Journal of Machine Learning Research 25 (2024) 1-49 Submitted 3/22; Revised 5/24; Published 8/24

Functional optimal transport: regularized map estimation and domain adaptation for functional data

Jiacheng Zhu* JZHU.ZJC@GMAIL.COM
Department of Mechanical Engineering, Carnegie Mellon University
Pittsburgh, PA 15213, USA

Aritra Guha* AG997X@ATT.COM
Data Science & AI Research, AT&T Chief Data Office
Bedminster, NJ 07921, USA

Dat Do* DODAT@UMICH.EDU
Department of Statistics, University of Michigan
Ann Arbor, MI 48105, USA

Mengdi Xu MENGDIXU@ANDREW.CMU.EDU
Department of Mechanical Engineering, Carnegie Mellon University
Pittsburgh, PA 15213, USA

XuanLong Nguyen XUANLONG@UMICH.EDU
Department of Statistics, University of Michigan
Ann Arbor, MI 48105, USA

Ding Zhao DINGZHAO@CMU.EDU
Department of Mechanical Engineering, Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Tommi Jaakkola

Abstract

We introduce a formulation of *regularized* optimal transport problem for distributions on function spaces, where the stochastic map between functional domains can be approximated in terms of an (infinite-dimensional) Hilbert-Schmidt operator mapping a Hilbert space of functions to another. For numerous machine learning applications, data can be naturally viewed as samples drawn from spaces of functions, such as curves and surfaces, in high dimensions. Optimal transport for functional data analysis provides a useful framework of treatment for such domains. Since probability measures in infinite dimensional spaces generally lack absolute continuity (i.e., with respect to non-degenerate Gaussian measures), the Monge map in the standard optimal transport theory for finite dimensional spaces typically does not exist in the functional settings arising in such machine learning applications. This necessitates a suitable notion of approximation for the best pushforward measure to be obtained via a transport map. Indeed, our approach to the transportation problem in functional spaces is by a suitable regularization technique — we restrict the class of transport maps to be a Hilbert-Schmidt space of operators. Within this regularization framework, we develop an efficient algorithm for finding the stochastic transport map between functional domains and provide theoretical guarantees on the existence, uniqueness, and consistency of our estimate for the Hilbert-Schmidt space of compact linear operators. We validate our method on synthetic datasets and examine the functional properties of the transport

map. Experiments on real-world datasets of robot arm trajectories further demonstrate the effectiveness of our method on applications in domain adaptation.

Keywords: Optimal transport, Optimal transport map estimation, Functional data analysis, Hilbert Schmidt operator, Domain adaptation

1. Introduction

Optimal transport (OT) is a formalism for finding and quantifying the movement of mass from one probability distribution to another (Villani, 2008). In recent years, it has been instrumental in the development of various new machine learning methods, including deep generative modeling (Arjovsky et al., 2017; Salimans et al., 2018), unsupervised learning (Ho et al., 2017; Mallasto and Feragen, 2017) and domain adaptation (Ganin and Lempitsky, 2015; Bhushan Damodaran et al., 2018). As statistical machine learning algorithms are applied to increasingly complex domains, it is of interest to develop optimal transport-based methods for complex data structures. A particularly common form of such structures arises from functional data — data that may be viewed as random samples of smooth functions, curves, or surfaces in high dimension spaces (Ferraty and Vieu, 2006; Ramsay and Silverman, 2005; Hsing and Eubank, 2015; Mirshani et al., 2019; Dupont et al., 2021). Examples of real-world applications involving functional data are numerous, ranging from robotics (Deisenroth et al., 2013) and natural language processing (Rodrigues et al., 2014) to economics (Horváth and Kokoszka, 2012) and healthcare (Cheng et al., 2020). Therefore, it is of interest to extend and develop a suitable optimal transport formulation to the *functional* data domains.

The goal of this paper is to provide a novel formulation of the optimal transport problem in function spaces, to develop an efficient learning algorithm for estimating a suitable notion of optimal stochastic mapping that transports samples from one functional domain to another, to provide theoretical guarantees regarding the *existence*, *uniqueness*, and *consistency* of our estimates, and to demonstrate the effectiveness of our approach to several application domains where the functional optimal transport (FOT) viewpoint proves natural and useful. There are several formidable challenges: both the source and the target function spaces can be quite complex and in general of *infinite dimensions*. One needs to deal with probability distributions over such spaces, which is difficult if one is to model them with data. Moreover, optimal coupling or optimal transport map between the two distributions on infinite dimensional spaces is generally hard to characterize and compute efficiently, except for very few specific cases. Yet, to be useful in practice, one must find an explicit transport map that can approximate the optimal coupling well, i.e., to find an approximate solution to the original Monge problem (Villani, 2008).

The primary technical challenge here, more specifically, is that the optimal Monge map may not exist in general. By a Brenier-type theorem, a sufficient condition for the existence of the Monge map under a suitable convex cost function is that the source distribution be absolutely continuous with respect to non-degenerate Gaussian measures in the source domain (see Ambrosio et al. (2005), Sec. 6.2). In finite dimensional (Euclidean) domains, absolute continuity with respect to non-degenerate Gaussian measures is equivalent to absolute continuity with respect to the Lebesgue measure. However, in infinite dimensional domains, probability measures tend to lack the requisite absolute continuity. In fact, distributions in infinite-dimensional space can be discrete and tend to be singular to each other (Kakutani,

1948). Historically, the lack of existence of the Monge map was a key motivation for Kantorovich's optimal coupling formulation, which is well-posed for probability distributions in Polish spaces, and the discovery of Brenier-type theorems linking the optimal coupling to that of Monge map helped to ignite renewed interest and fresh new developments in the field of optimal transport in the past several decades. Since we work in a setting where the Monge map is not expected to exist in general, and moreover, a direct application of the Kantorovich's optimal coupling formulation seems difficult, we shall take a rather natural approach based on regularization: we seek to find the best deterministic map within a class of operators acting on the space of functions in the source domain. As we shall see shortly, this approach can be viewed as a regularized form of the optimal coupling problem as well.

Our formulation is relevant to a growing interest, especially in the machine learning literature, in finding an explicit optimal transport map linked to the Monge problem, although such attempts were mainly confined to finite-dimensional domains. For discrete distributions, map estimation can be tackled by jointly learning the coupling and a transformation map (Perrot et al., 2016). This basic idea and extensions were shown to be useful for the alignment of multimodal distributions (Lee et al., 2019) and word embedding (Zhang et al., 2017; Grave et al., 2019); such joint optimization objective was shown (Alvarez-Melis et al., 2019) to be related to the softassign Procrustes method (Rangarajan et al., 1997). The learned map, although usually not the optimal Monge map, is particularly useful for applications such as domain adaptation and generative modeling. Meanwhile, a different strand of work focused on scaling up the computation of the approximation of Monge map (Genevay et al., 2016; Meng et al., 2019), including approximating transport maps with neural networks (Seguy et al., 2017; Makkuva et al., 2019), deep generative models (Xie et al., 2019), and flow models (Huang et al., 2020). It is emphasized that all these methods are not quite suitable for capturing the distributions on the space of functions. Recent developments on Gromov-Wasserstein distance enable the comparisons of distributions on space of different dimensions (Mémoli, 2011). An alternative approach to constructing distances between probability measures on (Euclidean) spaces of different dimensions was recently proposed by Cai and Lim (2022). These techniques are relevant but not immediately applicable to the functional domains, which are of infinite dimensions. In addition, a common feature of functional data analysis is that the function samples are typically observed at the different and possibly varying number of design points. A naive approach to functional data is to treat a function as a vector of components sampled at a number of design points in its domain. Such an approach fails to exploit the fine structures (e.g., continuity, regularity) present naturally in many functional domains. Moreover, non-functional approaches to functional data may be highly sensitive to the choice of design points as one moves from one domain to another.

Most known results and techniques on optimal transport between distributions on function spaces are related to Gaussian measures and Gaussian processes (Mallasto and Feragen, 2017; Masarotto et al., 2019; Knott and Smith, 1984; Pigoli et al., 2014). These results are natural generalization from those of the multivariate Gaussian distributions (Dowson and Landau, 1982; Givens and Shortt, 1984). Specifically, the 2-Wasserstein distance between Gaussian processes with certain covariance coincides with the Procrustes distance between the two covariance operators (cf. Section 2 of Masarotto et al. (2019)). Furthermore, there exists a linear subspace in Hilbert spaces where the optimal map between two centered

Gaussian processes is well-defined as a linear operator. In practice, the Gaussian distribution assumption is clearly too restrictive in many domains. Our work may be viewed as a first step at addressing optimal transport in the domains of functions that go beyond the Gaussian assumption, and with a particular focus on learning the explicit transport map for sampled functional data.

In our approach the mathematical machinery of functional data analysis (FDA) (Hsing and Eubank, 2015; Ramsay and Silverman, 2005), along with recent advances in computational optimal transport via regularization techniques will be brought to bear on the aforementioned problems. There are several ingredients in our work. First, we take a probabilistic model-free approach, by avoiding making assumptions on the source and target distributions of functional data. Instead, we aim to learn the (stochastic) transport map directly. Second, we follow the FDA perspective by assuming that both the source and target distributions be supported on suitable Hilbert spaces of functions H_1 and H_2, respectively. A map $T : H_1 \to H_2$ sending elements of H_1 to that of H_2 will be represented by a class of linear operators, namely the integral operators. In fact, we shall restrict ourselves to Hilbert-Schmidt operators, which are compact, computationally convenient to regularize and amenable to theoretical analysis. Finally, the optimal *deterministic* transport mapping between two probability measures on function spaces may not exist, due to the general lack of absolute continuity discussed earlier. To overcome this difficulty, we enlarge the space of transport maps by allowing for stochastic coupling Π between the two domains $T(H_1) \subseteq H_2$ and H_2, while the complexity of such coupling can be controlled via the entropic regularization technique (Cuturi, 2013).

It is quite interesting to note that our formulation for optimal transport in the functional domains has two complementary interpretations: it can be viewed as learning an integral operator T regularized by a transport plan (a coupling distribution Π) or it can also be seen as an optimal coupling problem for Π (the Kantorovich problem), which is associated with a cost matrix parameterized by the integral operator T. In any case, we take a joint optimization approach for the transport map$T$ and the coupling distribution Π in functional domains. Subject to suitable regularization on the space of transport maps, the existence of the optimal (T, Π) and the uniqueness of T can be established, which leads to a consistency result of our estimation procedure.

There are several advantages in our choice of bounded linear operators for modeling the transport map. First, in functional analysis and functional data analysis in particular, bounded linear operators (including the rich class of integral operators) are the main workhorse for representing transformation among spaces of functions (Yosida, 1995; Ramsay and Silverman, 2006; Hsing and Eubank, 2015). Using this class of operators allows us to draw from the principled machinery and tools from functional analysis to establish a solid theoretical foundation for optimal transport in infinite-dimensional function spaces. Second, compact linear operators are easily regularizable, which result in fast computational procedures for learning from the functional data. The third advantage is the interpretability of the representation of the transport map T, which can be helpful in applications. In fact, linearity in the representation space is the desired evaluation protocol for even the most sophisticated large-scale pre-trained models (Radford et al., 2021), which are powering applications across both natural language and image generation.

There are several limitations in our approach. First, bounded (or compact) linear operators may be too restrictive in some domains. Learning nonlinear operators in infinite

dimensional spaces is an interesting direction, but is beyond the scope of this paper. Moreover, as discussed earlier, for discrete probability measures in the source domain, even nonlinear operators are not enough since the Monge map generally does not exist anyway. When this is the case, a reasonable approach, in our opinion, is to revert to the Kantorovich's coupling formulation, where linear operators can still be efficiently utilized as a building block from a regularization viewpoint for the optimal coupling problem. Indeed, our modeling choice is sufficiently rich when coupled with the stochastic coupling to obtain the optimal (T, Π), which ensures both the existence and uniqueness of the solution to our formulation of the optimal coupling problem. The second limitation is more practical, as it is related to the fact that the learned linear operator T is represented in terms of its action on the eigenfunction basis of the function spaces of the source and target domains. This also highlights a key distinction for the functional optimal transport problem and our corresponding approach from the linear transformation techniques for fixed-dimensional vector spaces, which are relatively simpler and do not have to grapple with this issue (see, e.g. Perrot et al. (2016); Alvarez-Melis et al. (2019)). In many practical domains, the eigenfunction basis may not be available for certain types of datasets. In that case, we may rely on available methods to construct data-driven basis functions, such as functional principal component analysis (FPCA) (Shang, 2014; Liu et al., 2017).

In summary, the contributions presented in this paper are the following.

- We propose functional optimal transport (FOT), a formulation of regularized optimal transport in infinite-dimensional functional spaces. We take a probabilistic model-free approach, by avoiding making assumptions on the source and target distributions of functional data.

- An approximate optimization algorithm is developed to estimate the transport map from data. We follow the FDA perspective by assuming that both the source and target distributions be supported on suitable Hilbert spaces of functions H_1 and H_2, respectively. A map$T : H_1 \to H_2$ sending elements of H_1 to that of H_2 will be represented by a class of linear operators, namely the integral operators. Despite the infinite-dimensional nature of this problem, we propose an approximate estimator and solve it with an alternative minimization algorithm.

- We establish the existence and uniqueness for such a transport mapon empirical samples of functional data from both source and target domains, as well as consistency theorems providing the support for our estimator. Specifically, we show the asymptotic convergence in terms of the number of basis functions K, the number of observed sample functions n, and the number of design points d. To the best of our knowledge, this is the first work in which a rigorous statistical theory for optimal transport in the domains of functions is established.

- Simulation studies and experiments are conducted to validate our method and the associated theory. First, the convergence properties of our map estimation when the samples are synthesized with known basis functions are verified via simulations. In the task of estimating an explicit transport map for two sets of functional data, our proposed method displayed superior performance from both qualitative and quantitative perspectives in comparison to non-functional techniques. Next, our method is applied

to several real-world datasets. We conduct the optimal transport domain adaptation for predicting robot-arm motion from two different datasets.

1.1 Organization

The remainder of the paper is organized as follows. Section 2 contains some preliminary background of optimal transport and functional data analysis. Section 3 presents the formulation of functional optimal transport based on Hilbert-Schmidt operators, followed by theoretical results on the existence, uniqueness, and consistency of our map estimator. In Section 4, we describe an implementation of our estimation procedure by solving a block coordinate-wise convex optimization problem. The result is an efficient algorithm for finding explicit transport maps that can be applied on sampled functions. Then, in Section 5, the effectiveness of our approach is validated first on synthetic datasets of smooth functional data and then applied in a suite of experiments for mapping real-world 3D trajectories between robotic arms with different configurations. All proofs are given in Section 6. Finally, Section 7 provides further discussions of related work, as well as several directions for future work.

2. Preliminaries

This section provides some basic background of optimal transport and functional data analysis.

2.1 Optimal transport

The basic problem in optimal transport, also known as the Kantorovich problem (Villani, 2008; Kantorovitch, 1958), is to find an optimal coupling π of given measures μ on space $\mathcal{X}$ and ν on space $\mathcal{Y}$ to minimize

$$\inf_{\pi \in \Pi} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y), \text{ subject to } \Pi = \{\pi : \gamma^{\mathcal{X}}_{\#}\pi = \mu, \gamma^{\mathcal{Y}}_{\#}\pi = \nu\}. \tag{1}$$

In the above display, $c : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}^+$ is a cost function and $\gamma^{\mathcal{X}}$, $\gamma^{\mathcal{Y}}$ denote projections from $\mathcal{X} \times \mathcal{Y}$ onto $\mathcal{X}$ and $\mathcal{Y}$ respectively, and hence the corresponding pushforward measures denoted by $\gamma^{\mathcal{X}}_{\#}\pi, \gamma^{\mathcal{Y}}_{\#}\pi$ of π are its the marginal distributions on $\mathcal{X}$ and $\mathcal{Y}$. This optimization is well-defined and the optimal π exists under mild conditions (in particular, $\mathcal{X}, \mathcal{Y}$ are both separable and complete metric spaces, c is lower semi-continuous (Villani, 2008)). When $\mathcal{X} = \mathcal{Y}$ are metric spaces, $c(x, y)$ is the square of the distance between x and y, then the square root of the optimal cost given by (1) defines the Wasserstein metric $W_2(\mu, \nu)$ on the space of square integrable probability measures on $\mathcal{X}$. A related problem is Monge problem, where one finds a Borel map $T : \mathcal{X} \to \mathcal{Y}$ that realizes the infimum

$$\inf_{T} \int_{\mathcal{X}} c(x, T(x)) d\mu(x) \;\; \text{subject to} \;\; T_{\#}\mu = \nu. \tag{2}$$

Here the image $T_{\#}\pi$ denotes the pushforward measure of π by T. (Although T is traditionally referred to as a mapping, and the image $T_{\#}\pi$ as a map, here we use the two terms interchangably, following Villani (2008)).

By Brenier's theorem, the existence of the optimal deterministic map T is guaranteed when μ is an absolute continuous measure with respect to, say the Lebesgue measure on $\mathcal{X}$ for finite-dimensional source domain(Ambrosio et al., 2005; Villani, 2008), but T clearly does not exist in general when μ and ν are discrete probability measures . However, in various applications, it is of interest to find a deterministic map that approximates the optimal coupling to the Kantorovich problem. In many recent work, T is typically restricted to a family of maps $\mathcal{F}$ followed by joint optimization of T and π (Perrot et al., 2016; Alvarez-Melis et al., 2019; Grave et al., 2019; Seguy et al., 2017; Alvarez-Melis et al., 2020):

$$\inf_{\pi \in \Pi, T \in \mathcal{F}} \int_{\mathcal{X} \times \mathcal{Y}} c(T(x), y) d\pi(x, y), \tag{3}$$

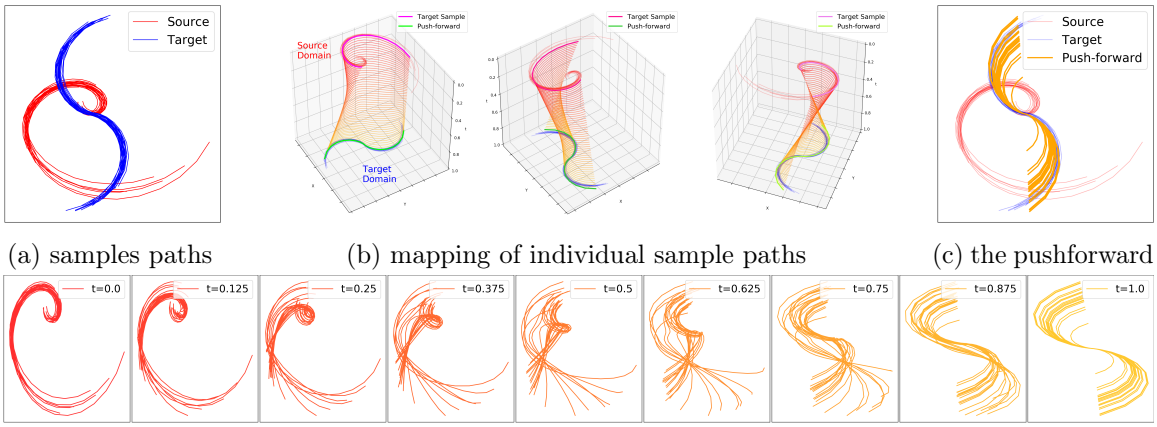
where $c : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$ is a cost function on $\mathcal{Y}$. On the one hand, the class of maps $\mathcal{F}$ may be chosen to be sufficiently rich to approximate the optimal transport maps for the measures μ and ν of interest defined on the respective spaces $\mathcal{X}$, $\mathcal{Y}$. On the other hand, $\mathcal{F}$ may be chosen to ease up the computational burden and facilitate meaningful interpretations. For instance, $\mathcal{F}$ may be a class of linear functions (e.g. rigid transformations) (Perrot et al., 2016; Alvarez-Melis et al., 2020), neural networks (Seguy et al., 2017).

At a high level, our approach will be analogous to (3), except that $\mathcal{X}$ and $\mathcal{Y}$ are taken to be spaces of functions, as we are motivated by applications in functional domains. As an illustration for a toy example, Figure 1 depicts a transport map that sends sample paths from the famous `Swiss-roll` curve data set to that of the target `Wave` curve data set. In a real-world application considered later in Section 5, $\mathcal{X}$ and $\mathcal{Y}$ represent the space of (smooth) trajectories of robot motions. A natural and powerful approach to such data domains is the framework of functional data analysis. In this framework, the data may be viewed as samples of random functions. In particular, we will be working with distributions on Hilbert spaces of functions, while $\mathcal{F}$ is a suitable class of operators acting on such Hilbert spaces. We proceed to a brief background of FDA in the sequel.

2.2 Functional data analysis

Functional data analysis adopts the perspective that certain types of data may be viewed as samples of random functions, which are viewed as random elements taking value in Hilbert spaces of functions (Hsing and Eubank, 2015). The data analysis techniques on functional data involve operators acting on Hilbert spaces. Let $A : H_1 \to H_2$ be a bounded linear operator, where H_1 (respectively, H_2) is a Hilbert space equipped with scalar product $\langle \cdot, \cdot \rangle_{H_1}$ (respectively, $\langle \cdot, \cdot \rangle_{H_2}$) and $\{U_i\}_{i \geq 1} (\{V_j\}_{j \geq 1})$ is the Hilbert basis in H_1 (H_2). We will focus on a class of compact integral operators, namely Hilbert-Schmidt operators, that are sufficiently rich for many applications and yet amenable to analysis and computation. A is said to be Hilbert-Schmidt if $\sum_{i \geq 1} \|AU_i\|_{H_2}^2 < \infty$ for any Hilbert basis $\{U_i\}_{i \geq 1}$. The space of Hilbert-Schmidt operators between H_1 and H_2, to be denoted by $\mathscr{B}_{HS}(H_1, H_2)$, is also a Hilbert space endowed with the scalar product $\langle A, B \rangle_{HS} = \sum_i \langle AU_i, BU_i \rangle_{H_2}$ and the corresponding Hilbert-Schmidt norm is denoted by $\| \cdot \|_{HS}$.

Denote the outer product operator between two elements $e_i \in H_i$ for $i = 1, 2$ by $e_1 \otimes e_2 : H_1 \to H_2$, which is defined by $(e_1 \otimes e_2) f = \langle e_1, f \rangle_{H_1} e_2$ for $f \in H_1$. An important fact of Hilbert-Schmidt operators is given as follows (see, e.g., Theorem 4.4.5 of (Hsing and Eubank, 2015)).



(a) samples paths (b) mapping of individual sample paths (c) the pushforward

(d) the geodesic parameterized by $t \in [0, 1]$ connecting the source and target domains.

Figure 1: Illustration of an estimated pushforward map that sends sample paths from the source of Swiss-roll curves to the target of Wave curves. (a) Datasets are a collection of continuous sample paths. (b) Three individual samples are mapped from source to target. (c) Resulting curves obtained by applying the pushforward map to the source's samples. (d) Illustration of the resulting geodesic between source and target distributions.

Theorem 1 *The linear space $\mathscr{B}_{HS}(H_1, H_2)$ is a separable Hilbert space when equipped with the HS inner product. For any choice of complete orthonormal basis system (CONS) $\{U_i\}$ and $\{V_j\}$ for H_1 and H_2 respectively, $\{U_i \otimes V_j\}$ forms a CONS for $\mathscr{B}_{HS}(H_1, H_2)$.*

As a result, the following representation of Hilbert-Schmidt operators and their norm will be useful.

Lemma 2 *Let $\{U_i\}_{i=1}^{\infty}, \{V_j\}_{j=1}^{\infty}$ be a CONS for H_1, H_2, respectively. Then any Hilbert-Schmidt operator $T \in \mathscr{B}_{HS}(H_1, H_2)$ can be decomposed as*

$$T = \sum_{i,j} \lambda_{ji} U_i \otimes V_j, \text{ where } \|T\|_{HS}^2 = \sum_{i,j} \lambda_{ji}^2. \tag{4}$$

3. Functional optimal transport: optimization and convergence analysis

We are ready to introduce a formulation for the functional optimal transport (FOT) problem, by reposing on the foundation of functional data analysis described earlier. Then we shall introduce estimators for solving the functional optimal transport problem from empirical data. Since we are formulating an infinite dimensional optimization problem, care must be taken to ensure the existence, uniqueness, and consistency of our proposed estimators, given sampled functions from source and target domains.

Given Hilbert spaces of functions H_1 and H_2, which are endowed with Borel probability measures μ and ν, respectively, we wish to find a Borel map $\Gamma : H_1 \mapsto H_2$ such that ν is the pushforward measure of μ by Γ. Expressing this statement probabilistically, if $f \sim \mu$ represents a random element of H_1, then Γf is a random element of H_2 and $\Gamma f \sim \nu$. As

noted in Section 2, such a map may not always exist, especially in the infinite-dimensional settings. Thus one is interested in finding a map by which the resulting pushforward measure approximates as well as possible the target distribution, although it is not necessarily a Monge map (Amos et al., 2022; Liu et al., 2022). This motivates the following formulation:

$$\Gamma := \arg \inf_{T \in \mathscr{B}_{HS}(H_1, H_2)} W_2(T_{\#}\mu, \nu), \tag{5}$$

where $T_{\#}\mu$ is the pushforward of μ by T, and W_2 is the 2-Wasserstein distance of probability measures on the metric space $(H_2, \|\cdot\|_{H_2})$ (we suppress the dependence on H_2 of the notation W_2 for ease of notations because we only consider 2-Wasserstein distance on H_2 in this work). The space of solutions of Eq. (5) may still be large and the problem itself might be ill-posed (the infimum is not achievable). Thus we consider imposing a shrinkage penalty, which leads to the problem of finding the infimum of the following objective function $J : \mathscr{B}_{HS} \to \mathbb{R}_+$:

$$\inf_{T \in \mathscr{B}_{HS}} J(T),\ J(T) := W_2^2(T_{\#}\mu, \nu) + \eta \|T\|_{HS}^2, \tag{6}$$

where $\eta > 0$ is a regularized hyperparameter. While regularization further reduces the approximation capabilities of the push-forward measure, it significantly enhances the robustness of estimating the map T. A contribution of this paper is to show that under suitably mild conditions, the aforementioned regularization technique ensures the existence and uniqueness of this map, and is amenable to an efficient computational procedure for the estimation problem.

To characterize this problem precisely, we shall place a mild condition on the moments of μ and ν, which are typically assumed for probability measures on Hilbert spaces (Lei, 2020).

(A.1) Assume

$$E_{f_1 \sim \mu} \|f_1\|_{H_1}^2 < \infty, \quad E_{f_2 \sim \nu} \|f_2\|_{H_2}^2 < \infty. \tag{7}$$

Several key properties of objective function (6) can be established as follows.

Lemma 3 *Under assumption (A.1), the following statements hold.*

(i) $W_2(T\#\mu, \nu)$ *is a Lipschitz continuous function of* $T \in \mathscr{B}_{HS}(H_1, H_2)$*, which implies that* $J : \mathscr{B}_{HS}(H_1, H_2) \to \mathbb{R}_+$ *is also continuous.*

(ii) J *is a strictly convex function.*

(iii) There are constants $C_1, C_2 > 0$ *such that* $J(T) \le C_1 \|T\|_{HS}^2 + C_2 \ \ \forall T \in \mathscr{B}_{HS}(H_1, H_2)$.

(iv) $\lim_{\|T\|_{HS} \to \infty} J(T) = \infty$.

Thanks to Lemma 3, the existence and uniqueness properties can be established.

Theorem 4 *Given (A.1), there exists a unique minimizer* T_0 *for problem* (6).

The challenge of solving (6) is that this is an optimization problem in the infinite-dimensional space of operators $\mathscr{B}_{HS}$. To alleviate this complexity, we reduce the problem to a suitable finite-dimensional approximation. We follow techniques in numerical functional analysis by taking a finite number of basis functions.

In particular, for some finite K_1, K_2, let $B_K = Span(\{U_i \otimes V_j : i = \overline{1, K_1}, j = \overline{1, K_2}\})$, where $K = (K_1, K_2)$. This yields the optimization problem of $J(T)$ over the space $T \in B_K$. The following result validates the choice of approximate optimization.

Lemma 5 *For each $K = (K_1, K_2)$, there exists a unique minimizer T_K of J over B_K. Moreover, $T_K \to T_0$ in $\|\cdot\|_{HS}$ as $K_1, K_2 \to \infty$.*

Consistency of M-estimator In practice, we are given i.i.d. samples $f_{11}, f_{12}, \ldots, f_{1n_1}$ from μ and $f_{21}, f_{22}, \ldots, f_{2n_2}$ from ν, the empirical version of our optimization problem becomes:

$$\inf_{T \in \mathscr{B}_{HS}} \hat{J}_n(T), \quad \hat{J}_n(T) := W_2^2(T_{\#}\hat{\mu}_{n_1}, \hat{\nu}_{n_2}) + \eta \|T\|_{HS}^2, \tag{8}$$

where $\hat{\mu}_{n_1} = \frac{1}{n_1}\sum_{l=1}^{n_1} \delta_{f_{1,l}}$ and $\hat{\nu}_{n_2} = \frac{1}{n_2}\sum_{k=1}^{n_2} \delta_{f_{2,k}}$ are the empirical measures, and $n = (n_1, n_2)$. We proceed to show that the minimizer of this problem exists and provides a consistent estimate of the minimizer of (6). The common technique to establish consistency of M-estimators is via the uniform convergence of objective functions $\hat{J}_n$ to J. The technical challenge here is that $\mathscr{B}_{HS}(H_1, H_2)$ is unbounded and locally non-compact. Thus care must be taken to ensure that the minimizer of (8) is eventually bounded so that a suitable uniform convergence behavior can be established, as explicated in the following key lemma:

Lemma 6 *Under assumption (A.1), the following hold.*

(i) For any fixed $C_0 > 0$,

$$\sup_{\|T\|_{HS} \le C_0} |\hat{J}_n(T) - J(T)| \xrightarrow{P} 0 \quad (n \to \infty). \tag{9}$$

(ii) For any n, K, $\hat{J}_n$ has a unique minimizer $\hat{T}_{K,n}$ over B_K. Moreover, there exists a finite constant D such that $P(\sup_K \|\hat{T}_{K,n}\|_{HS} < D) \to 1$ as $n \to \infty$.

Building upon the above results, we can establish consistency of our M-estimator when there are enough samples and the dimensions K_1, K_2 are allowed to grow with the sample size:

Theorem 7 *The minimizer of Eq. (8) for $\hat{T}_{K,n} \in B_K$ is a consistent estimate for the minimizer of Eq. (6). Specifically, $\hat{T}_{K,n} \xrightarrow{P} T_0$ in $\|\cdot\|_{HS}$ as $K_1, K_2, n_1, n_2 \to \infty$.*

It is worth emphasizing that the consistency of estimating $\hat{T}_{K,n}$ is ensured as long as sample sizes and approximate dimensions are allowed to grow. The specific schedule at which K_1, K_2 grows relatively to n_1, n_2 will determine the rate of convergence to T_0, which is also dependent on the choice of regularization parameter $\eta > 0$, the true probability measures μ, ν, and the choice of CONS. It is of great interest to have a refined understanding on this matter. In practice, we can choose K_1, K_2 by a simple cross-validation technique, which we shall discuss further in the sequel.

Finally, note that in Theorem 7, we assume that the data samples consist of the entire sample curves $\{f_{1,l}, f_{2,k}\}$ for $l = 1, \ldots, n_1$, $k = 1, \ldots, n_2$. In reality, the sampled functions $f_{1,l}$ and $f_{2,k}$ may be partially given only at selected design points on their domains. A consistency theorem, Theorem 8, for our estimator in this more realistic setting will be given in the following section.

4. Approximation, functional design and optimization

We will now translate the theoretical formulation for functional optimal transport and the regularized M-estimation framework presented in the preceding section into implementable algorithms. Recall that our FOT formulation is intrinsically an infinite dimensional problem: both source and target distributions are supported by infinite dimensional Hilbert spaces of functions, and so is the space of transport maps that we seek to estimate. On the other hand, we are given only a finite sample (n_1, n_2) of the source and target functions, and moreover such functions are observed only at a finite number of design points on their domains. Thus our approach is to derive approximate algorithms to approach the original objective of solving Eq's (6) and (8) by appropriate finite-dimensional approximations and by taking into considerations design choices for the space of functions and operators. Moreover, the consistency of our estimator in such practical forms of approximation is still maintained.

4.1 Approximation of the HS operator

Lemma 5 in the previous section paves the way for us to find an approximate solution to the original fully continuous infinite-dimensional problem, by utilizing finite sets of basis function, in the spirit of Galerkin method (Fletcher, 1984), which is justified by the consistency theorem (Theorem 7). Thus, we can focus on solving the optimization problem given in Eq. (8) instead of Eq. (6).

Choosing a basis $\{U_i\}_{i=1}^{\infty}$ of H_1 and a basis $\{V_j\}_{j=1}^{\infty}$ of H_2, and fixing K_1, K_2, we want to find T based on the $K_1 \times K_2$ dimensional subspace of $\mathscr{B}_{HS}(H_1, H_2)$ with the basis $\{U_i \otimes V_j\}_{i=\overline{1,K_1},j=\overline{1,K_2}}$. Lemma 2 gives us the following formula for T and its norm

$$T = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \lambda_{ji} U_i \otimes V_j, \quad \|T\|_{HS}^2 = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \lambda_{ji}^2. \tag{10}$$

As T is represented by matrix $\mathbf{\Lambda} := (\lambda_{ji})_{j,i=1}^{K_2,K_1}$, the cost to move function $f_{1,l}$ in H_1 to $f_{2,k}$ in H_2 is

$$\|Tf_{1,l} - f_{2,k}\|^2 = \left\| \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \lambda_{ji} V_j \langle f_{1,l}, U_i\rangle_{H_1} - f_{2,k} \right\|_{H_2}^2 =: C_{lk}(\mathbf{\Lambda}). \tag{11}$$

Hence, the optimization problem (8) as restricted to B_K can be written as

$$\min_{T\in B_K} \hat{J}_n(T) = \min_{\mathbf{\Lambda}\in\mathbb{R}^{K_2\times K_1}, \pi\in\hat{\Pi}} \sum_{l,k=1}^{n_1,n_2} \pi_{lk} C_{lk}(\mathbf{\Lambda}) + \eta\|\mathbf{\Lambda}\|_F^2, \tag{12}$$

where $\|\cdot\|_F$ is the Frobenius norm, and the empirical joint measure $\hat{\Pi} := \{\pi \in (\mathbb{R}^+)^{n_1\times n_2} |\ \pi \mathbf{1}_{n_2} = \mathbf{1}_{n_1}/n_1,\ \pi^T \mathbf{1}_{n_1} = \mathbf{1}_{n_2}/n_2\}$ with $\mathbf{1}_n$ a length n vector of ones. Eq.(12) indicates we need to simultaneously learn the HS operator T and the coupling distribution π.

For theoretical purposes we proceed to simplify the objective (12) to arrive at a finite-dimensional formulation specific to the basis $\{U_i\}_{i=1}^{\infty}$ of H_1 and basis $\{V_j\}_{j=1}^{\infty}$ of H_2. For

each $l = 1, \ldots, n_1$ and $k = 1, \ldots, n_2$, by Parseval's identity on H_2,

$$C_{lk}(\mathbf{\Lambda}) = \underbrace{\sum_{j=1}^{K_2} \left| \sum_{i=1}^{K_1} \lambda_{ji} \langle f_{1,l}, U_i \rangle_{H_1} - \langle f_{2,k}, V_j \rangle_{H_2} \right|^2}_{D_{lk}(\mathbf{\Lambda})} + \sum_{j=K_2+1}^{\infty} |\langle f_{2,k}, V_j \rangle_{H_2}|^2 . \tag{13}$$

Our optimization problem becomes

$$\begin{aligned} \sum_{l,k} \pi_{lk} C_{lk}(\mathbf{\Lambda}) + \eta \|\mathbf{\Lambda}\|_F^2 &= \sum_{l,k} \pi_{lk} D_{lk}(\mathbf{\Lambda}) + \sum_{j=K_2+1}^{\infty} \sum_{k=1}^{n_2} (\sum_{l=1}^{n_1} \pi_{lk}) \, |\langle f_{2,k}, V_j \rangle_{H_2}|^2 + \eta \|\mathbf{\Lambda}\|_F^2 \\ &= \sum_{l,k} \pi_{lk} D_{lk}(\mathbf{\Lambda}) + \sum_{j=K_2+1}^{\infty} \sum_{k=1}^{n_2} \frac{1}{n_2} \, |\langle f_{2,k}, V_j \rangle_{H_2}|^2 + \eta \|\mathbf{\Lambda}\|_F^2. \end{aligned}$$

Since the second term in the above display does not depend on $\mathbf{\Lambda}$ and π, it does not affect the optimization problem. The term $D_{lk}(\mathbf{\Lambda})$ can be further written as

$$D_{lk}(\mathbf{\Lambda}) = \|\mathbf{\Lambda} a_l - b_k\|_2^2, \tag{14}$$

where $a_{li} = \langle f_{1,l}, U_i \rangle_{H_1}$, and $a_l = (a_{li})_{i=1}^{K_1}$ are vectors in $\mathbb{R}^{K_1}$ (coordinates of $f_{1,l}$ in the first K_1 basis); $b_{kj} = \langle f_{2,k}, V_j \rangle_{H_2}$, and $b_k = (b_{kj})_{j=1}^{K_2}$ are vectors in $\mathbb{R}^{K_2}$ (coordinates of $f_{2,k}$ in the first K_2 basis). This leads to an equivalent presentation for (12)

$$\min_{T \in B_K} \hat{J}_n(T) = \min_{\mathbf{\Lambda} \in \mathbb{R}^{K_2 \times K_1}, \pi \in \hat{\Pi}} \sum_{l,k=1}^{n_1, n_2} \pi_{lk} D_{lk}(\mathbf{\Lambda}) + \eta \|\mathbf{\Lambda}\|_F^2. \tag{15}$$

In this way, we easily see a direct connection between a finite-dimensional approximation of the functional optimal transport relative to a pair of orthonormal bases $\{U_i\}_{H_1}$ and $\{V_j\}_{H_2}$ to a corresponding OT problem on fixed (finite) dimensional vectors.

4.2 Functional data computation via design points

In real-world applications with data in the functional domains, one typically does not directly observe functions $(f_{1,l})_{l=1}^{n_1}$ and $(f_{2,k})_{k=1}^{n_2}$ but only their values $(\mathbf{y}_{1,l})_{l=1}^{n_1}$ and $(\mathbf{y}_{2,k})_{k=1}^{n_2}$ at design points $(\mathbf{x}_{1,l})_{l=1}^{n_1}$ and $(\mathbf{x}_{2,k})_{k=1}^{n_2}$, respectively, where $\mathbf{x}_{1,l} \in X_1^{d_{1,l}}, \mathbf{y}_{1,l} \in \mathbb{R}^{d_{1,l}}, \mathbf{x}_{2,k} \in X_2^{d_{2,k}}, \mathbf{y}_{2,k} \in \mathbb{R}^{d_{2,k}} \; \forall \, l = 1, \ldots, n_1; k = 1, \ldots, n_2,$ and X_1 and X_2 denote the space of design points for the source and the target domain, respectively. In other words, $d_{1,l}$ and $d_{2,k}$ denote the possibly varying number of design points observed for the sampled function $f_{1,l}$ in the source and $f_{2,k}$ in the target domain, respectively. In order to evaluate the objective (15), the relevant inner products in H_1 and H_2 must be approximated using the observed values of sampled functions given at the design points.

As a concrete example, suppose that $H_1 = H_2 = L^2([0,1])$ so that the ordered design points $\mathbf{x}_{1,l} \in [0,1]^{d_{1,l}}$ and $\mathbf{x}_{2,k} \in [0,1]^{d_{2,k}}$ for all l, k. Moreover, assume that the support of μ and ν are contained in the subsets of continuous functions of L_2, and the basis functions $\{U_i\}$

and $\{V_j\}$ are continuous, then a simple numerical strategy that one can use to approximate $\langle f_{1,l}, U_i\rangle_{H_1}$ is by the Riemann sum approximation

$$\langle f_{1,l}, U_i\rangle_d := \sum_{j=2}^{d_{1,l}} (\mathbf{x}_{1l,j} - \mathbf{x}_{1l,j-1}) f(\mathbf{x}_{1l,j}) U_i(\mathbf{x}_{1l,j}). \tag{16}$$

In the above display, the subindex d is used to signify the fact our estimate of the relevant inner products for the sampled function is calculated using design points. We also say that $d \to \infty$ if $d_{1,l}, d_{2,k} \to \infty \, \forall \, l, k$. Thanks to the continuity of the sampled functions, for any functions f_1, f_2 being in the support of μ, ν respectively, we have as $d \to \infty$

$$\langle f_1, U_i\rangle_d \to \langle f_1, U_i\rangle_{H_1}, \quad \langle f_2, V_j\rangle_d \to \langle f_2, V_j\rangle_{H_2} \quad \forall i, j \in \mathbb{N}. \tag{17}$$

It is clear that the more design points where the function values are observed, the better the representation of continuous functions via a given orthonormal basis. In fact, this condition is sufficient for us to establish the consistency of our estimation procedure for the transport map as the number of design points increases. We formalize this intuition with the following consistency theorem, where we want to emphasize that this result holds for all approximation schemes satisfying Eq. (17), not only the approximation (16).

Theorem 8 *(i) For every n_1, n_2, K_1, K_2 and sequences of design points in source and target domains, the cost function*

$$\hat{J}_{n,K,d}(\mathbf{\Lambda}) = \min_{\pi \in \hat{\Pi}} \sum_{l,k=1}^{n_1,n_2} \pi_{lk} D_{lkd}(\mathbf{\Lambda}) + \eta \|\mathbf{\Lambda}\|_F^2, \tag{18}$$

where

$$D_{lkd}(\mathbf{\Lambda}) = \|\mathbf{\Lambda} a_{ld} - b_{kd}\|_2^2,$$

in which $a_{ld} = (\langle f_{1,l}, U_i\rangle_d)_{i=1}^{K_1}$ and $b_{kd} = (\langle f_{2,k}, V_j\rangle_d)_{i=1}^{K_2} \, \forall l, k$, has unique minimizer $\mathbf{\Lambda}_{n,K,d} \in \mathbb{R}^{K_2 \times K_1}$ that corresponds to operator $T_{n,K,d}$.

(ii) Suppose that for any natural index pair (i, j), there holds

$$\langle f, U_i\rangle_d \to \langle f, U_i\rangle_{H_1}, \langle g, V_j\rangle_d \to \langle g, V_j\rangle_{H_2}, \tag{19}$$

almost surely as $d \to \infty$, where $f \sim \mu$ and $g \sim \nu$. Then for any sequences of $n_1, n_2, K_1, K_2 \to \infty$ and $d \to \infty$ with a rate depends on n_1, n_2, K_1, K_2, we have $T_{n,K,d} \xrightarrow{P} T_0$ in $\|\cdot\|_{HS}$. Here, T_0 denotes the minimizer of the population version of FOT given in Eq. (6).

We make the following remarks regarding the evaluation of equivalent objectives given in Eq. (12), (15) and their estimate (18).

- It is worth noting that the objective function (18) can be computed easily from the sampled function observations. Moreover, our method works even in the case where different functions are observed at different design points (with possibly different numbers of design points). It is quite obvious that one *cannot* treat each function as a multidimensional vector to apply existing multivariate OT techniques in this case due to the dimensions mismatch.

- The objectives derived in the foregoing require the selection of basis functions for the Hilbert space in both the source and target domains. Since our method requires finite-dimensional approximations, a particular choice of orthonormal bases may have a substantial impact on the number of basis functions that one ends up using for approximating the support of the distributions (of the source and the target domain), and for the representation of the approximate pushforward map going from one domain to another. Note that increasing K_1 and K_2 can lower the objective function, but it may negatively affect the generalization of the estimate as we only observe a finite number of sampled functions (at a finite number of design points). Cross-validation is a simple and very effective technique for choosing K_1, K_2 and regularization parameters η, γ. This technique will be demonstrated via a simulation study in Section 5.1.

- As an example of the choice of basis functions for H_1 and H_2, we may take those that arise from a user-specified kernel via Mercer's theorem. Recall that if K is a continuous, symmetric, non-negative definite kernel on a measurable space $(E, \mathcal{B}, \mu)$ then it admits the following representation

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s) \phi_j(t), \tag{20}$$

  where the convergence is absolute and uniform, and $(\phi_j)_{j=1}^{\infty}$ forms an orthogonal basis for $L^2(E, \mathcal{B}, \mu)$. Examples of such bases can be found in (Wang, 2008; Zhu et al., 1997).

- When we discretize our objective from infinite-dimensional data, it is worth noting that the truncation of basis functions and sparse representation is necessary to eliminate computational challenges. For example, although certain continuous representations, such as Reproducing Kernel Hilbert Spaces (RKHS), allow for non-asymptotic generalization bounds (Maurer, 2008), their application is relatively restrictive. This is due to the fact that the Gaussian measure places zero probability on the corresponding RKHS, limiting its practical utility. On the other hand, in the formulation presented in this work, the support of distributions are general Hilbert spaces of functions, providing a considerably richer function space.

- A more adaptive approach for estimating basis functions is to use empirical samples. This can be achieved through the analysis of functional principal components (FPCA). Since FPCA is employed in some of our numerical experiments, we provide a brief introduction here. The literature on FPCA is extensive and covers a wide range of topics, such as incorporating smoothness (Foutz and Jank, 2010; Liu et al., 2017), robustness (Gervini, 2008), and sparsity (Kayano and Konishi, 2010). For theoretical perspectives of FPCA, see Dauxois et al. (1982); Yao et al. (2005); Hsing and Eubank (2015).

  A basic starting point of FPCA is a result of Karhunen and Loéve, who developed the optimal series expansion theory for continuous stochastic processes. Given N real valued functions $y_1, \ldots, y_N \in L^2$ defined on a closed interval $\mathcal{T} \subset \mathbb{R}$. We would like to select a collection of weight functions $\beta : \mathcal{T} \to \mathbb{R}$ that outlines the most significant

types of variation, which is captured by

$$f_i = \langle \beta, y_i \rangle_2 = \int \beta(t) y_i(t) dt, i = 1, ..., N.$$

Similar to the multivariate principal component analysis technique, one first finds a weight function β_1 by solving the maximization problem:

$$\max_{\beta} \frac{1}{N} \sum_{i=1}^{N} (\langle \beta_1, y_i \rangle_2)^2, \|\beta_1\|_2^2 = 1.$$

Then, inductively, for $m > 1$, one finds the next weight function β_m that maximizes $\frac{1}{N} \sum_{i=1}^{N} (\int \beta_m x_i)^2$ with the normalizing restriction $\|\beta_m\|_2^2 = 1$ and the orthogonality restrictions $\langle \beta_k, \beta_m \rangle_2 = 0, k < m$. The obtained weight functions are called the principal components and can serve as the basis functions. In practice, the maximization problem can be seen as an eigenvalue problem so that we can employ efficient methods (Ramos-Carreño et al., 2022) such as singular value decomposition (Section 2.5 of (Bie, 2021)).

4.3 Optimization algorithms

We are now ready to describe in detail the optimization algorithm for solving Eq. (12), or equivalently, Eq. (15). Recall that for functions $(f_{1,l})_{l=1}^{n_1}$ and $(f_{2,k})_{k=1}^{n_2}$ we observe their values $(\mathbf{y}_{1,l})_{l=1}^{n_1}$ and $(\mathbf{y}_{2,k})_{k=1}^{n_2}$ at design points $(\mathbf{x}_{1,l})_{l=1}^{n_1}$ and $(\mathbf{x}_{2,k})_{k=1}^{n_2}$, respectively, where $\mathbf{x}_{1,l}, \mathbf{y}_{1,l} \in \mathbb{R}^{d_{1,l}}, \mathbf{x}_{2,k}, \mathbf{y}_{2,k} \in \mathbb{R}^{d_{2,k}} \; \forall \, l, k$. So the objective function that we use for the optimization is

$$\underset{\mathbf{\Lambda} \in \mathbb{R}^{K_2 \times K_1}, \pi \in \hat{\Pi}}{\arg\min} \sum_{l,k=1}^{n_1,n_2} \pi_{lk} C_{lk}(\mathbf{\Lambda}) + \eta \|\mathbf{\Lambda}\|_F^2, \tag{21}$$

where

$$C_{lk}(\mathbf{\Lambda}) = \left\| \mathbf{V}_{2k} \mathbf{\Lambda} \mathbf{U}_{1l}^T \mathbf{y}_{1,l} - \mathbf{y}_{2,k} \right\|_2^2, \tag{22}$$

where $\mathbf{U}_{1l} = [U_1(\mathbf{x}_{1,l}), \ldots, U_{K_1}(\mathbf{x}_{1,l})] \in \mathbb{R}^{d_{1,l} \times K_1}, \mathbf{V}_{2k} = [V_1(\mathbf{x}_{2,k}), \ldots, V_{K_2}(\mathbf{x}_{2,k})] \in \mathbb{R}^{d_{2l} \times K_2}$ are basis functions that evaluated on the given design points (equivalently, we can work directly with (18)).

To speed up the computation of the classical optimal transport objective, a useful technique is to include a negative entropic penalty (Cuturi, 2013) defined as $\Omega_\gamma(\pi) = \gamma_h \sum_{l,k=1}^{n_1,n_2} \pi_{lk} \log \pi_{lk}$. However, entropic regularization keeps the probabilistic coupling dense and causes the lack of sparsity (Blondel et al., 2018). To promote sparsity, we can impose an ℓ_p penalty by taking $\Omega_\gamma(\pi) = \gamma_p \sum_{l,k=1}^{n_1,n_2} \pi_{lk}^p$, for some $p \geq 1, \gamma_p > 0$. This ensures that the optimal coupling (π_{lk}) has fewer active parameters thereby easing up the computing burden for large datasets. This can be considered as promoting a robustness criterion in addition to shrinkage, a similar behavior associated with the Huber loss (Huber, 1964). Hence, by imposing an additional regularization term to Eq. (21), we have our objective as

$$\underset{\mathbf{\Lambda} \in \mathbb{R}^{K_2 \times K_1}, \pi \in \hat{\Pi}}{\arg\min} \sum_{l,k=1}^{n_1,n_2} C_{lk}(\mathbf{\Lambda}) \pi_{lk} + \eta \|\mathbf{\Lambda}\|_F^2 + \Omega_\gamma(\pi), \tag{23}$$

Algorithm 1: Joint Learning of $\mathbf{\Lambda}$ and π

Input: Observed functional data $\{f_{1,l} = (\mathbf{x}_{1,l}, \mathbf{y}_{1,l})\}_{l=1}^{n_1}$ and $\{f_{2,k} = (\mathbf{x}_{2,k}, \mathbf{y}_{2,k})\}_{k=1}^{n_2}$, coefficient γ_h, γ_p, η, and learning rate l_r, source and target CONS $\{U_i(\cdot)\}_{i=1}^{K_1}$, $\{V_j(\cdot)\}_{j=1}^{K_2}$.
Initial value $\mathbf{\Lambda}_0 \leftarrow \mathbf{\Lambda}_{ini}$, $\pi_0 \leftarrow \pi_{ini}$.
$\mathbf{U}_{1l} = [U_1(\mathbf{x}_{1,l}), ..., U_{K_1}(\mathbf{x}_{1,l})]$, $\mathbf{V}_{2k} = [V_1(\mathbf{x}_{2,k}), ..., V_{K_2}(\mathbf{x}_{2,k})]$ # Evaluate eigenfunctions
for $t = 1$ **to** $T_{\max}$ **do**
  # Step 1. Update π_{t-1}
  $C_{lk} \leftarrow \|\mathbf{V}_{2k}\mathbf{\Lambda}_t\mathbf{U}_{1l}^T\mathbf{y}_{1,l} - \mathbf{y}_{2,k}\|_F^2$ # Cost matrix by Eq.(22)
  $\pi_t \leftarrow \text{argmin}_\pi \mathcal{L}(\pi, \lambda; \rho)$ # Fix $\mathbf{\Lambda}$ update π
  # Step 2. Update $\mathbf{\Lambda}_{t-1}$ with gradient descent
  Learn $\mathbf{\Lambda}_t$, solve Eq. (23) with fixed π_t using gradient descent
end for
Output: $\pi_{T_{max}}$, $\mathbf{\Lambda}_{T_{max}}$

where $\eta > 0$ is the regularization coefficient and $\Omega_\gamma(\pi)$ is the additional regularization term.

We provide a solution for local minima of this objective via an alternative minimization over $\mathbf{\Lambda}$ and π, whereby the first is fixed while the second is minimized, followed by the second fixed and the first minimized. The algorithm is described in Algorithm 1 and the explicit calculations are given below. Note that here we introduce our algorithm following the most general setting by using Eq. (22) as the transportation cost. Also we use the power regularization $\Omega_\gamma(\pi) = \gamma_p \sum_{l,k=1}^{n_1,n_2} \pi_{lk}^p$ in our objective. Later we will show that using the entropy regularization can let us utilize the Sinkhorn algorithm during the update.

Updating $\mathbf{\Lambda}$ with π fixed: Here we want to solve

$$\mathbf{\Lambda}_t = \underset{\mathbf{\Lambda} \in \mathbb{R}^{K_2 \times K_1}}{\arg\min}\ L(\mathbf{\Lambda}, \pi) = \underset{\mathbf{\Lambda} \in \mathbb{R}^{K_2 \times K_1}}{\arg\min} \sum_{l,k=1}^{n_1,n_2} \pi_{lk} C_{lk}(\mathbf{\Lambda}) + \eta \|\mathbf{\Lambda}\|_F^2. \tag{24}$$

The minimum is achieved by performing gradient descent updates, where the gradient is:

$$\nabla_{\mathbf{\Lambda}} L(\mathbf{\Lambda}, \pi) = 2 \sum_{l=1}^{n_1} \sum_{k=1}^{n_2} \pi_{lk} \left[(\mathbf{\Lambda}\mathbf{U}_{1l}^T\mathbf{y}_{1,l} - \mathbf{V}_{2k}^T\mathbf{y}_{2,k})\mathbf{y}_{1,l}^T\mathbf{U}_{1l} \right] + 2\eta\mathbf{\Lambda}. \tag{25}$$

Updating π with Λ fixed: Now we want to solve

$$\pi_t = \underset{\pi \in \hat{\Pi}}{\arg\min}\, L(\mathbf{\Lambda}, \pi) = \underset{\pi \in \hat{\Pi}}{\arg\min} \sum_{l,k=1}^{n_1,n_2} C_{lk}(\mathbf{\Lambda})\pi_{lk} + \gamma_p \sum_{l,k=1}^{n_1,n_2} \pi_{lk}^p. \tag{26}$$

To optimize for the probabilistic coupling π, we can consider this as a constrained linear programming problem and solve it through the augmented Lagrangian method (Afonso et al., 2010).

It is straightforward to extend the optimization framework to accommodate discrete source and target probability measures given by the (non-uniform) weights $p^s = (p_1^s, \ldots, p_{n_1}^s)$ and $p^t = (p_1^t, \ldots, p_{n_2}^t)$; the constraints are depicted as $\pi \in \hat{\Pi} := \{\sum_l^{n_1} \pi_{lk} = p_k^t, \forall k; \sum_k^{n_2} \pi_{lk} =$

$p_l^t, \forall l\}$. Here, we add a slack variable s to enforce the inequality constraints $\forall p_{ij} \geq 0$. Then the augmented Lagrangian takes the form

$$\begin{aligned}\mathcal{L}(\pi, s_{lk}, \lambda^a, \lambda^b, \lambda) &= \sum_{l,k=1}^{n_1,n_2} C_{lk}\pi_{lk} + \gamma_p \sum_{l,k=1}^{n_1,n_2} \pi_{lk}^p \\ &+ \sum_{k=1}^{n_2} \lambda_k^a (\sum_{l=1}^{n_1} \pi_{lk} - p_k^t) + \sum_{l=1}^{n_1} \lambda_l^b (\sum_{k=1}^{n_2} \pi_{lk} - p_l^s) + \frac{\rho_k}{2}(\sum_{l=1}^{n_1} \pi_{lk} - p_k^t)^2 + \frac{\rho_l}{2}(\sum_{k=1}^{n_2} \pi_{lk} - p_l^s)^2 \\ &+ \sum_{l,k=1}^{n_1,n_2} \lambda_{lk}(\pi_{lk} - s_{lk}) + \sum_{l,k=1}^{n_1,n_2} \frac{\rho_{lk}}{2}(\pi_{lk} - s_{lk})^2.\end{aligned} \tag{27}$$

In the above display, $\lambda^a \in \mathbb{R}^{n_1 \times 1}$, $\lambda^b \in \mathbb{R}^{n_2 \times 1}$, $\lambda \in \mathbb{R}^{n_1 \times n_2}$ are Lagrange multipliers, $s_{lk} \in \mathbb{R}^{n_1 \times n_2}$ are the slack variables. The sub-problem is

$$\begin{aligned}\pi_t, s_{lkt} &= \arg\min_{\pi, s_{lk}} \mathcal{L}(\pi, s_{lk}, \lambda^k, \lambda^l, \lambda^{lk}) \\ \lambda_k^a &\leftarrow \lambda_k^a + \rho_k(\sum_{l=1}^{n_1} \pi_{lk} - p_k^t) \\ \lambda_t^l &= \lambda_{t-1}^l + \rho_l(\sum_{k=1}^{n_2} \pi_{lk} - p_l^s) \\ \lambda_t^{lk} &= \lambda_{t-1}^{lk} + \rho_{lk}(\sum_{l,k=1}^{n_1,n_2} \pi_{lk} - s_{lk}).\end{aligned} \tag{28}$$

Entropic regularization: We may alternatively set the additional regularization term to be the negative entropy $\Omega_\gamma(\pi) = \gamma_h \sum_{l,k=1}^{n_1,n_2} \pi_{lk} \log \pi_{lk}$ to leverage the computational efficiency of the Sinkhorn algorithm. In that case, when updating π with $\mathbf{\Lambda}$ fixed, our problem reduces to an entropic regularized optimal transport problem:

$$\pi_t = \arg\min_{\pi \in \hat{\Pi}} L(\mathbf{\Lambda}, \pi) = \arg\min_{\pi \in \hat{\Pi}} \sum_{l,k=1}^{n_1,n_2} C_{lk}(\mathbf{\Lambda})\pi_{lk} + \gamma_h \sum_{l,k=1}^{n_1,n_2} \pi_{lk} \log \pi_{lk}. \tag{29}$$

This formulation reverts to a strictly convex optimization problem and we can efficiently obtain the solution via the Sinkhorn-Knopp algorithm (Cuturi, 2013). See Algorithm 2.

To summarize our learning scheme, during each iteration, our algorithm performs a gradient-type update for $\mathbf{\Lambda}$ with π fixed, followed by a step that updates the π. For the latter step, the algorithm either minimizes π following the Lagrangian multiplier method when using the power regularization, or invokes the Sinkhorn algorithm when using entropic regularization.

We end the description of the algorithms with the following additional remarks.

- In its final form in Eq. (23), our optimization formulation has a number of regularization terms. It is interesting to note how the different penalty terms play complementary roles in the final estimate of the joint parameter (T, Π) which represents the optimal

Algorithm 2: Sinkhorn algorithm

Input: Cost matrix $\mathbf{C} \in \mathbb{R}^{N \times n}$, entropy coefficient γ_h
$\mathbf{K} \leftarrow \exp(-\mathbf{C}/\gamma_h)$, $\nu \leftarrow \frac{\mathbf{1}_n}{n}$
while not converged **do**
  $\mu \leftarrow \frac{\mathbf{1}_N}{N} \oslash \mathbf{K}\nu$
  $\nu \leftarrow \frac{\mathbf{1}_n}{n} \oslash \mathbf{K}^T\mu$
end while
$\mathbf{\Pi} \leftarrow \mathrm{diag}(\mu)\mathbf{K}\mathrm{diag}(\nu)$
Output: $\mathbf{\Pi}$

coupling. Specifically, the penalty $\|T\|_{HS}$ (and equivalently, $\|\Lambda\|_F$) is required so as the overall optimization is well-posed (with unique solution for T in the space of Hilbert-Schmidt operators). The entropic regularization term for Π serves to speed up for the computation, while the powered penalty helps to induce a sparser representation for the coupling, in addition to typically reducing the variance in its estimate from empirical data. In the next section, the roles of these regularization terms will be assessed via a simulation study.

- Beyond the augmented Lagrangian method and Sinkhorn algorithm there are a variety of optimization approaches, e.g., the Alternating Direction Method of Multipliers (ADMM) (Ghadimi et al., 2014) and the Hungarian algorithm (Kuhn, 1955), which may be employed for solving the discretized optimal transport problem. Each method exhibits specific advantages from a computational standpoint; for instance, ADMM is adept at handling distributed computing contexts. A thorough investigation of the variety of aforementioned algorithms and their properties within the context of the functional optimal transport problem is of interest and a subject of future research.

Functional PCA: While the basis functions can be specified from a list of orthogonal basis families such as the Hermite polynomials or via Mercer kernels, a more data-driven approach is to estimate the basis from data using the functional PCA approach, as briefly introduced in the previous subsection.

Let $\{y_i(x)\}_{i=1}^N$ denote the function values observed at design points x so that from each function sample a vector y_i of length M is obtained. Then the data are presented by $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^\top$. The covariance operator associated with the function samples is approximated by matrix $\mathbf{V} = \frac{1}{N}\mathbf{Y}^\top\mathbf{Y}$. The integrals can be approximated as $\int f(x)dx \approx \sum_{m=1}^M w_m f(x_m) = \mathbf{w}^\top\mathbf{y}$ where $\mathbf{w}^\top = (w_1, \ldots, w_M)$ is the weight vector that characterizes the numerical quadrature. Using $\mathbf{W} = \mathrm{diag}(w)$ to denote the diagonal matrix of dimensions $M \times M$ whose elements in the diagonal are the elements of w, the eigen-equations for the discrete representation of functional data are

$$\mathbf{VW}\beta_j(x) = \lambda_j\beta_j(x); j = 1, \ldots, M.$$

Let $\nu_j(x) = \mathbf{W}^{1/2}\beta_j(x)$, then the above equations become $\mathbf{W}^{1/2}\left(\frac{1}{N}\mathbf{Y}^\top\mathbf{Y}\right)\mathbf{W}^{1/2}\nu_j(x) = \lambda_j\nu_j(x)$ for $j = 1, \ldots, M$. which can be solved by applying SVD to $\frac{1}{N}\mathbf{YW}^{1/2} = \mathbf{USV}^\top$, where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_N)$ is an orthonormal matrix and $\{\mathbf{u}_n\}_{n=1}^N$ is a basis in $\mathbb{R}^N$. The

corresponding eigenfunctions are represented by eigenvectors $\beta_j = \mathbf{W}^{-1/2}\mathbf{v}_j, j = 1, \ldots, M$. Therefore, we obtain an effective approximation of basis functions when the sample curves are observed at regularly spaced design points.

5. Experiments

In this section we present a thorough simulation study to demonstrate the viability and effectiveness of our method and to validate the theory presented above. We will also compare our functional optimal transport method to other existing domain adaptation techniques in the literature. Finally, we will describe an application of FOT to a real-world task of multivariate robot-arm motion prediction.

5.1 Simulation studies on the synthetic continuous functional dataset

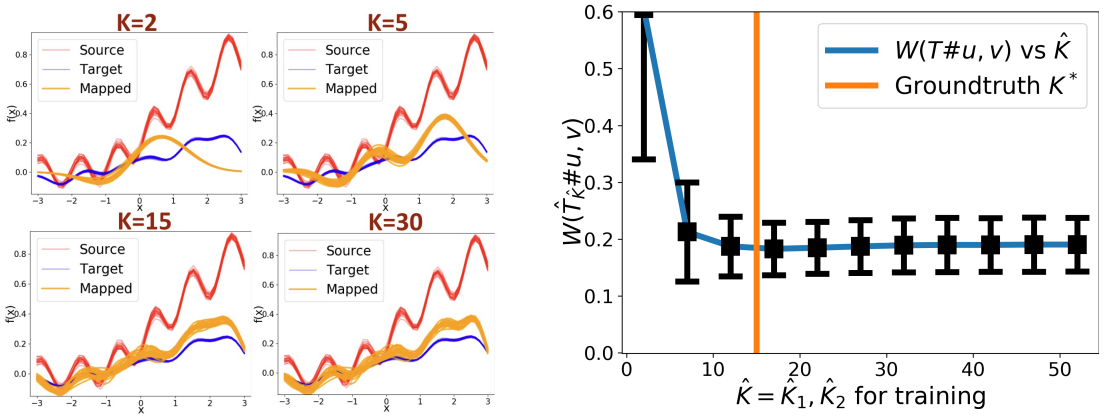
5.1.1 Verifying consistency and interpretation of the transport map

First, we present simulation studies to demonstrate that one can recover the "true" pushforward map via cross-validation. We explicitly constructed a ground-truth map T_0 that has finite intrinsic dimensions $K_1^* = K_2^* = 15$. Then we obtained the target curves by pushing forward source curves via T_0. The FOT algorithm is then applied to the data while $\hat{K}_1$ and $\hat{K}_2$ gradually being increased. The results are illustrated in Fig. 2. They demonstrate the effects of varying the number of basis eigenfunctions $\hat{K} = (\hat{K}_1, \hat{K}_2)$. We observed that the performance of the estimated map steadily improved as $\hat{K}$ increased until it exceeded K^*. As expected, further increasing the number of eigenfunctions did not reduce the learning objective.

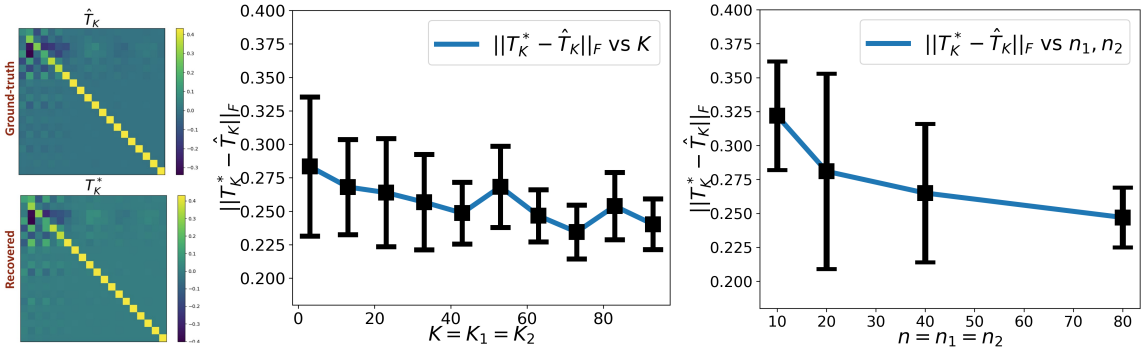
Next, we validate Lemma 5 by evaluating $\hat{T}_{\hat{K}}$ from an infinite dimensional map that transports sinusoidal functions. The Frobenius norm between the optimal T_K^* and estimated $\hat{T}_K$, $\|T_K^* - \hat{T}_K\|_F$, decreases as K increases. In both simulations, we set sample sizes $n_1 = n_2 = 30$. For hyperparameters, set $\gamma_h = 20$, $\eta = 1$. The results were found to be quite robust to other values of these hyperparameters. Finally, we verify Theorem 8, by varying the numbers of sample (n_1, n_2) in estimating the optimal transport (OT) problems between two empirical measures. It is well-known that for any absolutely continuous measure μ on $\mathbb{R}^d$, we have $\mathbb{E}W_1(\hat{\mu}_n, \mu) \lesssim n^{-1/d}$, where $\hat{\mu}_n$ is the empirical measure of μ with n samples (Dudley, 1969). Here, we provide quantitative results to investigate the convergence of estimation with regard to sample size. Similar to our previous setting, we gradually increased the sample size of both source and target ($n_1 = n_2 = n$). We set $K_1 = K_2 = 30$ and used the same hyperparameters as before. We repeated the experiment 10 times for each sample size. As shown in Fig. (2d), the Frobenius norm consistently decreases with increased sample sizes.

5.1.2 Map estimation

Synthetic data simulation: We evaluated our FOT method on a synthetic dataset in which the source and target data samples were generated from a mixture of sinusoidal functions. Each sample $\{y_i(x_i)\}_{i=1}^n$ is a realization evaluated from a (random) function $y_i = A_k \sin(\omega_k x_i + \phi_k) + m_k$ where the amplitude A_k, angular frequency ω_k, phase ϕ_k and translation m_k are random parameters generated from a probability distribution, i.e.,



(a) As $\hat{K}$ increases, $T_{\hat{K}}\#f_1$ moves toward f_2 and $W(T_{\hat{K}}\#\hat{u}, \hat{v})$ decreases until $\hat{K} \geq K^*$.



(b) Estimate T_K^* (c) Number of basis functions (d) Sample size

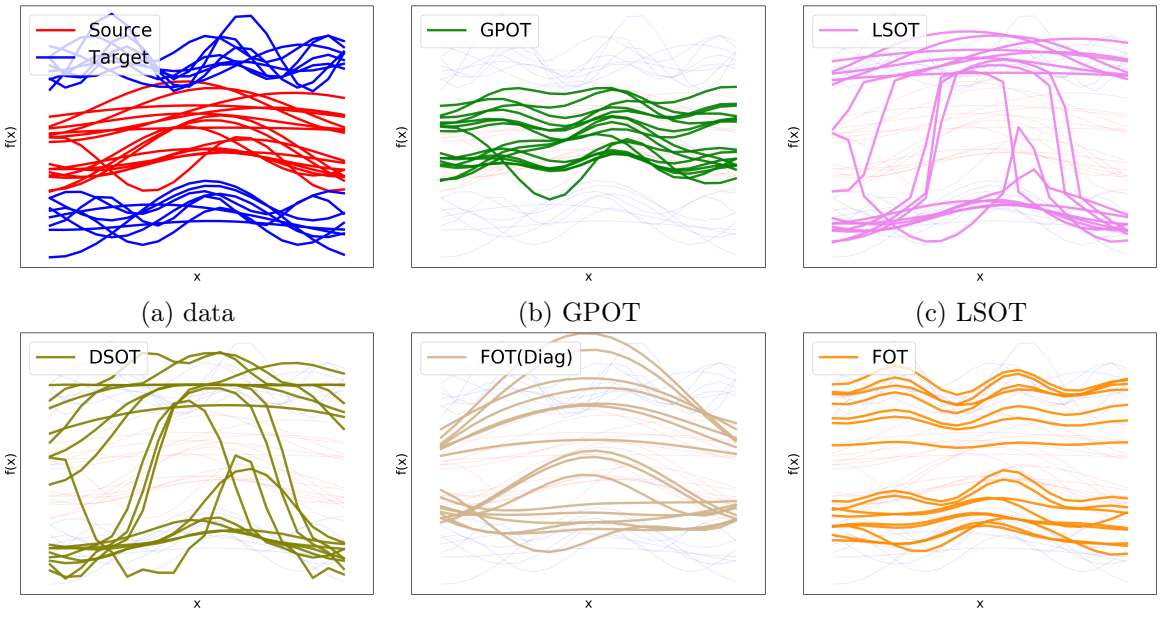
Figure 2: The experimental results that verify the convergence properties. The estimated linear operator $\hat{T}_K$ effectively recovers the groundtruth T_K^* as shown in Fig. (2b). The map estimation error improves (decreases) with increased number of basis functions (Fig. (2c)) and sample curves n_1, n_2 (Fig. (2d)).

$[A_k, \omega_k, \phi_k, m_k] \sim P(\theta_k)$, and θ_k represents the parameter vector associated with a mixture component.

Baseline comparison: We compared FOT method to several existing map estimation methods on the synthetic *mixture of sinusoidal functions* dataset. Sample paths were drawn from sinusoidal functions with random parameters. Then, curves were evaluated on random index sets. In Fig. 3, FOT was compared against the following baselines: (i) Transport map of Gaussian processes (Mallasto and Feragen, 2017; Masarotto et al., 2019), where a closed-form optimal transport map is available, (ii) Large-scale optimal transport (LSOT) (Seguy et al., 2017), and (iii) Mapping estimation for discrete OT (DSOT) (Perrot et al., 2016). For all discrete OT methods, which were not designed for functional data per se, the functional data were treated as point clouds of high dimensional vectors.

We observed that FOT did a remarkably good job at transporting source sample curves to match closely target samples. By contrast, GPOT only altered the oscillation of curves but failed to capture the target distribution's multi-modality. This failure is attributed to the GPOT method's Gaussian process (and thus unimodal distribution) assumption which

clearly did not hold in this example. The poor performance of LSOT and DSOT can be attributed to the fact these methods essentially ignored the smoothness of the sampled curves. In other words, these multivariate adaptations were not suitable for handling functional data.



(a) data (b) GPOT (c) LSOT

(d) DSOT (e) FOT with diagonal Λ (f) FOT

Figure 3: Pushforward measures of functions obtained by various approaches on mixtures of sinusoidal functions data: (a) Sample functions from source and target domain. The resulting pushforward measures obtained by (b) GPOT (Mallasto and Feragen, 2017); (c) LSOT (Ke, 2019); and (d) DSOT (Perrot et al., 2016); and (e)(f) our method FOT. In (e) we parameterized the $\mathbf{\Lambda}$ as only a diagonal function. A full matrix $\mathbf{\Lambda}$ (f) produces a more expressive pushforward.

For a quantitative comparison, we used the Wasserstein distance to measure how well the pushforward measure of (the empirical distribution of) source samples matches the target samples:

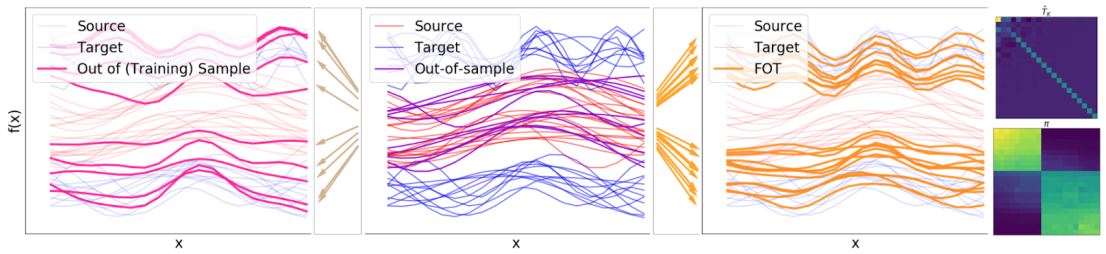
$$L = \min_{\mathbf{\Pi}} \frac{1}{n_L} \sum_{l,k} d(T(\mathbf{f_1}_l), \mathbf{f_2}_k)\Pi_{lk}. \tag{30}$$

Here, $d(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|_2^2$, $\{T(\mathbf{f_1}_i)\}_{i=1}^{n_l}$ and $\{\mathbf{f_2}_i\}_{i=1}^{n_k}$ are mapped samples and target samples, $T(\cdot)$ denotes the map given by different methods, n_L the length of each sample function and Π the probabilistic coupling. The experiments are labeled by $k_{source} \to k_{target}$, where k_{source} and k_{target} indicate the number of mixture components of the source and the target distribution, respectively. More mixture components typically entail more complex data distributions, and thus more complex transportation plan (more on this below). As shown in Table 1, the pushforward map obtained by FOT performed the best in matching target sample functions quantitatively using the Wasserstein distance based objective L.

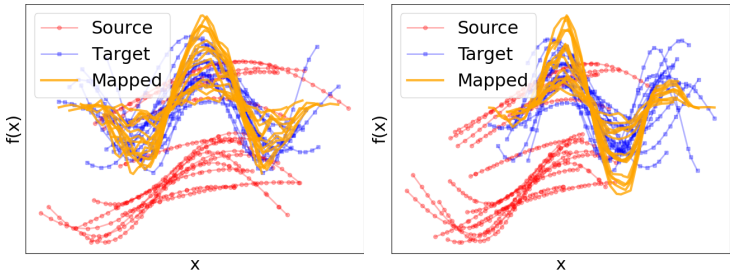
Continuity/ Multimodality preserving properties: As shown in Fig. (4a), the map learned by FOT does a good job at pushing forward out-of-sample curves that were not

Method	$1 \to 1$	$1 \to 2$	$2 \to 1$	$2 \to 2$	$2 \to 3$
GPOT	17.560	12.895	15.263	61.561	39.159
LSOT	133.434	94.229	117.832	929.108	663.461
DSOT	6.871	13.226	9.679	46.521	41.009
FOT	**2.873**	**11.982**	**3.316**	**44.071**	**32.547**

Table 1: Quantitative comparison on the mixture of sinusoidal functions data. The maps obtained by FOT method achieved the best performance under the Wasserstein distance objective.



(a) Out-of-sample curves. The rightmost and lower heatmap represents the coupling π and reveals the multimodality.

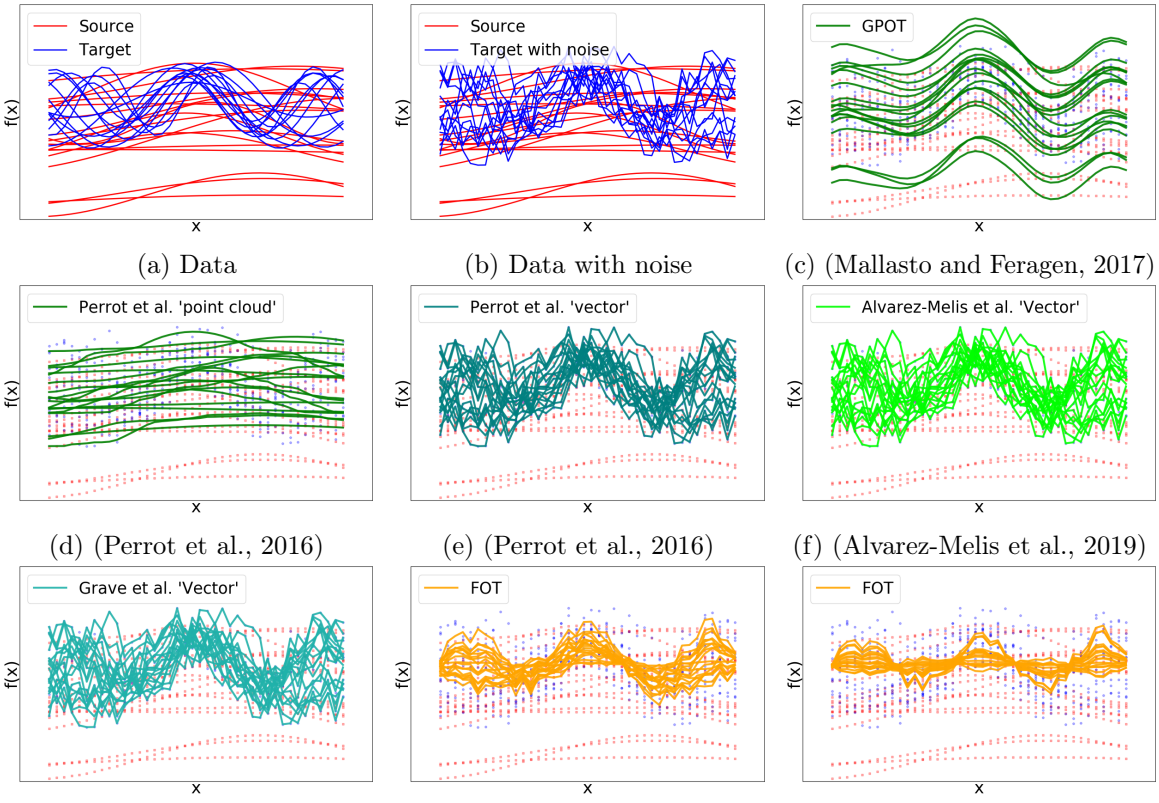


(b) Varying design points (c) Distinct design point sets

Figure 4: FOT performance on out-of-sample curves.

observed during training. Moreover, the coupling π reveals the multi-modality in the data: as the source distribution is unimodal but the target distribution is bimodal, there is a "splitting" behavior in how the sampled curves from the source are distributed and transported to the target. Finally, Fig. (4b) shows that FOT is very effective for functional data evaluated at different design points. On the upper right panel of Fig. (4a), the estimated integral operator $\hat{\mathbf{T}}_K$ is shown; note that it is close to an identity matrix while having some permutations around the first elements. We show the estimated coupling π on the lower right panel. The coupling clearly reveals the underlying cluster structure in the target function data.

Comparison to OT methods for finite-dimensional vectors: Although one can always apply existing OT map estimation methods such as that of Alvarez-Melis et al. (2019); Perrot et al. (2016); Grave et al. (2019) to functional data by simply discretizing continuous functions into fixed-dimension vector measurements, we shall demonstrate the ineffectiveness of such an approach due to its failure to properly account for the functional nature (e.g. smoothness) of the data in the source and/or target domain (see Fig. 5). In particular, we present



(a) Data (b) Data with noise (c) (Mallasto and Feragen, 2017)

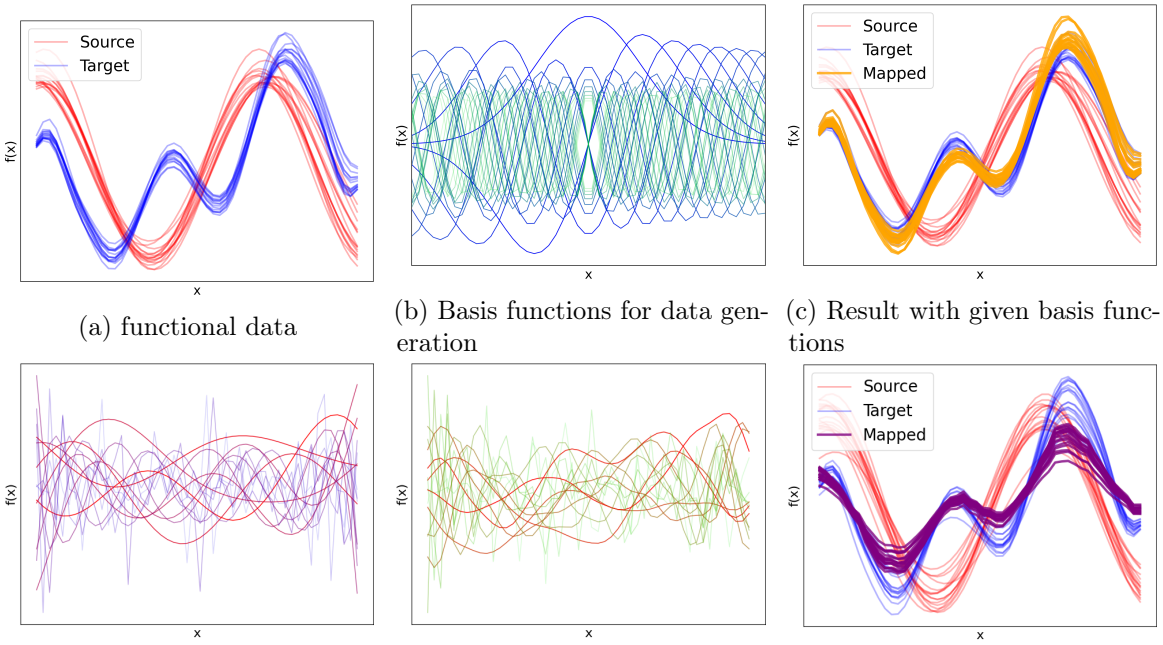
(d) (Perrot et al., 2016) (e) (Perrot et al., 2016) (f) (Alvarez-Melis et al., 2019)

(g) (Grave et al., 2019) (h) FOT with $\eta = 1$ (i) FOT with $\eta = 40$

Figure 5: Panel (b) depicts functional data generated by adding non-continuous noise to the smooth curves shown in panel (a). (c) depicts results obtained by applying the method of (Mallasto and Feragen, 2017). (d) and (e) depict results obtained by applying the method of (Perrot et al., 2016). (f) depicts results obtained by applying the method of (Alvarez-Melis et al., 2019). (g) depicts results obtained by the method of (Grave et al., 2019). Panels (h) and (i) depict results obtained by FOT using different regularization parameters.

experiments and comparisons with more baseline methods under the same settings considered in Section 5.1. In these experiments we assume all the functional data are evaluated on a set of fixed-size design points to apply conventional OT map estimation methods for fixed-dimensional vectors directly. In addition, the observed continuous function data is perturbed by non-continuous noise. Under this setting, all baseline OT formulations neglect the smooth nature of functional data and overfit the signals contaminated with noises. Only the pushforward of maps estimated with GPOT (Mallasto and Feragen, 2017) and our methods successfully recover the smoothness of the target curves. This suggests the necessity of treating data as sampled functions (rather than sampled vectors). Plot (h) and plot (i) of Fig. 5 show the role played by parameter η in controlling the smoothness of the map.

Selection of basis functions: We investigate the selection of basis functions using the FPCA approach introduced in Section 4.3. We generate the original data using Hermite polynomials as basis functions. As illustrated in Fig. 6, knowing the ground-truth basis



(a) functional data (b) Basis functions for data generation (c) Result with given basis functions

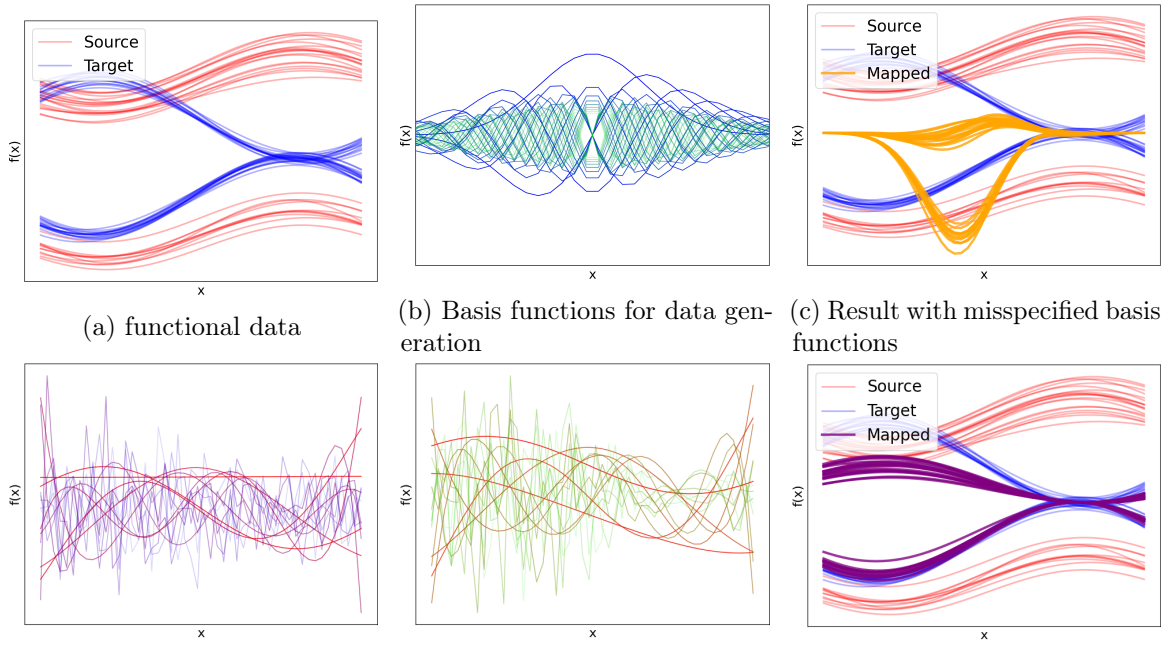
(d) Learned basis functions $\hat{V}$ (e) Learned basis functions $\hat{U}$ (f) Result with estimated basis

Figure 6: Estimate basis functions from data using FPCA. When the ground-truth basis functions are provided, the estimated map can achieve perfect results (Fig. (6c)) with given basis (Fig. (6b)). Using the estimated basis functions (Fig. (6d) and Fig. (6e)), the map can still have satisfactory results (Fig. (6f)).

function when estimating the transport map leads to a consistent pushforward map. We also display the estimated basis functions for both the source (Fig. 6d) and target (Fig. (6e)) sample curves. Following the FPCA procedure described in Sec. 4.3, we assembled the function values $y_i(x)$ into matrix $\mathbf{Y}$, with each row representing a function sample as a vector. Leveraging the shared index set, we applied Singular Value Decomposition (SVD) to $\frac{1}{N}\mathbf{Y}$, yielding $\mathbf{USV}^\top$ where the columns of $\mathbf{U}$ are the estimation of basis functions. The estimated basis functions are ordered by the degree of oscillation. As we can see, the estimated pushforward map does not match the pushforward curves perfectly but is still satisfactory, as illustrated in Fig. (6f).

In numerous situations, the basis functions for complex multimodal distributions may not be readily available; estimating a highly complex parametrized family of basis functions effectively can be challenging. Under these conditions, employing a data-driven approach such as FPCA for obtaining the most relevant basis functions can yield improved performance. The flexibility and usefulness of the FPCA approach are illustrated in Fig. 7 which demonstrates that our algorithm augmented with the FPCA-based basis functions can effectively adapt to previously unobserved data distributions.

Effects of regularization: The final set of simulations is designed to evaluate the effects of regularization terms in the FOT formulation. The results of this study are depicted in Fig. 8. We observe that the power regularizer finds sparser coupling distributions than entropic regularization. This phenomenon is expected as the entropic penalty keeps the coupling



(a) functional data

(b) Basis functions for data generation

(c) Result with misspecified basis functions

(d) Learned basis $\hat{V}$

(e) Learned basis $\hat{U}$

(f) Result with learned basis

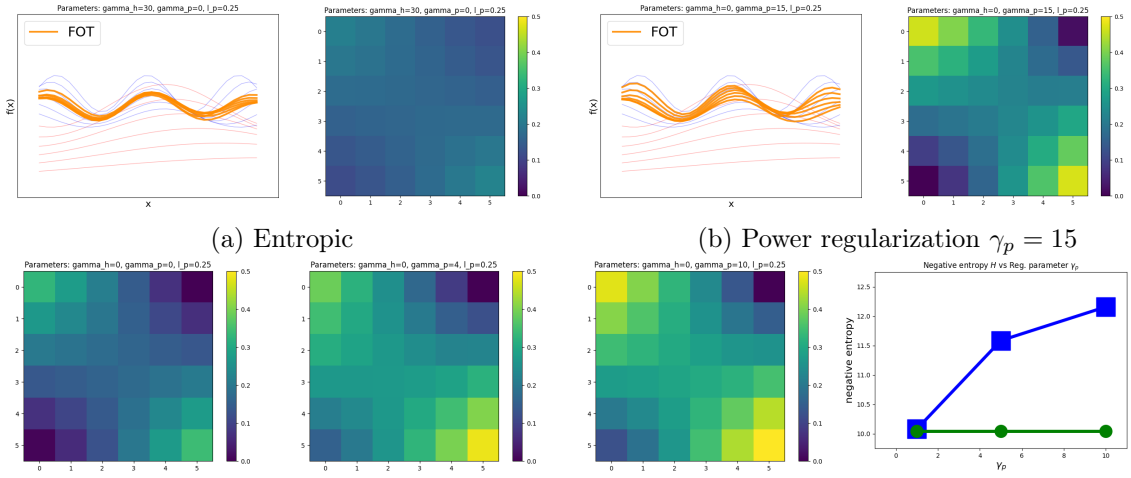
Figure 7: Defining appropriate basis functions (Fig. (7b)) for optimal map estimations itself is a challenging task. Employing FPCA (Fig. (7d) and Fig. (7e)) yields superior map estimation outcomes (Fig. (7f)). In contrast, misspecified basis functions from randomly selected basis function coefficients result in sub-optimal map pushforward samples (Fig. (7c)).

strictly positive in its support. The lack of sparsity can be problematic, especially in the case of iterative map estimation, where a sparse coupling distribution is crucial for learning a meaningful and expressive transport map between domains of functions. It is worth noting that, the estimated coupling represents the joint probability density matrix and it reveals the clustering structure of data (cf. Fig. (4a)).

5.2 Optimal transport domain adaptation for robot arm's multivariate sequences of motion

Recent advances in robotics include many novel data-driven approaches such as motion prediction (Jetchev and Toussaint, 2009), human-robot interaction (Liu et al., 2018), and others (Tompkins et al., 2020; Xu et al., 2020). However, generalizing knowledge across different automated tasks for a robot, and generalizing across robots, are considered challenging since data collection in the real world is expensive and time-consuming. A variety of approaches have been developed to tackle these problems, such as domain adaptation (Bousmalis et al., 2018), transfer learning (Weiss et al., 2016), and so on (Tobin et al., 2017; Finn et al., 2017).

Optimal transport based domain adaptation: We propose to eliminate the heterogeneity in robot learning datasets by following the formulation of optimal transport based domain adaptation (OTDA) (Courty et al., 2016). Specifically, the pipeline consists of the



(a) Entropic (b) Power regularization $\gamma_p = 15$

(c) $\gamma_p = 0$ (d) $\gamma_p = 4$ (e) $\gamma_p = 10$ (f) Entropy

Figure 8: Panel (a) and (b) depict the pushforward samples and coupling matrix obtained by entropic and power regularization, respectively. The entropic regularization leads to a smooth coupling while the corresponding pushforward samples concentrate near the average of the target samples, suggesting a rather poor transport map. Panels (b)–(f) show that the coupling becomes sparser as we increase the power regularization coefficient γ_p. Panel (f) confirms that the negative entropy of the coupling matrix increases with coefficient γ_p.

following three steps: 1) learn an optimal transport map, 2) apply the pushforward map on the observed source samples towards the target domain, and 3) train a motion predictor on the pushforwarded samples that lie in the target domain.

Although it might be possible to discretize and interpolate data to fixed-size vectors of observed measurements, trajectories of robot motion are intrinsically *continuous functions of time of various lengths*. Functional optimal transport provides a natural solution for this challenging task over existing OT map estimation methods for discrete samples.

Typically, a robot motion dataset $\{f_i\}_{i=0}^N$ contains multiple trails f_i for a specific task. Although these trials are somewhat similar, they differ slightly due to various real-world factors due to sensor and actuator noises, and human intervention. Each trail can be viewed as a multidimensional n_t length timeseries representing the joint or end effector location over time $f_i := \{(f_{i,1}, t_{i,1}), ..., (f_{i,n_t}, t_{i,n_t})\}$. Each robot motion dataset is viewed as samples for a distribution μ, via the empirical distribution $\hat{\mu} = \sum_{i=1}^N p_i \delta_{\hat{f}_i}$, where $\delta_{\hat{f}_i}$ is the Dirac measure at $\hat{f}_i$, which is the embedded function of f_i in the Hilbert space using basis functions, and p_i are probability masses associated to the i-th sample ($\sum_i p_i = 1$). Given a source and target robot-motion dataset D_s and D_t (with empirical measures $\hat{\mu}_s$ and $\hat{\mu}_t$, respectively), we assume that there are sufficient samples in the source $\{f_{s,i}\}_{i=1}^{N_s}$ but *only limited samples* in the target $\{f_{t,j}\}_{j=1}^{N_t}$, where $N_s \gg N_t$. To obtain an ML model in the target domain, we want to leverage the knowledge in the source by using a transport map that pushes forward all samples $\{f_i\}$ to match the distributions. At this point, the transport map can be estimated using the functional optimal transport formulation, by solving $T^* = \arg\min_T W_2(T_{\#}\hat{\mu}_s, \hat{\mu}_t) + \eta \|T\|_F^2$.

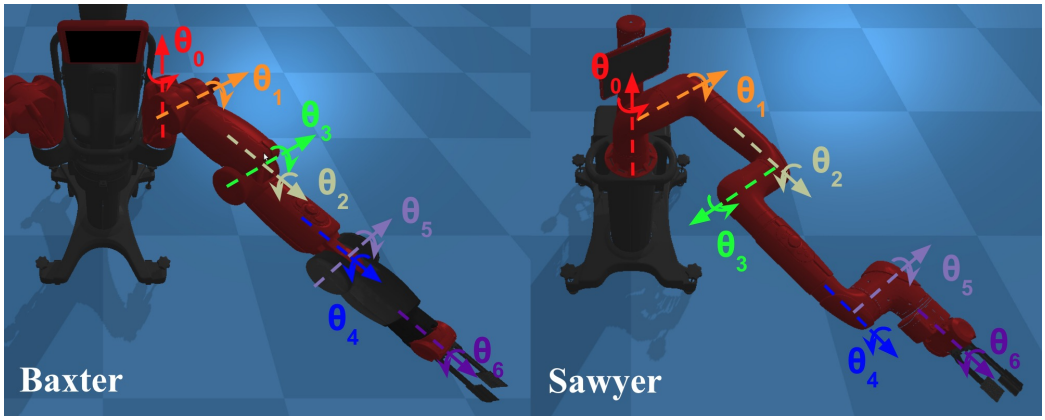
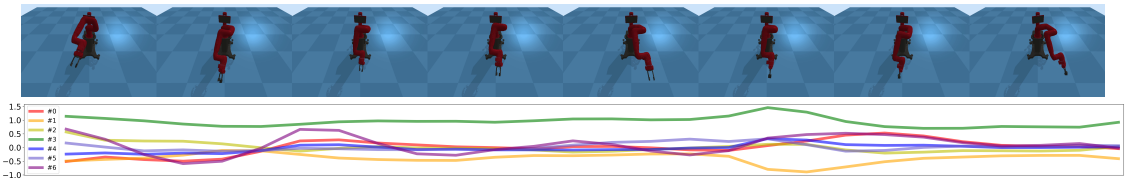
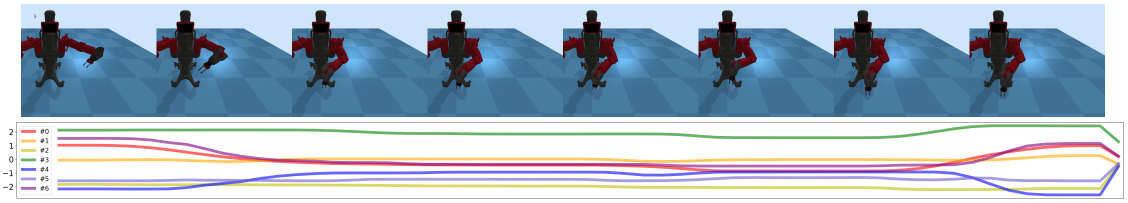


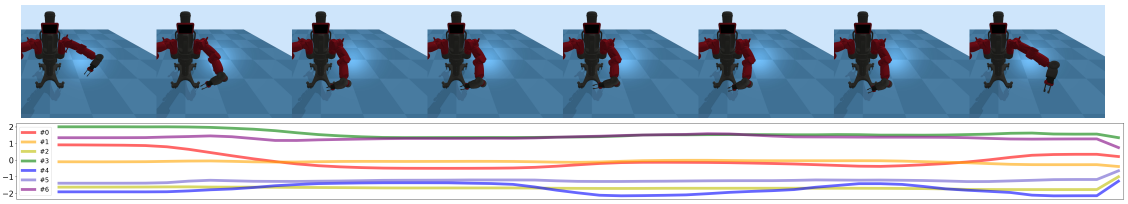
Figure 9: The structure of the Baxter robot and the Sawyer robot used in MIME dataset and Roboturk dataset. Their arms share a similar structure as they both have 7 joints and one end effector. This allows us to perform domain adaptation between these two datasets.



(a) Source motion: "Roboturk-bins-Bread" by Sawyer robot.



(b) Target motion: "MIME Picking (left-hand)" by Baxter robot.



(c) The pushforward motion of the transport map and the target motion look similar to each other but differ slightly.

Figure 10: Pushforward of robot arm motions: In each sub-figure, a robot-arm motion is visualized as image clips, while the robot-arm joint angles are plotted as a multivariate time series, the x-axis is time and are omitted for cleanliness.

Datasets: The **MIME** Dataset (Sharma et al., 2018) contains 8000+ motions across 20 tasks collected on a two-armed Baxter robot. The **Roboturk** Dataset (Mandlekar et al., 2018) is collected by a Sawyer robot over 111 hours. As shown in Fig. (9), both robot

Method	LSTM	ANP	RANP	MAML*	TL*	FOT_{LSTM}	FOT_{ANP}	FOT_{RANP}	FOT_{MAML}	FOT_{TL}
R1→M1	2.0217	1.3261	1.9874	0.0307	0.5743	0.0271	0.0963	0.0687	0.0165	0.0277
R1→M2	1.6821	1.0951	1.5681	0.0374	0.7083	0.0414	0.1642	0.1331	0.0191	0.0446
R2→M1	1.3963	0.6642	1.7256	0.0327	0.2491	0.0277	0.0951	0.0696	0.0202	0.0906
R2→M2	1.1952	0.6307	1.3659	0.0477	0.4020	0.0331	0.1620	0.1554	0.0167	0.0406

Table 2: MSE error results of different predictive models. R1: Roboturk-bins-bread, R2: Roboturk-pegs-RoundNut, M1:MIME1-Pour-left, M2: MIME12-Picking-left.

arms have 7 joints with similar but slightly different configurations, which enable us to learn domain adaptation between the two. We picked two tasks, *Pouring (left arm) and Picking (left arm)*, from MIME dataset and two tasks, *(bins-Bread, pegs-RoundNut)*, from Roboturk dataset. We considered each task as an individual domain.

Pushforward of robot motions: Our method successfully learns the transport map that pushes forward samples from one task domain to another. The source dataset contains motion records from task *bins-full* in the Roboturk dataset while the target includes motion records from task *Pour (left-arm)* in the MIME dataset. We visualize the motion by displaying the robot joint angles sequences in a physics-based robot simulation gym (Erickson et al., 2020). Animated motions can be found here[1]. In Fig. 10, we show image clips of each move along with a plot of time series of joint angles. We can see from the robot simulation that the pushforward sequence in Fig. 10c matches with the target motion in Fig. 10b while simultaneously preserving certain features of the source motion in Fig. 10a.

Motion prediction: For the *Robot Arm Motion Prediction* task, a motion trajectory $f_i = (f_{i,1}, t_{i,1}), ..., (f_{i,l}, t_{i,l})$ of length l consists of a set of vectors $f_{i,j} \in \mathbb{R}^d$ with associated timestamps $t_{i,j}$, where the time series trajectories are governed by continuous functions of time $f_S(t) : t \in \mathbb{R} \mapsto S \in \mathbb{R}^d$. Since the task is to predict the future l_f points based on the past l_p points, ignoring the index of individual trajectories, we arrange the data to have the format $X_t = \{(f_{t+1}, t+1), ..., (f_{t+l_p}, t+l_p)\}$, $Y_t = \{(f_{t+l_p+1}, t+l_p+1), ..., (f_{t+l_p+l_f}, t+l_p+l_f)\}$. Our task is learning a predictive model that minimizes the squared prediction error in the target domain $\arg\min_\theta \sum_{i=1}^{M} (F_\theta(X_i^t) - Y_i^t)^2$ where Y_i^t is the true label from target domain and $\hat{Y}_i^t = F_\theta(X_i^t)$ is the predictive label estimated by a model trained on source domain (X^s, Y^s) and a subset of target domain (X^{tm}, Y^{tm}). It is worth noting that if the distribution of the testing set differs from that of the target set, a predictive model trained solely on the training set will experience a significant performance drop on the target set. To address this issue, several paradigms have been proposed, including transfer learning, meta-learning, and few-shot learning.

Methods: We considered 5 baselines for this task, including

(1) a conventional baseline approach we employ is a vanilla LSTM that is trained exclusively on the source data samples. Despite the fact that LSTM models are recognized for their ability to capture temporal dependencies, their performance may still be restricted due to the distributional discrepancy;

(2) the Attentive Neural Process (ANP) (Kim et al., 2019), which is a deep Bayesian model that learns a distribution of functions. ANP can be seen as models that do few-shot learning since it looks for predictive distribution conditioning on context data;

1. More examples can be found here: https://sites.google.com/view/functional-optimal-transport.

(3) the RANP model (Qin et al., 2019), which is an extension of the ANP models, which incorporates the temporal dependency of data using an LSTM model. Similar to the ANP, the RANP is also adept at few-shot learning tasks;

(4) the MAML model (Model-Agnostic Meta-Learning), which is a popular meta-learning algorithm (Finn et al., 2017); It is designed to enable fast adaptation of a model's parameters to new tasks with limited data. To enable a fair comparison, we learn the meta parameters on a small set of various tasks and perform limited $t = 1$ adaptation step on the target domain;

(5) and finally, the conventional transfer learning (TL) (Weiss et al., 2016) method, where we first pre-trained the model on the source domain and then fine-tuned it on the target domain. To leverage our proposed FOT, we use the estimated map to pushforward all source samples to match the target distribution, and then use these samples for the training, adaptation, or the fine-tuning process for the above baseline methods.

Results: The results are given in Table 2. It is noted that the vanilla and few-shot training approaches are having large errors, as expected, while MAML and transfer learning have better generalization ability as they have the access to some target samples. However, we have also observed that, somewhat surprisingly, utilizing pushforward samples from the FOT transport map enhances the performance of LSTM, NP, and RANP to surpass the baseline performance of both meta-learning and transfer learning approaches. Additionally, even MAML and TL approaches can benefit from utilizing the mapped samples from the FOT. This is because the pushforward data offer additional samples that adhere to the target distribution, which helps mitigate the distributional gap due to the model misspecification.

All experiments were implemented with Numpy and PyTorch (matrix computation scaling) using one GTX2080TI GPU and a Linux desktop with 32GB memory. For all simulations, we set the optimization coefficients as $\rho_k = 800 \times \mathbf{1} \in \mathbb{R}^{N\times 1}$, $\rho_l = 800 \times \mathbf{1} \in \mathbb{R}^{n\times 1}$, $\eta = 0.001$, $\gamma_h = 40$, $\gamma_p = -10$, power $p = 3$. The learning rate for updating $\mathbf{\Lambda}$ is $lr_\Lambda = 4e-4$, the learning rate for updating π_{lk} is $lr_\pi = 1e-5$. The maximum iteration step is set as $T_{max} = 1000$. In the experimental results on the robot-arm datasets, the hyperparameters are set by $\gamma_h = 30$, $\gamma_p = -30$, and power $p = 3$. The same hyperparameters as in the simulation experiments were employed. We found that our algorithm's performance was not sensitive to varying hyperparameters. Specifically, when hyperparameters are perturbed around these values, the performance of the downstream domain adaptation experiments remains stable.

6. Proofs

In this section, we provide proofs for theoretical results in Section 3 and Section 4. We first define some notations regarding the proofs.

Notations. Fix Borel probability measures μ on H_1 and ν on H_2. We define the cost function (without regularization term) to be $\Phi(T) := W_2(T\#\mu, \nu)$ for $T \in \mathscr{B}_{HS}(H_1, H_2)$. For the ease of notation, as in the main text we write n for (n_1, n_2), K for (K_1, K_2), $\mathscr{B}_{HS}$ for $\mathscr{B}_{HS}(H_1, H_2)$ and B_K for its restriction on the space spanned by the first $K_1 \times K_2$ basis

operators. $\|\cdot\|_{HS}$ and $\|\cdot\|_{op}$ are used to denote the Hilbert-Schmidt norm and operator norm on operators, respectively. It is known that $\|T\|_{op} \leq \|T\|_{HS}$ for all operator T.

In this section we often deal with convergence of a sequence with multiple indices. Specifically, we say a tuple $m = (m_1, \ldots, m_p) \to \infty$ when $m_1 \to \infty, \ldots, m_p \to \infty$. By saying that a sequence $A(m_1, m_2, \ldots, m_p)$ of index $m = (m_1, \ldots, m_p)$ converge to a number a as $m \to \infty$, it is meant that for all $\epsilon > 0$, there exists $M_1, \ldots, M_p$ such that for all $m_1 > M_1, \ldots, m_p > M_p$, we have

$$|A(m_1, \ldots, m_p) - a| < \epsilon. \tag{31}$$

We write $(m_1, m_2, \ldots, m_p) > (m'_1, m'_2, \ldots, m'_p)$ if $m_1 > m'_1, \ldots, m_p > m'_p$.

We say a function $f : \mathscr{B}_{HS}(H_1, H_2) \to \mathbb{R}$ is *coercive* if

$$\lim_{\|T\|_{HS} \to \infty} f(T) = \infty, \tag{32}$$

and it is *(weakly) lower semi-continuous* if

$$f(T_0) \leq \liminf_{k \to \infty} f(T_k), \tag{33}$$

for all sequences T_k (weakly) converging to T_0. Further details on convergence in a strong and weak sense in Hilbert spaces can be found in standard texts on functional analysis, e.g., (Yosida, 1995).

Now we are going to prove the results presented in Section 3 of the main text. For ease of the readers, we recall all statements before proving them.

Existence and uniqueness First, we verify some properties of the objective function J.

Lemma 3 *The following statements hold.*

(i) *$W_2(T\#\mu, \nu)$ is a Lipschitz continuous function of $T \in \mathscr{B}_{HS}(H_1, H_2)$, which implies that $J : \mathscr{B}_{HS} \to \mathbb{R}_+$ is also continuous.*

(ii) *J is a strictly convex function.*

(iii) *There are constants $C_1, C_2 > 0$ such that $J(T) \leq C_1\|T\|^2_{HS} + C_2 \ \ \forall T \in \mathscr{B}_{HS}$.*

(iv) $\lim_{\|T\|_{HS} \to \infty} J(T) = \infty$.

Proof [Proof of Lemma 3]

(i) We first show that the cost function (without regularization term) $\Phi(T) = W_2(T\#\mu, \nu)$ for $T \in \mathscr{B}_{HS}(H_1, H_2)$ is Lipschitz continuous. Indeed, consider any $T_1, T_2 \in \mathscr{B}_{HS}$, by

the triangle inequality applied to Wasserstein metric,

$$\begin{aligned}
W_2(T_1\#\mu,\nu) - W_2(T_2\#\mu,\nu) &\le W_2(T_1\#\mu, T_2\#\mu) \\
&= \left(\inf_{\pi\in\Pi(\mu,\mu)} \int_{H_1\times H_1} \|T_1 f_1 - T_2 f_2\|_{H_2}^2 d\pi(f_1,f_2)\right)^{1/2} \\
&\le \left(\int_{H_1\times H_1} \|T_1 f_1 - T_2 f_2\|_{H_2}^2 d\pi'(f_1,f_2)\right)^{1/2} \\
&= \left(\int_{H_1} \|T_1 f_1 - T_2 f_1\|_{H_2}^2 d\mu(f_1)\right)^{1/2} \\
&\le \left(\int_{H_1} \|T_1 - T_2\|_{op}^2 \|f_1\|_{H_1}^2 d\mu(f_1)\right)^{1/2} \\
&\le \|T_1 - T_2\|_{HS} \left(\int_{H_1} \|f_1\|_{H_1}^2 d\mu(f_1)\right)^{1/2} \\
&= \|T_1 - T_2\|_{HS} (E_{f\sim\mu}\|f\|_{H_1}^2)^{1/2},
\end{aligned}$$

where π' is the identity coupling. Hence, both $\Phi^2(T)$ and $\eta\|T\|_{HS}^2$ are continuous, which entails continuity of J as well.

(ii) If we can prove that $\Phi^2(T)$ is convex with respect to T, then the conclusion is immediate from the strict convexity of $\eta\|T\|_{HS}^2$. We first observe that $W_2^2(\cdot,\nu)$ is convex, as for any measure ν_1,ν_2 on H_2 and $\lambda\in[0,1]$, if γ_1 is the optimal coupling of (ν_1,ν) and γ_2 is the optimal coupling of (ν_2,ν), then $\lambda\gamma_1+(1-\lambda)\gamma_2$ is a valid coupling of $(\lambda\nu_1+(1-\lambda)\nu_2,\nu)$, which yields

$$\begin{aligned}
W_2^2(\lambda\nu_1 + (1-\lambda)\nu_2,\nu) &\le \int_{H_1\times H_2} \|f-g\|_{H_2}^2 d(\lambda\gamma_1 + (1-\lambda)\gamma_2)(f,g) \\
&= \lambda W_2^2(\nu_1,\nu) + (1-\lambda) W_2^2(\nu_2,\nu).
\end{aligned}$$

Now the convexity of $\Phi^2(T)$ follows as for any $T_1,T_2\in\mathscr{B}_{HS}, \lambda\in[0,1]$,

$$\begin{aligned}
W_2^2(((1-\lambda)T_1 + \lambda T_2)\#\mu,\nu) &= W_2^2((1-\lambda)(T_1\#\mu) + \lambda(T_2\#\mu),\nu) \\
&\le (1-\lambda) W_2^2(T_1\#\mu,\nu) + \lambda W_2^2(T_2\#\mu,\nu).
\end{aligned}$$

(iii) This can be proved by an application of Cauchy-Schwarz inequality and the fact that the operator norm is bounded above by the Hilbert-Schmidt norm. Let π be any coupling of μ and ν,

$$\begin{aligned}
J(T) &= W_2^2(T\#\mu,\nu) + \eta\|T\|_{HS}^2 \\
&\le \int_{H_1\times H_2} \|Tf_1 - f_2\|_{H_2}^2 d\pi(f_1,f_2) + \eta\|T\|_{HS}^2 \\
&\le 2\int_{H_1\times H_2} (\|Tf_1\|_{H_2}^2 + \|f_2\|_{H_2}^2) d\pi(f_1,f_2) + \eta\|T\|_{HS}^2 \\
&\le 2\left(\|T\|_{HS}^2 \int_{H_1} \|f_1\|_{H_1}^2 d\mu(f_1) + \int_{H_2} \|f_2\|_{H_2}^2 d\mu(f_2)\right) + \eta\|T\|_{HS}^2 \\
&= C_1\|T\|_{HS}^2 + C_2,
\end{aligned}$$

for all $T \in B$, where $C_1 = 2E_{f_1\sim\mu}\|f_1\|_{H_1}^2 d\mu(f) + \eta, C_2 = 2E_{f_2\sim\nu}\|f_2\|_{H_2}^2 d\nu(f)$.

(iv) This follows from the fact that $\Phi^2(T) \geq 0$ for all T and $\eta\|T\|^2$ is coercive as in equation (32).

■

We are ready to establish existence and uniqueness of the minimizer of J. The technique being used is well-known in the theory of calculus of variations (e.g., cf. Theorem 5.25. in (Demengel and Demengel, 2012)).

Theorem 4 *There exists a unique minimizer T_0 for the problem* (6).

Proof [Proof of Theorem 4] As $J(T) \geq 0$ and is finite for all T, there exist $L_0 = \inf_{T\in\mathscr{B}_{HS}} J(T) \in [0,\infty)$. Consider any sequence $(T_k)_{k=1}^{\infty}$ such that $J(T_k) \to L_0$. We see that this sequence is bounded, as otherwise, there exists a subsequence $(T_{k_h})_{h=1}^{\infty}$ such that $\|T_{k_h}\|_{HS} \to \infty$. But this means $L_0 = \lim J(T_{k_h}) = \infty$ (due to the coercivity), which is a contradiction. Now, because (T_k) is bounded, by Banach-Alaoglu theorem, there exists a subsequence $(T_{k_p})_{p=1}^{\infty}$ converges weakly to some T_0.

Next, we will prove that J is weakly lower semi-continuous. Indeed, we can readily verify that J is (weakly) lower semi-continuous if and only if the epigraph $\{(T,y) : y \geq J(T)\}$ is (weakly) closed. Because J is convex and continuous, we have the epigraph is convex and closed. Recall a theorem of Mazur (page 292 of Royden and Fitzpatrick (1988)), which states that a convex, closed subset of a Banach space is weakly closed. This result implies that the epigraph $\{(T,y) : y \geq J(T)\}$ is also weakly closed. Hence, J is weakly lower semi-continuous. Thus,

$$J(T_0) \leq \liminf_{p\to\infty} J(T_{k_p}) = L_0. \tag{34}$$

Therefore the infimum of J is attained at some T_0. The uniqueness of T_0 follows from the strict convexity of J. ■

Approximation analysis Next, we proceed to analyze the convergence of the minimizers of finite dimensional approximations to the original problem (6). The proof is valid thanks to the presence of the regularization term $\eta\|T\|_{HS}^2$.

Lemma 5 *For each $K = (K_1, K_2)$, there exists a unique minimizer T_K of J over B_K. Moreover, $T_K \to T_0$ in $\|\cdot\|_{HS}$ as $K_1, K_2 \to \infty$.*

Proof [Proof of Lemma 5] Similar to the proof above, for every $K = (K_1, K_2)$ there exists uniquely a minimizer T_K for J on B_K as B_K is closed and convex. Denote $T_{0,K}$ the projection of T_0 to B_K. As $K \to \infty$, we have $T_{0,K} \to T_0$, which yields $J(T_{0,K}) \to J(T_0)$. From the definition of minimizers, we have

$$J(T_{0,K}) \geq J(T_K) \geq J(T_0), \quad \forall\, K. \tag{35}$$

Now let $K \to \infty$, we have $\lim_{K\to\infty} J(T_K) = J(T_0)$ thanks to the Sandwich rule. Since J is convex,

$$J(T_0) + J(T_K) \geq 2J\left(\frac{1}{2}(T_0 + T_K)\right), \tag{36}$$

passing this through the limit, we also have

$$\lim_{K\to\infty} J\left(\frac{1}{2}(T_0+T_K)\right) = J(T_0). \tag{37}$$

Now using the parallelogram rule,

$$\begin{aligned}
\frac{\eta}{2}\|T_K - T_0\|_{HS}^2 &= \eta\left(\|T_K\|_{HS}^2 + \|T_0\|_{HS}^2 - 2\left\|\frac{1}{2}(T_0+T_K)\right\|_{HS}^2\right) \\
&= \left(J(T_K) + J(T_0) - 2J\left(\frac{1}{2}(T_0+T_K)\right)\right) \\
&\quad - \left(\Phi^2(T_K) + \Phi^2(T_0) - 2\Phi^2\left(\frac{1}{2}(T_0+T_K)\right)\right) \\
&\leq \left(J(T_K) + J(T_0) - 2J\left(\frac{1}{2}(T_0+T_K)\right)\right),
\end{aligned}$$

as Φ^2 is convex. Let $K\to\infty$, we have the last expression goes to 0. Hence, $\|T_K - T_0\|_{HS} \to 0$. ■

What is remarkable in the proof above is that it works for any sequence $(T_m)_{m=1}^{\infty}$: whenever we have $J(T_m)\to J(T_0)$ then we must have $T_m \to T_0$.

Uniform convergence and consistency analysis Now we turn our discussion to the convergence of empirical minimizers. Using the technique above, there exists uniquely minimizer $\hat{T}_{K,n}$ for $\hat{J}_n$ over B_K. We want to prove that $\hat{T}_{K,n} \xrightarrow{P} T_K$ uniformly in K in a suitable sense and then combine with the result above to have the convergence of $\hat{T}_{K,n}$ to T_0. A standard technique in the analysis of M-estimator is to establish uniform convergence of $\hat{J}_n$ to J in the space of T (Keener, 2010). Note that the spaces $\mathscr{B}_{HS}$ and all B_K's are not bounded, so care must be taken to show that $(\hat{T}_{K,n})_{K,n}$ will eventually reside in a bounded subset and then uniform convergence is attained in that subset. The following auxiliary result presents that idea.

Lemma 6

(i) For any fixed $C_0 > 0$,

$$\sup_{\|T\|_{HS}\leq C_0} |\hat{J}_n(T) - J(T)| \xrightarrow{P} 0 \quad (n\to\infty). \tag{38}$$

(ii) Let $\hat{T}_{K,n}$ be the unique minimizer of $\hat{J}_n$ over B_K. There exists a constant D such that $P(\sup_K \|\hat{T}_{K,n}\|_{HS} < D) \to 1$ as $n\to\infty$.

Proof

(i) The proof proceeds in a few small steps.

Step 1. By triangle inequality of Wasserstein distances,

$$\begin{aligned}|W_2(T\#\mu,\nu) - W_2(T\#\hat{\mu}_{n_1},\hat{\nu}_{n_2})| &\le W_2(T\#\hat{\mu}_{n_1},T\#\mu) + W_2(\hat{\nu}_{n_2},\nu)\\ &\le \|T\|_{op} W_2(\hat{\mu}_{n_1},\mu) + W_2(\hat{\nu}_{n_2},\nu)\\ &\le \|T\|_{HS} W_2(\hat{\mu}_{n_1},\mu) + W_2(\hat{\nu}_{n_2},\nu). \end{aligned} \tag{39}$$

Therefore,

$$\sup_{\|T\|_{HS}\le C_0} |\hat{\Phi}_n(T) - \Phi(T)| \le C_0 W_2(\hat{\mu}_{n_1},\mu) + W_2(\hat{\nu}_{n_2},\nu) \tag{40}$$

By Proposition 2.2.6. of Panaretos and Zemel (2020) and with our assumption of bounded second moments of μ and ν, we have $W_2(\hat{\mu}_{n_1},\mu)$ and $W_2(\hat{\nu}_{n_2},\nu)$ converge almost surely to 0 as $n\to\infty$. As almost surely convergence implies convergence in probability, we have

$$\sup_{\|T\|_{HS}\le C_0} |\hat{\Phi}_n(T) - \Phi(T)| \xrightarrow{P} 0, \tag{41}$$

which means for all $\epsilon > 0$,

$$P\left(\sup_{\|T\|_{HS}\le C_0} |\hat{\Phi}_n(T) - \Phi(T)| < \epsilon\right) \to 1, \tag{42}$$

Step 2. Combining $\sup_{\|T\|_{HS}\le C_0} |\hat{\Phi}_n(T) - \Phi(T)| < \epsilon$ with the fact that $\Phi^2(T) \le C_1\|T\|_{HS} + C_2$ implies that for all T such that $\|T\|_{HS} \le C_0$, we have $\Phi^2(T) \le C_1C_0 + C_2 =: C$

$$\begin{aligned}|\hat{J}_n(T) - J(T)| &= |\hat{\Phi}_n^2(T) - \Phi^2(T)|\\ &= |\hat{\Phi}_n(T) - \Phi(T)||\hat{\Phi}_n(T) + \Phi(T)|\\ &\le \epsilon(2\sqrt{C} + \epsilon).\end{aligned}$$

Hence

$$P\left(\sup_{\|T\|_{HS}\le C_0} |\hat{J}_n(T) - J(T)| < \epsilon(2\sqrt{C}+\epsilon)\right) \ge P\left(\sup_{\|T\|_{HS}\le C_0} |\hat{\Phi}_n(T) - \Phi(T)| < \epsilon\right) \to 1. \tag{43}$$

Noticing that for all $\delta > 0$, there exists an $\epsilon > 0$ such that $\epsilon(2\sqrt{C}+\epsilon) = \delta$, we arrive at the convergence in probability to 0 of $\sup_{\|T\|_{HS}\le C_0} |\hat{J}_n(T) - J(T)|$.

(ii) We also organize the proof in a few steps.

Step 1. Denote $\hat{\Phi}_n(T) = W_2(T\#\hat{\mu}_{n_1},\hat{\nu}_{n_2})$. We first show that for any fixed C_0,

$$\sup_{\|T\|_{HS}\ge C_0} \frac{|\hat{\Phi}_n(T) - \Phi(T)|}{\|T\|_{HS}} \xrightarrow{P} 0 \quad (n\to\infty). \tag{44}$$

Indeed, from (39),

$$\sup_{\|T\|_{HS}\ge C_0} \frac{|\hat{\Phi}_n(T) - \Phi(T)|}{\|T\|_{HS}} \le W_2(\hat{\mu}_{n_1},\mu) + \frac{W_2(\hat{\nu}_{n_2},\nu)}{C_0}. \tag{45}$$

As above, we have $W_2(\hat{\mu}_{n_1}, \mu)$ and $W_2(\hat{\nu}_{n_2}, \nu)$ converge to 0 almost surely as $n \to \infty$. Hence, $\sup_{\|T\|_{HS} \geq C_0} \frac{|\hat{\Phi}_n(T) - \Phi(T)|}{\|T\|_{HS}} \to 0$ almost surely, and therefore in probability.

Step 2. For any fixed C_0 and δ,

$$P\left(\sup_{\|T\|_{HS} \geq C_0} \frac{|\hat{\Phi}_n(T) - \Phi(T)|}{\|T\|_{HS}} < \delta\right) \to 1 \quad (n \to \infty). \tag{46}$$

The event $\sup_{\|T\|_{HS} \geq C_0} \frac{|\hat{\Phi}_n(T) - \Phi(T)|}{\|T\|} < \delta$ implies that for all T such that $\|T\|_{HS} \geq C_0$, from Lemma 3 we have

$$\hat{J}_n(T) \leq (\Phi(T) + \delta\|T\|_{HS})^2 + \eta\|T\|_{HS}^2 \leq (\sqrt{C_1\|T\|_{HS}^2 + C_2} + \delta\|T\|_{HS})^2 + \eta\|T\|_{HS}^2.$$

Now for each K, we can choose a $\tilde{T}_K \in B_K$ such that $\|\tilde{T}_K\|_{HS} = C_0$. Thus,

$$\begin{aligned}\inf_{T \in B_K} \hat{J}_n(T) \leq \hat{J}_n(\tilde{T}_K) &\leq (\sqrt{C_1\|\tilde{T}_K\|_{HS}^2 + C_2} + \delta\|\tilde{T}_K\|_{HS})^2 + \eta\|\tilde{T}_K\|_{HS}^2 \\ &= (\sqrt{C_1 C_0^2 + C_2} + \delta C_0)^2 + \eta C_0^2 =: C,\end{aligned}$$

which is a constant.

On the other hand, choose $D = \sqrt{C/\eta}$, we have for all T such that $\|T\|_{HS} > D$

$$\hat{J}_n(T) \geq \eta\|T\|_{HS}^2 > C, \tag{47}$$

which means $\inf_{T \in B_K : \|T\|_{HS} > D} \hat{\Phi}_n(T) > C$ for all K.

Combining two facts above, we have $\|\hat{T}_{K,n}\|_{HS} \leq D$ for all K.

Step 3. It follows from the previous step that

$$P\left(\sup_K \|\hat{T}_{K,n}\|_{HS} \leq D\right) \geq P\left(\sup_{\|T\|_{HS} \geq C_0} \frac{|\hat{\Phi}_n(T) - \Phi(T)|}{\|T\|_{HS}} < \delta\right), \tag{48}$$

which means this probability also goes to 1 as $n \to \infty$.

■

We are ready to tackle the consistency of our estimation procedure.

Theorem 7 *There exists a unique minimizer $\hat{T}_{K,n}$ of $\hat{J}_n$ over B_K for all n and K. Moreover, $\hat{T}_{K,n} \xrightarrow{P} T_0$ in $\|\cdot\|_{HS}$ as $K_1, K_2, n_1, n_2 \to \infty$.*

Proof [Proof of Theorem 7] The proof proceeds in several smaller steps.

Step 1. Take any $\epsilon > 0$. As $T_K \to T_0$ when $K \to \infty$, there exist $\kappa = (\kappa_1, \kappa_2)$ such that $\|T_K - T_0\|_{HS} \leq \epsilon$ for all $K_1 > \kappa_1, K_2 > \kappa_2$. Let

$$L_\epsilon = \inf_{T \in \mathscr{B}_{HS} \setminus B(T_0, \epsilon)} J(T), \tag{49}$$

where $B(T, \epsilon)$ is the Hilbert-Schmidt open ball centered at T having radius ϵ. It can be seen that $L_\epsilon > J(T_0)$, as otherwise, there exists a sequence $(T_p)_p \notin B(T, \epsilon)$ such that $J(T_p) \to J(T_0)$, which implies $T_p \to T_0$, a contradiction.

Step 2. Let $\delta = L_\epsilon - J(T_0) > 0$. By Lemma 5, we can choose κ large enough so that we also have $|J(T_K) - J(T_0)| < \delta/2 \; \forall \, K_1 > \kappa_1, K_2 > \kappa_2$. Let

$$L_{K,\epsilon} = \inf_{B_K \setminus B(T_K, 2\epsilon)} J(T).$$

As $B(T_0, \epsilon) \subset B(T_K, 2\epsilon)$ and $B_K \subset \mathscr{B}_{HS}$, we have

$$L_{K,\epsilon} = \inf_{B_K \setminus B(T_K, 2\epsilon)} J(T) \geq \inf_{T \in \mathscr{B}_{HS} \setminus B(T_0, \epsilon)} J(T) = L_\epsilon. \tag{50}$$

Therefore,

$$L_{K,\epsilon} - J(T_K) \geq L_\epsilon - J(T_0) - \delta/2 = \delta/2. \tag{51}$$

for all $K > \kappa$.

Step 3. Now, if we have

$$\sup_{\|T\| \leq D} |\hat{J}_n(T) - J(T)| \leq \delta/4, \quad \sup_K |\hat{T}_{K,n}| \leq D, \tag{52}$$

where D is a constant as in Lemma 6, then

$$\hat{J}_n(T_K) \leq J(T_K) + \delta/4, \tag{53}$$

and

$$\hat{J}_n(T) \geq J(T) - \delta/4 \geq J(T_K) + \delta/4, \tag{54}$$

for all $\|T\|_{HS} \leq D$ *and* $T \in B_K \setminus B(T_K, 2\epsilon)$, where the last inequality is due to Step 2.

Combining with $|\hat{T}_{K,n}| \leq D$, we have $\hat{T}_{K,n}$ must lie inside $B(T_K, 2\epsilon) \cap B_K$ because it is the minimizer of $\hat{J}_n$ over B_K. Hence $\|\hat{T}_{K,n} - T_K\|_{HS} \leq 2\epsilon$, which deduces that $\|\hat{T}_{K,n} - T_0\|_{HS} \leq \|\hat{T}_{k,n} - T_K\|_{HS} + \|T_k - T_0\|_{HS} \leq 2\epsilon + \epsilon = 3\epsilon$.

Step 4. Continuing from the previous step, for all κ large enough, we have the following inclusive relation of events

$$\{\sup_{\|T\| \leq D} |\hat{J}_n(T) - J(T)| \leq \delta/4\} \cap \{\sup_K |\hat{T}_{K,n}| \leq D\} \subset \{\sup_{K > \kappa} \|\hat{T}_{K,n} - T_0\|_{HS} \leq 3\epsilon\} \tag{55}$$

Using the inequality that for any event A, B, $P(A \cap B) \geq P(A) + P(B) - 1$, we obtain

$$P(\sup_{K > \kappa} \|\hat{T}_{K,n} - T_K\|_{HS} \leq 3\epsilon) \geq P(\sup_{\|T\|_{HS} \leq D} |\hat{J}_n(T) - J(T)| \leq \delta/4) + P(\sup_K |\hat{T}_{K,n}| \leq D) - 1, \tag{56}$$

which goes to 1 as $n \to \infty$ due to Lemma 6. Because this is true for all $\epsilon > 0$, we have

$$\hat{T}_{K,n} \xrightarrow{P} T_0, \tag{57}$$

as $K, n \to \infty$. ■

Consistency when the functional data are observed only at design points Finally, we are ready to prove Theorem 8, which is re-stated herein.

Theorem 8 *(i) For every n_1, n_2, K_1, K_2 and sequences of design points in source and target domains, the cost function*

$$\hat{J}_{n,K,d}(\mathbf{\Lambda}) = \min_{\pi \in \hat{\Pi}} \sum_{l,k=1}^{n_1,n_2} \pi_{lk} D_{lkd}(\mathbf{\Lambda}) + \eta \|\mathbf{\Lambda}\|_F^2, \tag{58}$$

where

$$D_{lkd}(\mathbf{\Lambda}) = \|\mathbf{\Lambda} a_{ld} - b_{kd}\|_2^2,$$

in which $a_{ld} = (\langle f_{1,l}, U_i \rangle_d)_{i=1}^{K_1}$ and $b_{kd} = (\langle f_{2,k}, V_j \rangle_d)_{i=1}^{K_2}$ $\forall l, k$, has unique minimizer $\mathbf{\Lambda}_{n,K,d} \in \mathbb{R}^{K_2 \times K_1}$ that corresponds to operator $T_{n,K,d}$.

(ii) Suppose that for any natural index pair (i, j), there holds

$$\langle f, U_i \rangle_d \to \langle f, U_i \rangle_{H_1}, \langle g, V_j \rangle_d \to \langle g, V_j \rangle_{H_2}, \tag{59}$$

almost surely as $d \to \infty$, where $f \sim \mu$ and $g \sim \nu$. Then for any sequence $n_1, n_2, K_1, K_2 \to \infty$ and $d \to \infty$ with a rate depends on n_1, n_2, K_1, K_2, we have $T_{n,K,d} \xrightarrow{P} T_0$ in $\|\cdot\|_{HS}$. Here T_0 denotes the minimizer of the population version of FOT given in Eq. (6).

Proof For each n_1, n_2, K_1, K_2 and sequences of design points, the existence and uniqueness of $\mathbf{\Lambda}_{n,K,d}$ follows from Theorem 4. Thus, part (i) is immediate. To establish part (ii), we rewrite the objective function (58) as

$$W_2^2(T_{\#}\mu_{n,K,d}, \nu_{n,K,d}) + \eta \|T\|_{HS}^2, \tag{60}$$

where

$$\mu_{n,K,d} = \frac{1}{n_1} \sum_{l=1}^{n_1} \delta_{f_{l,K,d}}, \quad f_{l,K,d} = \sum_{i=1}^{K_1} \langle f_l, U_i \rangle_d U_i, \quad (f_l)_{l=1}^{n_1} \overset{iid}{\sim} \mu,$$

and

$$\nu_{n,K,d} = \frac{1}{n_2} \sum_{k=1}^{n_2} \delta_{g_{k,K,d}}, \quad g_{k,K,d} = \sum_{i=1}^{K_2} \langle g_k, V_j \rangle_d V_j, \quad (g_k)_{k=1}^{n_2} \overset{iid}{\sim} \nu.$$

As a consequence of Lemma 6 and Theorem 7, the conclusion of the theorem can be achieved if we can show that

$$W_2(\mu_{n,K,d}, \mu) \xrightarrow{P} 0, \quad W_2(\nu_{n,K,d}, \nu) \xrightarrow{P} 0. \tag{61}$$

It suffices to establish the convergence for μ, as the work for ν can be done in the same way. Note that

$$W_2(\mu_n, \mu) \xrightarrow{a.s.} 0, \tag{62}$$

as $n_1 \to \infty$, where $\mu_n = \frac{1}{n_1} \sum_{l=1}^{n_1} \delta_{f_l}$, so that we only need to show

$$W_2(\mu_{n,K,d}, \mu_n) \xrightarrow{P} 0 \quad (\text{as } n \to \infty). \tag{63}$$

Consider the coupling that places mass $\frac{1}{n_1}$ on $(f_l, f_{l.K,d})$ for all $l = 1, \ldots, d_1$, then we have

$$W_2^2(\mu_{n,K,d}, \mu_n) \leq \frac{1}{n_1}\sum_{l=1}^{n_1} \|f_l - f_{l,K,d}\|_{H_1}^2$$
$$= \frac{1}{n_1}\sum_{l=1}^{n_1}\left(\sum_{i=1}^{K_1}(\langle f_l, U_i\rangle_d - \langle f_l, U_i\rangle_{H_1})^2 + \sum_{i=K_1+1}^{\infty}\langle f_l, U_i\rangle_{H_1}^2\right). \tag{64}$$

As $n_1, K_1 \to \infty$, by an application of Fubini's theorem, we have

$$\lim_{n_1,K_1\to\infty}\sum_{l=1}^{n_1}\sum_{i=1}^{K_1}\frac{1}{n_1}\langle f_l, U_i\rangle_{H_1}^2 = \lim_{n_1\to\infty}\sum_{l=1}^{n_1}\frac{1}{n_1}\|f_l\|_{H_1}^2 = E_{f\sim\mu}\|f\|_{H_1}^2, \tag{65}$$

almost surely. Hence,

$$\lim_{n_1,K_1\to\infty}\sum_{l=1}^{n_1}\sum_{i=K_1+1}^{\infty}\frac{1}{n_1}\langle f_l, U_i\rangle_{H_1}^2 = 0, \tag{66}$$

almost surely. Now consider the first sum of the right-hand side of (64), for each $l = 1, \ldots, n_1$ and $i = 1, \ldots, K_1$, we have from the assumption of the theorem that

$$(\langle f_l, U_i\rangle_d - \langle f_l, U_i\rangle_{H_1})^2 \to 0, \tag{67}$$

almost surely for $f_l \sim \mu$ as $d \to \infty$, and almost surely convergence implies convergence in probability. So, for every $\delta, \epsilon > 0$, we can choose $D = D(\delta, \epsilon, K_1, n_1)$ such that

$$P\left(\sum_{i=1}^{K_1}(\langle f_l, U_i\rangle_d - \langle f_l, U_i\rangle_{H_1})^2 > \epsilon\right) \leq \frac{\delta}{n_1}, \quad \forall d > D. \tag{68}$$

Hence,

$$P\left(\frac{1}{n_1}\sum_{l=1}^{n_1}\sum_{i=1}^{K_1}(\langle f_l, U_i\rangle_d - \langle f_l, U_i\rangle_{H_1})^2 > \epsilon\right) \leq P\left(\bigcup_{l=1}^{n_1}\left\{\sum_{i=1}^{K_1}(\langle f_l, U_i\rangle_d - \langle f_l, U_i\rangle_{H_1})^2 > \epsilon\right\}\right)$$
$$\leq \sum_{l=1}^{n_1} P\left(\sum_{i=1}^{K_1}(\langle f_l, U_i\rangle_d - \langle f_l, U_i\rangle_{H_1})^2 > \epsilon\right)$$
$$\leq \delta,$$

for all $d > D(\delta, \epsilon, K_1, n_1)$. It means that $W_2(\mu_{n,K,d}, \mu_n)$ converges to 0 in probability as $n_1, K_1 \to \infty$ and the numbers of design points grow to infinity with a rate depending on n_1 and K. Thus, the RHS of (64) vanishes in probability, so Eq. (63) is established, and the rest of the proof follows similarly to the proofs of Lemma 6 and Theorem 7. ■

7. Discussions and Future Work

We proposed a formulation of optimal transport for probability distributions on domains of functions, where the stochastic map between functional domains can be represented by an infinite dimensional Hilbert-Schmidt operator mapping a Hilbert space of functions to another. We proposed a learning method for transport maps based on subspace approximations of Hilbert-Schmidt operators, and implemented an efficient algorithm that involves joint optimization of such operators and a suitable space of stochastic couplings. Theoretical guarantees on the existence, uniqueness, and consistency of our estimator were achieved. Through simulation studies, we validated our theory and demonstrated the effectiveness of our method of approximation, and that of the learning algorithm, by taking into account the functional nature of the data domains. The effectiveness of our approach was further demonstrated in a couple of real-world domain adaptation applications involving complex and realistic robot arm movements.

Work	measure	functional	transport map
(Genevay et al., 2016)	discrete, semi-discrete	no	n/a
(Seguy et al., 2017)	continuous	no	neural network
(Alvarez-Melis et al., 2019; Grave et al., 2019), (Meng et al., 2019)	discrete	no	rigid transformation
(Perrot et al., 2016)	discrete	no	linear & kernel
(Xie et al., 2019)	empirical	no	GAN
(Makkuva et al., 2019)	empirical	no	ICNN
(Mallasto and Feragen, 2017)	single Gaussian process	**yes**	n/a*
This work	measures on Hilbert spaces	**yes**	Hilbert-Schmidt operator

Table 3: Related works on optimal transport map estimation.

7.1 Related work

Optimal transport-based applications have made significant strides in the field of machine learning (Arjovsky et al., 2017; Ho et al., 2017; Li et al., 2019; Chen et al., 2019; Alvarez-Melis et al., 2019; Fan et al., 2020; Alvarez-Melis and Fusi, 2020; Fan et al., 2021; Korotin et al., 2021; Fatras et al., 2021; Nguyen et al., 2022; Bunne et al., 2022b). A suite of OT based approaches involve solving the Kantorovich problem through linear programming or the Sinkhorn algorithm (Cuturi, 2013; Courty et al., 2016; Pooladian and Niles-Weed, 2021). Typically, they look for an optimal coupling between empirical measures while the objectives can be extended depending on the problems at hand, such as the ones that gave rise to the Gromov-Wasserstein distance (Mémoli, 2011), Sliced Wasserstein (Nguyen et al., 2022), partial minibatch OT (Fatras et al., 2021), outlier robustness criterion (Mukherjee et al., 2020), Orlicz-Wasserstein distance (Guha et al., 2023), and so on. The results obtained can be extended to vector spaces with high dimensions, but scaling the algorithms for larger sample sizes is typically challenging. Also, managing continuous measures within vector

spaces is complicated and the most significant difficulty lies in generalizing to out-of-sample data that has not been seen before.

From an applied viewpoint, the problem of (Monge) map estimation has always been of interest, because the Monge map represents the simplest form of the optimal coupling distribution *if* such a map exists. The question of existence and almost sure uniqueness of the Monge map was settled by Brenier (1987) in what is now known as the celebrated Brenier theorem, a result which has been extended and generalized by many other authors (see Chapter 6 of Ambrosio et al. (2005) and Chapters 9–11 of Villani (2008)). Brenier-type theorems helped to rejuvenate the development of optimal transport, in theory and subsequently in practice. In practice, early attempts at *learning* a transport map from data include approaches that represent transport maps as linear transformation or kernel based function classes (Perrot et al., 2016). A more ambitious approach was made in the work of Seguy et al. (2017), who proposed to find the optimal transport map parameterized by a rich neural network model, following the regularized Kantorovich dual formulation. Exploiting the characterization of the Monge map as the gradient of a convex potential, various authors have proposed to exploit computational convexity (Makkuva et al., 2019; Korotin et al., 2021; Fan et al., 2020; Korotin et al., 2022; Bunne et al., 2022a) by leveraging input convex neural networks (ICNNs) (Amos et al., 2017). A summary of map estimation work is given in Table 3. While most of the aforementioned literature still centers at the situation where the support of the distributions are subsets of finite dimensional vector spaces, the growing implementation of machine learning algorithms in various real-world applications, such as time series, have motivated the formulation of optimal transport problem in the domains of functions.

As mentioned in the Introduction, most known results and techniques on optimal transport between distributions on function spaces are related to Gaussian processes and Gaussian measures on normed spaces (Mallasto and Feragen, 2017; Masarotto et al., 2019; Knott and Smith, 1984; Pigoli et al., 2014). These results are natural generalization from those of the multivariate Gaussian distributions (Dowson and Landau, 1982; Givens and Shortt, 1984). Specifically, the 2-Wasserstein distance between Gaussian processes with certain covariance coincides with the Procrustes distance between the two covariance operators, which can be approximated arbitrarily well via finite-dimensional approximation (Masarotto et al., 2019). Additional advances on optimal transport for Gaussian measures have been made by other authors, e.g., (Takatsu et al., 2011; Agueh and Carlier, 2011; Álvarez-Esteban et al., 2016). In particular, for centered Gaussian measures supported by Hilbert spaces, there exists a linear *subspace* of the (source) Hilbert space where the optimal map is explicit and well-defined as a linear operator. Unfortunately, such a linear map is unbounded so it cannot be extended to the whole (source and target) domains. Such theoretical advances notwithstanding, in practice, the Gaussian distribution assumption is clearly too restrictive in many domains (recall our comparison to GPOT of Mallasto and Feragen (2017) in Section 5). Our work may be viewed as a first step at addressing optimal transport in the domains of functions that go beyond the Gaussian assumption, and with a particular focus on learning the explicit transport map for sampled functional data.

The formulation presented in this paper can be viewed as a regularization approach to the optimal coupling problem with respect to the source and target distributions on Hilbert spaces of functions. The overall optimal coupling is therefore represented jointly by

a compact linear operator T which transports functions in the source domain to functions in the target domain, and by a stochastic coupling distribution Π. The compactness of T and the sparseness of Π induced by suitable penalty techniques offer several theoretical advantages (i.e., existence, uniqueness and consistency of the estimates), as well as practical advantages (i.e., efficient computation and interpretability).

The primary limitation of our approach, as discussed in the Introduction, is in the situation where the deterministic optimal transport map exists but is either an unbounded linear operator or a nonlinear operator. Developing a theory and methods to accommodate unbounded or nonlinear optimal transport map in the infinite dimensional setting will be challenging. From a practical and data-driven perspective, one needs to balance the cost of estimating a nonlinear map with the possibility that a single nonlinear map may not offer a very accurate pushforward probability measure for the target domain. That is because in infinite dimensional settings, it is more likely than not that such a deterministic map does not exist, unless certain restrictions are in place.

7.2 Future work

Within our formulation of linear operator regularized optimal transport in the functional domains, there are a number of interesting questions and approaches that are worthy of further investigation.

- It is of interest to characterize the convergence behavior of the FOT estimator with respect to the number of function samples, basis functions, and design points in the infinite-dimensional setting. Very recently, several groups of researchers have started to study rates of convergence of optimal transport map estimators. However, their work focuses on finite-dimensional domain settings (e.g., (Hütter and Rigollet, 2021; Manole et al., 2021; Gunsilius, 2021; Deb et al., 2021)). New ideas and more powerful techniques will probably be required to address analogous questions in the functional domains.

- There is a vast opportunity to expand the scope of real-world applications by bridging functional data analysis techniques with the optimal transport formalism, where both functional data analysis and optimal transport viewpoints play complementary roles toward achieving effective solutions. For example, in domain adaptation tasks in robotics, healthcare, and autonomous driving, data are intrinsically associated with physical processes and functions. One may also be interested in learning a transport map that pushes forward the samples across diverse yet related domains, such as assembling in manufacturing. For these tasks, it would be critical to investigate the choice of basis functions for a specific machine learning problem, which is further related to functional PCA or representational learning.

- The proposed FOT framework should be useful toward tackling the domain generalization problem. A principled way is to pushforward the predictive function from a source domain towards a target domain by directly working on the predictive function's parameters. This idea can be generalized to prevalent deep learning methods by considering the basis functions as deep feature extractors. Thus, an FOT based approach may provide a more *interpretable* solution for domain generalization with

high-dimensional data, such as those that arise in natural language processing and computer vision.

References

Manya V Afonso, José M Bioucas-Dias, and Mário AT Figueiredo. An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems. *IEEE transactions on image processing*, 20(3):681–695, 2010.

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. Towards optimal transport with global invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1870–1879. PMLR, 2019.

David Alvarez-Melis, Youssef Mroueh, and Tommi Jaakkola. Unsupervised hierarchy matching with optimal transport over hyperbolic spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 1606–1617. PMLR, 2020.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

Brandon Amos, Samuel Cohen, Giulia Luise, and Ievgen Redko. Meta optimal transport. *arXiv preprint arXiv:2206.05262*, 2022.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018.

Yifeng Bie. *Functional principal component analysis based machine learning algorithms for spectral analysis*. PhD thesis, 2021.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 880–889. PMLR, 2018.

Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, et al. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4243–4250. IEEE, 2018.

Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional monge maps. *arXiv preprint arXiv:2206.14262*, 2022a.

Charlotte Bunne, Laetitia Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal optimal transport modeling of population dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR, 2022b.

Yuhang Cai and Lek-Heng Lim. Distances between probability distributions of different dimensions. *IEEE Transactions on Information Theory*, 68(6):4020–4031, Jun 2022. ISSN 1557-9654. doi: 10.1109/tit.2022.3148923. URL http://dx.doi.org/10.1109/tit.2022.3148923.

Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. Improving sequence-to-sequence learning via optimal transport. *arXiv preprint arXiv:1901.06283*, 2019.

Li-Fang Cheng, Bianca Dumitrascu, Gregory Darnell, Corey Chivers, Michael Draugelis, Kai Li, and Barbara E Engelhardt. Sparse multi-output gaussian processes for online medical time series prediction. *BMC medical informatics and decision making*, 20(1):1–23, 2020.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

Jacques Dauxois, Alain Pousse, and Yves Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of multivariate analysis*, 12(1):136–154, 1982.

Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *arXiv preprint arXiv:2107.01718*, 2021.

Marc Peter Deisenroth, Dieter Fox, and Carl Edward Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):408–423, 2013.

Françoise Demengel and Gilbert Demengel. *Functional Spaces for the Theory of Elliptic Partial Differential Equations*. Springer London, 2012. doi: 10.1007/978-1-4471-2807-6. URL https://doi.org/10.1007%2F978-1-4471-2807-6.

D. Dowson and B. Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Emilien Dupont, Yee Whye Teh, and Arnaud Doucet. Generative models as distributions of functions. *arXiv preprint arXiv:2102.04776*, 2021.

Zackory Erickson, Vamsee Gangaram, Ariel Kapusta, C. Karen Liu, and Charles C. Kemp. Assistive gym: A physics simulation framework for assistive robotics. *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. Scalable computations of wasserstein barycenter via input convex neural networks. *arXiv preprint arXiv:2007.04462*, 2020.

Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. Variational wasserstein gradient flow. *arXiv preprint arXiv:2112.02424*, 2021.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR, 2021.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, 2006.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

Clive AJ Fletcher. Computational galerkin methods. In *Computational galerkin methods*, pages 72–85. Springer, 1984.

Natasha Zhang Foutz and Wolfgang Jank. Research note—prerelease demand forecasting for motion pictures using functional shape analysis of virtual stock markets. *Marketing Science*, 29(3):568–579, 2010.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pages 3440–3448, 2016.

Daniel Gervini. Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600, 2008.

Euhanna Ghadimi, André Teixeira, Iman Shames, and Mikael Johansson. Optimal parameter selection for the alternating direction method of multipliers (admm): quadratic problems. *IEEE Transactions on Automatic Control*, 60(3):644–658, 2014.

Clark R Givens and Rae Michael Shortt. A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, 2019.

Aritra Guha, Nhat Ho, and XuanLong Nguyen. On excess mass behavior in gaussian mixture models with orlicz-wasserstein distances. *Proceedings of the ICML*, 2023.

Florian F. Gunsilius. On the convergence rate of potentials of brenier maps. *Econometric Theory*, page 1–37, 2021. doi: 10.1017/S0266466621000037.

Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via wasserstein means. In *International Conference on Machine Learning*, pages 1501–1509. PMLR, 2017.

Lajos Horváth and Piotr Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

Tailen Hsing and Randall Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.

Chin-Wei Huang, Ricky TQ Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. *arXiv preprint arXiv:2012.05942*, 2020.

Peter J. Huber. Robust estimation of a location parameter. *Annals of Statistics*, 53 (1): 73–101, 1964.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.

Nikolay Jetchev and Marc Toussaint. Trajectory prediction: learning to map situations to robot trajectories. In *Proceedings of the 26th annual international conference on machine learning*, pages 449–456, 2009.

Shizuo Kakutani. On equivalence of infinite product measures. *Annals of Mathematics*, pages 214–224, 1948.

Leonid Kantorovitch. On the translocation of masses. *Management Science*, 5(1):1–4, 1958.

Mitsunori Kayano and Sadanori Konishi. Sparse functional principal component analysis via regularized basis expansions and its application. *Communications in Statistics-Simulation and Computation*, 39(7):1318–1333, 2010.

Y Ke. Large-scale optimal transport map estimation using projection pursuit. *NeurIPS 2019*, 2019.

R.W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer New York, 2010. ISBN 9780387938394. URL https://books.google.com.vn/books?id=aVJmcega44cC.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

Martin Knott and Cyril S Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49, 1984.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *Advances in Neural Information Processing Systems*, 34:14593–14605, 2021.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

John Lee, Max Dabagia, Eva L Dyer, and Christopher J Rozell. Hierarchical optimal transport for multimodal distribution alignment. *arXiv preprint arXiv:1906.11768*, 2019.

Jing Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767 – 798, 2020. doi: 10.3150/19-BEJ1151. URL https://doi.org/10.3150/19-BEJ1151.

Ruilin Li, Xiaojing Ye, Haomin Zhou, and Hongyuan Zha. Learning to match via inverse optimal transport. *Journal of machine learning research*, 20, 2019.

Changliu Liu, Te Tang, Hsien-Chung Lin, Yujiao Cheng, and Masayoshi Tomizuka. Serocs: Safe and efficient robot collaborative systems for next generation intelligent industrial co-robots. *arXiv preprint arXiv:1809.08215*, 2018.

Chong Liu, Surajit Ray, and Giles Hooker. Functional principal component analysis of spatially correlated data. *Statistics and Computing*, 27(6):1639–1654, 2017.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2022.

Ashok Vardhan Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason D Lee. Optimal transport mapping via input convex neural networks. *arXiv preprint arXiv:1908.10962*, 2019.

Anton Mallasto and Aasa Feragen. Learning from uncertain curves: The 2-wasserstein metric for gaussian processes. In *Advances in Neural Information Processing Systems*, pages 5660–5670, 2017.

Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. *arXiv preprint arXiv:1811.02790*, 2018.

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.

Valentina Masarotto, Victor M Panaretos, and Yoav Zemel. Procrustes metrics on covariance operators and optimal transportation of gaussian processes. *Sankhya A*, 81(1):172–213, 2019.

Andreas Maurer. Learning similarity with operator-valued large-margin classifiers. *The Journal of Machine Learning Research*, 9:1049–1082, 2008.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Cheng Meng, Yuan Ke, Jingyi Zhang, Mengrui Zhang, Wenxuan Zhong, and Ping Ma. Large-scale optimal transport map estimation using projection pursuit. *Advances in Neural Information Processing Systems*, 32:8118–8129, 2019.

Ardalan Mirshani, Matthew Reimherr, and Aleksandra Slavković. Formal privacy for functional data with gaussian perturbations. In *International Conference on Machine Learning*, pages 4595–4604. PMLR, 2019.

Debarghya Mukherjee, Aritra Guha, Justin Solomon, Yuekai Sun, and Mikhail Yurochkin. Outlier-robust optimal transport. *arXiv preprint arXiv:2012.07363*, 2020.

Khai Nguyen, Tongzheng Ren, Huy Nguyen, Litu Rout, Tan Nguyen, and Nhat Ho. Hierarchical sliced wasserstein distance. *arXiv preprint arXiv:2209.13570*, 2022.

Victor M Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. *Advances in Neural Information Processing Systems*, 29: 4197–4205, 2016.

Davide Pigoli, John AD Aston, Ian L Dryden, and Piercesare Secchi. Distances and inference for covariance operators. *Biometrika*, 101(2):409–422, 2014.

Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

Shenghao Qin, Jiacheng Zhu, Jimmy Qin, Wenshuo Wang, and Ding Zhao. Recurrent attentive neural process for sequential data. *arXiv preprint arXiv:1910.09323*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Carlos Ramos-Carreño, José Luis Torrecilla, Miguel Carbajo-Berrocal, Pablo Marcos, and Alberto Suárez. scikit-fda: a python package for functional data analysis. *arXiv preprint arXiv:2211.02566*, 2022.

J. O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, 2 edition, 2006.

JO Ramsay and BW Silverman. *Functional data analysis*. Springer, 2005.

Anand Rangarajan, Haili Chui, and Fred L Bookstein. The softassign procrustes matching algorithm. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 29–42. Springer, 1997.

Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Gaussian process classification and active learning with multiple annotators. In *International conference on machine learning*, pages 433–441. PMLR, 2014.

Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*, volume 32. Macmillan New York, 1988.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

Han Lin Shang. A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, 98:121–142, 2014.

Pratyusha Sharma, Lekha Mohan, Lerrel Pinto, and Abhinav Gupta. Multiple interactions made easy (mime): Large scale demonstrations data for imitation. *arXiv preprint arXiv:1810.07121*, 2018.

Asuka Takatsu et al. Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

Anthony Tompkins, Ransalu Senanayake, and Fabio Ramos. Online domain adaptation for occupancy mapping. *arXiv preprint arXiv:2007.00164*, 2020.

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

Limin Wang. *Karhunen-Loeve expansions and their applications.* PhD thesis, London School of Economics and Political Science (United Kingdom), 2008.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Yujia Xie, Minshuo Chen, Haoming Jiang, Tuo Zhao, and Hongyuan Zha. On scalable and efficient computation of large scale optimal transport. *arXiv preprint arXiv:1905.00158*, 2019.

Mengdi Xu, Wenhao Ding, Jiacheng Zhu, ZUXIN LIU, Baiming Chen, and Ding Zhao. Task-agnostic online reinforcement learning with an infinite mixture of gaussian processes. *Advances in Neural Information Processing Systems*, 33, 2020.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903, 2005.

K. Yosida. *Functional Analysis*. Classics in Mathematics. Springer Berlin Heidelberg, 1995. ISBN 9783540586548. URL https://books.google.com.vn/books?id=QqNpbTQwKXMC.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, 2017.

Huaiyu Zhu, Christopher KI Williams, Richard Rohwer, and Michal Morciniec. Gaussian regression and optimal finite dimensional linear models. 1997.