(T_2\#\mu, \nu). \end{aligned} \tag{4.6}$$

Also, by strict convexity of squared norm we have

$$\left\| aT_1 + (1-a)T_2 - T_{\mu\to\nu} \right\|^2_{L^2(\mu)} < a \left\| T_1 - T_{\mu\to\nu} \right\|^2_{L^2(\mu)} + (1-a) \left\| T_2 - T_{\mu\to\nu} \right\|^2_{L^2(\mu)}. \tag{4.7}$$

As we observe from the above equation, the middle value in inequality (4.5) *is strictly smaller than the middle value in inequality* (4.6)*. Moreover, if λ approaches 1, both inequalities' right-hand sides converge to their respective left-hand sides. Consequently, for small values of λ, the left-hand side of* (4.5) *is majorised by the left-hand side of* (4.6)*, which establishes convexity. It is worth mentioning that when λ approaches 1, $T_{\mu\to\nu}$ converges to the identity map, and ν converges to μ, providing a perspective on why convexity occurs when ν is close to μ.*

To evaluate the estimator's quality, we will use the *Fréchet mean squared error*, which is the natural risk function in this context:

$$R(T) := \mathbb{E}_{\mu\sim P_M} d^2_{\mathcal{W}}(T_0\#\mu, T\#\mu) = \int d^2_{\mathcal{W}}(T_0\#\mu, T\#\mu)\, \mathrm{d}P_M(\mu).$$

The value on the right-hand side defines a semi-metric on the set of maps $\mathcal{T}$. More specifically, for any two maps $T_1, T_2 \in \mathcal{T}$, we can define

$$\rho(T_1, T_2) = \left[\int d^2_{\mathcal{W}}(T_1\#\mu, T_2\#\mu)\, \mathrm{d}P_M(\mu) \right]^{1/2}.$$

Lemma 4.3.6 (Semi-metric property of ρ)**.** *The map $\rho(\cdot,\cdot)$ satisfies all the properties of a metric, except that there may exist pairs $T_1 \neq T_2$ such that $\rho(T_1, T_2) = 0$.*

Since $\mathcal{W}_2(\mathbb{R}^d)$ is non-negatively curved (equation (2.4)), for any two maps $T_1, T_2 \in \mathcal{T}$ we have

$$\begin{aligned}\rho^2(T_1, T_2) &= \int d^2_{\mathcal{W}}(T_1\#\mu, T_2\#\mu)\,\mathrm{d}P_M(\mu)\\ &\le \int \|T_1 - T_2\|^2_{L^2(\mu)}\,\mathrm{d}P_M(\mu)\\ &= \int\int |(T_1 - T_2)(x)|^2\,\mathrm{d}\mu(x)\,\mathrm{d}P_M(\mu)\\ &= \|T_1 - T_2\|^2_{L^2(Q)},\end{aligned} \tag{4.8}$$

with equality when $d = 1$. We use this to derive an upper bound for the rate of convergence of $\hat{T}_N$ with respect to semi-metric ρ.

Theorem 4.3.7. *(Rate of Convergence) Suppose the Assumptions 4.2.1, 4.2.2, 4.2.3 and 4.3.2 are satisfied with some $\beta > \max\{\frac{d}{2}, 2\}$ and $\gamma, R > 0$. Then*

$$N^{\frac{\beta}{2\beta+d}}\rho(\hat{T}_{N,(\beta,\gamma,R)}, T_0) = O_{\mathbb{P}}(1).$$

In particular, for $d > 4$, and depending on $\frac{d}{2} < \beta < \infty$, the rate is between $N^{-1/4}$ and $N^{-1/2}$.

Remark 4.3.8 (The case $d = 1$)**.** *In chapter 3, we showed the Fréchet least square estimator achieves a convergence rate of $N^{-1/3}$ using the $L^2(Q)$-norm, and also we demonstrated that this rate is minimax optimal. It's worth noting that, for $d = 1$, the $L^2(Q)$-norm is equivalent to the semi-metric ρ under the assumption that T_0 is a non-decreasing map. As a result, we can compare their rates with the one provided by Theorem 4.3.7. When $d = 1$, Theorem 4.3.7 suggests that the convergence rate lies between $N^{-2/5}$ and $N^{-1/2}$, depending on the assumed degree of smoothness ($2 < \beta \le \infty$). The faster convergence rate is a result of the additional smoothness assumption, and thus, there is no inconsistency between the two results.*

4.4 Differentiability

In this section, we establish an additional result about Gateaux-differentiability of the functional M and M_N. While this result does not contribute directly to the technical results in this chapter, it might be relevant for computation of the estimator.

Lemma 4.4.1. *For any $\alpha > 0$, the functionals M and M_N are Gateaux-differentiable at any maps in $\mathcal{T}$ and there exist couplings $\gamma_{\mu,\nu} \in \Gamma(\mu, \nu)$ and $\gamma_{\mu_i,\nu_i} \in \Gamma(\mu_i, \nu_i)$ such that:*

$$
\begin{aligned}
D_\eta M(T) &= \int\int < \eta(x), T(x) - y > \mathrm{d}\gamma_{\mu,\nu}(x,y)\, \mathrm{d}P(\mu,\nu), \\
d^2_{\mathcal{W}}(T\#\mu,\nu) &= \int |T(x)-y|^2\, \mathrm{d}\gamma_{\mu,\nu}(x,y)
\end{aligned}
\tag{4.9}
$$

and

$$
\begin{aligned}
D_\eta M_N(T) &= \frac{1}{N}\sum_{i=1}^{N}\int < \eta(x), T(x) - y > \mathrm{d}\gamma_{\mu_i,\nu_i}(x,y), \\
d^2_{\mathcal{W}}(T\#\mu_i,\nu_i) &= \int |T(x)-y|^2\, \mathrm{d}\gamma_{\mu_i,\nu_i}(x,y).
\end{aligned}
\tag{4.10}
$$

4.5 Proofs

In the next statement, we restate a Theorem from Ponomarev [61] that will be used to prove Lemma 4.5.2.

Theorem 4.5.1. *(Theorem 3 of Ponomarev [61]) When $f : \Omega \subset \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is continuous and almost everywhere differentiable, the following properties are equivalent:*

- $\operatorname{rank}\{f'(x)\} = d$ *for almost all* $x \in \Omega$,
- *the f-preimage of any set of measure zero is set of measure zero, i.e. if $E \subset \mathbb{R}^d$ such that $\lambda(E) = 0$ then $\lambda(f^{-1}(E)) = 0$.*

Lemma 4.5.2. *$T\#\mu$ is absolutely continuous, when μ is absolutely continuous and $T \in \mathcal{T}$.*

Proof. We need to show that if A is a measurable set with Lebesgue measure $\lambda(A) = 0$, then $T\#\mu(A) = 0$. We begin by noting that $T \in \mathcal{T}$ implies the Jacobian of T is full rank. Using Theorem 4.5.1, we conclude that $\lambda(T^{-1}(A)) = 0$. Next, recall that $T\#\mu(A) = \mu(T^{-1}(A))$. Since μ is absolutely continuous and $\lambda(T^{-1}(A)) = 0$, it follows that $\mu(T^{-1}(A)) = 0$. Therefore, we have $T\#\mu(A) = \mu(T^{-1}(A)) = 0$, as desired. □

Our argument for the identifiability of T_0 relies on a result from Chewi et al. [18], which was originally employed to establish quadratic growth of the Fréchet functional (2.6) around its minimiser. In addition to identifiability, this finding is also applied to exhibit the quadratic growth of functional M around its minimizer, crucial for determining the convergence rate of our proposed estimator. We will start by revisiting the notion of variance inequality, introduced by Ahidar-Coutrix et al. [2].

A distribution P conforms to a variance inequality with a positive constant C_{var}, if for any absolutely continuous measure $b \in \mathcal{W}_2(\mathbb{R}^d)$, the following inequality holds:

$$F(b) - F(b^*) \geq \frac{C_{var}}{2} d^2_{\mathcal{W}}(b, b^*),$$

where b^* is the minimizer of F defined by equation (2.6). Now we restate the following result from Chewi et al. [18]:

Theorem 4.5.3. *(Theorem 6 of Chewi et al. [18]) Let P be the law of a random measure in $\mathcal{W}_{2,\mathrm{ac}}(\Omega)$ with barycenter $b^* \in \mathcal{W}_{2,\mathrm{ac}}(\Omega)$. Assume there exists a measurable map $\varphi : \mathcal{W}_{2,\mathrm{ac}}(\Omega) \times \mathbb{R}^d \to \mathbb{R}$ such that for P-almost all μ, φ_μ is an optimal Kantorovich potential for b^* to μ and is $\alpha(T_{b^* \to \mu})$-strongly convex, where $T_{b^* \to \mu} = \nabla \varphi_\mu$ is the corresponding optimal map. Moreover, assume that for almost all $x \in \mathbb{R}^d$*

$$\mathbb{E}_{\mu \sim P}[\varphi_\mu(x)] = \frac{1}{2} \|x\|^2 . \tag{4.11}$$

Then, P satisfies a variance inequality for all $b \in \mathcal{W}_{2,\mathrm{ac}}(\Omega)$ with constant

$$C_{var} = \int \alpha(T_{b^* \to \mu}) \, \mathrm{d}P(\mu).$$

Proof of Theorem 4.2.4. To prove that T_0 is the unique minimizer of the population functional in $\mathcal{T}$ up to Q-null sets, fix a measure μ_0 in the support of P_M, and let ν be a random measure such that $\nu = T_\epsilon \# (T_0 \# \mu_0)$ (where we recall that $\mathbb{E}(T_\epsilon) = \mathrm{id}$). According to Lemma 4.5.2, $T_0 \# \mu_0$ is absolutely continuous, since μ_0 is absolutely continuous, and $T_0 \in \mathcal{T}_\alpha$. Similarly, we can argue that ν is also absolutely continuous because $T_\epsilon \in \mathcal{T}$. Therefore according to Proposition 3.2.7 of Panaretos and Zemel [55], for any μ_0, the induced random measure ν has a unique Fréchet mean. By Theorem 4.2.4 of Panaretos and Zemel [55] we conclude this unique Fréchet mean is $T_0 \# \mu_0$. Therefore, for any μ_0,

$$\arg\inf_{b \in \mathcal{W}_2(\Omega)} \int_{\mathcal{W}_2(\Omega)} d^2_{\mathcal{W}}(b, \nu) \, \mathrm{d}P(\nu | \mu_0) = T_0 \# \mu_0,$$

where P is the joint distribution of (μ, ν) induced by Model (4.1). Now we show that T_0 is a minimiser of M:

$$\begin{aligned}M(T) &= \int d^2_{\mathcal{W}}(T\#\mu,\nu)\,\mathrm{d}P(\mu,\nu)\\ &= \int\int d^2_{\mathcal{W}}(T\#\mu_0,\nu)\,\mathrm{d}P(\nu|\mu_0)\,\mathrm{d}P(\mu_0)\\ &\geq \int\int d^2_{\mathcal{W}}(T_0\#\mu_0,\nu)\,\mathrm{d}P(\nu|\mu_0)\,\mathrm{d}P(\mu_0)\\ &= \int d^2_{\mathcal{W}}(T_0\#\mu,\nu)\,\mathrm{d}P(\mu,\nu).\end{aligned}$$

We will show that T_0 is the unique minimizer up to $L^2(Q)$-norm. We apply Theorem 4.5.3 in the following way: Fix a measure μ_0 in the support of P_M. As we stated, $T_0\#\mu_0$ is the conditional Fréchet mean of the random measure ν given μ_0 (i.e. the baruycentre of the law $P(\nu|\mu_0)$). Define a mapping $\varphi : \mathcal{W}_{2,\mathrm{ac}}(\Omega)\times\mathbb{R}^d \to \mathbb{R}$ such that $\varphi_{\nu|\mu_0}$ is equal to the Kantorovich potential of the optimal map T_ϵ, where we abuse notation for tidiness and write $\nu|\mu_0 \equiv T_\epsilon\#T_0\#\mu_0$ (the existence of such map T_ϵ is guaranteed by the model assumptions). The mapping φ satisfies the assumptions of Theorem 4.5.3 because the Kantorovich potential of T_ϵ is indeed an optimal Kantorovich potential from $T_0\#\mu_0$ to $\nu|\mu_0$ and each T_ϵ is the gradient of an α-strongly convex function by assumption 4.2.3. Furthermore, based on the same assumption, $\mathbb{E}(T_\epsilon) = \mathrm{id}$, and as a result, equation (4.11) holds. We can deduce that the mapping $\varphi_{\nu|\mu_0}$ is $\alpha(T_\epsilon)$-strongly convex. We can also observe that the value $\mathbb{E}\alpha(T_\epsilon)$ no longer depends on μ_0. Additionally, we can deduce $\mathbb{E}\alpha(T_\epsilon) > 0$ since each map T_ϵ is strongly convex.

Collecting and combining the statements above, we can apply Theorem 4.5.3 and show that for any $T \in \mathcal{T}$:

$$\begin{aligned}M(T) - M(T_0) &= \frac{1}{2}\int\int [d^2_{\mathcal{W}}(T\#\mu_0,\nu) - d^2_{\mathcal{W}}(T_0\#\mu_0,\nu)]\,\mathrm{d}P(\nu|\mu_0)\,\mathrm{d}P_M(\mu_0)\\ &\geq \frac{\mathbb{E}\alpha(T_\epsilon)}{2}\int d^2_{\mathcal{W}}(T\#\mu_0,T_0\#\mu_0)\,\mathrm{d}P_M(\mu_0).\end{aligned} \tag{4.12}$$

Consequently, if for some T we have that $M(T) = M(T_0)$, inequality (4.12) implies that

$$P_M\{\mu \text{ such that } d^2_{\mathcal{W}}(T\#\mu, T_0\#\mu) = 0\} = 1.$$

Since both T_0 and T are optimal maps, whenever $d^2_{\mathcal{W}}(T\#\mu, T_0\#\mu) = 0$ we can infer that $\|T - T_0\|^2_{L^2(\mu)} = 0$. Therefore

$$P_M\{\mu \text{ such that } \|T - T_0\|^2_{L^2(\mu)} = 0\} = 1,$$

which is equivalent to $\|T - T_0\|^2_{L^2(Q)} = 0$. Therefore, T_0 is the unique minimizer of M up to $L^2(Q)$-norm, and hence identifiable up to Q-null sets.

□

Lemma 4.5.4. *If $\beta, \gamma > 0$ are such that $\lfloor \beta + \gamma \rfloor = \lfloor \beta \rfloor$, then for any $f \in C^{\beta+\gamma}$ we have*

$$\|f\|_{C^\beta} \leq 3 \|f\|_{C^{\beta+\gamma}} .$$

Proof. Recall that

$$\|f\|_{C^\beta} := \max_{|\boldsymbol{k}| \leq \lfloor \beta \rfloor} \left\| D^{\boldsymbol{k}} \right\|_\infty + \max_{|\boldsymbol{k}| = \lfloor \beta \rfloor} \sup_{x \neq y} \frac{|D^{\boldsymbol{k}} f(x) - D^{\boldsymbol{k}} f(y)|}{\|x - y\|^{\beta - \lfloor \beta \rfloor}}. \tag{4.13}$$

First, let's compare the expressions for $\|f\|_{C^\beta}$ and $\|f\|_{C^{\beta+\gamma}}$. For a function g, we have

$$\sup_{x,y\in\Omega, x\neq y} \frac{|g(x) - g(y)|}{\|x - y\|^{\beta - b}} \leq \sup_{x,y\in\Omega, |x-y|<1} \frac{|g(x) - g(y)|}{\|x - y\|^{\beta - b}} + \sup_{x,y\in\Omega, |x-y|\geq 1} \frac{|g(x) - g(y)|}{\|x - y\|^{\beta - b}}.$$

When $\|x - y\| \geq 1$, we have $\|x - y\|^{\beta - b} \geq 1$, therefore

$$\sup_{x,y\in\Omega, |x-y|\geq 1} \frac{|g(x) - g(y)|}{\|x - y\|^{\beta - b}} \leq \sup_{x,y\in\Omega, |x-y|\geq 1} |g(x) - g(y)| \leq 2\|g\|_\infty,$$

but whenever $\|x - y\| < 1$, we have $\|x - y\|^\gamma < 1$, therefore $\|x - y\|^{\beta+\gamma-b} \leq \|x - y\|^{\beta - b}$, so we obtain,

$$\sup_{x,y\in\Omega, |x-y|<1} \frac{|g(x) - g(y)|}{\|x - y\|^{\beta - b}} \leq \sup_{x,y\in\Omega, |x-y|<1} \frac{|g(x) - g(y)|}{\|x - y\|^{\beta+\gamma-b}} \leq \sup_{x,y\in\Omega, x\neq y} \frac{|g(x) - g(y)|}{\|x - y\|^{\beta+\gamma-b}}.$$

It follows that

$$\sup_{x,y\in\Omega, x\neq y} \frac{|g(x) - g(y)|}{\|x - y\|^{\beta - b}} \leq 2 \|g\|_\infty + \sup_{x,y\in\Omega, x\neq y} \frac{|g(x) - g(y)|}{\|x - y\|^{\beta+\gamma-b}}. \tag{4.14}$$

Moreover, as $\lfloor \beta + \gamma \rfloor = \lfloor \beta \rfloor = b$, we have

$$\max_{|\boldsymbol{k}| \leq \lfloor \beta+\gamma \rfloor} \left\| D^{\boldsymbol{k}} f \right\|_\infty = \max_{|\boldsymbol{k}| \leq \lfloor \beta \rfloor} \left\| D^{\boldsymbol{k}} f \right\|_\infty \geq \max_{|\boldsymbol{b}| = b} \left\| D^{\boldsymbol{b}} f \right\|_\infty . \tag{4.15}$$

Therefore, using inequality (4.15), and inequality (4.14) for a function g of the form $D^{\boldsymbol{b}}f$, where $\boldsymbol{b}$ is a vector such that $|\boldsymbol{b}| = b$, we obtain

$$\begin{aligned}
\|f\|_{C^\beta} &= \max_{|\boldsymbol{k}|\leq\lfloor\beta\rfloor}\left\|D^{\boldsymbol{k}}f\right\|_\infty + \max_{|\boldsymbol{b}|=b}\sup_{x\neq y}\frac{|D^{\boldsymbol{b}}f(x)-D^{\boldsymbol{b}}f(y)|}{\|x-y\|^{\beta-b}}\\
&\leq \max_{|\boldsymbol{k}|\leq\lfloor\beta+\gamma\rfloor}\left\|D^{\boldsymbol{k}}f\right\|_\infty + \max_{|\boldsymbol{b}|=b}\left[2\left\|D^{\boldsymbol{b}}f\right\|_\infty + \sup_{x,y\in\Omega,x\neq y}\frac{|D^{\boldsymbol{b}}f(x)-D^{\boldsymbol{b}}f(y)|}{\|x-y\|^{\beta+\gamma-b}}\right]\\
&\leq 3\,\|f\|_{C^{\beta+\gamma}}\,.
\end{aligned} \tag{4.16}$$

□

Proof of Theorem 4.3.4. Suppose we have a sequence $\{T_n \in \mathcal{T}_{\beta,\gamma,R}\}$ that converges to a minimizer of M_N. Let us consider the corresponding sequence of convex potential functions,

$$\left\{\varphi_n : T_n = \nabla\varphi_n, \varphi_n \in \Phi \cap \overline{C_R^{\beta+\gamma}}^{\|\cdot\|_{C^\beta}}\right\}.$$

Since $C_R^{\beta+\gamma}$ is precompact in C^β (as shown in Lemma 6.33 of Gilbarg et al. [28]), there exists a subsequence φ_{n_k} converging to a function φ in C^β. Moreover, since $\beta > 2$, and convergence in β-Hölder norm implies convergence of second-order derivatives, and since $\{\varphi_n\} \subset \Phi$, we conclude $\varphi \in \Phi_0$.

Since the norm $\|\cdot\|_{C^\beta}$ is continuous with respect to its own induced topology, and $\|\varphi_n\|_{C^\beta} \leq 3\,\|\varphi_n\|_{C^{\beta+\gamma}} \leq 3R$, we can also infer that $\|\varphi\|_{C^\beta} \leq 3R$.

Next, let's consider the map $T = \nabla\varphi$ which consequently is in $\mathcal{T}_{\beta,3R}$. Given that the functional M is continuous in T and with respect to $L^2(Q)$-topology (which can be deduced using the triangle inequality and inequality (4.8)), and since convergence in Hölder norm is stronger than convergence in $L^2(Q)$, we can conclude that the initial sequence of maps minimizing M_N, converges to T.

□

Proof of Lemma 4.3.6. It is trivial that $\rho(T,T) = 0$ for any T. We can show that ρ

satisfies the triangle inequality as follows:

$$
\begin{aligned}
&\big(\rho(T_1, T_2) + \rho(T_2, T_3)\big)^2 \\
&= \rho(T_1, T_2)^2 + \rho(T_2, T_3)^2 + 2\rho(T_1, T_2)\rho(T_2, T_3) \\
&= \rho(T_1, T_2)^2 + \rho(T_2, T_3)^2 \\
&+ 2\Big[\int d^2_{\mathcal{W}}(T_1\#\mu, T_2\#\mu)\,\mathrm{d}P_M(\mu)\Big]^{1/2}\Big[\int d^2_{\mathcal{W}}(T_1\#\mu, T_2\#\mu)\,\mathrm{d}P_M(\mu)\Big]^{1/2} \\
&\geq \rho(T_1, T_2)^2 + \rho(T_2, T_3)^2 \\
&+ 2\int d_{\mathcal{W}}(T_1\#\mu, T_2\#\mu) d_{\mathcal{W}}(T_1\#\mu, T_2\#\mu)\,\mathrm{d}P_M(\mu) \quad \text{(Cauchy Schwarz)} \\
&= \int d^2_{\mathcal{W}}(T_1\#\mu, T_2\#\mu) + d^2_{\mathcal{W}}(T_2\#\mu, T_3\#\mu) + 2d_{\mathcal{W}}(T_1\#\mu, T_2\#\mu) d_{\mathcal{W}}(T_1\#\mu, T_2\#\mu)\,\mathrm{d}P_M(\mu) \\
&\geq \int d^2_{\mathcal{W}}(T_1\#\mu, T_3\#\mu)\,\mathrm{d}P_M(\mu) \\
&= \rho^2(T_1, T_3),
\end{aligned}
$$

□

To derive the convergence rate for the estimator, we will make use of some theorems from M-estimation [72]. For the reader's convenience, we will first restate these theorems (with minor modifications to more easily relate to our context). Furthermore, we restate a theorem from Gunsilius [31] that will be essential for determining the rate.

Lemma 4.5.5 (Bracketing Entropy of Hölder Class, Corollary 2.7.4 [72])**.** *Let Ω be a bounded, convex subset of $\mathbb{R}^d$ with a nonempty interior. There exists a constant K, depending only on β, vol(Ω), r and ρ such that,*

$$\log N_{[\,]}(\epsilon, C^\beta_R(\Omega), L^r(Q)) \leq KR^{d/\beta}\epsilon^{-d/\beta},$$

for every $r \geq 1$,$\epsilon > 0$, and probability measure Q on $\mathbb{R}^d$.

Lemma 4.5.6. *(Lemma 5 of Gunsilius [31]) Let φ_1, φ_2 be proper strictly convex and bounded potential functions on every compact subset of Ω^0 with Lipschitz-continuous gradients $\nabla\varphi_1$ and $\nabla\varphi_2$ satisfying $\nabla\varphi_1(\Omega^0) = \nabla\varphi_2(\Omega^0)$. Then it holds for all $x \in \Omega^0$*

$$\|\nabla\varphi_1(x) - \nabla\varphi_2(x)\|^2 \leq c(1 + \max\{L_1, L_2\})^2|\varphi_1(x) - \varphi_2(x)|$$

where $0 \leq L_1, L_2 < +\infty$ are the Lipschitz constants of $\nabla\varphi_1$ and $\nabla\varphi_2$, respectively, $c < +\infty$ is a constant.

Proof of Theorem 4.3.7. By observing that the right-hand side of inequality (4.12) is equal to $\frac{\alpha}{2}\rho(T, T_0)$, it follows that the functional M demonstrates quadratic growth in the vicinity of its minimizer T_0 with respect to the semi-metric ρ. Therefore we can use the empirical process approach to obtain an upper bound for the rate of convergence.

First, we find a function $\phi_N(\delta)$ such that

$$\mathbb{E} \sup_{\rho(T,T_0)\leq\delta, T\in\mathcal{T}_{\beta,\gamma,R}} \sqrt{N}\Big|(M_N - M)(T) - (M_N - M)(T_0)\Big| \leq \phi_N(\delta).$$

Given that $\mathcal{T}_{\beta,\gamma,R} \subset \mathcal{T}_{\beta,3R}$ according to Lemma 4.5.4, we can instead find $\phi_N(\delta)$ such that

$$E \sup_{\rho(T,T_0)\leq\delta, T\in\mathcal{T}_{\beta,3R}} \sqrt{N}\Big|(M_N - M)(T) - (M_N - M)(T_0)\Big| \leq \phi_N(\delta). \tag{4.17}$$

We define a class of functions indexed by T as follows:

$$\mathcal{F}_u := \{f_T(\mu,\nu) = d^2_{\mathcal{W}}(T\#\mu,\nu) - d^2_{\mathcal{W}}(T_0\#\mu,\nu), \text{ s.t. } T\in\mathcal{T}_{\beta,3R} \text{ and } \rho(T,T_0)\leq u\},$$

with the domain of each function $f_T \in \mathcal{F}_u$ being $\mathcal{W}_2(\Omega)\times\mathcal{W}_2(\Omega)$. We can see that (4.17) is equivalent to

$$\mathbb{E} \sup_{f\in\mathcal{F}_\delta} \sqrt{N}|(P_N - P)f| \leq \phi_N(\delta).$$

Denote by $\log N_{[\,]}(\epsilon, \mathcal{F}_u, L^2(P))$, the bracketing number of $\mathcal{F}_u$. We find an upper bound for this bracketing entropy using the bracketing entropy of the class of functions C^β_{3R}. To do this, note that any $f_T \in \mathcal{F}_u$ is induced by a map T such that $T = \nabla\varphi$, for a convex function $\varphi \in C^\beta_{3R}$. Thus, for any $f_{T_1}, f_{T_2} \in \mathcal{F}_u$, we have:

$$\begin{aligned}
\left\|f_{T_1} - f_{T_2}\right\|^2_{L^2(P)} &\leq \int |f_{T_1}(\mu,\nu) - f_{T_2}(\mu,\nu)|^2 \,\mathrm{d}P(\mu,\nu)\\
&\leq \int |d^2_{\mathcal{W}}(T_1\#\mu,\nu) - d^2_{\mathcal{W}}(T_2\#\mu,\nu)|^2 \,\mathrm{d}P(\mu,\nu)\\
&\leq 4\mathrm{diam}(\Omega)^2 \int \|T_1 - T_2\|^2_{L^2(\mu)} \,\mathrm{d}P_M(\mu)\\
&\leq 4\mathrm{diam}(\Omega)^2 \|T_1 - T_2\|^2_{L^2(Q)}\\
&\leq 4\mathrm{diam}(\Omega)^2 C' \|\varphi_1 - \varphi_2\|_{L^1(Q)}\,.
\end{aligned} \tag{4.18}$$

To see why the last inequality holds, note that any $T \in \mathcal{T}_{\beta,3R}$ is the gradient of an L-smooth convex function, making it L-Lipschitz. Therefore, by applying Lemma 4.5.6 (our re-statement of [31, Lemma 5]), we can establish that for $T_1, T_2 \in \mathcal{T}_{\beta,3R}$

with corresponding potential functions φ_1, φ_2, the inequality $\|T_1 - T_2\|^2_{L^2(Q)} \leq c(1 + L)^2 \|\varphi_1 - \varphi_2\|_{L^1(Q)}$ holds for some constant c.

Using inequality (4.18) and Lemma 4.5.5, we obtain

$$\log N_{[\,]}(\epsilon, \mathcal{F}_u, L^2(P)) \lesssim \log N_{[\,]}(\epsilon/(4\text{diam}(\Omega)^2 C'), C^{\beta}_{3L}(\Omega), L^1(Q)) \lesssim \left(\frac{1}{\epsilon}\right)^{d/\beta}.$$

By inequality (4.18), we can also show that

$$Pf_T^2 \leq P\,\|T - T_0\|^2_{L^2(\mu)} = \|T - T_0\|^2_{L^2(Q)} \leq u^2,$$

for all $f_T \in \mathcal{F}_u$.

Given that the functions in $\mathcal{F}_\delta$ are uniformly bounded over (μ, ν) and T, and $Pf^2 \leq \delta^2$ for all $f \in \mathcal{F}_\delta$, the conditions of Theorem 3.6.5 are satisfied. This being the case, we can choose:

$$\phi_N(\delta) = \tilde{J}_{[\,]}(\delta, \mathcal{F}_\delta, L^2(P))\left(1 + \frac{\tilde{J}_{[\,]}(\delta, \mathcal{F}_\delta, L^2(P))}{\delta^2\sqrt{N}}\bar{c}\right),$$

where the constant $\bar{c} = 2\text{diam}(\Omega)^2$ is a uniform upper bound for the functions in the class $\mathcal{F}_\delta$. Using Lemma 4.5.5 and when $\beta > \frac{d}{2}$, we can write

$$\begin{aligned}
\tilde{J}_{[\,]}(\delta, \mathcal{F}_\delta, L^2(P)) &:= \int_0^\delta \sqrt{1 + \log N_{[\,]}(\epsilon, \mathcal{F}_\delta, L^2(P))}\,\mathrm{d}\epsilon \\
&\lesssim \int_0^\delta \sqrt{1 + C\left(\frac{1}{\epsilon}\right)^{d/\beta}}\,\mathrm{d}\epsilon \\
&\lesssim \delta + \int_0^\delta \sqrt{\left(\frac{1}{\epsilon}\right)^{(d/\beta)}}\,\mathrm{d}\epsilon \qquad \text{since } \sqrt{1+a} \leq 1 + \sqrt{a} \quad \text{for } a \geq 0 \\
&\lesssim \delta + \frac{1}{1 - \frac{d}{2\beta}}\delta^{(1-\frac{d}{2\beta})} \\
&\lesssim \delta^{1-\frac{d}{2\beta}}.
\end{aligned} \tag{4.19}$$

Now, we can apply Theorem 3.6.3 to conclude the proof. The rate of convergence r_N is obtained by the requirement $r_N^2\phi_N\left(\frac{1}{r_N}\right) \lesssim \sqrt{N}$. This gives us the rate $r_N = N^{\frac{\beta}{2\beta+d}}$ when $\beta > \frac{d}{2}$.

□

Proof of Lemma 4.4.1. Let $T \in \mathcal{T}$ and take any continuous function η with domain Ω. For $\epsilon > 0$ sufficiently small, $T + \epsilon\eta$ is also in $\mathcal{T}$. If it exists, the Gateaux derivative of

$d^2_{\mathcal{W}}(T\#\mu, \nu)$ is

$$D_\eta d^2_{\mathcal{W}}(T\#\mu, \nu) = \lim_{\epsilon \to 0} \frac{d^2_{\mathcal{W}}((T + \epsilon\eta)\#\mu, \nu) - d^2_{\mathcal{W}}(T\#\mu, \nu)}{\epsilon}.$$

In the following, we show that $D_\eta d^2_{\mathcal{W}}(T\#\mu, \nu)$ exists and we calculate the limit. To do so, we will construct a coupling γ between μ and ν with the property that $d^2_{\mathcal{W}}(T\#\mu, \nu) = \int |T(x) - y|^2 \, \mathrm{d}\gamma(x, y)$. To this aim, let S be an optimal map such that $S\#(T\#\mu) = \nu$. According to Brenier's theorem, such an optimal map exists and is unique if $T\#\mu$ is absolutely continuous. Since $T \in \mathcal{T}$ and μ is absolutely continuous, according to Lemma 4.5.2, so is $T\#\mu$. Let $X \sim \mu$ and define $Z = T(X)$ and $Y = S(Z)$. Denote by γ the induced joint distribution of the pair (X, Y). Note that

$$d^2_{\mathcal{W}}(T\#\mu, \nu) = \mathbb{E}\,\|Z - S(Z)\|^2 = \mathbb{E}\,\|Z - Y\|^2 = \mathbb{E}\,\|T(X) - Y\|^2 = \int |T(x) - y|^2 \, \mathrm{d}\gamma(x, y).$$

Note that from the construction we can infer that $\gamma = (\mathrm{id}, S \circ T)\#\mu$ and thus the coupling γ is independent of the random variable X. Thus:

$$\begin{aligned} d^2_{\mathcal{W}}((T + \epsilon\eta)\#\mu, \nu) &\leq \int |(T + \epsilon\eta)(x) - y|^2 \, \mathrm{d}\gamma(x, y) \\ &= \int (|T(x) - y|^2 + 2\epsilon\langle \eta(x), T(x) - y\rangle) \, \mathrm{d}\gamma(x, y) + o(\epsilon^2) \\ &= d^2_{\mathcal{W}}(T\#\mu, \nu) + 2\epsilon \int \langle \eta(x), T(x) - y\rangle \, \mathrm{d}\gamma(x, y) + o(\epsilon^2), \end{aligned}$$

and therefore

$$\lim_{\epsilon \downarrow 0} \frac{d^2_{\mathcal{W}}((T + \epsilon\eta)\#\mu, \nu) - d^2_{\mathcal{W}}(T\#\mu, \nu)}{\epsilon} \leq 2 \int \langle \eta(x), T(x) - y\rangle \, \mathrm{d}\gamma(x, y).$$

Now define γ_ϵ using the same procedure as above, but such that γ couples μ and ν while satisfying $d^2_{\mathcal{W}}((T + \epsilon\eta)\#\mu, \nu) = \int |T(x) + \epsilon\eta(x) - y|^2 \, \mathrm{d}\gamma_\epsilon(x, y)$. Therefore similar to above we can see that $\gamma_\epsilon = (\mathrm{id}, (S_\epsilon \circ (T + \epsilon\eta))\#\mu$, where S_ϵ is the optimal map between $(T + \epsilon\eta)\#\mu$ and ν. Thus

$$\begin{aligned} d^2_{\mathcal{W}}((T + \epsilon\eta)\#\mu, \nu) &= \int |(T + \epsilon\eta)(x) - y|^2 \, \mathrm{d}\gamma_\epsilon(x, y) \\ &= \int (|T(x) - y|^2 + 2\epsilon\langle \eta(x), T(x) - y\rangle) \, \mathrm{d}\gamma_\epsilon(x, y) + o(\epsilon^2) \\ &\geq d^2_{\mathcal{W}}(T\#\mu, \nu) + 2\epsilon \int \langle \eta(x), T(x) - y\rangle \, \mathrm{d}\gamma_\epsilon(x, y) + o(\epsilon^2), \end{aligned}$$

and

$$\lim_{\epsilon\downarrow 0}\frac{d^2_{\mathcal{W}}((T+\epsilon\eta)\#\mu,\nu)-d^2_{\mathcal{W}}(T\#\mu,\nu)}{\epsilon}\geq \liminf_{\epsilon\downarrow 0} 2\int\langle\eta(x),T(x)-y\rangle\,\mathrm{d}\gamma_\epsilon(x,y).$$

To prove the existence of the limit, it is enough to show the integral with respect to γ_ϵ converges to the integral with respect to γ. We show the convergence of γ_ϵ to γ in the Wasserstein metric. Using the inequality (2.4) one can control the Wasserstein distance between the two measures by controlling $\|S_\epsilon\circ(T+\epsilon\eta)-S\circ T\|^2_{L^2(\mu)}$. First note that as ϵ converges to zero, $(T+\epsilon\eta)\#\mu$ converges to $T\#\mu$, and again using the inequality (2.4) one can control the Wasserstein distance between the two measures by $\|\epsilon\eta\|^2_{L^2(\mu)}$. As convergence in the Wasserstein distance results in narrow convergence, one can show that S_ϵ converges to S using Theorem 1.7.7 in Panaretos & Zemel [55]. By Lemma 4.5.2, $(T+\epsilon\eta)\#\mu$ and $T\#\mu$ are absolutely continuous, thus using Theorem 5.20 in Villani [74] S_ϵ and S are continuous. Therefore $S_\epsilon\circ(T+\epsilon\eta)\longrightarrow S\circ T$ and we can conclude $\gamma_\epsilon\longrightarrow\gamma$. Additionally, T is bounded and continuous, therefore we can conclude the integral with respect to γ_ϵ converges to the integral with respect to γ. The two inequalities prove the existence of the derivative and:

$$D_\eta d^2_{\mathcal{W}}(T\#\mu,\nu)=\lim_{\epsilon\to 0}\frac{d^2_{\mathcal{W}}((T+\epsilon\eta)\#\mu,\nu)-d^2_{\mathcal{W}}(T\#\mu,\nu)}{\epsilon}=2\int\langle\eta(x),T(x)-y\rangle\,\mathrm{d}\gamma(x,y).$$

Thus

$$\begin{aligned}&D_\eta M_N(T)=\frac{1}{N}\sum_{i=1}^{N}\int\langle\eta(x),T(x)-y\rangle\,\mathrm{d}\gamma_{\mu_i,\nu_i}(x,y)\\&\text{for }\gamma_{\mu_i,\nu_i}\in\Gamma(\mu_i,\nu_i)\quad\text{s.t.}\quad d^2_{\mathcal{W}}(T\#\mu_i,\nu_i)=\int|T(x)-y|^2\,\mathrm{d}\gamma_{\mu_i,\nu_i}(x,y),\end{aligned}\tag{4.20}$$

and

$$\begin{aligned}&D_\eta M(T)=\int\int<\eta(x),T(x)-y>\mathrm{d}\gamma_{\mu,\nu}(x,y)\,\mathrm{d}P(\mu,\nu)\\&\text{for }\gamma_{\mu,\nu}\in\Gamma(\mu,\nu)\quad\text{s.t.}\quad d^2_{\mathcal{W}}(T\#\mu_i,\nu_i)=\int|T(x)-y|^2\,\mathrm{d}\gamma_{\mu,\nu}(x,y).\end{aligned}\tag{4.21}$$

□

4.6 Generating random convex functions

The purpose of this section is to briefly discuss how one might numerically generate random functions $\varphi(x,y)$ that are convex on a domain $U=[x_0,x_1]^2$ and average to

$(x^2+y^2)/2$ (therefore their gradient is an optimal map and their average is the id map, as required in our Model). This can be easily done with linear maps, but we are interested in more variability and complex maps. To this aim, we take functions of the form

$$\varphi = \frac{x^2+y^2}{2} + \sum_{d=2}^{D}\sum_{k=0}^{d} \frac{1}{k!(d-k)!}\left[a_{d,k}x^k y^{d-k} + b_{d,k}x^{1/k}y^{1/(d-k)}\right], \tag{4.22}$$

where $a_{d,k}$ and $b_{d,k}$ are random coefficients. To avoid the singularity of the second term, we take $(x_0, x_1) = (0.5, 1.5)$. We require that $\mathbb{E}[\nabla\varphi] = (x,y)^\top$ which implies $\mathbb{E}[a_{d,k}] = \mathbb{E}[b_{d,k}] = 0$. Our considerations in this section being practical, we shall probe for convexity numerically and approximately (at least in probability). Namely, we set $D = 8$ and we take the coefficients $a_{d,k}$ and $b_{d,k}$ to be independent random variables with a centred normal distribution of variance σ^2. Using the `sympy` library of Python, we explicitly calculate the expectation of the determinant of the Hessian matrix, and we calculate the minimum in the domain at $(1.5, 1.5)$, specifically

$$\min_{x,y\in U} \mathbb{E}\left[\det(\nabla^2\varphi)\right] \approx 1 - 11\sigma^2. \tag{4.23}$$

Similarly, we estimate the maximum variance of the Hessian determinant to be at most

$$\mathrm{Var}\left[\det(\nabla^2\varphi)\right] \lesssim 10\sigma^2, \tag{4.24}$$

for small enough σ. Assuming the determinant follows a normal distribution, the probability of generating a non-convex function $\mathbb{P}\left\{\det(\nabla^2\varphi) < 0\right\}$ decreases exponentially below $\sigma \lesssim 0.2$, being around 0.1 for $\sigma = 0.15$ and $\sim 10^{-25}$ for $\sigma = 0.03$, which is the parameter used to generate the noise maps in Fig. 4.1. A way to limit the probability of generating a non-convex function even further could be to use a distribution with support on a finite domain like a beta distribution adjusted to the domain U, instead of a normal distribution.

Chapter 5

Autoregressive Models via Iterated Transportation

The work in this chapter was done in collaboration with my supervisor Victor Panaretos and is publicly available as a preprint [25]. This chapter follows the priprint with slight changes.

5.1 Introduction

Distributional autoregression is a natural next-step for distributional regression models – indeed, it is arguably the setting where most distributional regression data sets arise. Rather than i.i.d. covariate/response distributions, one observes a dependent sequence of probability distributions $\{\mu_n\}_{n=1}^N$. When viewed as a Markov chain in the Wasserstein space, this sequence can be modeled autoregressively by specifying a relationship between the conditional Fréchet mean at time $n+1$ and the chain at time n. Once again, this can be done geometrically (as indeed was already explored in [15] and [79]), or by way of optimal transport maps, with similar advantages/disadvantages.

A first contribution based directly on transport maps was made in Zhu and Müller [80], where random perturbations of the identity were iteratively contracted/composed to form a time-dependent sequence. This was subsequently used either as "increments" between consecutive distributions or as "deviations" from the marginal Fréchet mean, to produce autoregressive models. Key in this approach was the use of iterated random function systems and a canny definition of a contraction operation on the space of transport maps, allowing to mimic the contractive effect of a correlation operator in usual autoregression. Jiang [36] subsequently generalised this approach to autoregressive modeling to the case of vector-valued distributional chains, i.e. time-evolving vectors with distributions as coordinates.

A salient limitation of this approach is that the entire dynamics of the process reduce to a single scalar quantity $|\alpha| \leq 1$, regulating the "strength" of the contraction. While this resembles real-valued autoregressive processes, it is likely too rigid in a functional context (or even a multivariate context), and can have undesirable consequences when asserting stationarity (see Section 5.2.3 for a more extensive discussion). Ideally, a genuinely functional model would allow for a *functional* specification of the dynamics, thus capable of expressing more complex dependencies. In response to this drawback, Zhu and Müller [80] also defined a model where the scalar contraction coefficient is replaced by a *functional contraction coefficient*, contracting variably across the domain. This comes with the caveat of a more complicated theory, including cumbersome technical assumptions, as well as a more involved interpretation.

The purpose of this Chapter is to introduce and develop transportation-based autoregressive models with genuinely functional dynamics, yielding easily interpretable yet rich classes of distributional autoregressions. To do so, we extend to the autoregressive case the functional structure of the model in Chapter 3, where the regression operator is a monotone rearrangement, making use of the scalar "contractive effect" introduced by Zhu and Müller [80] – intuitively, we posit a model where the *shape* of the dynamics is captured by a monotone map, modulated by a contractive parameter α regulating the degree of non-degeneracy of the model. In its simplest form, this approach can be interpreted as positing that

$$\mu_{n+1} = \theta_n \# [\alpha \mu_n], \quad n \in \mathbb{Z},$$

for i.i.d. random increasing maps θ_n with $\mathbb{E}[\theta_n(x)] = S(x)$; S a deterministic monotone map; and $\mu_n \mapsto [\alpha\mu_n]$ a barycentric contraction operation, suitably defined at the level of quantile functions (see Equation (5.1) for a precise definition). Intuitively, the model suggests that step $n+1$ in the chain is obtained by pushing forward the nth step ("shrunken" slightly to allow for temporal stationarity) via a random perturbation of the deterministic deformation S. This is a direct autoregressive extension of the model in Chapter 3, employing the contractive device of Zhu and Müller [80] to assure temporal stability in law. However, more modeling possibilities are available in our approach, and this is just the motivating one (see Section 5.2.2).

The rest of the Chapter is organised as follows. We first revisit the problem of defining iterated random function systems of increasing maps. In particular, Section 5.2.1 presents a functional extension of the iterated system employed in Zhu and Müller [80]. This extension is then used in Section 5.2.2 in order to define three different possible notions of autoregression – in each case, the iterated transport map system serves to model a different characteristic of the distributional time series (e.g. the increments, the quantiles, or the generalised quantiles). We compare the resulting models to existing approaches in Section 5.2.3 and determine conditions for stationarity in Section 5.2.4. We then show in Section 5.2.5 that all three models can be fitted and analysed using the same estimation theory – albeit applied to optimal maps that represent a different characteristic in each case. In particular, we establish identifiability, consistency, and rates of convergence. Finally, the finite sample performance of our methodology is illustrated on some simulated and real data (Sections 5.3 and 5.4). The proofs are collected in a separate Section, and we conclude with a discussion of some further possible generalisations.

5.2 Autoregressive Models via Iterated Transportation

5.2.1 Random Iterated Transport

Our definition of autoregressive models for distributions will hinge on appropriately defined iterated random systems of transport maps (following the approach of Zhu and Müller [80], to whom we compare below). This is a special case of a framework for studying questions about Markov chains via iterated random functions, going back to at least Diaconis and Freedman [23]. They define an iterated random function system on a state space $\mathcal{T}$ as

$$T_i = f(T_{i-1}\,;\,\theta_i)$$

for a family of transformations $\{f(\,\cdot\,;\,\theta) : \theta \in \Theta\}$ acting on $\mathcal{T}$, and random elements θ_i in some parameter space Θ, independent of $T_i \in \mathcal{T}$. By suitable choice of the family $f(\,\cdot\,;\,\theta)$ and some distribution on Θ they show how a plethora of Markov chains can be cast in this light.

Let $\Omega = [\omega_0, \omega_1]$ be a closed interval of $\mathbb{R}$. In our case, the state space $\mathcal{T}$ will be the set of optimal transport maps

$$\mathcal{T} := \{T : \Omega \to \Omega | T(\omega_1) = \omega_1, T(\omega_2) = \omega_2, T \text{ is strictly increasing and continuous}\},$$

viewed as a closed and complete subset of the Lebesgue space $L^p(\Omega)$ equipped with the corresponding p-distance $\|\cdot\|_p$, for some $1 \leq p < \infty$ (we will mostly focus on $p = 2$). And, the question is how to define f and θ_i to generate an iterated random system that is sufficiently rich to serve as a basis for interesting autoregressive models, yet remains tractable and admits a non-degenerate stationary solution. Naively, one might simply posit that $\Theta = \mathcal{T}$ and $f(T;\theta) = \theta \circ T$, as increasing maps form a transformation group under composition. However, $f(.;\theta)$ needs to be a contraction "on average" (in a precise sense) for the Diaconis and Freedman [23] results to be applicable.

This motivates forms of f that are "contractive compositions". To this aim, given $|\alpha| \leq 1$, define the α-contraction of an optimal transport map to be the operator $T \mapsto [\alpha T]$ defined pointwise via

$$[\alpha T](x) = \begin{cases} x + \alpha(T(x) - x) & 0 < \alpha \leq 1 \\ x & \alpha = 0 \\ x + \alpha(x - T^{-1}(x)) & -1 \leq \alpha < 0. \end{cases} \tag{5.1}$$

This definition is due to Zhu and Müller [80], under slightly different terminology/notation, and mimics the operation of contracting an unconstrained function by a scalar, but conforming to the constraints elicited by working in $\mathcal{T}$. Notice that $T \mapsto [\alpha T]$ is indeed a contraction on $\mathcal{T}$ with respect to L^1 norm, with the identity as its fixed point – any other fixed point must equal the identity almost everywhere by the Banach fixed-point theorem.

Finally, given $|\alpha| < 1$ and $\theta \in \mathcal{T}$ we can now make precise the notion of f being a "contractive composition" map by defining

$$f(T;\theta) = \theta \circ [\alpha T].$$

To define an iterated random system, it suffices to put a probability distribution Q on $\mathcal{T}$, and make i.i.d. draws $\theta_i \sim Q$ yielding

$$T_i = f(T_{i-1};\theta_i). \tag{5.2}$$

Our proposal is to draw i.i.d. elements of $\mathcal{T}$ with a specified expectation $S \in \mathcal{T}$, say $\theta_i = T_{\epsilon_i} \circ S$, for $\{T_{\epsilon_i}\}_{i=1}^N$ a collection of independent and identically distributed random optimal maps satisfying $\mathbb{E}\{T_{\epsilon_i}(x)\} = x$ almost everywhere on Ω. Explicitly, our iteration is now

$$T_i = f(T_{i-1};\underbrace{T_{\epsilon_i} \circ S}_{\theta_i}) = \underbrace{T_{\epsilon_i} \circ S}_{\theta_i} \circ [\alpha T_{i-1}]. \tag{5.3}$$

The degrees of freedom in this iteration are the choice of $S \in \mathcal{T}$ and $\alpha \in [-1,1]$. In a statistical setting, these would be the targets of estimation. This definition extends the iteration of Zhu and Müller [80] where S was a priori fixed to be the identity. Our extension seems natural and conceptually straightforward: it iterates contracted composition with perturbations of an arbitrary element of the transformation group, rather than with perturbations of the neutral element. Yet, it substantially complicates the subsequent probabilistic analysis and estimation theory. In exchange, we get a richer class of autoregressive models that exhibit advantages in the context of modeling and data analysis. We elaborate on the relationship and the nature of the extension in a subsequent paragraph. We then show that the iteration admits a unique stationary solution (under some additional assumptions). First, though, we explore how such an iterated random system of optimal maps could be used as a basis for distributional autoregression.

5.2.2 Autoregressive Models

The main purpose of a random iteration (5.2) is the construction of a Markov chain model for a dependent sequence of probability distributions $\mu_i \in \mathcal{W}_2(\Omega)$, that will always be taken to possess a continuous cumulative distribution function. The models we seek are of autoregressive type, and so should ultimately be interpretable as a structural specification of the one-step conditional mean. Given stationary random sequence $\{T_i\}$ of optimal maps, there appear to be (at least) three different ways of doing so, by relating the T_i to some suitable feature of $\{\mu_i\}$:

(I) Modeling the "increments" $T_{\mu_{i-1}}^{\mu_i} := F_{\mu_i}^{-1} \circ F_{\mu_{i-1}}$ as being equal to T_i (we call these increments, as $T_{\mu_{i-1}}^{\mu_i}$ is the optimal map pushing μ_{i-1} forward to μ_i), or equivalently modeling the quantiles as

$$F_{\mu_i}^{-1} := T_i \circ F_{\mu_{i-1}}^{-1}.$$

When $\{T_i\}$ is stationary, this yields a process with stationary increments, but the process could be non-stationary (if so, it's interesting to understand if there is "drift"). This chain corresponds to specifying that the (usual) conditional expectation of $F_{\mu_i}^{-1}$ given $F_{\mu_{i-1}}^{-1}$ as

$$\mathbb{E}[F_{\mu_i}^{-1}|F_{\mu_{i-1}}^{-1}] = \mathbb{E}\{T_i\} \circ F_{\mu_{i-1}}^{-1} = \mathbb{E}\{f(T_{i-1};\theta_i)\} \circ F_{\mu_{i-1}}^{-1} = \mathbb{E}\{\theta_i \circ [\alpha T_{i-1}]\} \circ F_{\mu_{i-1}}^{-1}.$$

The form of $\mathbb{E}[T_i]$ will depend on the stationary solution of $T_i = f(T_{i-1};\theta_i)$.

(UQ) Modeling the (uniform) quantiles $F_{\mu_i}^{-1}$ as being equal to T_i,

$$F_{\mu_i}^{-1} := T_i.$$

This automatically yields a stationary process when $\{T_i\}$ is stationary, directly interpretable at the level of quantiles, and corresponds to specifying the (usual) conditional expectation of $F_{\mu_i}^{-1}$ given $F_{\mu_{i-1}}^{-1}$ as

$$\mathbb{E}[F_{\mu_i}^{-1}|F_{\mu_{i-1}}^{-1}] = (\mathbb{E}\theta_i) \circ [\alpha F_{\mu_{i-1}}^{-1}] = S \circ [\alpha F_{\mu_{i-1}}^{-1}] = f(F_{\mu_{i-1}}^{-1}; S).$$

This model corresponds to an autoregressive extension of model (3.1).

(GQ) Modeling the generalised quantiles [17] or μ-quantiles $F_{\mu_i}^{-1} \circ F_\mu$ with respect to some measure μ as being equal to T_i. This also immediately yields stationarity and (under regularity conditions) is equivalent to stating $\mu_i = T_i\#\mu$, in effect modeling the μ_i as serially dependent "perturbations" of a fixed μ. This corresponds to specifying the (usual) conditional expectation of $F_{\mu_i}^{-1}$ given $F_{\mu_{i-1}}^{-1}$ as

$$\mathbb{E}[F_{\mu_i}^{-1}|F_{\mu_{i-1}}^{-1}] = (\mathbb{E}\theta_i) \circ [\alpha[F_{\mu_{i-1}}^{-1} \circ F_\mu]] = S \circ [\alpha[F_{\mu_{i-1}}^{-1} \circ F_\mu]] = f(F_{\mu_{i-1}}^{-1} \circ F_\mu; S).$$

Note that setting $\alpha = 1$ in (UQ) yields the same model as setting $\alpha = 0$ in (I), interpretable as a random walk, and this we shall revisit. In Section 5.4 we will focus on (UQ) and (I) to model sequential distributional data and discuss the merits/drawbacks of each approach. Model (GQ) can actually be seen to be a variant of the model (UQ) albeit under a modification of the definition of the contraction operator itself – see Section (5.5.2), and especially Remark (5.5.11) for an equivalent characterization of the model (GQ)

5.2.3 Comparison with Related Work

Our iteration (5.3) represents a generalization of the iteration in Zhu and Müller [80], by combining their notion of α-contraction (which they call *distributional scalar multiplication*), with the functional structure of the model (3.1). Specifically, Zhu and Müller [80] considered autoregressive models for distributional time series, based on the iterative system of optimal transport maps

$$T_i = T_{\epsilon_i} \circ [\alpha T_{i-1}]. \tag{5.4}$$

This is a special case of our system (5.3) when S is fixed to be the identity map $\mathrm{id}(x) = x$. Their clever α-contraction, combined with classical results on iterated random function theory, allows one to deduce the existence of a unique stationary solution to the iteration (5.4) thanks to the contracting effect of α for $-1 < \alpha < 1$ (and some additional technical assumptions).

However, basing a distributional autoregressive model on this system is restrictive in two important ways:

1. As a stochastic model, the system (5.4) is parametric and univariate: the only unknown is the scalar coefficient $\alpha \in (-1, 1)$. Correspondingly, when basing our model on that iteration (with any of the three interpretations specified in the previous section), the temporal dependence of μ_i on μ_{i-1} will be completely specified up to an unknown scalar parameter. This is reminiscent of autoregressive models on the real line but is arguably overly restrictive in a functional data analysis (or even multivariate analysis) setting, where the temporal dependence is very likely more complex. A genuinely functional model would replace the scalar coefficient with a suitable *functional coefficient*, e.g. a

non-linear operator.

2. If a stationary solution to system (5.4) exists, then it must satisfy $\mathbb{E}(T_i) = \mathrm{id}$. To see this, recall the definition of the scalar multiplication (5.1) and observe that

$$\mathbb{E}[T_i] = \mathbb{E}[T_{i+1}] = \mathbb{E}[\mathbb{E}[T_{i+1}|T_i]] = \mathbb{E}[\alpha T_i].$$

   This is consequential if using the sequence T_i to induce a distributional time series $\{\mu_i\}$. In the (I) model, where T_i models the increments between consecutive μ_i, this implies that the conditional Fréchet mean (in the Wasserstein metric) of μ_i given μ_{i-1} is exactly equal to μ_{i-1}, a sort of 'Fréchet martingale'. Effectively this trivializes the regressor relationship to be an identity – there is no modeling flexibility for the conditional mean, only the conditional variance (via α). In the (UQ) model, where $T_i \equiv F_{\mu_i}^{-1}$ is taken as the quantile function of μ_i, the fact that $\mathbb{E}(T_i) = \mathrm{id}$ implies that the distributional autoregression model can only admit the uniform distribution as its Fréchet mean (with respect to the Wasserstein metric). There is no flexibility in the modeling of the marginal mean.

By contrast, models based on our system (5.3) are genuinely functional, since on account of the unknown transport map S. Furthermore, our model can accommodate *any* distribution as its Fréchet mean: given any optimal map $T \in \mathcal{T}$, there exist S and α such that $\mathbb{E}(T_i) = T$.

The optimal map interpretation of our system (5.3) is an auto-regressive modification of the distributional optimal transport regression model (3.1). By direct analogy, an autoregressive model (optimal map interpretation) for a time series of distributions $\{\mu_i\}$ would be defined as

$$\mu_i = T_{\epsilon_i}\#(S\#\mu_{i-1}),$$

which is equivalent to model (5.3) when $\alpha = 0$ and when we interpret T_i such that $\mu_i = T_i\#\mu_{i-1}$, i.e. the optimal map interpretation. If we take the quantile interpretation, the two models are again related for $\alpha = 1$ since model (5.3) is equivalent to

$$F_i^{-1} = T_{\epsilon_i} \circ S \circ F_{i-1}^{-1}.$$

However, assuming the noise maps T_{ϵ_i} are close to identity, one observes that the series of CDFs F_i^{-1} would stabilize around a step function where the position of the jumps coincide with fixed points of the map S, and therefore the distribution μ_i would oscillate around a mixture of Dirac measures. This is where we combine the functional structure of model (3.1) with the scalar "contractive effect" introduced by Zhu and Müller [80] – intuitively, the magnitude of α regulates the non-degeneracy of the model. The next Section demonstrates that this combined extension does indeed yield a unique stationary solution.

5.2.4 Existence of Unique Stationary Solution

We now turn to establish the existence of a unique stationary solution for the system (5.3). We will use the results of Wu and Shao [76], extending to our iteration (5.3) the steps follows by Zhu and Müller [80] in the context of iteration (5.4). Let $\{T_{\epsilon_i}\}_{i=1}^{N}$ be a collection of independent and identically distributed random optimal maps satisfying $\mathbb{E}\{T_{\epsilon_i}(x)\} = x$ almost everywhere on Ω. Define $\Phi_i, \tilde{\Phi}_{i,m} : \mathcal{T} \to \mathcal{T}$ by

$$\begin{aligned} \Phi_i(T) &= f(T; T_{\epsilon_i} \circ S) = T_{\epsilon_i} \circ S \circ [\alpha T] \\ \tilde{\Phi}_{i,m}(T) &= \Phi_i \circ \Phi_{i-1} \circ \cdots \circ \Phi_{i-m+1}(T). \end{aligned} \tag{5.5}$$

The following assumption stipulates

Assumption 5.2.1. *(Moment Contracting Condition [76]) Suppose there exists $\eta > 0, Q_0 \in \mathcal{T}, C > 0$ and $r \in (0,1)$ such that*

$$\mathbb{E}\left\|\tilde{\Phi}_{i,m}(Q_0) - \tilde{\Phi}_{i,m}(T)\right\|_2^{\eta} \leq Cr^m \|Q_0 - T\|_2^{\eta} \tag{5.6}$$

holds for all $i \in \mathbb{Z}$, $m \in \mathbb{N}$ and all $T \in \mathcal{T}$.

Lemma 5.2.2. *Assume the parameters of the model* (5.3) *satisfy the Assumption 5.2.1. Then for all $T \in \mathcal{T}$, $\tilde{T}_i := \lim_{m\to\infty} \tilde{\Phi}_{i,m}(T) \in \mathcal{T}$ exists almost surely and does not depend on T. In addition, $\tilde{T}_i$ is a stationary solution to the following system of stochastic transport equations:*

$$T_i = T_{\epsilon_i} \circ S \circ [\alpha T_{i-1}], \quad i \in \mathbb{Z},$$

and is unique almost surely.

Remark 5.2.3. *Zhu and Müller [80] proposed a specific parameter condition for their model that ensures Assumption 5.2.1 is satisfied. We provide a similar sufficient condition for the parameters of Model* (5.3) *that also guarantees the satisfaction of Assumption 5.2.1. Let L_ϵ be constant such that $\mathbb{E}|T_\epsilon(x) - T_\epsilon(y)|^2 \leq L_\epsilon^2|x-y|^2$. Assuming $\alpha \geq 0$, if $|S(x) - S(y)| \leq L_S|x-y|$ and $\alpha L_S L_\epsilon < 1$, then Model* (5.3) *satisfies Assumption 5.2.1 with $\eta = 2$ and $r = \sqrt{\alpha L_S L_\epsilon}$. Similarly, if $\alpha < 0$, suppose the aforementioned conditions are met and define $\mathcal{T}_{l,u} = \{T \in \mathcal{T} : 0 < L_l \leq T' \leq L_u < \infty\}$ and assume $\{T_i\} \subset \mathcal{T}_{l,u} \subset \mathcal{T}$ (see Lemma 5.5.1). Then Model* (5.3) *also satisfies Assumption 5.2.1 with $\eta = 2$ and $r = \sqrt{\alpha L_S L_\epsilon}$.*

5.2.5 Estimation and Statistical Analysis

We consider a time series of continuous distributions $\mu_i \in \mathcal{W}_2(\Omega)$ and corresponding time series $T_i \in \mathcal{T}$, which are related by one of the models from section 3.2. Although the methods to obtain T_i may differ for each model, we can always obtain T_i by observing μ_i. Our analysis is thus applicable to all three models studied, but in each different model, the T_i will represent a different feature of the distributional time series. For the remainder of our analysis, we assume that T_i is a (the) stationary solution obtained from system (5.3).

As discussed in Section 5.2.3, when S is fixed a priori to be the identity, our iteration (5.3) will reduce to that of Zhu and Müller [80]. In this simplified setting, Zhu and Müller [80] use the fact that α is the minimizer of $\mathbb{E}\,\|T_{i+1} - [\alpha T_i]\|_2^2$ to obtain a closed form expression for α as

$$\frac{\displaystyle\int_\Omega \mathbb{E}[(T_{i+1}(x) - x)(T_i(x) - x)]\,\mathrm{d}x}{\displaystyle\int_\Omega \mathbb{E}[(T_i(x) - x)^2]\,\mathrm{d}x}$$

when $\alpha \in [0, 1)$ or

$$\frac{\displaystyle\int_\Omega \mathbb{E}[(T_{i+1}(x) - x)(x - T_i^{-1}(x))]\,\mathrm{d}x}{\displaystyle\int_\Omega \mathbb{E}[(x - T_i^{-1}(x))^2]\,\mathrm{d}x}$$

when $\alpha \in (-1, 0)$. These show that α can be interpreted as the autocorrelation coefficient, and can be estimated by its empirical version, which allows for a straightforward path to consistency and parametric rates of convergence.

However, our more general iteration (5.3), involves an arbitrary non-decreasing map S that also needs to be estimated. Consequently, not only are there no closed forms for the estimands (α, S) but the estimation problem becomes distinctly non-linear.

To motivate our estimators, we note that if S were known, then α could be estimated by non-linear least squares, as the minimiser of $\frac{1}{N}\sum_{i=1}^N \|S \circ [\alpha T_{i-1}] - T_i\|_2^2$. On the other hand, if α were known, then a natural candidate to estimate S would be the ergodic average

$$S_{N,\alpha} := \frac{1}{N}\sum_{j=1}^N T_j \circ [\alpha T_{j-1}]^{-1}.$$

This is because the definition of the iteration $T_j = f(T_{j-1}; T_{\epsilon_i} \circ S) = T_{\epsilon_i} \circ S \circ [\alpha T_{j-1}]$, combined with the assumption that $\mathbb{E}[T_{\epsilon_j}(x)] = x$, yields that

$$\mathbb{E}\{T_j \circ [\alpha T_{j-1}]^{-1}\} = \mathbb{E}\{T_{\epsilon_j} \circ S\} = S.$$

Since $S_{N,\alpha}$ is available in closed form for any choice of α, this suggests plugging the expression $S_{N,\alpha}$ for S into the sum of squares, to obtain an objective that depends only on α. Minimising the said objective over α one obtains an estimator $\hat{\alpha}$, which automatically induces an estimator of S in the form of $S_{N,\hat{\alpha}}$.

Formally, we define the estimators $(\hat{\alpha}_N, S_{N,\hat{\alpha}_N})$ of (α, S) as follows:

$$\hat{\alpha}_N := \arg\min_{\alpha} M_N(\alpha), \tag{5.7}$$

where

$$\begin{aligned} M_N(\alpha) &:= \frac{1}{N}\sum_{i=1}^{N} g_\alpha(T_{i-1}, T_i, S_{N,\alpha}) \\ g_\alpha(T_{i-1}, T_i, S) &:= \|S \circ [\alpha T_{i-1}] - T_i\|_2^2 \\ S_{N,\alpha} &:= \frac{1}{N}\sum_{j=1}^{N} T_j \circ [\alpha T_{j-1}]^{-1}. \end{aligned} \tag{5.8}$$

To analyse the behaviour of our estimators, we also define the following population quantities:

$$\begin{aligned} S_\alpha &:= \mathbb{E}[T_j \circ [\alpha T_{j-1}]^{-1}] \\ M(\alpha) &:= \mathbb{E} g_\alpha(T_{i-1}, T_i, S_\alpha). \end{aligned} \tag{5.9}$$

The left-hand sides do not depend on j due to stationarity, which will be assumed throughout.

For the sake of clarity, we will henceforth denote the true parameters of the model using boldface fonts, namely as $(\boldsymbol{\alpha}, \mathrm{S})$.

Theorem 5.2.4. *If the true parameters of the model are* $(\boldsymbol{\alpha}, \mathrm{S})$*, then* $S_{\boldsymbol{\alpha}} = \mathrm{S}$.

Proof. For the true $\boldsymbol{\alpha}$, we have

$$\begin{aligned} S_{\boldsymbol{\alpha}} &= \mathbb{E}[T_j \circ [\boldsymbol{\alpha} T_{j-1}]^{-1}] \\ &= \mathbb{E}[T_{\epsilon_j} \circ \mathrm{S} \circ [\boldsymbol{\alpha} T_{j-1}] \circ [\boldsymbol{\alpha} T_{j-1}]^{-1}] \\ &= \mathbb{E}[T_{\epsilon_j} \circ \mathrm{S}] = \mathrm{S} \end{aligned} \tag{5.10}$$

□

We show the consistency of the estimators $(\hat{\alpha}_N, S_{N,\hat{\alpha}_N})$ in the following 4 steps corresponding to the lemmas 5.2.5, 5.2.7, 5.2.8 and Theorem 5.2.9 respectively:

- $\boldsymbol{\alpha}$ is the unique minimizer of $M(\alpha)$.

- $S_{N,\alpha}$ converges uniformly (with respect to α) in probability to S_α in L^2.

- $M_N(\alpha)$ converges uniformly in probability to $M(\alpha)$.

- we conclude the consistency (and identifiability) using the M-estimation theory.

Lemma 5.2.5. *(Unique Minimizer of $M(\alpha)$) For any $\alpha \neq \boldsymbol{\alpha}$ we have*

$$M(\boldsymbol{\alpha}) = \mathbb{E}g_{\boldsymbol{\alpha}}(T_{i-1}, T_i, S_{\boldsymbol{\alpha}}) < \mathbb{E}g_\alpha(T_{i-1}, T_i, S_\alpha) = M(\alpha),$$

where $\boldsymbol{\alpha}$ is the true α.

Now we show that $S_{N,\alpha}$ converges to S_α in probability for any α and also prove a central limit theorem (CLT) for $S_{N,\alpha}$.

If $\alpha = \boldsymbol{\alpha}$, then it is straightforward to argue that $S_{N,\boldsymbol{\alpha}}$ converges to $S_{\boldsymbol{\alpha}}$: first note that for any $x \in [0,1]$, the strong law of large numbers yields that

$$S_{N,\boldsymbol{\alpha}}(x) = \frac{1}{N}\sum_{j=1}^{N} T_j \circ [\boldsymbol{\alpha} T_{j-1}]^{-1}(x) = \frac{1}{N}\sum_{j=1}^{N} T_{\epsilon_j} \circ \mathrm{S}(x) \longrightarrow \mathbb{E}[T_{\epsilon_j} \circ \mathrm{S}](x) = \mathrm{S}(x).$$

Therefore the terms in the expression are independent and identically distributed with mean S. From Theorem 5.2.4, we know that the true $\mathrm{S} = S_{\boldsymbol{\alpha}}$. Therefore in this case that $\alpha = \boldsymbol{\alpha}$, $S_{N,\alpha}$ converges in probability to $S_{\boldsymbol{\alpha}} = \mathrm{S}$. However, in general, when $\alpha \neq \boldsymbol{\alpha}$ the terms $T_j \circ [\alpha T_{j-1}]^{-1}$ are not independent for different j. Therefore, we first show that since $\{T_j\}$ satisfies the moment generating condition, we can quantify the dependency between the terms in the sequence $T_j \circ [\alpha T_{j-1}]^{-1}$ and apply CLT methods developed for functional time series.

Lemma 5.2.6. *A sequence $\{T_n\}_{i=-\infty}^{\infty}$ that satisfies the geometric moment contracting condition (5.2.1) for $\eta \geq 2$, also satisfies the conditions (1.1),(1.2),(2.1) and (2.2) of Horváth et al. [33]. Namely, assume*

$$T_n = f(\epsilon_n, \epsilon_{n-1}, \cdots),$$

where $\{\epsilon_i'\}$ is an independent copy of $\{\epsilon_i\}$ defined in the same probability space. Then, letting

$$T_{n,m}' = f(\epsilon_n, \epsilon_{n-1}, \cdots, \epsilon_{n-m+1}, \epsilon_{n-m}', \cdots), \tag{5.11}$$

for any $0 < \delta < 1$ *we have*

$$\sum_{m=1}^{\infty} (\mathbb{E} \left\| T_n - T'_{n,m} \right\|_2^2)^{1/2} < \infty. \tag{5.12}$$

Lemma 5.2.7. *(Central limit for* $S_{N,\alpha}$*) Suppose the parameters of the iteration* (5.3) *satisfy the Assumption 5.2.1. Then for any* α*, there is a Gaussian process* Γ_α *such that*

$$\sqrt{N}(S_{N,\alpha} - S_\alpha) \xrightarrow{d} \Gamma_\alpha, \quad \text{in } L^2.$$

Also,

$$\sup_\alpha \left\| S_{N,\alpha} - S_\alpha \right\|_2 = o_{\mathbb{P}}(1)$$

Lemma 5.2.8. *Suppose the parameters of the iteration* (5.3) *satisfy the Assumption 5.2.1. Then for any* α*, there is a* $\sigma_\alpha \geq 0$ *such that*

$$\sqrt{N}[M_N(\alpha) - M(\alpha)] \to N(0, \sigma_\alpha^2).$$

Moreover,

$$\sup_\alpha |M_N(\alpha) - M(\alpha)| = o_{\mathbb{P}}(1).$$

Theorem 5.2.9. *(Identifiability and Consistency) Under Assumption 5.2.1, the parameters of the iteration* (5.3) *are identifiable and* $(\hat{\alpha}_N, S_{N,\hat{\alpha}_N})$ *are consistent estimators for* $(\boldsymbol{\alpha}, \mathrm{S})$.

Theorem 5.2.10. *(Rate of Convergence) Let* $\mathcal{T}_{l,u} = \{T \in \mathcal{T} : 0 < L_l \leq T' \leq L_u < \infty\}$ *and suppose* $\{T_i\} \subset \mathcal{T}_{l,u} \subset \mathcal{T}$*. Under Assumption 5.2.1 and twice differentiability of the* T_i*, we have*

$$N^{\frac{1}{2}} |\hat{\alpha}_N - \boldsymbol{\alpha}| = O_{\mathbb{P}}(1),$$
$$N^{\frac{1}{2}} \left\| S_{N,\hat{\alpha}_N} - \mathrm{S} \right\|_2 = O_{\mathbb{P}}(1).$$

5.3 Simulation Experiments

In this section, we probe the behaviour of our models, and the finite sample performance of our estimation framework, via simulation. To generate the noise maps T_{ϵ_i}, we use the class of random optimal maps introduced in Section 3.3.

Each plot in Figure 5.1 corresponds to a time series simulation with a different combination of S and $\boldsymbol{\alpha}$. Each column corresponds to a different value of $\boldsymbol{\alpha} \in \{-0.9, -0.5, 0, 0.5, 0.9\}$ from left to right. In the three top rows, S is chosen to be ζ_K (see Section 3.3 for the definition) for $K = \{-6, -4, -2\}$ from top to bottom. In row four, S is the average of ζ_1 and an instance of T_ϵ. Rows five and six exemplify the method on non-differentiable and discontinuous maps S respectively.

Plots that fall within the bounding red rectangle correspond to settings where our theory is guaranteed to apply. Plots outside of that rectangle are not guaranteed to be covered by our theory: they either distinctly violate our assumptions (such as the last row where the true map is not continuous, as required) or we cannot confirm whether the assumption 5.2.1 holds true. Starting from the identity map, we generate a time series with 300 iterations and discard the first 100 maps of the series. The remaining 200 maps $\{T_i\}$ are shown in light blue, the true map S is in dark blue, and the estimated map is in orange. For each time series, we show the estimated $\hat{\alpha}$ and the error between the estimator and true map in $||.||_2$-norm.

As expected from Remark 5.2.3, smaller values of $|\boldsymbol{\alpha}|$ lead to time series which apparently oscillate around the mean of the stationary time series, which in turn leads to the convergence of our estimator with respect to the true map. In particular, good agreement is seen between the estimator and true map for values of $|\boldsymbol{\alpha}|$ up to 0.5 at least, only noticeably failing for $\boldsymbol{\alpha} = 0.5$ in the discontinuous map case (where our theoretical guarantee does not apply due to the discontinuity).

Larger values of $|\boldsymbol{\alpha}|$ can still lead to similar stationary state time series (sometimes even outside of the red rectangle, where our theoretical guarantees apply) but with naturally larger oscillations. Still, a good agreement between the estimator and ground truth is observed. This can depend on the choice of map S and the precise value of $\boldsymbol{\alpha}$. For instance, in the third, fourth, and fifth rows, when $\boldsymbol{\alpha} = -0.9$. In the remaining rows of the first column, the stationary state behavior changes to a period-two time series (with noise) where the maps oscillate alternatively between two maps related by inversion (recall that negative values of $\boldsymbol{\alpha}$ imply an inversion of the map T_{i-1} at each time step). Nevertheless, the estimator is able to capture features of the S map that are not visible in the time series itself: notably, the discontinuous step in row six is present in the estimated map.

In the other extreme of $\boldsymbol{\alpha} = 0.9$, the time series maps are close to step-like functions with some variation in the step height. The maps are in fact still oscillating around the mean of the stationary time series that is very close to the step-like map S^{∞}, which is the mean of the solution to the model (5.3) when $\alpha \longrightarrow 1$, that is $T_s = \mathrm{S} \circ T_s$. However, the performance of the estimator is the worst in this limit.

Do note that the family of maps $\zeta_K(x)$ is not symmetric with respect to inversion in the sense that the derivative of $\zeta_K(x)$ is 0 at some fixed points ($\zeta_K(x) = x$) but is never infinite, and therefore the random maps T_ϵ, which are derived from $\zeta_K(x)$, are biased in this way. For this reason, the vertical variance observed in most maps is much more pronounced than the horizontal one, which is very clear in the case $\boldsymbol{\alpha} = 0.9$.

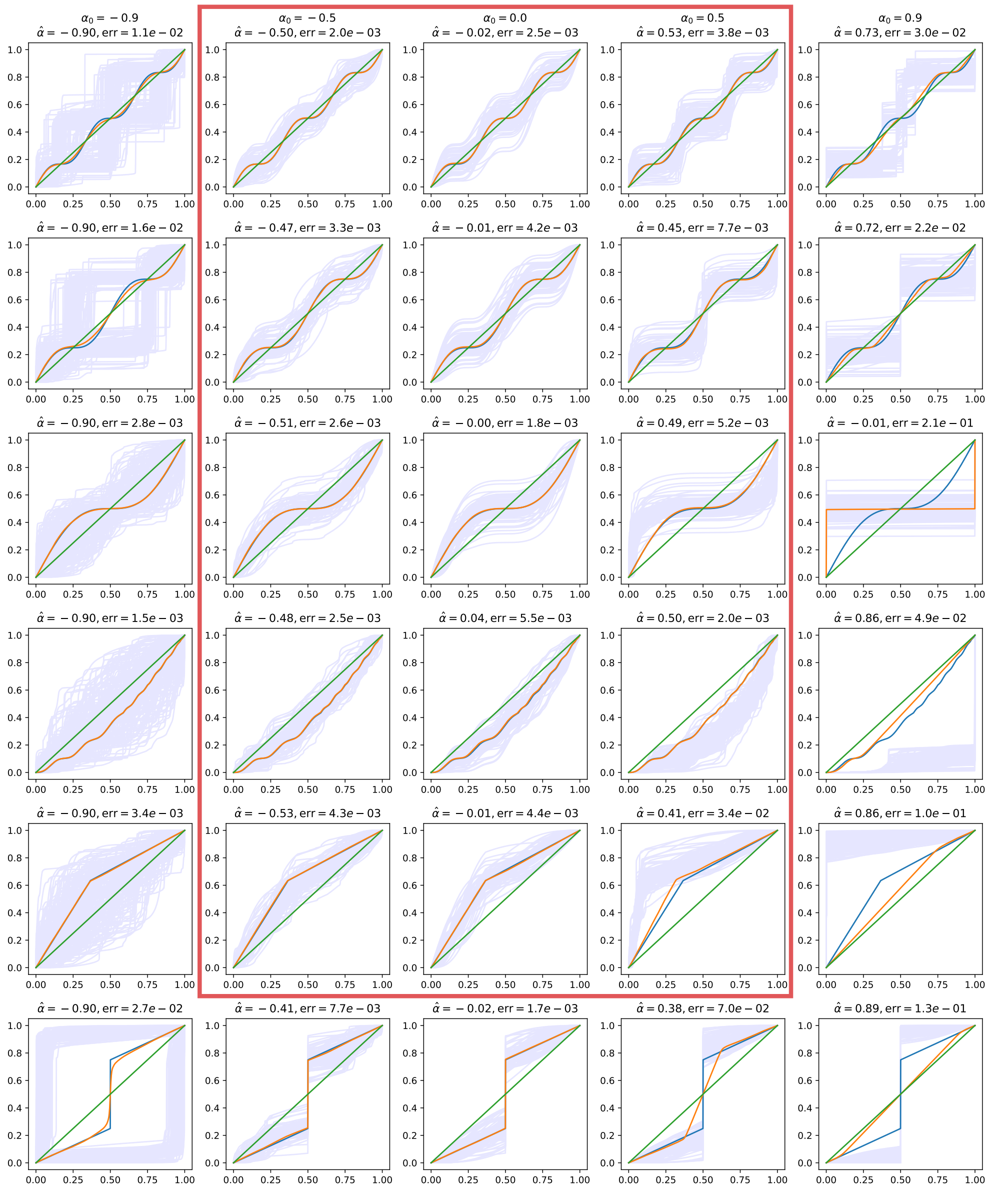


Figure 5.1: Estimated map (orange) versus the true map (blue) for different combinations of $\boldsymbol{\alpha}$ and S. The light blue line represents the simulated time series, while the green line represents the id map. Each column corresponds to a different value of $\boldsymbol{\alpha}$, ranging from -0.9 on the left to 0.9 on the right. The top three rows show results for $\mathrm{S} = \zeta_K$ where K is chosen from $\{-6, -4, -2\}$ from top to bottom. In the fourth row, S is the average of ζ_1 and an instance of T_ϵ. The fifth and sixth rows demonstrate the method on non-differentiable and discontinuous maps, respectively. The cases within the red rectangle are covered by our theoretical guarantees.

5.4 Illustrative Data Analysis

In this section, we consider the distribution of minimum daily temperatures recorded in the summer of the years from 1960 to 2020 from several airports in the USA (available at `www.ncei.noaa.gov`). That is, the years are taken as the time index, and for any given time index we observe a distribution over the temperature scale (representing the distribution of minimal temperatures over that year's summer). Thus, each airport gives rise to a distributional time series. This data set has been also analysed by Zhu and Müller [80] to demonstrate their own distributional autoregressive model, which allows for constructive comparison.

We examine the daily minimum temperature for June, July, August, and September from 1960 to 2020 in four locations: Chicago O'Hare International Airport, Atlanta Hartsfield-Jackson International Airport, Phoenix Airport, and New Orleans Airport. The corresponding distributions are displayed in Figure 5.2.

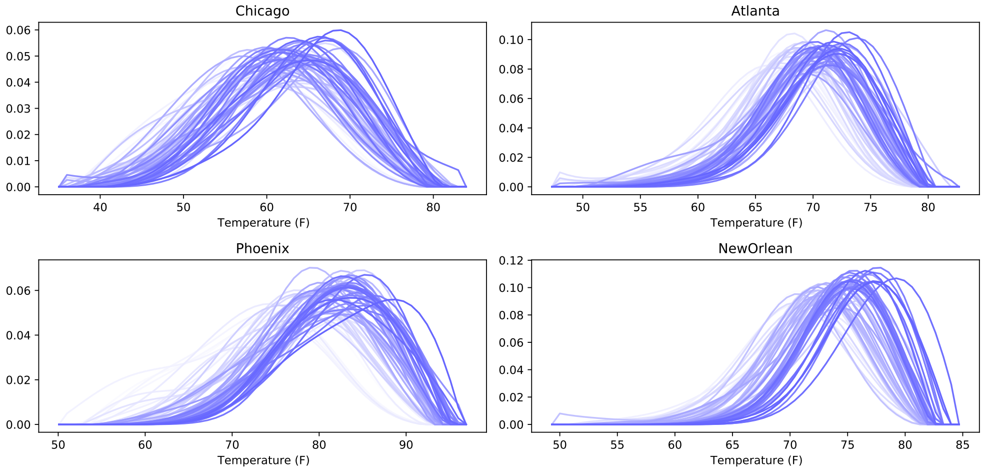
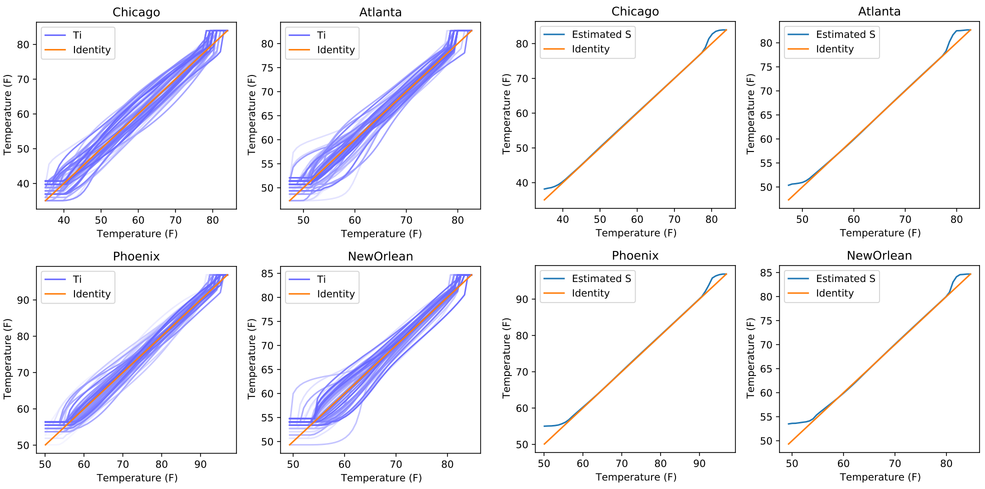


Figure 5.2: Time series of distribution of daily minimum temperature in summer from 1960 to 2020 at Chicago Ohare international airport, Atlanta Hartsfield Jackson international airport, Phoenix airport, and New Orleans airport. The shading reflects the time index: the fainter the curve, the earlier in time it corresponds to.

The map sequence elicited by adopting the increment model (Model (I)) is shown in Figure 5.3a. These maps are obtained by calculating the optimal maps between consecutive annual temperature distributions for each location. These maps exhibit oscillations around the identity, except in the subdomains corresponding to extreme temperature values. In the lower extreme, the maps impose a cutoff on the lower end of the support of the temperature distribution, while the higher end is pushed towards higher values and eventually reaches the extreme of the support. This implies that extreme temperatures are increasing, indicating that the coldest and hottest nights in summer are becoming hotter.

Figure 5.3b presents the estimates of S obtained using Model (I), where the estimated α was found to be 0 up to three decimal points for all airports. This suggests that the optimal maps T_i are independent from each other and, on average, they are equal to the estimated maps S presented. The estimated S maps are very similar across all airports, effectively being the identity map in the middle portion of the support and above the identity at the extreme points.

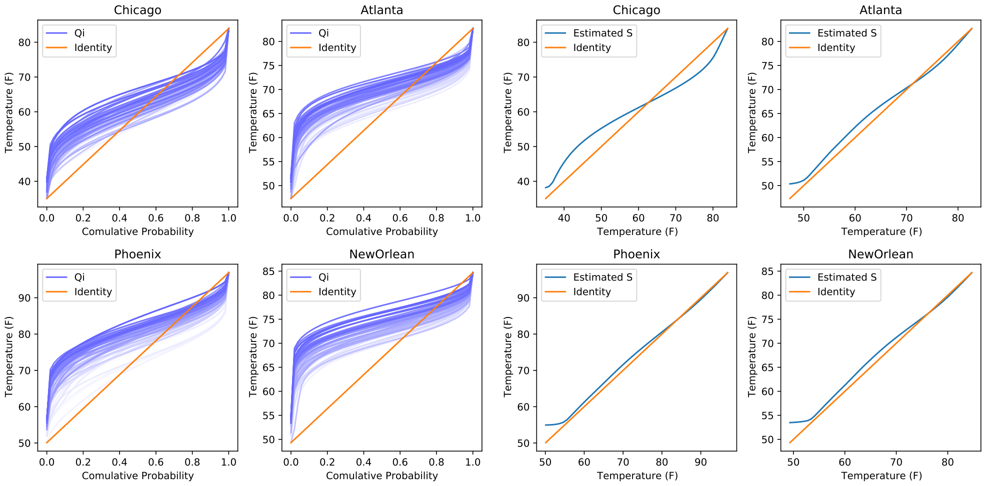


(a) (b)

Figure 5.3: (a): Time series $\{T_i\}$ (blue) based on model (I) and identity map (orange) for the four locations. (b): Estimated map S (blue) and identity map (orange). Faint blue shading corresponds to early years, and bold shading to later years.

Examining the maps generated by fitting Model (I), i.e. computing the optimal maps between consecutive annual distributions in Figure 5.3a, we can observe an increasing trend in the cutoff value of the lower endpoint over time. This implies that the time series of optimal maps may not be stationary. Of course, the S maps are not able to capture the overall increase in the cutoff value of the lower end over time: the plateaus of the S maps are just the averages of the optimal maps T_i and don't show this trend. Indeed, a problem of modeling such data is that the system may be dynamically evolving due to factors like global warming, and it is not obvious a priori if stationary regimes exist that can be captured by our models.

However, using the uniform quantile model (Model (UQ)), the resulting maps are more interpretable and reveal more refined dynamics beyond the cutoffs at the extremes. To obtain these maps, we fitted iteration (5.3) to the time series of quantile functions of the temperature distributions. The quantile functions are shown in figure 5.4a. The resulting estimated maps S are in figure 5.4b, and the estimated $\hat{\alpha}$ for the four airports are $\{0.39, 0.80, 0.89, 0.89\}$. All the maps show a cutoff at the lower end and a fixed point in the second half of the support where the derivative is smaller than 1. The fixed point implies a point of stability, and the derivative means there is a trend towards a concentration of weight around this point, that is, if we start the time series at a Gaussian-like distribution of mean different from the S fixed point, the distributions in the time series will progress towards Gaussian-like distributions of mean approaching the fixed point. Again, the model may be failing to capture a trend of ever-increasing temperature, or it may be implying a stabilization at temperatures given by the fixed points, which will become the new norm.



(a)

(b)

Figure 5.4: (a): Time series of quantile functions (blue) based on model (UQ) and identity map (orange) for the four locations. (b): Estimated map S (blue) and identity map (orange). Faint blue shading corresponds to early years, and bold shading to later years.

Even if the model is possibly misspecified, the estimated maps S are still able to condense several features of the time series of distributions. Namely, the reduction of extreme cold events and the progression toward higher modal temperatures which may or may not be static.

There is an interesting observation to be made given that the estimated α when fitting the intercept model (I) is numerically 0 while it is in (0,1) when fitting the quantile model (UQ). Specifically, in combination, these results suggest that the quantile model is, in a certain sense, a better fit to the data. The reasoning is as follows. Recall that the increment model (I) with $\alpha = 0$ is equivalent to the quantile model (UQ) when $\alpha = 1$, and corresponds to “trivial dynamics” (random walk). Therefore, whenever fitting model (I) results in an estimated α that is nearly zero, then the best fitting model of type (I) is in fact a (UQ) model. In which case we have evidence to prefer a (UQ) modeling approach instead, which will correspond to non-trivial dynamics. Conversely, if fitting model (UQ) yields an estimated α near 1, it may be preferable to use model (I) instead.

5.5 Proofs

Proof of Lemma 5.2.2. The proof is directly analogous to that of Theorem 2 in Wu and Shao [76] and theorem 1 in Zhu and Müller [80]. □

Proof of Lemma 5.2.5. We prove the theorem in the following 4 steps:

1. Given a function $f \in L^2$, and a random function ϵ such that $\mathbb{E}(\epsilon) = \mathrm{id}$, we can show that
$$\arg\min_h \mathbb{E}_\epsilon \|h - \epsilon \circ f\|_2^2 = f.$$
To do so, we can apply Fubini’s theorem and rewrite the expression as follows:
$$\int\int |h(x) - \epsilon(f(x))|^2 \,\mathrm{d}x \,\mathrm{d}\epsilon = \int\int |h(x) - \epsilon(f(x))|^2 \,\mathrm{d}\epsilon \,\mathrm{d}x$$
Since $\mathbb{E}_\epsilon[\epsilon(f(x))] = f(x)$ for any x, the minimizer of the inner integral on the left-hand side is $h(x) = \mathbb{E}[\epsilon(f(x))] = f(x)$.

2. We will now demonstrate that for any fixed T_i and T_{i-1}, as well as for all α, the following inequality holds:
$$\mathbb{E}_\epsilon[g_{\boldsymbol{\alpha}}(T_{i-1}, T_i, S_{\boldsymbol{\alpha}})] \le \mathbb{E}_\epsilon[g_\alpha(T_{i-1}, T_i, S_\alpha)].$$
Let us define $f(\alpha, T) = S_\alpha \circ [\alpha T]$. Note that for all indices i, we have
$$\begin{aligned} g_\alpha(T_i, T_{i-1}, S_\alpha) &= \left\|S_\alpha \circ [\alpha T_{i-1}] - T_{\epsilon_i} \circ S_{\boldsymbol{\alpha}} \circ [\boldsymbol{\alpha} T_{i-1}]\right\|_2^2 \\ &= \left\|f(\alpha, T_{i-1}) - T_{\epsilon_i} \circ f(\boldsymbol{\alpha}, T_{i-1})\right\|_2^2. \end{aligned} \tag{5.13}$$
Using the result from part 1 and the equation (5.13), we can conclude that if there exists an α such that $f(\alpha, T_{i-1})$ minimizes the expression $\mathbb{E}_\epsilon[g_{\boldsymbol{\alpha}}(T_{i-1}, T_i, S_{\boldsymbol{\alpha}})]$ in equation 5.13, then we must have $f(\alpha, T_{i-1}) = f(\boldsymbol{\alpha}, T_{i-1})$. Therefore, we obtain the desired inequality.

3. We now aim to prove that for any α, we have

$$\mathbb{E}[g_{\boldsymbol{\alpha}}(T_{i-1}, T_i, S_{\boldsymbol{\alpha}})] \leq \mathbb{E}[g_{\alpha}(T_{i-1}, T_i, S_{\alpha})].$$

We start by denoting by π the marginal distribution of T_i, and Q the marginal distribution of the pair (T_{i-1}, T_i). Then, we can express the expectation of $g_\alpha(T_{i-1}, T_i, S_\alpha)$ as follows:

$$\mathbb{E}_Q[g_\alpha(T_{i-1}, T_i, S_\alpha)] = \mathbb{E}_\pi[\mathbb{E}_\epsilon g_\alpha(T_{i-1}, T_i, S_\alpha)|T_{i-1}].$$

By using part 2 of the proof, we know that $\boldsymbol{\alpha}$ is a minimizer for the inner expectation of the right-hand side, i.e.,

$$\mathbb{E}_\epsilon\{g_{\boldsymbol{\alpha}}(T_{i-1}, T_i, S_{\boldsymbol{\alpha}})|T_{i-1}\} \leq \mathbb{E}_\epsilon\{g_\alpha(T_{i-1}, T_i, S_\alpha)|T_{i-1}\},$$

and this for all T_{i-1}. Therefore, taking the expectation over T_{i-1}, we get

$$\mathbb{E}_Q[g_{\boldsymbol{\alpha}}(T_{i-1}, T_i, S_{\boldsymbol{\alpha}})] \leq \mathbb{E}_Q[g_\alpha(T_{i-1}, T_i, S_\alpha)].$$

4. Finally we can conclude that $\boldsymbol{\alpha}$ is the unique minimizer of $M(\alpha)$. Suppose there exists an α such that $\mathbb{E}[g_{\boldsymbol{\alpha}}(T_{i-1}, T_i, S_{\boldsymbol{\alpha}})] = \mathbb{E}[g_\alpha(T_{i-1}, T_i, S_\alpha)]$. Using parts 2 and 3, we can deduce that for each fixed T_i, T_{i-1}, $\mathbb{E}_\epsilon[g_{\boldsymbol{\alpha}}(T_{i-1}, T_i, S_{\boldsymbol{\alpha}})] = \mathbb{E}_\epsilon[g_\alpha(T_{i-1}, T_i, S_\alpha)]$. Then using equation 5.13 we can conclude that, for all indices j,

$$S_\alpha \circ [\alpha T_j] = S_{\boldsymbol{\alpha}} \circ [\boldsymbol{\alpha} T_j].$$

If $S_\alpha \circ \alpha T_j = S_{\boldsymbol{\alpha}} \circ [\boldsymbol{\alpha} T_j]$ for all j, we can deduce $S_\alpha = S_{\boldsymbol{\alpha}} \circ [\boldsymbol{\alpha} T_j] \circ [\alpha T_j]^{-1}$ for all j. However, note that while S_α is deterministic, the right-hand side is deterministic (and not random) if and only if $\alpha = \boldsymbol{\alpha}$. This is because if $\alpha \neq \boldsymbol{\alpha}$, then the right-hand side depends on T_j, which is a random variable.

□

Lemma 5.5.1. *For any $T, S \in \mathcal{T}$ we have $\left\|T^{-1} - S^{-1}\right\|_2 \lesssim \sqrt{\|T - S\|_2}$. Moreover, let $\mathcal{T}_{l,u} = \{T \in \mathcal{T} : 0 < L_l \leq T' \leq L_u < \infty\}$. For any $T, S \in \mathcal{T}_{l,u} \subset \mathcal{T}$ we have $\left\|T^{-1} - S^{-1}\right\|_2 \lesssim \|T - S\|_2$. In summary, there exists $b \in [\frac{1}{2}, 1]$ such that $\left\|T^{-1} - S^{-1}\right\|_2 \lesssim \|T - S\|_2^b$ for any $T, S \in \mathcal{T}$.*

Proof. Let $T, S \in \mathcal{T}$. For some constant C', we have: $\left\|T^{-1} - S^{-1}\right\|_2 \leq C' \left\|T^{-1} - S^{-1}\right\|_1$, because the functions are bounded. Moreover, $\left\|T^{-1} - S^{-1}\right\|_1 = \|T - S\|_1$. And, finally, by applying the Cauchy-Schwarz inequality, we get $\|T - S\|_1 \leq C\sqrt{\|T - S\|_2}$, where C is a constant. Therefore, we conclude $\left\|T^{-1} - S^{-1}\right\|_2 \leq C\sqrt{\|T - S\|_2}$.

When $T, S \in \mathcal{T}_{l,u}$ we can write

$$\begin{aligned}
\left\|T^{-1} - S^{-1}\right\|_2^2 &= \int_0^1 |T^{-1}(x) - S^{-1}(x)|^2 \,\mathrm{d}x \\
&= \int_0^1 |T^{-1} \circ S(y) - y|^2 S'(y) \,\mathrm{d}y \qquad (S^{-1}(x) = y) \\
&\leq L_u \int |T^{-1} \circ S(y) - y|^2 \,\mathrm{d}y \\
&\leq L_u \int |z - S^{-1} \circ T(z)|^2 \frac{1}{S'(S^{-1} \circ T(z))} T'(z) \,\mathrm{d}z \qquad (T^{-1} \circ S(y) = z) \\
&\leq L_u \frac{L_u}{L_l} \int |z - S^{-1} \circ T(z)|^2 \,\mathrm{d}z \\
&\leq L_u \frac{L_u}{L_l} \frac{1}{L_l} \int |S(z) - T(z)|^2 \,\mathrm{d}z \qquad (\forall x, y \quad |x - y| \leq \frac{1}{L_l} |S(x) - S(y)|) \\
&\leq \frac{L_u^2}{L_l^2} \|S - T\|_2^2 .
\end{aligned} \tag{5.14}$$

□

Lemma 5.5.2. *There exists a constant $\frac{1}{2} \leq b \leq 1$ such that the following inequalities hold:*

$$\left\|S_{\alpha_1} - S_{\alpha_2}\right\|_2 \lesssim |\alpha_1 - \alpha_2|^b,$$

and

$$\left\|S_{N,\alpha_1} - S_{N,\alpha_2}\right\|_2 \lesssim |\alpha_1 - \alpha_2|^b,$$

and

$$g_{\alpha_1}(T_{i-1}, T_i, S_{N,\alpha_1}) - g_{\alpha_2}(T_{i-1}, T_i, S_{N,\alpha_2}) \leq C(T_i)|\alpha_1 - \alpha_2|^b,$$

where $\frac{1}{n} \sum_i \mathbb{E}[C(T_i)] = O(1)$.

Define $\mathcal{T}_{l,u} = \{T \in \mathcal{T} : 0 < L_l \leq T' \leq L_u < \infty\}$. If $\{T_i\} \subset \mathcal{T}_{l,u}$, then $b = 1$ in the above inequalities.

Proof. To begin with, it should be noted that given any two real numbers $\alpha_1, \alpha_2 \in (-1, 1)$ with the same sign, and for any given map T, we have the following inequality:

$$\|[\alpha_1 T] - [\alpha_2 T]\|_2 \leq C|\alpha_1 - \alpha_2|,$$

where C is a constant. In fact, it suffices to consider the definition of $[\alpha T]$ for the cases when $\alpha \geq 0$ and $\alpha < 0$ separately. Using Lemma 5.5.1 we can write that for some $\frac{1}{2} \leq b \leq 1$,

$$\begin{aligned} \left\|S_{\alpha_1} - S_{\alpha_2}\right\|_2 &= \left\|\mathbb{E}[T_j \circ [\alpha_1 T_{j-1}]^{-1}] - \mathbb{E}[T_j \circ [\alpha_2 T_{j-1}]^{-1}]\right\|_2 \\ &\leq LC'|\alpha_1 - \alpha_2|^b, \end{aligned} \tag{5.15}$$

where L is the common Lipschitz constant for all T_j. Similarly

$$\left\|S_{N,\alpha_1} - S_{N,\alpha_2}\right\|_2 \leq L|\alpha_1 - \alpha_2|^b,$$

We now proceed to show that both $S_{N,\alpha}$ and S_α are Lipschitz functions of x. To do this, we observe that the inverse of a Lipschitz function is Lipschitz, and also the composition of two Lipschitz functions is Lipschitz. Since all T_j are Lipschitz and $S_{N,\alpha}$ and S_α are defined as compositions, they are also Lipschitz with respect to x.

We will now show that g is Lipschitz function of α:

$$\begin{aligned} &g_{\alpha_1}(T_{i-1}, T_i, S_{N,\alpha_1}) - g_{\alpha_2}(T_{i-1}, T_i, S_{N,\alpha_2}) \\ &\qquad = \left\|S_{N,\alpha_1} \circ [\alpha_1 T_i] - T_{i+1}\right\|^2 - \left\|S_{N,\alpha_2} \circ [\alpha_2 T_i] - T_{i+1}\right\|_2^2 \\ &\qquad \lesssim \left\|S_{N,\alpha_1} \circ [\alpha_1 T_i] - S_{N,\alpha_2} \circ [\alpha_2 T_i]\right\|_2 \\ &\qquad \lesssim \left\|S_{N,\alpha_1} \circ [\alpha_1 T_i] - S_{N,\alpha_1} \circ [\alpha_2 T_i]\right\|_2 \\ &\qquad\quad + \left\|S_{N,\alpha_1} \circ [\alpha_2 T_i] - S_{N,\alpha_2} \circ [\alpha_2 T_i]\right\|_2 \\ &\qquad \leq D(T_i)|\alpha_1 - \alpha_2| + C(T_i)|\alpha_1 - \alpha_2|^b \\ &\qquad \lesssim C(T_i)|\alpha_1 - \alpha_2|^b \end{aligned} \tag{5.16}$$

where $D(T_i)$ and $C(T_i)$ are constants that depend on T_i and $\sum_i \mathbb{E}[C(T_i)]/n = O(1)$. □

Proof of Lemma 5.2.6. Let $T_{n-m} = f(\epsilon_{n-m}, \epsilon_{n-m-1}, \cdots,)$ and $T'_{n-m} = f(\epsilon'_{n-m}, \epsilon'_{n-m-1}, \cdots,)$. Thus we can write $T_n = \Phi_{n,m}(T_{n-m})$ and $T'_{n,m} = \Phi_{n,m}(T'_{n-m})$.

$$\begin{aligned}
\sum_{m=1}^{\infty} (\mathbb{E}\left\|T_n - T'_{n,m}\right\|_2^2)^{1/2} &= \sum_{m=1}^{\infty} (\mathbb{E}\left\|\Phi_{n,m}(T_{n-m}) - \Phi_{n,m}(T'_{n-m})\right\|_2^2)^{1/2} \\
&\leq \sum_{m=1}^{\infty} (\mathbb{E}\left\|\Phi_{n,m}(T_{n-m}) - \Phi_{n,m}(Q_0)\right\|_2^2)^{1/2} \\
&\quad + (\mathbb{E}\left\|\Phi_{n,m}(T'_{n-m}) - \Phi_{n,m}(Q_0)\right\|_2^2)^{1/2} \\
\text{(Lyapunov's inequality)} \quad &\leq \sum_{m=1}^{\infty} (\mathbb{E}\left\|\Phi_{n,m}(T_{n-m}) - \Phi_{n,m}(Q_0)\right\|_2^{\eta})^{1/\eta} \\
&\quad + (\mathbb{E}\left\|\Phi_{n,m}(T'_{n-m}) - \Phi_{n,m}(Q_0)\right\|_2^{\eta})^{1/\eta} \\
\text{(Assumption 5.2.1)} \quad &\leq \sum_{m=1}^{\infty} Cr^{m/\eta}(\|T_{n-m} - Q_0\|_2 \vee \left\|T'_{n-m} - Q_0\right\|_2) \\
&< \infty
\end{aligned} \tag{5.17}$$

□

The following statement is virtually obvious, but is used multiple times in the proofs below and so is most easily quoted directly:

Lemma 5.5.3. *Let $\{X_i\}$ be a sequence of random variables, and suppose that $W_n := \sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu) \xrightarrow{d} W$ for some (almost surely finite) random variable W. Then, $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to μ.*

Proof. By Slutsky's Theorem, we get $n^{-1/2}W_n \xrightarrow{d} 0$, which also implies convergence in probability to zero. □

Proof of Lemma 5.2.7. We start by using Lemma 5.2.6 to conclude that the series $\{T_i - \mathbb{E}T_i\}_{i=-\infty}^{\infty}$ satisfies the assumptions (1.1),(1.2),(2.1) and (2.2) of Horváth et al. [33]. From this, we can argue that the series $\{T_j \circ [\alpha T_{j-1}]^{-1} - \mathbb{E}T_j \circ [\alpha T_{j-1}]^{-1}\}_{i=-\infty}^{\infty}$ also satisfies those assumptions and therefore we obtain the following central limit theorem for $S_{N,\alpha}$: for any α, there is a Gaussian process Γ_α such that

$$\sqrt{N}(S_{N,\alpha} - S_\alpha) \xrightarrow{d} \Gamma_\alpha, \quad \text{in } L^2.$$

Using the central limit theorem and Lemma 5.5.3, we can infer the convergence in probability of $S_{N,\alpha}$ to S_α for any α (in L^2). Since both S_α and $S_{N,\alpha}$ are globally Lipschitz with respect to α, in the sense of Lemma 5.5.2, we can use Corollary 3.1 of Newey [49] to obtain uniform convergence in probability:

$$\sup_\alpha \left\|S_{N,\alpha} - S_\alpha\right\|_2 \to 0 \quad \text{in probability.}$$

□

5.5.1 Overview of Wu and Shao [76]

In their work, Wu and Shao [76] investigated the properties of nonlinear time series expressed in terms of iterated random functions and established a central limit theorem for additive functionals of such systems. The construction involves a sequence of functions of the form $X_n(x) = F_{\theta_n} \circ F_{\theta_{n-1}} \circ \cdots F_{\theta_1}(x)$. The authors assume that X_n satisfies a geometric moment condition, which requires the existence of $\beta > 0$, $C = C(\alpha) > 0$, and $r = r(\alpha) \in (0,1)$ such that, for all $n \in N$,

$$\mathbb{E}\{\rho(X_n(X_0'), X_n(X_0))^\beta\} \leq Cr^n. \tag{5.18}$$

In addition, they define the l-dimensional vector $Y_i = (X_{i-l+1}, X_{i-l+2}, \cdots, X_i)$ and for any $\delta > 0$, they introduce the functional $\Delta_g(\delta)$ as

$$\Delta_g(\delta) = \sup\{\left\|[g(Y) - g(Y_1)]1_{\rho(Y,Y_1)} \leq \delta\right\| : \quad Y, Y_1 \quad \text{are identically distributed}\},$$

Where $\rho(.,.)$ is the product metric and is defined as

$$\rho(z,z') = \sqrt{\sum_{i=1}^{l} \rho(z_i, z_i')^2} \quad \text{for } z = (z_1, \cdots, z_l), z' = (z_1', \cdots, z_l').$$

Finally, the functional $S_{n,l}(g) = \sum_{i=1}^{n} g(X_{i-l+1}, X_{i-l+2}, \cdots, X_i)$ is defined. The authors establish the following central limit theorem for this functional:

Theorem 5.5.4. *(Wu and Shao [76, Theorem 3]) Assume that (5.18) holds, that $X_1 \sim \pi$, $E\{g(Y_1)\} = 0$, and $E\{|g(Y_1)|^p\} < \infty$ for some $p > 2$, and that*

$$\int_0^1 \frac{\Delta_g(t)}{t} < \infty. \tag{5.19}$$

Then there exists a $\sigma_g \geq 0$ such that, for π-almost x, $\{S_{\lfloor nu \rfloor, l}(g)/\sqrt{n}, 0 \leq u \leq 1\}$ conditional on $X_0 = x$, converges to $\sigma_g B$, where B is a standard Brownian motion.

A function that satisfies (5.19) is referred to as *stochastic Dini continuous.* Using Theorem 5.5.4 to derive a central limit theorem for M_N poses a problem: Theorem 5.5.4 uses fixed-length sub-sequences of the time series, i.e., $(X_{i-l+1}, X_{i-l+2}, \cdots, X_i)$, as arguments for the function g, however the arguments of the function g that appears in the expression of M_N in 5.7, include not only (T_{i-1}, T_i), but also $S_{N,\alpha}$, thus making it dependent on the entire time series. Therefore, Theorem 5.5.4 cannot be applied directly, and a modified version is required. We present a modified version of

Theorem 5.5.4 that is specifically tailored for functions of finite dimensional random variables, followed by another modification that is suitable for functionals of infinite dimensional variables.

Corollary 5.5.5. *(Modified version of Wu and Shao [76, Theorem 3] for finite dimensional arguments) Suppose $\overline{Z}_n$ is a measurable function of $(X_1, X_2, \cdots, X_n)$ such that $\overline{Z}_n$ converges in probability to some constant μ. Let $Y_i = (X_{i-l+1}, X_{i-l+2}, \cdots, X_i)$, and assume that $g(Y_i, \mu)$ is differentiable with respect to its second argument and that both $g(Y_i, \mu)$ and the derivative of $g(Y_i, \mu)$ with respect to its second argument satisfy the conditions of Theorem 5.5.4. Then there exists $\sigma_g \geq 0$ such that*

$$\frac{S_n}{\sqrt{n}} \longrightarrow N(0, \sigma_g^2),$$

where $S_n = \sum_{i=1}^n g(Y_i, \overline{Z}_n)$.

Proof. By Taylor expansion, we write

$$g(Y_i, \overline{Z}_n) = g(Y_i, \mu) + g^{(0,1)}(Y_i, \mu)(\overline{Z}_n - \mu) + \text{higher order terms.}$$

Since $g^{(0,1)}(Y_i, \mu)$ is only a function of Y_i and of a constant μ, by Theorem 5.5.4 we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \left[g^{(0,1)}(Y_i, \mu) - \mathbb{E}g^{(0,1)}(Y, \mu)\right] \longrightarrow N(0, \sigma_{g'}^2),$$

where $Y \overset{D}{\sim} Y_i$. This implies that if $N \overset{D}{\sim} N(0, \sigma_{g'}^2)$,we then have

$$\begin{aligned} \frac{1}{\sqrt{n}} S_n &= [\frac{1}{\sqrt{n}} \sum_{i=1}^n g(Y_i, \mu)] + (N + \sqrt{n}\mathbb{E}g^{(0,1)}(Y, \mu))(\overline{Z}_n - \mu) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(Y_i, \mu) + \mathbb{E}g^{(0,1)}(Y, \mu)(Z_i - \mu)] + N(\overline{Z}_n - \mu). \end{aligned} \tag{5.20}$$

Since $N(\overline{Z}_n - \mu) = o_{\mathbb{P}}(1)$, applying Theorem 5.5.4 we get

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n [g(Y_i, \mu) + \mathbb{E}g^{(0,1)}(Y, \mu)(Z_i - \mu)] \longrightarrow N(0, \sigma_g^2)$$

□

Corollary 5.5.6. *(Modified version of Wu and Shao [76, Theorem 3] for infinite dimensional arguments) Suppose $\overline{Z}_n$ is a measurable function of $(X_1, X_2, \cdots, X_n)$ such that $\overline{Z}_n$ converges in probability to some constant μ. Let $Y_i = (X_{i-l+1}, X_{i-l+2}, \cdots, X_i)$, and assume that $g(Y_i, \mu)$ is Fréchet differentiable with respect to its second argument, and* *that both $g(Y, \mu)$ and the Fréchet derivative of g with respect to its second argument satisfy the conditions of Theorem 5.5.4. Then there exists $\sigma_g \geq 0$ such that*

$$\frac{S_n}{\sqrt{n}} \longrightarrow N(0, \sigma_g^2),$$

where $S_n = \sum_{i=1}^n g(Y_i, \overline{Z}_n)$.

Remark 5.5.7. *The proof of this Corollary can be understood by following the same steps as in the proof of Corollary 5.5.5, without the added technical complexities that arise when dealing with the Fréchet derivative.*

Proof of Corollary 5.5.6. Let $D_g(Y_i, u, v)$ denote the Fréchet derivative of g with respect to its second argument at u in the direction v. Assume $\overline{Z}_n = \mu + v_n$, and apply the Taylor formula for the Fréchet derivative (Kurdila and Zabarankin [40]) to get

$$g(Y_i, \overline{Z}_n) = g(Y_i, \mu) + D_g(Y_i, \mu, v_n) + R(Y_i, \mu, v_n),$$

where

$$\lim_{\|v_n\| \to 0} \frac{|R(Y_i, \mu, v_n)|}{\|v_n\|} = 0.$$

Note that we can identify the Fréchet derivative with a bounded linear operator as

$$D_g(Y_i, \mu, v_n) = \langle D_g(Y_i, \mu), v_n \rangle.$$

Furthermore, as the Fréchet derivative is also stochastic Dini continuous, we can apply Theorem 5.5.4 to obtain

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\langle D_g(Y_i, \mu), v_n \rangle - \mathbb{E} \langle D_g(Y, \mu), v_n \rangle \right] \longrightarrow N(0, \sigma_{g'}^2),$$

where $Y \overset{D}{\sim} Y_i$.

This implies that if $N \overset{D}{\sim} N(0, \sigma_{g'}^2)$, using the fact that the mapping $D_g(Y_i, \mu, .)$ is linear, we get:

$$\begin{aligned}
\frac{S_n}{\sqrt{n}} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(Y_i, \mu) + D_g(Y_i, \mu, \overline{Z}_n - \mu)] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(Y_i, \mu) + \langle D_g(Y_i, \mu), \overline{Z}_n - \mu \rangle] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(Y_i, \mu)] + \langle N \times \mathrm{id} + \sqrt{n} \mathbb{E} D_g(Y, \mu), \overline{Z}_n - \mu \rangle
\end{aligned}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [g(Y_i, \mu) + \langle \mathbb{E} D_g(Y, \mu), Z_i - \mu \rangle] + N\langle \mathrm{id}, \overline{Z}_n - \mu \rangle \tag{5.21}$$

Since $N\langle \mathrm{id}, \overline{Z}_n - \mu \rangle = o_P(1)$, and $\mathbb{E}\langle \mathbb{E} D_g(Y, \mu), Z_i - \mu \rangle = 0$, we can apply Theorem 5.5.4 and conclude $\frac{S_n}{\sqrt{n}} \longrightarrow N(0, \sigma^2)$ for some σ. □

Lemma 5.5.8. *The function $g_\alpha(T_{i-1}, T_i, S) = \|S \circ [\alpha T_{i-1}] - T_i\|_2^2$ is Fréchet differentiable with respect to S and satisfies the Taylor formula*

$$g_\alpha(T_{i-1}, T_i, S + v) = g_\alpha(T_{i-1}, T_i, S) + D_g(T_{i-1}, T_i, S, v) + R(T_{i-1}, T_i, S, v),$$

where $D_g(T_{i-1}, T_i, S, v)$ is the Fréchet derivative of g_α with respect to S in the direction v, and

$$\lim_{\|v\| \to 0} \frac{|R(T_{i-1}, T_i, S, v)|}{\|v\|} = 0.$$

Furthermore, the mapping $D_g(T_{i-1}, T_i, \mu, .)$ is both linear and bounded.

Proof. To begin, we show that $g_\alpha(T_{i-1}, T_i, S)$ is Gateaux differentiable.

$$\begin{aligned}
&\lim_{\epsilon \to 0} \frac{g_\alpha(T_{i-1}, T_i, S + \epsilon v) - g_\alpha(T_{i-1}, T_i, S)}{\epsilon} \\
&\quad = \lim_{\epsilon \to 0} \frac{\|(S + \epsilon v) \circ [\alpha T_{i-1}] - T_i\|_2^2 - \|S \circ [\alpha T_{i-1}] - T_i\|_2^2}{\epsilon} \\
&\quad = \lim_{\epsilon \to 0} \frac{\epsilon^2 \|v \circ [\alpha T_{i-1}]\|_2^2 + \epsilon \langle v \circ \alpha T_{i-1}, S \circ [\alpha T_{i-1}] - T_i \rangle}{\epsilon} \\
&\quad = \langle v \circ [\alpha T_{i-1}], S \circ [\alpha T_{i-1}] - T_i \rangle \\
&\quad = D_g((T_i, T_{i-1}), S, v).
\end{aligned} \tag{5.22}$$

As the above expression is linear and bounded with respect to v, it serves as the Gateaux differential. As $D_g(T_{i-1}, T_i, S)$ is Gateaux differentiable for every T and the mapping $T \longrightarrow D_g(T_{i-1}, T_i, S)$ is continuous, Corollary 4.1.1. of [40] guarantees that D_g is also the Fréchet derivative. □

Lemma 5.5.9. *The stochastic Dini continuity condition* (5.19) *is satisfied by the function g_α.*

Proof. We want to show $\int_0^1 \frac{\Delta_g(t)}{t} < \infty$, where

$$\begin{aligned}
\Delta_g(\delta) = \sup \Big\{ \Big\| [g_\alpha(T_{i-1}, T_i, S) - g_\alpha(T'_{i-1}, T'_i, S)] 1_{\rho((T_{i-1}, T_i), (T'_{i-1}, T'_i)) \le \delta} \Big\| \\
: T_i, T'_i \text{ are identically distributed} \Big\}.
\end{aligned} \tag{5.23}$$

and $\rho((T_1, T_2), (T'_1, T'_2)) = \sqrt{\|T_1 - T'_1\|_2^2 + \|T_2 - T'_2\|_2^2}$.

First, recall that $g_\alpha(T_{i-1}, T_i, S) = \|S \circ [\alpha T_{i-1}] - T_i\|_2^2$. When $\alpha \geq 0$, we have $\left\|[\alpha T_{i-1}] - [\alpha T'_{i-1}]\right\|_2 \leq \alpha \left\|T_{i-1} - T'_{i-1}\right\|_2$. When $\alpha < 0$, we can use Lemma 5.5.1 to conclude that $\left\|[\alpha T_{i-1}] - [\alpha T'_{i-1}]\right\|_2 \leq \alpha \left\|T_{i-1} - T'_{i-1}\right\|_2^b$ for some $b \geq \frac{1}{2}$. As S is Lipschitz, we can deduce that $\Delta_g(t) \leq C\alpha t^b$, for some $b > 0$. Therefore the integral is finite. □

Proof of Theorem 5.2.8. From Lemma 5.2.7, we see that $S_{N,\alpha}$ converges in probability to S_α and we also obtained a central limit theorem for $S_{N,\alpha}$. Then Lemma 5.5.8 and 5.5.9 show that g_α is Fréchet differentiable and stochastically Dini continuous, which are sufficient conditions for Corollary 5.5.6 to be applicable, and yield a central limit theorem for $M_N(\alpha) = \frac{1}{N}\sum_{i=1}^N g_\alpha(T_{i-1}, T_i, S_{N,\alpha})$:

$$\sqrt{N}[M_N(\alpha) - M(\alpha)] \to N(0, \sigma_\alpha^2).$$

Thus for any α, $M_N(\alpha)$ converges in probability to $M(\alpha)$. By applying Corollary 3.1 from Newey [49] and utilizing Lemma 5.5.2, which establishes that g_α satisfies Lipschitz continuity with respect to α, we can achieve uniform convergence in probability of M_N to M with respect to α:

$$\sup_\alpha |M_N(\alpha) - M(\alpha)| \to 0 \quad \text{in probability.}$$

□

Proof of Theorem 5.2.9 (Consistency). Lemma 5.2.8 implies that M_N converges uniformly in probability to M with respect to α, and Lemma 5.2.5 shows that $\boldsymbol{\alpha}$ is the unique minimizer of M. By applying Van Der Vaart and Wellner [72, Theorem 3.2.3], we can conclude that the estimator $\hat{\alpha}_N = \arg\min_\alpha M_N(\alpha)$ converges to the true parameter $\arg\min_\alpha M(\alpha) = \boldsymbol{\alpha}$. □

Lemma 5.5.10. *Let $\mathcal{T}_{l,u} = \{T \in \mathcal{T} : 0 < L_l \leq T' \leq L_u < \infty\}$ and suppose $\{T_i\} \subset \mathcal{T}_{l,u}$. Then*

$$\mathbb{E}|M'_N(\boldsymbol{\alpha})| \lesssim \frac{1}{\sqrt{N}}.$$

Proof. Note that

$$M'_N(\alpha) = \frac{1}{N}\sum_{j=1}^N \frac{\partial g_\alpha(T_{j-1}, T_j, S_{N,\alpha})}{\partial \alpha},$$

and

$$\begin{aligned}\frac{\partial g_\alpha(T_{j-1}, T_j, S_{N,\alpha})}{\partial \alpha} &= \frac{\partial \left\|S_{N,\alpha} \circ [\alpha T_{j-1}] - T_j\right\|_2^2}{\partial \alpha} \\ &= \int 2|S_{N,\alpha} \circ [\alpha T_{j-1}](x) - T_j(x)| \times \frac{\partial}{\partial \alpha} S_{N,\alpha} \circ [\alpha T_j](x)\, \mathrm{d}x\end{aligned}$$

The expression $|S_{N,\alpha} \circ [\alpha T_{j-1}](x) - T_j(x)|$ can be uniformly bounded. In what follows we will explicitly calculate $\frac{\partial}{\partial\alpha} S_{N,\alpha} \circ [\alpha T_j](x)$ for a fixed j. The calculation is tedious but elementary. To calculate the derivative we use the following fact: if $f(\alpha, x) = C(\alpha, y(x,\alpha))$, then

$$\frac{\partial f}{\partial \alpha} = \frac{\partial C(\alpha, y(x,\alpha'))}{\partial \alpha}|_{\alpha'=\alpha} + \frac{\partial C(\alpha, y)}{\partial y} \times \frac{\partial y(x,\alpha)}{\partial \alpha}.$$

Using the above equation we can write:

$$\begin{aligned}\frac{\partial}{\partial\alpha} S_{N,\alpha} \circ [\alpha T_j](x) &= \frac{\partial}{\partial\alpha} S_{N,\alpha}([\alpha' T_j](x))|_{\alpha'=\alpha} + \frac{\partial S_{N,\alpha}([\alpha T_j](x))}{\partial([\alpha T_j](x))} \times \frac{\partial[\alpha T_j](x)}{\partial\alpha} \\ &= \frac{\partial S_{N,\alpha}(y)}{\partial\alpha}|_{y=[\alpha T_j](x)} + \frac{\partial S_{N,\alpha}(y)}{\partial y}|_{y=[\alpha T_j](x)} \times \frac{\partial[\alpha T_j](x)}{\partial\alpha}.\end{aligned} \tag{5.24}$$

First, we derive the first term on the LHS of (5.24):

$$\frac{\partial S_{N,\alpha}(y)}{\partial\alpha} = \sum_{i=1}^{N} \frac{\partial}{\partial\alpha} T_i \circ [\alpha T_{i-1}]^{-1}(y)$$

If we consider one of the terms in this summation we have

$$\begin{aligned}\frac{\partial}{\partial\alpha} T_i \circ [\alpha T_{i-1}]^{-1}(y) &= \frac{\partial T_i([\alpha T_{i-1}]^{-1}(y))}{\partial[\alpha T_{i-1}]^{-1}(y)} \times \frac{\partial[\alpha T_{i-1}]^{-1}(y)}{\partial\alpha} \\ &= T_i'(z_i)|_{z_i=[\alpha T_{i-1}]^{-1}(y)} \times \frac{\partial}{\partial\alpha}[\alpha T_{i-1}]^{-1}(y)\end{aligned} \tag{5.25}$$

Now to calculate $\frac{\partial}{\partial\alpha}[\alpha T_{i-1}]^{-1}(y)$ note that:

$$\begin{aligned}0 = \frac{\partial}{\partial\alpha} y &= \frac{\partial}{\partial\alpha}[\alpha T_{i-1}]([\alpha T_{i-1}]^{-1}(y)) \\ &= \frac{\partial}{\partial\alpha}[\alpha T_{i-1}]([\alpha' T_{i-1}]^{-1}(y))|_{\alpha'=\alpha} + \frac{\partial[\alpha T_{i-1}]([\alpha T_{i-1}]^{-1}(y))}{\partial[\alpha T_{i-1}]^{-1}(y)} \times \frac{\partial[\alpha T_{i-1}]^{-1}(y)}{\partial\alpha} \\ &= \frac{\partial}{\partial\alpha}[\alpha T_{i-1}](z_i)|_{z_i=[\alpha T_{i-1}]^{-1}(y)} + \frac{\partial[\alpha T_{i-1}](z_i)}{\partial z_i}|_{z_i=[\alpha T_{i-1}]^{-1}(y)} \times \frac{\partial[\alpha T_{i-1}]^{-1}(y)}{\partial\alpha}\end{aligned} \tag{5.26}$$

Thus

$$\begin{aligned}\frac{\partial[\alpha T_{i-1}]^{-1}(y)}{\partial\alpha} &= (-1)\times\frac{\partial}{\partial\alpha}[\alpha T_{i-1}](z_i)|_{z_i=[\alpha T_{i-1}]^{-1}(y)} \times \frac{1}{\frac{\partial[\alpha T_{i-1}](z_i)}{\partial z_i}|_{z_i=[\alpha T_{i-1}]^{-1}(y)}}\\ &= \begin{cases} (z_i - T_{i-1}(z_i))\times \frac{1}{\alpha(T'_{i-1}(z_i)-1)+1}\Big|_{z_i=[\alpha T_{i-1}]^{-1}(y)}, & \text{for } 0<\alpha\le 1\\ (T^{-1}_{i-1}(z_i)-z_i)\times\frac{1}{\alpha(1-(T^{-1}_{i-1})'(z_i))+1}\Big|_{z_i=[\alpha T_{i-1}]^{-1}(y)}, & \text{for } -1\le\alpha<0,\end{cases}\end{aligned} \tag{5.27}$$

And we can conclude that

$$\begin{aligned}&\frac{\partial}{\partial\alpha}T_i\circ[\alpha T_{i-1}]^{-1}(y)\\ &\quad= T'_i(z_i)\times\begin{cases} (z_i - T_{i-1}(z_i))\times \frac{1}{\alpha(T'_{i-1}(z_i)-1)+1}\Big|_{z_i=[\alpha T_{i-1}]^{-1}(y)}, & \text{for } 0<\alpha\le 1\\ (T^{-1}_{i-1}(z_i)-z_i)\times\frac{1}{\alpha(1-(T^{-1}_{i-1})'(z_i))+1}\Big|_{z_i=[\alpha T_{i-1}]^{-1}(y)}, & \text{for } -1\le\alpha<0\end{cases}\end{aligned} \tag{5.28}$$

With this, we have all the needed terms to calculate the left terms of (5.24). Now we calculate the right term of (5.24):

$$\begin{aligned}\frac{\partial S_{N,\alpha}(y)}{\partial y} &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial y}T_i\circ[\alpha T_{i-1}]^{-1}(y)\\ &= \frac{1}{N}\sum_{i=1}^{N}T'_i(z_i)|_{z_i=[\alpha T_{i-1}]^{-1}(y)}\times\frac{1}{\frac{\partial[\alpha T_{i-1}](z_i)}{\partial z_i}|_{z_i=[\alpha T_{i-1}]^{-1}(y)}}\\ &= \begin{cases}\frac{1}{N}\sum_{i=1}^{N}T'_i(z_i)\times\frac{1}{\alpha(T'_{i-1}(z_i)-1)+1}\Big|_{z_i=[\alpha T_{i-1}]^{-1}(y)}, & \text{for } 0<\alpha\le 1\\ \frac{1}{N}\sum_{i=1}^{N}T'_i(z_i)\times\frac{1}{\alpha(1-(T^{-1}_{i-1})'(z_i))+1}\Big|_{z_i=[\alpha T_{i-1}]^{-1}(y)}, & \text{for } -1\le\alpha<0\end{cases}\end{aligned} \tag{5.29}$$

By plugging all the terms calculated above in (5.24) we get

$$\begin{aligned}&\frac{\partial}{\partial\alpha}S_{N,\alpha}\circ[\alpha T_j](x)\\ &= \begin{cases}\frac{1}{N}\sum_{i=1}^{N}T'_i(z_i)\times\frac{z_i-T_{i-1}(z_i)+T_j(x)-x}{\alpha(T'_{i-1}(z_i)-1)+1}\Big|_{z_i=[\alpha T_{i-1}]^{-1}(y)}, & \text{for } 0<\alpha\le 1\\ \frac{1}{N}\sum_{i=1}^{N}T'_i(z_i)\times\frac{T^{-1}_{i-1}(z_i)-z_i+x-T^{-1}_j(x)}{\alpha(1-(T^{-1}_{i-1})'(z_i))+1}\Big|_{z_i=[\alpha T_{i-1}]^{-1}(y)}, & \text{for } -1\le\alpha<0,\end{cases}\end{aligned} \tag{5.30}$$

where $y = [\alpha T_j](x)$.

The differentiability of M_N with respect to α follows from the equation above. Similarly, if we replace $S_{N,\alpha}$ with S_α, the summations can be replaced by an integral, and we can see that M is also differentiable with respect to α. Let

$$g'(T_{j-1}, T_j, S_{N,\alpha}, \alpha) = \frac{\partial g_\alpha(T_{j-1}, T_j, S_{N,\alpha})}{\partial \alpha}, \quad g'(T_{j-1}, T_j, S_\alpha, \alpha) = \frac{\partial g_\alpha(T_{j-1}, T_j, S_\alpha)}{\partial \alpha}.$$

Since $\boldsymbol{\alpha}$ is the minimizer of M, we must have $\mathbb{E}g'(T_{j-1}, T_j, S_\alpha, \alpha)|_{\alpha=\boldsymbol{\alpha}} = M'(\boldsymbol{\alpha}) = 0$. Additionally, We can argue g' is stochastically Dini-continuous when $\{T_i\} \subset \mathcal{T}_{l,u}$ (similar to the arguments in the proof of Lemma 5.5.9). Therefore the assumptions of Corollary 5.5.6 (CLT) are satisfied for g', and we have $\mathbb{E}|M_N'(\boldsymbol{\alpha})| \lesssim \frac{1}{\sqrt{N}}$. □

Proof of Theorem 5.2.10 (Convergence Rate). Using Theorem 3.6.3, we can obtain a rate of convergence for our estimator. First, it should be noted that the functional M is twice differentiable with respect to α since it is a composition of twice differentiable functions. As $\boldsymbol{\alpha}$ is the unique minimizer of M, its first derivative vanishes at $\boldsymbol{\alpha}$, which implies that M has quadratic growth around $\boldsymbol{\alpha}$. Next, we need to find a function $\phi_N(\delta)$ such that

$$\mathbb{E} \sup_{|\alpha-\boldsymbol{\alpha}|\leq\delta} \sqrt{N}\Big|(M_N - M)(\alpha) - (M_N - M)(\boldsymbol{\alpha})\Big| \leq \phi_N(\delta). \tag{5.31}$$

Taylor expanding, we can write:

$$(M_N - M)(\alpha) = (M_N - M)(\boldsymbol{\alpha}) + (M_N - M)'(\boldsymbol{\alpha}) \times (\alpha - \boldsymbol{\alpha}) + \text{higher order terms} \tag{5.32}$$

Since $\boldsymbol{\alpha}$ is the minimiser of M, yielding $M'(\boldsymbol{\alpha}) = 0$, we only need to calculate $M_N'(\boldsymbol{\alpha})$. But by Lemma 5.5.10 we can see that

$$\mathbb{E}|M_N'(\boldsymbol{\alpha})| \lesssim \frac{1}{\sqrt{N}}.$$

By plugging the inequality into the expression (5.31) we obtain

$$\mathbb{E}\sqrt{N}\Big|(M_N - M)(\alpha) - (M_N - M)(\boldsymbol{\alpha})\Big| \leq |\alpha - \boldsymbol{\alpha}|.$$

And, we conclude $\phi_N(\delta) = \delta$ and the rate of convergence for $\hat{\alpha}_N$ is $N^{-\frac{1}{2}}$. Using Lemma 5.5.1 we can see

$$\left\|S_{N,\hat{\alpha}_N} - \mathrm{S}\right\|_2 \leq \left\|S_{N,\hat{\alpha}_N} - S_{N,\boldsymbol{\alpha}}\right\|_2 + \left\|S_{N,\boldsymbol{\alpha}} - \mathrm{S}\right\|_2 \lesssim N^{-\frac{b}{2}} + N^{-\frac{1}{2}} \lesssim N^{-\frac{b}{2}},$$

and since $\{T_i\} \subset \mathcal{T}_{l,u}$, $b = 1$ according to Lemma 5.5.1.

□

5.5.2 Generalization of Iterated System (5.4)

The definition of the iterated system (5.3) is based on the contraction of maps around the identity map. It extends system (5.4) by introducing the map S. However, we could alternatively generalise (5.4) by introducing S not at the level of the iteration itself, but rather at the level of the contraction itself: contracting around an arbitrary map S, instead of the identity. Specifically, define the α-contraction of a map T around an arbitrary map S as follows:

$$\alpha[T,S](x) := \begin{cases} S(x) + \alpha(T(x) - S(x)) & 0 < \alpha \leq 1 \\ S(x) & \alpha = 0 \\ S(x) + \alpha(S(x) - T^{-1}(x)) & -1 \leq \alpha < 0. \end{cases} \tag{5.33}$$

With this definition, the original contraction operation (5.1) now corresponds to $\alpha[T, \mathrm{id}]$, for $\mathrm{id}(x) = x$ the identity map. Definition (5.33) directly leads to the following extension of system (5.4)

$$T_i = T_{\epsilon_i} \circ \alpha[T_{i-1}, S], \tag{5.34}$$

where $\{T_{\epsilon_i}\}_{i=1}^N$ is again a collection of independent and identically distributed random optimal maps satisfying $\mathbb{E}\{T_{\epsilon_i}(x)\} = x$ almost everywhere on Ω. Compared to system (5.3),

$$T_i = T_{\epsilon_i} \circ S \circ \alpha[T_{i-1}, \mathrm{id}].$$

this system interjects S at the level of the contraction and not at the level of the random perturbation (note that for identifiability reasons it does not make sense to do both). Of course, either is more general than system (5.4)

$$T_i = T_{\epsilon_i} \circ \alpha[T_{i-1}, \mathrm{id}].$$

Remark 5.5.11. *Suppose we use the contraction definition* (5.33)*, and define the iteration* (5.34)*. Then, the quantile model (UQ) with $S = F_\mu^{-1}$ (i.e. where we contract around the quantile function of a measure μ) is equivalent to the generalised quantile model (GQ) with $S = \mathrm{id}$; that is, they produce the same stationary time series. To demonstrate this equivalence, consider the model (GQ) with $S = \mathrm{id}$. We then have:*

$$\mathbb{E}(F_{\mu_i}^{-1} \circ F_\mu(x) | F_{\mu_{i-1}}^{-1} \circ F_\mu) = \mathbb{E}(F_{\mu_i}^{-1} | F_{\mu_{i-1}}^{-1} \circ F_\mu) \circ F_\mu(x) = x + \alpha(F_{\mu_{i-1}}^{-1}(F_\mu(x)) - x)$$

Thus,

$$\mathbb{E}(F_{\mu_i}^{-1} | F_{\mu_{i-1}}^{-1} \circ F_\mu) = F_\mu^{-1}(x) + \alpha(F_{\mu_{i-1}}^{-1}(x) - F_\mu^{-1}(x)),$$

which is equal to the conditional expectation of $\mathbb{E}(F_{\mu_i}^{-1} | F_{\mu_{i-1}}^{-1})$ when we use model (5.33) *for $F_{\mu_i}^{-1}$ and contract around $S = F_{\mu}^{-1}$.*

Remark 5.5.12. *Note that $\alpha[T, S] = T$ when $\alpha = 1$, and $\alpha[T, S] = T^{-1}$ when $\alpha = -1$. Therefore in either of these cases, the time series T_i does not provide any information about S and it would impossible to estimate the map S. Therefore we assume $-1 < \alpha < 1$. This is in contrast with system* (5.3), *where consistent estimation is possible for all values of* $0 \leq \alpha \leq 1$,

If a stationary solution to system (5.34) exists, then

$$\mathbb{E}[T_i] = \mathbb{E}[T_{i+1}] = \mathbb{E}[E[T_{i+1}|T_i]] = \mathbb{E}[\alpha[T_i, S]],$$

and therefore $\mathbb{E}[T_i] = S$, when $-1 < \alpha < 1$.

We define the estimators $(\hat{\alpha}_N, S_N)$ of (α, S) as follows:

$$\hat{\alpha}_N := \arg\min_{\alpha} M_N(\alpha),$$

where

$$\begin{aligned} M_N(\alpha) &:= \frac{1}{N}\sum_{i=1}^{N} g(T_{i-1}, T_i, S_N) \\ g_\alpha(T_{i-1}, T_i, S) &:= \|\alpha[T_{i-1}, S] - T_i\|_2^2 \\ S_N &:= \frac{1}{N}\sum_{j=1}^{N} T_j \end{aligned}$$

It is worth noting that unlike in system (5.3), where the estimation of the map S depends on the estimator of α, in this system, the estimator of the map S is simply the average of the maps T_i. Consequently, the statistical analysis of the estimators is somewhat easier in this case. Similar procedures to those used for model (5.3) can be used to demonstrate the existence of a unique stationary solution, the consistency of the estimator, and obtain the rate of convergence.

Assuming that system (5.34) satisfies the moment contracting condition 5.2.1, a unique stationary solution for this system exists, and $\mathbb{E}[T_i] = S$, as in the previous case. We can then use Lemma 5.2.6 to obtain the central limit theorem (CLT) for S_N and show that S_N converges in probability to the true S.

It is worth noting that the Lipschitz continuity property of the new function g with respect to α can be shown using the fact that $\|\alpha_1[T, S] - \alpha_2[T, S]\|_2 \lesssim |\alpha_1 - \alpha_2|$. Using this property and following a similar proof technique as in Theorem 5.2.10, we can argue that the rate of convergence is $N^{-1/2}$.

Remark 5.5.13. *Once again we can use the system* (5.34) *to construct a Markov chain model for a dependent sequence of probability distributions $\mu_i \in \mathcal{W}_2(\Omega)$ by either in-*

terpreting the maps as consecutive optimal maps between a time series of probability distributions or directly using the maps to model the quantile functions. While using system (5.3)*, the increment interpretation using $\alpha = 0$ is equivalent to quantile interpretation using $\alpha = 1$, a similar straightforward relationship does not appear to exist when using system* (5.34).

Chapter 6

Outlook

In this thesis, we have explored several aspects of distributional regression and autoregression. However, there are still many questions and potential directions for future work that have not been addressed in this thesis. In this chapter, we outline some of the potential inquiries stemming from chapters 3,4 and 5, that could be further examined. Additionally, we introduce an alternative model for distributional regression and present some preliminary results.

Chapter 3

Convergence Rate For fully observed distributions, we derived the $N^{-1/3}$ rate without imposing extra regularity conditions on the covariate measures and demonstrated its minimax optimality. It remains an open question how the rate would be affected by imposing additional constraints on the input measures, such as absolute continuity. Can a faster optimal rate than $N^{-1/3}$ be achieved? Is the current estimator able to reach the optimal rate under these conditions? If not, can an optimal estimator be identified?

We could begin by examining specific situations. For instance, when all input measures are the same, the estimator $\hat{T}_N$ is equal to $\frac{1}{N}\sum_{i=1}^{N} T_i$ and converges at a rate of $N^{-1/2}$. If the input measures differ but all fully supported on Ω, the estimator $\frac{1}{N}\sum_{i=1}^{N} T_i$ also converges at a rate of $N^{-1/2}$. What about our estimator, which is equal to $\hat{T}_N = \arg\min_T \frac{1}{N}\sum_{i=1}^{N} \|T - T_i\|^2_{L^2(\mu_i)}$ according to equation (3.11)?

Incorporating Additional Covariates When analyzing mortality data with our model, we observed noticeable differences in the residual maps between Eastern European and Western European countries. It seems plausible that a more accurate model for this problem would not treat the optimal map in the regression operator as fixed, but rather allow it to depend on another covariate, such as GDP.

Consequently, another question to explore is how to extend the model to include extra covariates, such as a scalar covariate. We can consider the situation where we want to incorporate a scalar covariate c. We can define a regression operator $\Gamma : \mathcal{W}_2(\Omega) \times \mathbb{R} \to \mathcal{W}_2(\mathbb{R})$ as the minimizer of the conditional Fréchet functional, considered as a function of both μ and c,

$$\operatorname*{argmin}_{b} \int_{\mathcal{W}_2(\Omega)} d^2_{\mathcal{W}}(b, \nu)\, \mathrm{d}P(\nu \mid \mu, c) = \Gamma(\mu, c),$$

and we can impose $\Gamma(\mu) = T_c\#\mu$, where T_c is an increasing map that depends on the covariate c. We could consider a specific partial ordering $\preceq$ of the optimal maps and stipulate that if $c < c'$, then $T_c \preceq T_{c'}$. Since optimal maps are non-decreasing functions when $d = 1$, we can consider them as quantile functions and choose $\preceq$ to be a certain partial ordering of distributions.

A similar problem is considered in the machine learning community by Bunne et al. [11]. This paper considers a setting where a data set of the form $\{c_i, (\mu_i, \nu_i)\}$ is observed, where μ_i and ν_i are distributions and c_i is a scalar. And the goal is to learn a mapping T_θ such that for any c, $T_\theta(c)$ is an optimal map and $T_\theta(c_i)\#\mu_i$ is close to ν_i for each i. They use a particular neural network architecture to parameterize the optimal map T_θ (based on the paper by Amos et al. [4]) which guarantees that T_θ is the gradient of a convex function. The focus of the paper is not on developing a model or methodology with theoretical guarantees but rather developing certain computational techniques for a similar problem.

Chapter 4

Computational Method The proposed estimator minimizes M_N over a set of optimal maps with certain smoothness. A major challenge in our theoretical framework is to numerically compute the proposed infinite-dimensional estimator. Are there efficient computational methods, possibly using its directional derivative for gradient descent to find the minimizer? Another possibility is to apply ideas from the machine learning community such as Amos et al. [4], Bunne et al. [11] and parameterize the optimal map T as a neural network with a specific architecture such that the output of the network is the gradient of a convex function of (some of) the inputs.

Minimax Optimality We demonstrated the minimax optimality of our estimator for $d = 1$ using Fano's method, so another direction could be determining if the estimator is minimax optimal for $d > 1$, possibly starting by examining the applicability of Fano's method.

Fano's method is based on reducing the problem to multiple hypothesis testing and identifying a set of optimal maps $\{T_1, \cdots, T_m\}$ that are sufficiently separated yet challenging to distinguish given the samples. More formally, let's consider a map T, and denote by P_T the distribution induced on the response variable. Suppose an index J is uniformly drawn at random from $\{1, \cdots, m\}$, and a sample Z is generated from

P_{T_J}. Fano's method provides a lower bound on the estimation error in terms of an upper bound on the mutual information between Z and J (see Theorem 3.6.7). One approach is to use the following inequality:

$$I(Z;J) \leq \frac{1}{m^2} \sum_{j,k=1}^{m} KL(\mathbb{P}_{T_j} || \mathbb{P}_{T_k}),$$

(see inequality 15.34 of Wainwright [75]). Typically, an upper bound on the mutual information is derived by finding an upper bound on such KL divergence quantities or by bounding the KL divergence with other divergences or distances.

However, recall that the response variable in our regression model is a distribution itself, so P_T is a distribution in $\mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^d))$ and finding KL divergence between such distributions could be challenging. It might be possible to find an upper bound for the Wasserstein distance $d_{\mathcal{W}}(P_{T_1}, P_{T_2})$, but typically, such an upper bound cannot be used to find an upper bound for KL-divergence. It is generally easier to find an upper bound for Wasserstein distance rather than a lower bound.

In this specific situation, determining minimax optimality using Fano's method might be difficult due to the complexity of the problem and the nature of the induced distributions being in $\mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^d))$. It may be necessary to explore alternative methods or develop new techniques tailored to this problem's unique structure that establish the minimax optimality of the existing rate.

Adaptive Approach We have examined an estimator that minimizes the functional M_N within a particular Hölder ball, dependent on β, γ, and R. This proposed estimator calls for knowledge of these parameters. Could there exist an estimator that doesn't require such information and might provide a rate that adapts to the real intrinsic smoothness of the parameters?

I hypothesize that, under the same conditions as Theorem 4.3.7, specifically Assumptions 4.2.1, 4.2.2, 4.2.3, and 4.3.2, the following estimator would also be consistent:

$$\hat{T}_N := \operatorname*{argmin}_{T \in \mathcal{T}} M_N(T). \tag{6.1}$$

This estimator's definition no longer hinges on the knowledge of the true smoothness parameters. However, this might result in a slower convergence rate.

A possible strategy for determining the rate of this estimator might involve initially creating an approximation of $\hat{T}_N$ by convolution of $\hat{T}_N$ with a kernel η_{ϵ_N}, where ϵ_N is a bandwidth that depends on the sample size N. It can be demonstrated that this approximation sits within the Hölder ball C_R^β for any given β and an adequately large radius R that depends on ϵ_N. By obtaining some bounds on the growth rate of R, we can compute the metric entropy of the class C_R^β. Thus, by using a method analogous to the one we employed to derive the rate of convergence of our estimator, we can establish the rate of convergence for the approximation of $\hat{T}_N$. This could then be used to define upper bounds for the estimator rate of $\hat{T}_N$.

Chapter 5

Higher-Dimensional Extensions Our autoregression models were based on the iterated random function system (5.3). That system relies on compositions of optimal maps being optimal for $d = 1$, and hence it is not applicable to higher dimensions. Are there any alternative approaches?

Potential avenues of research could involve constructing iterated random functions of optimal maps, by exploring transformations on the space of optimal maps that preserve optimality. Considering transformations of convex functions first, we can construct transformations for optimal maps, given their characterization as the gradient of convex functions.

Alternative Model for Distribution-on-Distribution Regression

In this section, we examine a model where the noise component is a random operator acting on the space of distributions, characterized as a Markov kernel. Similar to Model (4.1), where the specific form of the noise component leads to interpreting the model as specifying the conditional Fréchet mean of the response distribution, the form of the noise component in this model leads to interpreting the model as specifying a conditional weak barycenter (or weak Fréchet mean), which is introduced by [13].

We briefly discuss some preliminary results on the identifiability of the model parameters and the estimation procedure.

Markov Regression

A Markov kernel from $\mathbb{R}^d$ to $\mathbb{R}^d$ is a map of the form $\kappa : (x, B) \longrightarrow \kappa(x, B)$ such $\kappa(x, .)$ is a probability measure for all $x \in \mathbb{R}^d$ and $\kappa(., B) : \mathbb{R}^d \longrightarrow \mathbb{R}$ is measurable for every Borel set $B \subset \mathbb{R}^d$.

Given a Markov kernel κ and a probability distribution μ on the same measurable space, the Markov operator M_κ induced by κ is defined as:

$$(M_\kappa \mu)(A) = \int \kappa(x, A)\, \mathrm{d}\mu(x),$$

where A is a measurable set.

We consider a specific set of Markov kernels called Dilatations. A dilatation p is a Markov kernel such that $\int y \, \mathrm{d}p(x, y) = x$ for all x. We use Dilatations to model the noise which leads to the following model for distribution-to-distribution regression:

$$\nu_i = M_{p_\epsilon}(T_0 \# \mu_i), \tag{6.2}$$

where T_0 is an optimal transport map, and M_{p_ϵ} is the Markov kernel induces by a random dilatation p_ϵ. This model is equivalent to $\nu_i = M_{\kappa_\epsilon} \mu_i$, where M_{κ_ϵ} is a Markov operator induced by a random Markov kernel κ_ϵ, such that $\int y \, \mathrm{d}\kappa_\epsilon(x, y) = T_0(x)$ for all ϵ.

Remark 6.0.1. *The Model* (6.2) *implies the existence of a coupling* (X, Y) *of* (μ, ν), *where* $\mathbb{E}[Y|X] = T_0(X)$ *and* $(E[Y|X], Y)$ *forms a martingale. This coupling is called a mixture of Brenier and Strassen as described by [30].*

Let's recall the definition of the convex ordering of distributions. For two distributions $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, we say μ is dominated by ν in the convex order if $\int \psi \, \mathrm{d}\mu \leq \int \psi \, \mathrm{d}\nu$ for all convex functions $\psi : \mathbb{R}^d \to \mathbb{R}$. This is denoted by $\mu \preceq_c \nu$.

Strassen's theorem (Theorem 8 in Strassen [67]) establishes the following: $\mu \preceq_c \nu$ if and only if there exists a dilatation p such that $\nu = M_p \mu$. From Strassen's theorem and Theorem 2 of Cazelles et al. [13], we can deduce that for any fixed measure μ, $T_0 \# \mu$ is a weak Barycenter of the random measure $M_{p_\epsilon}(T_0 \# \mu)$.

Before stating a theorem for identifiability of T_0, we mention a Lemma:

Lemma 6.0.2. *The convex functions uniquely determine a measure with compact support* Ω, *in the sense that if*

$$\int_\Omega \varphi(x) \, \mathrm{d}\mu(x) = \int_\Omega \varphi(x) \, \mathrm{d}\nu(x),$$

for all convex functions $\varphi : \mathbb{R}^d \to R$, *then* $\mu = \nu$. *(see [71] for a proof)*

Theorem 6.0.3. *(Identifiability) Assuming the kernel* p_ϵ *is equal to the identity kernel* $\mathrm{id}(x, .) = x$ *with positive probability, the map* T_0 *can be identified up to the measure* Q.

Proof. First note that for any fixed measure μ, $T_0 \# \mu \preceq_c \nu$, where $\nu = M_{p_\epsilon}(T_0 \# \mu)$. Also, since p_ϵ is equal to the identity kernel $\mathrm{id}(x, .) = x$ with positive probability, it means that $T_0 \# \mu = \nu$, with positive probability. Now if any other map T, satisfies these two conditions, we can infer that $T \# \mu \preceq_c T_0 \# \mu$ and $T_0 \# \mu \preceq_c T \# \mu$. Therefore for any convex function, the integral of this convex function with respect to measures $T_0 \# \mu$ and $T \# \mu$ are equal. By Lemma 6.0.2, convex functions uniquely determined a measure, hence $T_0 \# \mu = T \# \mu$. Therefore, we can infer that $\|T - T_0\|_{L^2(\mu)} = 0$ for any covariate measure μ and hence $\|T - T_0\|_{L^2(Q)}$. □

Remark 6.0.4. *(Ideas for the estimator) We consider an estimator $\hat{T}_N$ such that $\hat{T}_N \# \mu_i \preceq_c \nu_i$. Note that such a map always exists, as T_0 is a solution by the model assumption.*

In the following, we will show that when $d = 1$, an estimator $\hat{T}_N$ can be computed using linear programming solvers.

Lemma 6.0.5. *(Theorem 3.A.5 of [65]) When $d = 1$, $\mu \preceq_c \nu$ if and only if the corresponding distributions have the same mean and*

$$\int_p^1 F^{-1}(u)\,\mathrm{d}u \leq \int_p^1 G^{-1}(u)\,\mathrm{d}u,$$

for all $p \in [0,1]$, where $F^{-1}(u)$ and $G^{-1}(u)$ are quantile functions of μ and ν respectively.

Remark 6.0.6. *From Lemma 6.0.5 we deduce that when $d = 1$, $T\#\mu \preceq_c \nu$ is equivalent to distributions $T\#\mu$ and ν have the same mean and:*

$$\int_p^1 T(F^{-1}(u))\,\mathrm{d}u \leq \int_p^1 G^{-1}(u)\,\mathrm{d}u,$$

for all $p \in [0,1]$.

Since $\int_p^1 T(F^{-1}(u))\,\mathrm{d}u = \int_{F^{-1}(p)}^1 T(y)f(y)\,\mathrm{d}y$, from the previous inequality we deduce that for all $p \in [0,1]$:

$$\int_{F^{-1}(p)}^1 T(y)f(y)\,\mathrm{d}y \leq \int_p^1 G^{-1}(u)\,\mathrm{d}u.$$

Moreover, $T\#\mu$ and ν having the same mean is equivalent to

$$\int_0^1 f(y)T(y)\,\mathrm{d}y = \int_0^1 xg(x)\,\mathrm{d}x,$$

Since these two conditions depend linearly on T, the computation of the estimator $\hat{T}_N$, can be done by a linear programming solver.

Bibliography

[1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] Adil Ahidar-Coutrix, Thibaut Le Gouic, and Quentin Paris. Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics. *Probability theory and related fields*, 177(1-2):323–368, 2020.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[4] Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

[5] Germán Aneiros, Ricardo Cao, Ricardo Fraiman, Christian Genest, and Philippe Vieu. Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis*, 170:3–9, 2019.

[6] François Bachoc, Fabrice Gamboa, Jean-Michel Loubes, and Nil Venet. A gaussian process regression model for distribution inputs. *IEEE Transactions on Information Theory*, 64(10):6620–6637, 2017.

[7] Andreas Berzel, Gillian Z Heller, and Walter Zucchini. Estimating the number of visits to the doctor. *Australian & New Zealand Journal of Statistics*, 48(2): 213–224, 2006.

[8] Jérémie Bigot, Raúl Gouet, Thierry Klein, Alfredo Lopez, et al. Upper and lower risk bounds for estimating the wasserstein barycenter of random measures on the real line. *Electronic journal of statistics*, 12(2):2253–2289, 2018.

[9] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[10] Élodie Brunel, André Mas, and Angelina Roche. Non-asymptotic adaptive prediction in functional linear models. *Journal of Multivariate Analysis*, 143:208–232, 2016.

[11] Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional monge maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[12] Alexandra Carpentier and Teresa Schlüter. Learning relationships between data obtained independently. In *Artificial Intelligence and Statistics*, pages 658–666. PMLR, 2016.

[13] Elsa Cazelles, Felipe Tobar, and Joaquin Fontbona. A novel notion of barycenter for probability distributions based on optimal weak mass transport. *Advances in Neural Information Processing Systems*, 34, 2021.

[14] Gaëlle Chagny and Angelina Roche. Adaptive and minimax estimation of the cumulative distribution function given a functional covariate. *Electronic Journal of Statistics*, 8(2):2352–2404, 2014.

[15] Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, pages 1–14, 2021.

[16] Zhicheng Chen, Yuequan Bao, Hui Li, and Billie F Spencer Jr. A novel distribution regression approach for data loss compensation in structural health monitoring. *Structural Health Monitoring*, 17(6):1473–1490, 2018.

[17] Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1): 223–256, 2017.

[18] Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for bures-wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.

[19] Antonio Cuevas. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23, 2014.

[20] Michael John Priestley Cullen. *A mathematical theory of large-scale atmosphere/ocean flow*. World Scientific, 2006.

[21] Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753, 2021.

[22] Pedro Delicado. Dimensionality reduction when data are density functions. *Computational Statistics & Data Analysis*, 55(1):401–420, 2011.

[23] Persi Diaconis and David Freedman. Iterated random functions. *SIAM review*, 41(1):45–76, 1999.

[24] Laya Ghodrati and Victor M Panaretos. Distribution-on-distribution regression via optimal transport maps. *Biometrika*, 109(4):957–974, 2022.

[25] Laya Ghodrati and Victor M Panaretos. On distributional autoregression and iterated transportation. *arXiv preprint arXiv:2303.09469*, 2023.

[26] Laya Ghodrati and Victor M. Panaretos. Minimax rate for optimal transport regression between distributions. *Statistics & Probability Letters*, 194:109758, 2023.

[27] Laya Ghodrati and Victor M Panaretos. Transportation of measure regression in higher dimensions. *arXiv preprint arXiv:2305.17503*, 2023.

[28] David Gilbarg, Neil S Trudinger, David Gilbarg, and NS Trudinger. *Elliptic partial differential equations of second order*, volume 224. Springer, 1977.

[29] Aldo Goia and Philippe Vieu. An introduction to recent advances in high/infinite dimensional statistics. *Journal of Multivariate Analysis*, 146:1–6, 2016.

[30] Nathael Gozlan and Nicolas Juillet. On a mixture of brenier and strassen theorems. *Proceedings of the London Mathematical Society*, 120(3):434–463, 2020.

[31] Florian F Gunsilius. On the convergence rate of potentials of brenier maps. *Econometric Theory*, 38(2):381–417, 2022.

[32] Peter Hall, Joel L Horowitz, et al. Methodology and convergence rates for functional linear regression. *Annals of Statistics*, 35(1):70–91, 2007.

[33] Lajos Horváth, Piotr Kokoszka, and Ron Reeder. Estimation of the mean of functional time series and a two-sample problem. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):103–122, 2013.

[34] Tailen Hsing and Randall Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.

[35] Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166 – 1194, 2021.

[36] Yiye Jiang. Wasserstein multivariate auto-regressive models for modeling distributional time series and its application in graph learning. *arXiv preprint arXiv:2207.05442*, 2022.

[37] Lanchakorn Kittiratanawasin and Supa Hannongbua. The effect of edges and shapes on band gap energy in graphene quantum dots. *Integrated Ferroelectrics*, 175(1):211–219, 2016.

[38] Alois Kneip and Klaus J Utikal. Inference for density families using functional principal component analysis. *Journal of the American Statistical Association*, 96 (454):519–542, 2001.

[39] Piotr Kokoszka, Hong Miao, Alexander Petersen, and Han Lin Shang. Forecasting of density functions with an application to cross-sectional and intraday returns. *International Journal of Forecasting*, 35(4):1304–1317, 2019.

[40] Andrew J Kurdila and Michael Zabarankin. *Convex functional analysis*. Springer Science & Business Media, 2006.

[41] Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of wasserstein barycenters. *Probability Theory and Related Fields*, 168:901–917, 2017.

[42] Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J Stromme. Fast convergence of empirical barycenters in alexandrov spaces and the wasserstein space. *Journal of the European Mathematical Society*, 2022.

[43] Nengxiang Ling and Philippe Vieu. Nonparametric modelling for functional data: selected survey and tracks for future. *Statistics*, 52(4):934–949, 2018.

[44] John W Lynch, George Davey Smith, George A Kaplan, and James S House. Income inequality and mortality: importance to health of individual income, psychosocial environment, or material conditions. *Bmj*, 320(7243):1200–1204, 2000.

[45] Patrick Mair, Kurt Hornik, and Jan de Leeuw. Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32(5):1–24, 2009.

[46] Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.

[47] Jeffrey S Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359, 2015.

[48] Boris Muzellec, Adrien Vacher, Francis Bach, François-Xavier Vialard, and Alessandro Rudi. Near-optimal estimation of smooth transport maps with kernel sums-of-squares. *arXiv preprint arXiv:2112.01907*, 2021.

[49] Whitney K Newey. Uniform convergence in probability and stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 1161–1167, 1991.

[50] Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in wasserstein distance. *The Annals of Statistics*, 50(3):1519–1540, 2022.

[51] Junier Oliva, Barnabás Póczos, and Jeff Schneider. Distribution to distribution regression. In *International Conference on Machine Learning*, pages 1049–1057. PMLR, 2013.

[52] Junier Oliva, Willie Neiswanger, Barnabás Póczos, Jeff Schneider, and Eric Xing. Fast distribution to real regression. In *Artificial Intelligence and Statistics*, pages 706–714. PMLR, 2014.

[53] Victor M Panaretos and Yoav Zemel. Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771–812, 2016.

[54] Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019.

[55] Victor M Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

[56] Victor Patrangenaru and Leif Ellingson. *Nonparametric statistics on manifolds and their applications to object data analysis*. CRC Press, 2015.

[57] Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with euclidean predictors. *Annals of Statistics*, 47(2):691–719, 2019.

[58] Alexander Petersen, Hans-Georg Müller, et al. Functional data analysis for density functions by transformation to a hilbert space. *Annals of Statistics*, 44(1): 183–218, 2016.

[59] Alexander Petersen, Chao Zhang, and Piotr Kokoszka. Modeling probability density functions as data objects. *Econometrics and Statistics*, 21:159–178, 2022.

[60] Barnabás Póczos, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman. Distribution-free distribution regression. In *Artificial Intelligence and Statistics*, pages 507–515. PMLR, 2013.

[61] Stanislav P Ponomarev. Submersions and preimages of sets of measure zero. *Siberian Mathematical Journal*, 28(1):153–163, 1987.

[62] Philippe Rigollet and Jonathan Weed. Uncoupled isotonic regression via minimum wasserstein deconvolution. *Information and Inference: A Journal of the IMA*, 8(4):691–717, 2019.

[63] Magnus Röding, Siobhan J Bradley, Nathan H Williamson, Melissa R Dewi, Thomas Nann, and Magnus Nydén. The power of heterogeneity: Parameter relationships from distributions. *Plos one*, 11(5):e0155718, 2016.

[64] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[65] Moshe Shaked and J George Shanthikumar. *Stochastic orders.* Springer, 2007.

[66] Martin Slawski and Bodhisattva Sen. Permuted and unlinked monotone regression in $\mathbb{R}^d$: an approach based on mixture modeling and optimal transport. *arXiv preprint arXiv:2201.03528*, 2022.

[67] Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

[68] Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

[69] Chengliang Tang, Nathan Lenssen, Ying Wei, and Tian Zheng. Wasserstein distributional learning. *arXiv preprint arXiv:2209.04991*, 2022.

[70] Beth C Truesdale and Christopher Jencks. The health effects of income inequality: averages and disparities. *Annual Review of Public Health*, 37:413–430, 2016.

[71] user940. Are convex functions enough to determine a measure? Mathematics Stack Exchange, 2015. URL:https://math.stackexchange.com/q/1376262 (version: 2015-07-28).

[72] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[73] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[74] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[75] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[76] Wei Biao Wu and Xiaofeng Shao. Limit theorems for iterated random functions. *Journal of Applied Probability*, 41(2):425–436, 2004.

[77] Qilin Yuan, Ting Wang, Panlong Yu, Hanzhuang Zhang, Han Zhang, and Wenyu Ji. A review on the electroluminescence properties of quantum-dot light-emitting diodes. *Organic Electronics*, 90:106086, 2021.

[78] Yoav Zemel and Victor M. Panaretos. Fréchet means and procrustes analysis in wasserstein space. *Bernoulli*, 25(2):932–976, 2019.

[79] Chao Zhang, Piotr Kokoszka, and Alexander Petersen. Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis*, 43(1):30–52, 2022.

[80] Changbo Zhu and Hans-Georg Müller. Autoregressive optimal transport models. *arXiv preprint arXiv:2105.05439*, 2021.

Laya Ghodrati

Contact Information

EPFL SB MATH SMAT
MA B1 493 (Bâtiment MA)
Station 8
CH-1015 Lausanne

Linkedin: www.linkedin.com/in/laya-ghodrati
✉ E-mail: laya71@gmail.com

Education

- **Ph.D. in Mathematical Statistics** July 2018–2023 (expected)
  - Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
  - Supervisor : Prof. Victor Panaretos
- **Master (M2) in Applied Mathematics and Statistics** September 2016–August 2017
  - Paris Descartes University, France.
  - **Highest Distinction**
- **B.Sc. in Computer Science and**
  B.Sc. in Mathematics (Double Major Program) September 2011–June 2016
  - Sharif University of Technology, Tehran, Iran.
  - **Ranked 2nd**, based on GPA among all Computer Science entrants of 2011
- **Diploma in Mathematics and Physics** September 2007–June 2011
  - Farzanegan High School (Branch of NODET*), Tehran, Iran.
    ***National Organization for Development of Exceptional Talent**

Internships

- **Quantitative Research Intern** June 2022–August 2022
  - Susquehanna International Group (SIG), Dublin
  - Assisted in optimizing speed and efficiency of a trading algorithm in the futures market.
- **Research Assistant Intern** March 2017–December 2017
  - Statistical Physics Laboratory, École Normale Supérieure (ENS), Paris
  - Supervisor: Prof. Vincent Hakim
  - Participated in modeling stochastic gradient descent in a biological network (Cerebellum).
- **Research Assistant Intern** July 2014–September 2014
  - Laboratory for cryptologic algorithms, EPFL
  - Supervisor: Prof. Arjen Lenstra
  - Contributed to the implementation and parallelization of the sieve algorithm on Xeon Phi coprocessor for solving the Euclidean shortest vector problem (SVP) on lattices.

Publications

- **L. Ghodrati and V.M. Panaretos** "Distribution-on-Distribution Regression via Optimal Transport Maps", published in Biometrika.
- **L. Ghodrati and V.M. Panaretos** "Minimax Rate for Optimal Transport Regression Between Distributions", published in Statistics and Probability Letters
- **L. Ghodrati and V.M. Panaretos** "Distributional Autoregression and Iterated Transportation", arXiv
- **L. Ghodrati and V.M. Panaretos** "Transportation of Measure Regression in Higher Dimensions", in preparation

Honors and Awards

- **Scholarship** from École doctorale Cerveau-Cognition-Comportement 2017–2018
- **Scholarship** from Fondation Sciences Mathématiques de Paris (FSMP) 2016–2017
- Recipient of the grant from **Iran's National Elites Foundation** 2011–2016
- **Ranked 2nd and 3rd in ACM** programming competitions
  in Sharif University-Mathematics section 2014, 2015
- Summer **internship grant** from EPFL 2014
- **Bronze medal** in **Iranian National Olympiad on Informatics** 2010

Academic Services

- **Teacher Assistant**, EPFL

○ Statistics (Bachelor)	Spring 2019, 2020 and 2022
○ Statistics for Data Science (Master)	Fall 2018, 2020 and 2021
○ Randomization and Causation (Bachelor)	Spring 2021
○ Statistics and Probability for Life Science (Bachelor)	Fall 2019

- **Teacher Assistant**, Sharif University of Technology

○ Probability and Applications	Spring 2014
○ Design and Analysis of Algorithms	Spring 2014
○ Data Structures and Algorithms	Fall 2013

- **Olympiad Teacher**, Farzanegan-1 High School — September 2012–August 2013
  - Teaching Combinatorics and Algorithms to prepare students for participating in the national Informatics Olympiad.

- **Student Co-Supervisor**, EPFL

○ Sandro Barissi, Master semester project on **Fairness in Decision Making**	Spring 2022
○ Mehdi Guelzim, Master thesis project on **Statistical Optimal Transport**	Fall 2021
○ Lorraine Jacot-Descombes, Master semester project on **Causal Inference**	Spring 2019

Programming

- Python, C++, Java , R

Languages

- English (fluent), Persian (native), French (B2/intermediate), Arabic (beginner)