UNIVERSITY OF CALIFORNIA,
IRVINE

Data-driven approximation of transfer operators: DMD, Perron–Frobenius, and statistical learning in Wasserstein space

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mechanical and Aerospace Engineering

by

Amirhossein Karimi

Dissertation Committee:
Professor Tryphon T. Georgiou, Chair
Professor Solmaz S. Kia
Professor Faryar Jabbari

2022

DEDICATION

To my family

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

First of all, I would like to express my deepest gratitude and thanks to my advisor Professor Tryphon Georgiou for his valuable guidance and continuous support during my PhD study.

I would like to express my sincere gratitude to my committee members Professor Solmaz Kia, professor Katya Krupchyk, Professor Faryar Jabbari, and Professor Haithem Taha for their help and insightful comments on my research.

Many thanks to my research colleague, Luigia Ripani for very helpful comments and ideas to write the paper "Statistical learning in Wasserstein space" [1].

Chapter 2 is taken from the paper that is published as "The challenge of small data: dynamic mode decomposition, Redux" by Amirhossein Karimi and Tryphon Georgiou in IEEE Conference on Decision and Control, 2021 [2]. The dissertation author is the primary investigator and author of this paper.

Chapter 4 is taken from the paper "Regression analysis of distributional data through Multi-Marginal Optimal transport" by Amirhossein Karimi and Tryphon Georgiou which is submitted to SIAM Journal on Mathematics of Data Science, 2021 [3] and the paper "Statistical learning in Wasserstein space" by Amirhossein Karimi, Luigia Ripani and Tryphon Georgiou published in IEEE Control Systems Letters, 2020 [1]. The dissertation author is the primary investigator and author of these papers.

Chapter 5 is taken from the paper "Data-Driven Approximation of the Perron-Frobenius Operator Using the Wasserstein Metric" by Amirhossein Karimi and Tryphon Georgiou, that is submitted to Automatica, 2021. The dissertation author is the primary investigator and author of this paper.

VITA

Amirhossein Karimi

EDUCATION

PhD student, in Mechanical and Aerospace Engineering **2016-2021**
University of California, Irvine *Irvine, USA*

M.S., in Mechatronics **2015**
Universty of Tehran *Tehran, Iran*

B.S., in Mechanical Engineering **2012**
Isfahan University of Technology *Isfahan, Iran*

RESEARCH EXPERIENCE

Graduate Research Assistant **2017–2021**
University of California, Irvine *Irvine, USA*

REFEREED JOURNAL PUBLICATIONS

Statistical learning in Wasserstein space **2020**
IEEE Control Systems Letters

Regression analysis of distributional data through Multi-Marginal Optimal transport **2021**
SIAM Journal on Mathematics of Data Science

Data-Driven Approximation of the Perron-Frobenius Operator Using the Wasserstein Metric **2021**
Automatica

REFEREED CONFERENCE PUBLICATIONS

Statistical learning in Wasserstein space **2020**
IEEE Conference of Decision and Control, 2020 (CDC)

The Challenge of Small Data: Dynamic Mode Decomposition, Redux **2021**
IEEE Conference of Decision and Control, 2021 (CDC)

ABSTRACT OF THE DISSERTATION

Data-driven approximation of transfer operators:
DMD, Perron–Frobenius, and statistical learning in Wasserstein space

By

Amirhossein Karimi

Doctor of Philosophy in Mechanical and Aerospace Engineering

University of California, Irvine, 2022

Professor Tryphon T. Georgiou, Chair

The Perron–Frobenius and Koopman operators provide natural dual settings to investigate the dynamics of complex systems. In this thesis we focus on certain pertinent concepts and strategies for obtaining dynamical models and approximating the transfer operators from data.

First, we explain the setting and the assumptions that underlie the so-called Dynamic Mode Decomposition (DMD); this methodology relates to Koopman-operator approximation through full-state observables. The goal is to highlight caveats as well as to suggest metrics that indicate that the use of DMD on specific dataset is warranted. In many applications it is often the case that only a limited number of data samples is available for modeling an otherwise exceedingly high dimensional process. The dimensionality of the process, which may represent visual or distributional fields, in conjunction with the limited observation record requires careful analysis. It is precisely this regime of "small data," i.e., "few samples," that has been a challenge in traditional signal analysis since its inception. DMD is a recent development that aims to identify suitable linear dynamics that can explain the data. We show how the concept of the gap metric can be used as a tool to quantify how subspaces spanned by data impact modeling assumptions. Also, the gap metric provides guidance in selecting appropriate dimensionality for models for such processes.

Next, we formulate and solve a regression problem with time-stamped distributional data. Distributions are considered as points in the Wasserstein space of probability measures, metrized by the 2-Wasserstein metric, and may represent images, power spectra, point clouds of particles, and so on. The data sets may be thought to represent densities of particles whose precise trajectories are not be available (e.g., partially observed). The regression seeks a curve in the Wasserstein space that passes closest to the dataset. Our regression problem allows utilizing general curves in a Euclidean setting (linear, quadratic, sinusoidal, and so on), lifted to corresponding measure-valued curves in the Wasserstein space. It represents a relaxation of geodesic regression in Wasserstein space. The apparently nonlinear primal problem can be recast as a multi-marginal optimal transport, leading to a formulation as a linear program. Entropic regularization and a generalized Sinkhorn algorithm can be effectively employed to solve this multi-marginal problem.

The proposed framework can be used to estimate correlation between given distributional snapshots. Potential applications of the theory are envisioned to aggregate data inference, estimating meta-population dynamics, power spectra tracking, and more generally, system identification.

Finally, we introduce a regression-type formulation for approximating the Perron-Frobenius operator by relying on distributional snapshots of data. The Wasserstein metric is leveraged to define a suitable functional optimization in the space of distributions, weighing in distances between successive distributional snapshots. The formulation allows seeking suitable dynamics so as to interpolate the distributional flow in function space. A first-order necessary condition for optimality is derived and utilized to construct a gradient flow approximating algorithm. It should be noted that we assume no information on statistical dependence between successive pairs of distributions. The method extends to search for nonlinear dynamics assuming a suitable parametrization of the nonlinear state transition map in terms of selected basis functions.

Chapter 1

Introduction

1.1 Background and Motivation

Whereas the topic of "big data" dominates current headlines in research publications and popular news analyses alike, the perennial challenge of obtaining reliable models with only limited observation records persists in a wide range of time series applications. Indeed, one often hears the admission from practitioners that the problem is not "big data" but "small data." A case in point is that of time series of flow fields where a exceedingly high-dimension state is observed, or partially observed, albeit over a relatively short time window. It is precisely for these types of applications that Dynamic Mode Decomposition (DMD) and related frameworks were conceived to address [4–6].

DMD, as introduced by Schmid [7], is a formalism to identify dominant modes in a high-dimensional time series $x_t \in \mathbb{R}^N$, $t \in \{1, 2, \ldots, L\}$, where the dimensionality N of the time series is much larger than the number L of available observations. In its original formulation, DMD takes x_t as a convenient state of an underlying process and thereby dispenses of higher order dynamics that may be hidden in differences between the time series data. The more

general situation of higher order dynamics can be treated similarly [8]. The main issue that we discuss in this thesis is the pertinence of the assumption in seeking such a state model, and whether a reliable estimate of state dynamics should be expected to reflect the structure of the data. Dynamic mode decomposition relates to the approximation of Koopman operator through full-state observables. The Perron–Frobenius and Koopman operators provide natural dual settings to investigate the dynamics of complex systems. The focus of this thesis is on certain pertinent concepts and strategies for obtaining dynamical models and approximating the transfer operators from data. There are many realistic scenarios in which the available data are time-stamped distributions which necessitates the extension of system identification strategies such as regression to the space of distributions.

Regression analysis seeks a (non)linear correspondence between two sets of variables by minimizing a suitable function of residuals. In case the data are probability distributions lying on a nonlinear manifold, finding an appropriate metric, defining and interpreting the structure of fitting model, and finding a tractable solution can be challenging. This is pervasive for various applications such as in longitudinal image study [9] where the brain development or tumor growth patterns need to be studied, power spectral tracking [10], traffic control [11] and so on.

The Wasserstein metric is becoming increasingly popular in recent years due to a number of natural and useful properties (e.g., being weakly continuous, allowing efficient computation via entropic regularization) [12–14] which is the rationale behind its extensive use in statistical learning of distributional data. Most studies in the context of distribution learning reckon upon their projections onto the tangent space at some reference point (usually the barycenter of the dataset) [15–17]. This invokes the pseudo-Riemannian structure of Wasserstein space [18] which amounts to statistical learning in a Hilbert space. Another approach is based on generalized geodesics as proposed in [19]. In [10], an approximation method is presented for the measures supported on $\mathbb{R}$ to track the behavior of the power spectral

densities of non-stationary time series. There are some issues with these methods accuracy, reasonable interpretation and computational complexity which might lead to undesirable results (see the introduction in [16]). In the following we touch upon some of the potential applications of regression analysis in the space of distributions.

State tracking of individuals form one population (or sub-population) to another one plays an important role in many areas, such as target tracking or (meta)population dynamics (e.g. see [20] and [21]). In many practical scenarios the trajectories of individuals may not be accessible due to different reasons, e.g. the population is huge or tracking the individuals requires a prohibitive number of sensory equipment. For instance, in animal ecology, one is often interested in movement of a group of animals based on studies of unmarked individuals [20]. In other words, the identity of individual animals are unavailable, and any analysis should be based on the aggregate data. The regression-type methods in the space of distributions can be employed to track the evolution of an ensemble of indistinguishable individuals (mass particles, agents or so on). Namely, we can estimate the flow of individuals (or mass particles) for which the one-time marginals resemble the distributional snapshots in the sense of Wasserstein distance. In [20], optimal transport theory is utilized to estimate the transition probabilities associated with indistinguishable populations moving among multiple spatial locations at two points in time. Some other studies deal with steering the states of a linear or non-linear dynamical system from an initial density to a target one. This can be stated in both cases of interacting or non-interacting particles [22, 23].

There are other studies conducted in order to account for the cases where more than two (initail and final) one-time marginals are available. For instance, in [24] hidden Markov models (HMMs) are used to describe the particle flows and aggregate observations where the most likely paths that the agents have taken are sought. At the very high level, unlike the problem of interest in this thesis which is regression-type, those studied in [21,24,25], are interpolation-type problems where a flow is sought such that the one-time marginals meet

a sequence of probability densities precisely. Despite being built on interesting ideas, this might be problematic in cases of noisy data or a large number of snapshots where overfitting may be inevitable.

This can also be useful to identify the coherent sets in the flow. These are non-dispersing or minimally-dispersing regions in the flow's domain, namely, the particles in these regions are less likely to leave them [26]. The study in [27] deals with the case where the initial and final densities of particles are the only available data, i.e., the underlying flow map is not accessible. The unbalanced regularized optimal transport plans are used to find coherent sets in evolving particle ensembles.

It is often the case that dynamics are to be inferred by the collective response of dynamical systems (particles, agents, and so on) recorded as distributional snapshots of observables [28]. Regardless of whether the underlying dynamics is linear or not, provided there is no interaction between particles, the distributional data on observables evolve under the action of a linear operator. The two broadly-studied alternatives for this purpose are the Perron-Frobenius and the Koopman operators, both known as transfer operators. They are indeed linear, but defined on infinite-dimensional spaces of distributions and of observable (functions), respectively, and are adjoint to one another [29].

Modeling and approximation of transfer operators often relies on samples of along collections of trajectories, e.g., see [4, 30–32]. This, in fluid mechanical systems, can be effected via recording the motion tracers seeded in the flow; such tracers provide pointwise correspondence among particles at different snapshots. However, perhaps equally often, in many real-world situations, complete trajectories may not available. Labeling and tracking particles individually is simply not feasible. In such cases, distributions of ensembles at different time instances is the only accessible data. This may also be the case in applications, as in modeling flow/traffic, when density, average speed, and other parameters quantifying congestion are being recorded and available, and not the path of individual drivers.

Besides applicactions related to flows of particles and collections of dynamical systems, the problems we consider are relevant in image registration, tumor growth monitoring, and system identification from visual data [33]. Another instance is domain adaptation, which aims at finding a model on a *target data* distribution, by training on a *source data* distribution [34, 35].

As mentioned before, a popular and effective method in identifying dynamics using snapshots of data is DMD which relates to Koopman-operator approximation through full-state observables. The Liouville operator [36] is another example of a linear operator associated with non-linear dynamics; this is the infinitesimal generator for the Koopman operator [37]. In this context, we also mention the concept of occupation kernels which allows for the embedding of a dynamical system into a *Reproducing Kernel Hilbert Space* (RKHS). For further studies and taxonomy of the substantial and rapidly expanding literature we refer to [38].

A well-known method for the approximation of Perron-Frobenius operator is Ulam's method, in which the evolution of a set of test points within the discretized state-space under the action of dynamics leads to a probability matrix in the discretized state-space [39, 40]. There are other methods to approximate Perron-Frobenius operator, most of which rely on Petrov-Galerkin projections of infinite-dimensional operators onto some finite-dimensional subspace (see for example [29, 41, 42]). Also, one can utilize one of the aforementioned techniques to approximate the Koopman operator and use the duality property to find an approximate representation for the Perron-Frobenius operator [28]. These approaches hypothesize the existence of pointwise correspondence among the distributions at different snapshots as the data are collected along one or several trajectories of the dynamics.

The long-term behavior of a dynamical system can be characterized by its associated invariant measure supported on an invariant set. Indeed, this is the fixed point of Perron-Frobenius operator which pushes forward distributions under the action of dynamics [43]. The invariant sets, for example, can represent equilibrium points, periodic and quasi-periodic orbits

sitting on some lower-dimensional manifolds [44]. There are a plethora of numerical methods to compute invariant measure and sets, most of which conducted for known dynamics, or where the pointwise correspondence between the successive points in time is available (See [45, 46] and references therein).

1.2 Objectives and Contributions

In this thesis we focus on certain pertinent concepts and strategies for obtaining dynamical models and approximating the transfer operators from data. First we explain the setting and the assumptions that underlie the so-called Dynamic Mode Decomposition (DMD); this methodology relates to Koopman-operator approximation through full-state observables. Next we will discuss a regression-type formulation for the approximation of Perron-Frobenius Operator using distributional snapshots of data. The data sets may be thought to represent densities of particles whose precise trajectories are not be available (e.g., partially observed). The Wasserstein metric is leveraged to define a suitable functional optimization in the space of distributions. The formulation allows for seeking suitable dynamics so as to interpolate the distributional flow in function space. The objectives and contributions of each chapter are as follows.

- Chapter 2: The goal is to highlight caveats as well as to suggest metrics that indicate that the use of DMD on specific dataset is warranted. We show how the concept of the gap metric can be used as a tool to quantify how subspaces spanned by data impact modeling assumptions. Also, the gap metric provides guidance in selecting appropriate dimensionality for models for such processes.

- Chapter 3: We provide background on the theory of optimal mass transport (OMT) that underlies the developments in the body of this thesis.

- Chapter 4: We present a regression model for the probability distributions indexed by

timestamps which enjoys a reasonable geometric interpretation and solution. To do so, we define a probability measure on the space of curves with a specified complexity (linear, quadratic or so on). A probability measure over this space is sought such that its one-time marginals replicate the distributional snapshots at each timestamp. This probability measure represents a flux, i.e., how much mass is flowing along each path. This approach represents a least-squares minimization in Wasserstein space. Then, using multi-marginal optimal transportation [47], this problem is recast as a linear programming. Generalized Sinkhorn's algorithm can be employed to solve efficiently the entropy-regularized version of this problem.

Furthermore, to study metapopulation dynamics which consist of a group of local populations, we will describe how our approach can be carried over to mixture distributions especially Gaussian mixture models. The space of Gaussian mixtures can be equipped with a Wasserstein-type distance [48,49] which we employ to lift our regression strategy to this submanifold of probability distributions.

Also, We will delineate how to use the method of this paper to estimate the Perron-Frobenius operators and invariant measures associated with dynamical systems. We estimate the Perron-Frobenius operator and its corresponding invariant measure without hypothesizing any information on the dynamics and by relying solely on a few available distributional snapshots.

- Chapter 5: We deal with the problems where dynamics are to be inferred from data on density flows. We advance a viewpoint that leverages the geometry of optimal mass transport and the Wasserstein metric on distributions, to identify underlying dynamics. Data are assumed to be probability distributions over a suitable state-space, and that any statistical dependence between pairs of distributions is not available. These observations (one-time marginal distributions) are the successive projections of the flow generated by the underlying dynamics. We seek a suitable approximation of

Perron-Frobenius operator and, thereby, an embedding of the dynamics into a function space based on these distributional snapshots. The Wasserstein metric is employed to define an appropriate cost, by minimization of which, a desirable embedding can be achieved. This notion of distance, which represents cost of transport, compares two probability distributions based on the ground metric of the underlying state-space.

Chapter 2

The Challenge of Small Data: Dynamic Mode Decomposition, Redux

2.1 The basic DMD rationale

Consider the basic linear dynamical model,

$$x_{t+1} = Ax_t + v_t, \text{ for } 1 \leq t \leq L-1, \tag{2.1}$$

where $A \in \mathbb{R}^{N \times N}$, while $v_t \in \mathbb{R}^N$ signifies deviation from linear deterministic dynamics (via the input term v_t that may represent stochastic excitation or contribution of nonlinear terms). The standard formulation of DMD is based on the assumption that the time series under consideration, herein x_t, is dominated by the linear transition mechanism and that, moreover, the dimension of x_t is much larger than the size of the observation window $t \in \{1, 2, \ldots, L\}$.

The underlying premise of the DMD methodology is that the state vector x_t concentrates

along the directions that correspond to the dominant eigendirections of A and, thereby, DMD aims (and has a viable chance) to identify the dynamics that are manifested by restricting the recurrence relation in (2.1) onto the range of a data matrix

$$X_{1:n-1} := [x_1,\, x_2,\, \ldots, x_{n-1}]\,,$$

for n possibly $n \leq L$. Thereby, the dynamics are sought in a matrix $A \in \mathbb{R}^{N\times N}$ to satisfy

$$X_{2:n} \simeq AX_{1:n-1}. \tag{2.2}$$

One readily observes that the operator A, restricted onto the orthogonal complement of $\text{range}(X_{1:n-1})$, namely,

$$A|_{\text{range}(X_{1:n-1})^\perp},$$

is undefined, i.e., it cannot be determined from the data. DMD sets out to determine the action of A precisely on the range of $X_{1:n-1}$. To this end, complete the columns of $X_{1:n-1}$ into a basis $\mathcal{B} := \{x_1, \ldots, x_{n-1}, y_n, \ldots, y_N\}$ for $\mathbb{R}^N$. We tacitly assume that $\{x_1, \ldots, x_{n-1}\}$ are linearly independent. The matrix $Y_{n:N} = [y_n, \ldots, y_N]$ formed out of the added (column) vectors is such that

$$T = [X_{1:n-1}, Y_{n:N}]$$

is an invertible matrix. Selection of $Y_{n:N}$ can be accomplished by taking the singular value decomposition

$$X_{1:n-1} = U\Sigma V^T,$$

of $X_{1:n-1}$, where $U \in O(N)$, $V \in O(n-1)$, and

$$\Sigma = \begin{bmatrix} \sigma_1(X_{1:n-1}) & 0 & 0 & \dots \\ 0 & \sigma_2(X_{1:n-1}) & 0 & \dots \\ 0 & 0 & \ddots & \\ \vdots & \vdots & & \end{bmatrix}$$

is the $N \times (n-1)$ matrix with the (non-increasing sequence of) singular values of $X_{1:n-1}$ on the main diagonal, and where $O(k)$ denotes the group of $k \times k$ orthogonal matrices. Then, if after partitioning

$$U = [U_{1:n-1}, U_{n:N}],$$

the selection $Y_{n:N} = U_{n:N}$ presents a convenient option.

Similarity transformation with T bring A into the form

$$\begin{bmatrix} S & S_{12} \\ S_{21} & S_{22} \end{bmatrix},$$

since

$$A\,[X_{1:n-1}, Y_{n:N}] = [X_{1:n-1}, Y_{n:N}] \begin{bmatrix} S & S_{12} \\ S_{21} & S_{22} \end{bmatrix}.$$

Thus,

$$AX_{1:n-1} = X_{1:n-1}S + Y_{n:N}S_{21}.$$

Assuming that A leaves range$(X_{1:n-1})$ invariant, the intertwining relation

$$AX_{1:n-1} = X_{1:n-1}S,$$

holds and S represents the restriction of A onto the range$(X_{1:n-1})$. Thus, assuming that (2.2) holds with equality,

$$X_{2:n} = X_{1:n-1}S, \tag{2.3}$$

captures the action of A on the range of $X_{1:n-1}$ and can be used to determine S. Finally, because the columns of $X_{2:n-1}$ are shared with a shift between $X_{1:n-1}$ and $X_{2:n}$, S has the companion structure

$$S = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & -s_{n-1} \\ 1 & 0 & 0 & \dots & 0 & -s_{n-2} \\ 0 & 1 & 0 & \dots & 0 & -s_{n-3} \\ \vdots & & \ddots & & & \vdots \\ & & & & & \\ 0 & 0 & 0 & \dots & 1 & -s_1 \end{bmatrix},$$

where the last column can be easily identified by solving (2.3).

Since in general the linear transformation A does not leave range$(X_{1:n-1})$ entirely invariant, and thereby (2.2) does not hold with equality, suitable approximation is carried out to obtain S. For instance, the vector $s = (s_{n-1}, \dots, s_1)^T$ can be obtained as

$$\text{argmin}\{\|x_n - X_{1:n-1}s\| \mid s \in \mathbb{R}^{n-1}\}, \tag{2.4}$$

with $\|\cdot\|$ denoting (typically, and herein) the Euclidean norm, and to this end several

alternative numerical schemes have been proposed (such as Arnoldi and SVD based) [4, 7]. This is the typical scenario for DMD applications.

2.1.1 Regularizations

An alternative approach is to regularize the problem by penalizing perturbation from the recorded values in data matrix $X_{1:n-1}$ as well, e.g., by solving instead (the nonlinear problem)

$$\text{argmin}\{\|x_n - \hat{X}_{1:n-1}s\| + \epsilon\|\hat{X}_{1:n-1} - X_{1:n-1}\|\},$$

over $s \in \mathbb{R}^{n-1}$ and $\hat{X}_{1:n-1} \in \mathbb{R}^{N\times(n-1)}$, for a choice of regularizing parameter $\epsilon > 0$. This option is especially reasonable in case (2.2) fails to hold with equality due to stochastic noise or the (small) effect of nonlinear dynamics, or in cases where prior information dictates specific structural features, e.g., see [50, 51].

2.1.2 Higher order dynamics

We note that in cases when higher order dynamics are at play and $\mathbb{R}^N$ is insufficient as a choice of state-space, an option is to account for lagged values of x_t and thereby select as a candidate state vector, e.g., for the case of one lag,

$$\xi_t = [x_t^T,\ x_{t-1}^T]^T.$$

Very little changes in the basic setting [8]. In this case, one seeks an A matrix of twice the size to now satisfy $\Xi_{3:n} \simeq A\Xi_{2:n-1}$, cf. (2.2). Thence, a matrix S as before, with companion structure, such that

$$\Xi_{3:n} = \Xi_{2:n-1}S,$$

with $\Xi_{k:\ell} := [\xi_k, \xi_{k+1}, \ldots, \xi_\ell]$, assuming $k < \ell$, cf. (2.3). Thus, without loss of generality we will only discuss the basic setting without further expanding neither on higher order dynamics nor on the relevance of various choices for regularization.

2.1.3 Recap & concluding thoughts

The goal of DMD is to identify dominant modes that capture the relation between successive vectors of the time series. These are the roots of the polynomial

$$\mathbf{s}(\lambda) = \lambda^{n-1} + s_1\lambda^{n-2} + \ldots + s_{n-1}.$$

An underlying premise of the framework is that the time series does not depart significantly from being quasi-stationary. This can only hold if the observed dynamics result from a "tug-of-war" mechanism that provides excitation and saturation at the same time (*a la* fluctuation-dissipation). Such a dynamical mechanism can be based in either or both, a stochastic excitation or nonlinear contributions, as in (2.1), where v_t may represent either. This understanding suggests that the effectiveness of DMD and relevance of the underlying dynamical structure may be quantified by the geometric relation between subspaces spanned by successive collections of time series samples x_t. From a more practical perspective, the effectiveness of DMD, by necessity, rests on how close the subspaces spanned by $X_{2:n}$ and $X_{1:n-1}$ are.

To this end, below, we explore the use of geometric concepts that quantify how well the above expectations are reflected in the data. Specifically, we introduce the analogue of partial autocorrelation coefficients that can serve to identify the size of the state-space that can usefully be exploited to identify dominant dynamics.

2.2 Innovation parameters (IP's)

The effectiveness of DMD in modeling the underlying dynamics rests on the relation between the subspaces spanned by $\{x_\ell, x_{\ell+1}, \ldots, x_m\}$, over a progression of intervals $[\ell, m]$ of indices and over varying window sizes.

Consider first intervals $[1, k-1]$ and $[2, k]$. We seek to quantify the new information that is contained in the last vector x_k as compared to the previous ones. Specifically, we consider how introducing these new data point x_k impacts the distance (angle) between the subspaces spanned by $X_{1:k-1}$ and $X_{2:k}$. Evidently, the angle between these subspaces relates to the discrepancy in (2.2) from holding with identity.

We will similarly consider relations between subspaces corresponding to adjacent windows $[\ell, \ell+k-1]$ and $[\ell+1, \ell+k]$, and how angles between such subspaces change with the indices ℓ and k.

2.2.1 Angles and the gap metric

The distance between subspaces $\mathcal{X}_1, \mathcal{X}_2 \subseteq \mathcal{X}$, of a Hilbert space $\mathcal{X}$, is naturally quantified by the angle operator

$$R_{12} := \Pi_{\mathcal{X}_1}|_{\mathcal{X}_2^\perp},$$

where $\Pi_{\mathcal{X}_1}$ denotes orthogonal projection onto $\mathcal{X}_1$ and $|_{\mathcal{X}_2^\perp}$ the restriction onto the orthogonal complement of $\mathcal{X}_2$. Herein, we will be concerned with finite dimensional Euclidean spaces. In this case, provided the subspaces have equal dimension,

$$\|\Pi_{\mathcal{X}_1}|_{\mathcal{X}_2^\perp}\| = \|\Pi_{\mathcal{X}_2}|_{\mathcal{X}_1^\perp}\|.$$

This common value is equal to $\|\Pi_{\mathcal{X}_1} - \Pi_{\mathcal{X}_2}\|$ and defines a *bona fide* metric between subspaces [52, 53]. This is referred to as the *gap metric*

$$d(\mathcal{X}_1, \mathcal{X}_2) := \|\Pi_{\mathcal{X}_1} - \Pi_{\mathcal{X}_2}\|.$$

Thence,

$$\theta(\mathcal{X}_1, \mathcal{X}_2) := \arcsin(d(\mathcal{X}_1, \mathcal{X}_2))$$

represents an angular distance between the two subspaces. In case their dimensions do not match, the gap is the maximal norm of the two angle operators, and equals $d(\mathcal{X}_1, \mathcal{X}_2) = 1$, giving $\theta(\mathcal{X}_1, \mathcal{X}_2) = \frac{\pi}{2}$.

We remark that the gap metric between the graphs (infinite dimensional subspaces) of dynamical systems is a natural metric to quantify uncertainty in the context of feedback theory, and as such has been a chapter in modern robust control [54–56]. Herein we are only concerned with the geometry of finite dimensional subspaces spanned by the vectorial entries of a time series.

2.2.2 Innovation parameters and PARCOR's

In order to assess the consistency of successive measurements of the time series we consider gaps between subspaces spanned by successive segments, e.g., range($X_{1:k}$) and range($X_{2:k+1}$) for different values of k. We refer to these as *innovation parameters (IP)*

$$\boxed{r_k := d(\text{range}(X_{1:k}), \text{range}(X_{2:k+1}))}$$

In geometric terms, r_k is the sine of the angular distance

$$\theta_k := \arcsin(r_k)$$

between $\Pi_{\text{span}(x_2,\ldots,x_k)^\perp} x_1$ and $\Pi_{\text{span}(x_2,\ldots,x_k)^\perp} x_{k+1}$, i.e., between the projections of x_1, x_{k+1} onto the orthogonal complement of the span of the intermediate vectors $\{x_1, \ldots, x_k\}$. Similarly, we define

$$\boxed{r_{\ell,k} := d(\text{range}(X_{\ell:\ell+k-1}), \text{range}(X_{\ell+1:\ell+k}))}$$

to capture the same dependence between successive subspaces from a different starting point ℓ.

The innovation parameters relate to the partial correlation coefficients (PARCOR) in time-series analysis [57]. Specifically, if $\mathbf{X}_k$, for $k \in \mathbb{Z}$, denotes a stationary time series, the PARCOR's are the cosines of the angles between

$$\mathbf{X}_\ell - \mathbb{E}\{\mathbf{X}_\ell | \mathbf{X}_{\ell+1}, \ldots, \mathbf{X}_{\ell+k-1}\} \text{ and}$$
$$\mathbf{X}_{\ell+k} - \mathbb{E}\{\mathbf{X}_{\ell+k} | \mathbf{X}_{\ell+1}, \ldots, \mathbf{X}_{\ell+k-1}\},$$

where in the conditioning, for $k = 1$, we define the set $\{\mathbf{X}_{\ell+1}, \ldots, \mathbf{X}_{\ell+k-1}\}$ as empty. Thus, these also coincide with the cosines of the angles between the spans of the random variables $\{\mathbf{X}_\ell, \ldots, \mathbf{X}_{\ell+k-1}\}$ and $\{\mathbf{X}_{\ell+1}, \ldots, \mathbf{X}_{\ell+k}\}$.

Besides one set of parameters corresponding to sines and the other to cosines, the main difference between IP's and PARCORs is that the latter are typically defined for stationary stochastic processes, in that the kernel

$$\mathcal{K}(i,j) := \langle x_i, x_j \rangle$$

has a Toeplitz structure [57], unlike the case of IP's which do not have necessarily a Toeplitz structure, as the geometric relations in the data sequence $x_1, x_2, \ldots$ are not shift-invariant in general, which often necessitates exploring the double indexing in $r_{\ell,k}$.

2.2.3 Recursive computation of innovation parameters

Efficient code for computing the innovation parameters for large data sets and size of vectors can be devised based on a recursive scheme that orthonormalizes successive vectors in the data base.

Specifically, consider a basis for the span of $X_{1:k-1}$ to consist of x_1 and the orthonormal columns of a matrix $U_{2:k-1}$. Likewise, the span of $X_{2:k}$ consist of x_k and the orthonormal columns of a matrix $U_{2:k-1}$. Define the orthogonal projection onto the orthogonal complement of the range of $U_{2:k-1}$

$$\Pi_{\text{range}(U_{2:k-1})^\perp} = I - U_{2:k-1}U_{2:k-1}^T.$$

Then the angle between the span of $X_{1:k-1}$ and that of $X_{2:k}$ coincides with the angle between

$$(I - \Pi_{\text{range}(U_{2:k-1})^\perp})x_1 \text{ and } (I - \Pi_{\text{range}(U_{2:k-1})^\perp})x_k.$$

The computation of the innovation parameters can be carried our recursively as follows:

Algorithm 1: Recursive computation of IP's

Data: Given $X_{1:n} \in \mathbb{R}^{N\times n}$
Initialization: $k = 1$, $u_1 = x_1/\|x_1\|$, $u_2 = x_2/\|x_2\|$,
$u_{\text{first}} = u_1 - \langle u_1, u_2\rangle u_2$, $u_{\text{last}} = u_2$, $U = [u_2]$;
while $k < n - 1$ **do**
  $u_{\text{last}} = x_{k+2}$;
  $u_{\text{last}} = u_{\text{last}} - UU'u_{\text{last}}$;
  $u_{\text{last}} = u_{\text{last}}/\|u_{\text{last}}\|$;
  $r_k = \sin(\text{acos}(\langle u_{\text{first}}, u_{\text{last}}\rangle))$;
  $U = [U \;\; u_{\text{last}}]$;
  $u_{\text{first}} = u_{\text{first}} - \langle u_{\text{first}}, u_{\text{last}}\rangle u_{\text{last}}$;
  $u_{\text{first}} = u_{\text{first}}/\|u_{\text{first}}\|$;
  $k = k + 1$;
end

Alternatively, the same computation can be carried out in Matlab utilizing the "econ" feature that optimizes computations for large data sets. E.g., in order to compute r_n set $Y_1 = X_{1:n-1}$ and $Y_2 = X_{2:n}$, and compute U_i for $i \in \{1, 2\}$ with the command $[U_i, \Sigma_i, V_i] = \text{svd}(Y_i, '\,\text{econ}')$. Since,

$$\Pi_{\text{range}(Y_i)} = U_i U_i',$$

with U_i an isometry, the gap between the two subspaces is

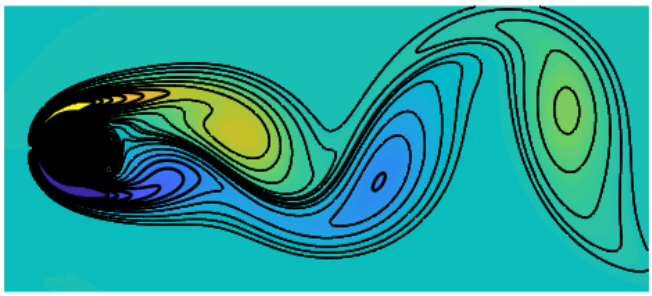
$$\begin{aligned}
\|U_1 U_1'(I - U_2 U_2')\|^2 &= \|U_1' - \underbrace{(U_1' U_2)}_{M} U_2'\|^2 \\
&= \|(U_1' - M U_2')(U_1 - U_2 M')\| \\
&= \|I - MM' - MM' + MM'\| \\
&= \|I - MM'\|
\end{aligned}$$

Therefore, the gap between range(Y_1) and range(Y_2) is

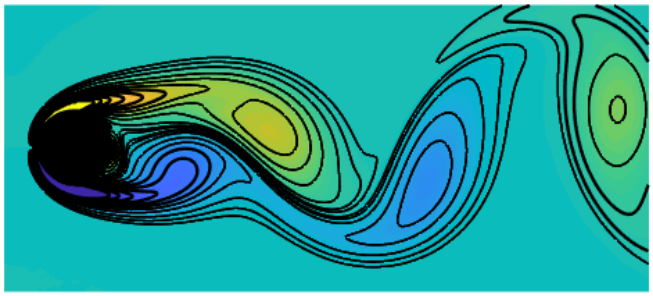
$$\sqrt{1 - \sigma_{\min}(M)^2}$$

with $M = U_1' U_2$ and $\sigma_{\min}(M)$ denotes the smallest singular value of M.

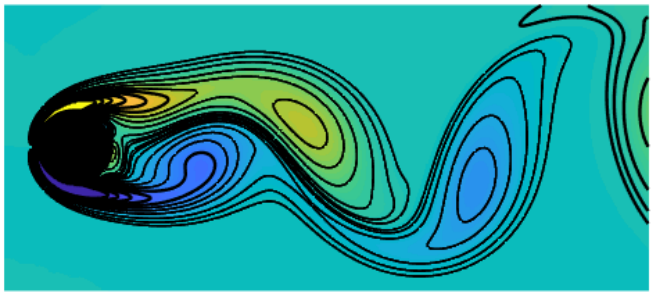
We proceed to motivate and explain the use and relevance of the IP's in selecting a suitable size n for the dynamics sought via DMD on a case study. An additional technical result will be presented along with the example, which highlights the fact that under- or over-estimating the value for n leads to significant errors in identifying the correct dynamics. The example we consider is that of an almost periodic series.



(a) $t = 1$



(b) $t = 5$



(c) $t = 9$

Figure 2.1: Vorticity field around a cylinder wake.

2.3 A case study

We consider time series data that represent a persistent vorticity of a periodically fluctuating fluid flow field in the wake behind a circular cylinder. This dataset can be generated by publicly accessible code in [58]. The two-dimensional Navier–Stokes equations are numerically solved at Reynolds number 100, to obtain these data. At this Reynolds number the flow undergoes a laminar vortex shedding which can be thought of as a stable limit cycle. The data are collected after simulations converge to steady-state vortex shedding. The reader is referred to [4] for more details on how these data set is extracted. At each of 151 snapshots, the values of vorticity are stacked up in a column of a data matrix X which is of size 89351×151. The images of the vorticity field at successive timestamps $t \in \{1, 5, 9\}$ are depicted in Fig. 2.1. The color-coded velocity fluctuations reveal the mechanism of vortex shedding.

The DMD formalism, and specifically (2.4), is applied to identify the apparent modes of oscillation. The resulting modes are dramatically affected by the choice of n in (2.4). Important points that are highlighted below by this example are as follows:

i) The time series is very close to being periodic. This can been seen in a variety of ways, including standard spectral or Fourier analysis. However, here, we compute the sequence of innovation parameters that quantify how far the subspaces spanned by sliding windows of data, of varying width, are from each other in the gap metric.

   Fig. 2.2 shows $r_k = r_{1,k}$ as a function of k. A dimple that repeats with period 30 indicates periodicity. It turns out that exact periodicity of the r_k's, even when the time series is very close to being periodic is masked by numerical sensitivity that we will comment later on (discussion leading to, and Proposition 2.3.1).

ii) Fig. 2.3 shows the color-coded values of $r_{\ell,k}$ as a function of ℓ vs. k. Specificaly, 50 snapshots are drawn as rows. The ℓth row corresponds to the gap $d(\text{range}(X_{\ell:k}), \text{range}(X_{\ell:k+1}))$,

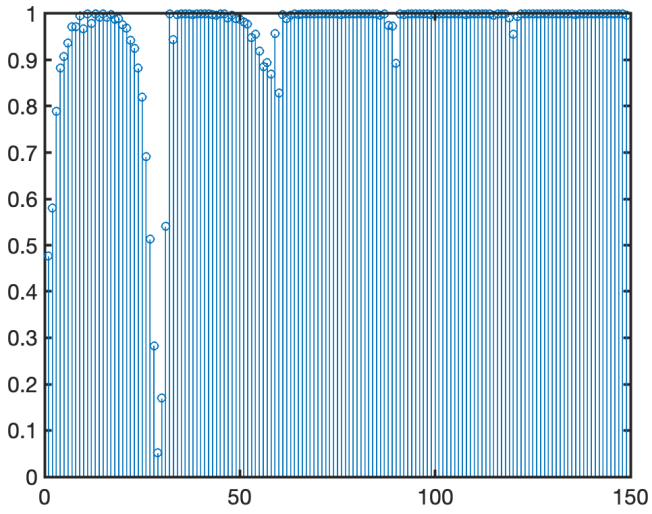


Figure 2.2: $r_k = d(\text{range}(X_{1:k}), \text{range}(X_{2:k+1}))$ vs. k

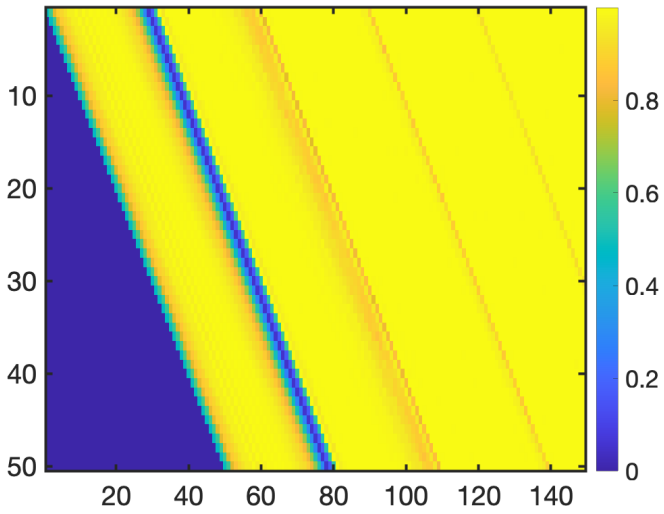
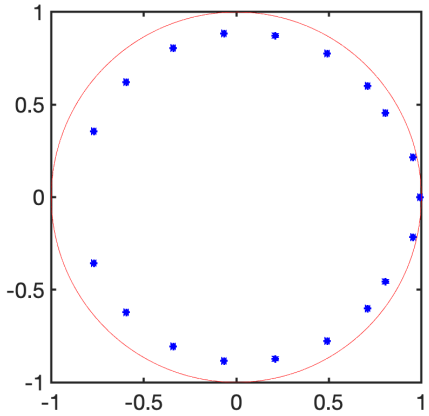
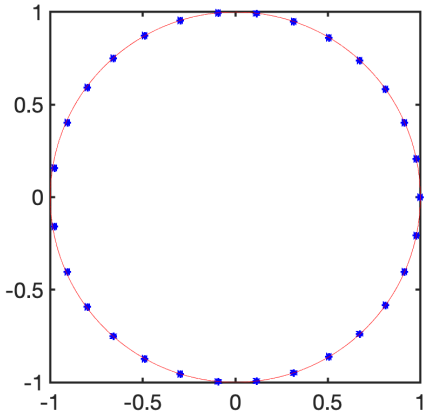


Figure 2.3: $d(\text{range}(X_{\ell:k}), \text{range}(X_{\ell:k+1}))$ color-coded as function of starting time ℓ and window size k

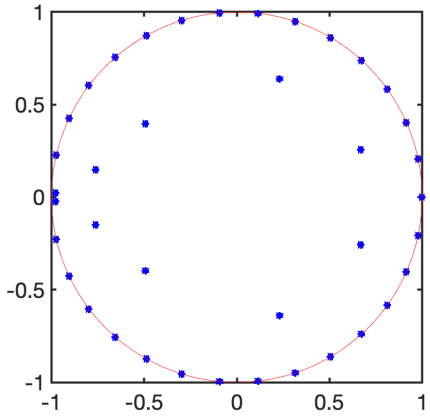
where k sweeps from $\ell + 1$ to the last one. The first row, for instance, corresponds to the values illustrated in Fig. 2.2. One can observe that at each row the minimum gap occurs at the 30th timestamp. This strongly suggests the use of a time-window of size $n = 30$ to find the DMD modes. Periodicity is evident in Fig. 2.3; the decreasing dimples with period 30 are repeated with regularity starting from any chosen starting point ℓ (cf. discussion leading to Proposition 2.3.1).



(a) DMD eigenvalues for $n = 20$



(b) DMD eigenvalues for $n = 30$



(c) DMD eigenvalues for $n = 40$

Figure 2.4: DMD eigenvalues from vorticity field data.

iii) The eigenvalues of S (DMD eigenvalues of the sought dynamics) are shown in Fig. 2.4 for $n \in \{20, 30, 40\}$. It is observed that their distribution is dramatically affected by the chosen size of the subspaces to compare in (2.2), namely n.

iv) For $n = 30$ the eigenvalues of S shown in Fig. 2.4 have modulus $\simeq 1$, in agreement with the observed periodic structure of the flow field. Exact periodicity of the time series results in equispaced eigenvalues, and this is (almost) the case here.

At this point we would like to explain the source of the apparent diminishing of periodic dimples in Fig. 2.2 with period 30. As noted earlier, the gap

$$r_k = d(\text{range}(X_{1:k-1}), \text{range}(X_{2:k}))$$

is the sine of the angle between

$$\xi_1 := \Pi_{\text{span}(x_2,\ldots,x_{k-1})^\perp} x_1, \text{ and}$$
$$\xi_k := \Pi_{\text{span}(x_2,\ldots,x_{k-1})^\perp} x_k.$$

Assuming that the series is k-periodic, the angle between ξ_1 and ξ_k is zero and $x_k \in \text{span}(x_1, \ldots, x_{k-1})$. Likewise,

$$x_{k+1} \in \text{span}(x_2, \ldots, x_k) = \text{span}(x_1, \ldots, x_{k-1}).$$

Denote

$$\xi_{\text{next}} := \Pi_{\text{span}(x_2,\ldots,x_{k-1})^\perp} x_{k+1},$$

and observe that the angle to ξ_k, and therefore ξ_1 too, is zero. Then

$$\begin{aligned} r_{k+1} &= d(\text{range}(X_{1:k}), \text{range}(X_{2:k+1})) \\ &= d(\text{span}(\xi_1, X_{2:k-1}, \xi_k), \text{span}(X_{2:k-1}, \xi_k, \xi_{\text{next}})) \\ &= d(\text{span}(\xi_1, \xi_k), \text{span}(\xi_k, \xi_{\text{next}})) = 0, \end{aligned}$$

with all three vectors $\xi_1, \xi_k, \xi_{\text{next}}$ co-linear. However, a small perturbation in each has a significant effect. Indeed, for arbitrarily small δ's,

$$\begin{aligned} &d(\text{span}(\xi_1 + \delta_1, \xi_k + \delta_k), \text{span}(\xi_k + \delta_k, \xi_{\text{next}} + \delta)) \\ &= d(\Pi_{\text{span}(\xi_k+\delta_k)^\perp}(\xi_1 + \delta_1), \Pi_{\text{span}(\xi_k+\delta)^\perp}(\xi_{\text{next}} + \delta)) \end{aligned}$$

can take any value on $[0, 1]$. We recast the claim as follows.

Proposition 2.3.1. *Consider a vector $\xi \in \mathbb{R}^N$ and perturbations $\xi_i = \xi + \delta_i$, for $i \in \{1, 2\}$, with $\delta_i \perp \xi$. Then*

$$d(\text{span}(\xi + \delta_1, \xi), \text{span}(\xi + \delta_2, \xi) = d(\text{span}(\delta_1), \text{span}(\delta_2)).$$

The proof is elementary. What this statement helps exemplify (and prove) is that in cases where elements that determine the span of interest are almost co-linear, the angles between the subspaces are very sensitive to errors. A more precise mathematical statement can be worked out that involves the conditioning number of the matrix $X_{1:k}$ in our earlier setting.

Chapter 3

Preliminaries on optimal mass transport

We herein provide background on the theory of optimal mass transport (OMT) that underlies the developments in the body of this thesis, and refer to [12, 59, 60] for more detailed exposition.

Let $\mathbb{X} = \mathbb{R}^d$ be equipped with the Borel σ-algebra $\mathcal{B}(\mathbb{X})$. Let μ_0 and μ_1 be two probability measures in $\mathcal{P}_2(\mathbb{X})$, the space of probability measures with finite second moments. We consider the problem to minimize the quadratic cost

$$\int_{\mathbb{X}} \|T(x) - x\|_2^2 \, \mathrm{d}\mu_0(x)$$

over the space of transport maps

$$\begin{aligned} T: \quad & \mathbb{X} \to \mathbb{X} \\ & x \mapsto T(x) \end{aligned}$$

that are measurable and "push forward" μ_0 to μ_1, a property written as $T_{\#}\mu_0 = \mu_1$. This means that for all $A \in \mathcal{B}(\mathbb{X})$, we have $\mu_1(A) = \mu_0(T^{-1}(A))$ or, equivalently, that for all integrable functions $f(x)$ with respect to μ_1,

$$\int_{\mathbb{X}} f(x)\mathrm{d}\mu_1(x) = \int_{\mathbb{X}} f(T(x))\mathrm{d}\mu_0(x). \tag{3.1}$$

If μ_0 is absolutely continuous with respect to the Lebesgue measure, it is known that the optimal transport problem has a unique solution $\hat{T}(x)$ which turns out to be the gradient of a convex function $\phi(x)$, i.e., $\hat{T}(x) = \nabla\phi(x)$.

The problem is nonlinear and, in general, the optimal transport map may not exist. To this end, in 1942, Kantorovich introduced a relaxed formulation in which, instead of a transportation map T, one seeks a joint distribution (referred to as coupling) π on $\mathbb{X} \times \mathbb{X}$, having marginals μ_0 and μ_1 along the two coordinates. The Kantorovich formulation is

$$\inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{\mathbb{X}\times\mathbb{X}} \|x - y\|^2 \mathrm{d}\pi(x, y)$$

where $\Pi(\mu_0, \mu_1)$ is the space of all couplings with the marginals μ_0 and μ_1. In case the optimal transport map exists, the optimal coupling coincides with $\hat{\pi} = (\mathrm{Id} \times \hat{T})_{\#}\mu_0$, where Id denotes the identity map.

The square root of the quadratic transportation cost provides a metric on $\mathcal{P}_2(\mathbb{X})$, known as the Wasserstein 2-metric and denoted by W_2, which makes $\mathcal{P}_2(\mathbb{X})$ a geoesic space and induces a formal Riemannian structure on $\mathcal{P}_2(\mathbb{X})$ as discussed in [12, 18]. Specifically, a constant-speed geodesic between μ_0 and μ_1 is given by

$$\mu_t = \{(1-t)x + t\hat{T}(x)\}_{\#}\mu_0, \quad 0 \le t \le 1 \tag{3.2}$$

and is known as displacement interpolation or a McCann geodesic; i.e., it satisfies

$$W_2(\mu_s, \mu_t) = (t-s)W_2(\mu_0, \mu_1), \quad 0 \leq s < t \leq 1.$$

In the Kantorovich formulation, the geodesic reads

$$\mu_t = \{(1-t)x + ty\}_{\#}\hat{\pi}, \quad 0 \leq t \leq 1. \tag{3.3}$$

We recall the definition of weak convergence of probability measures: A sequence $\{\mu_k\}_{k\in\mathbb{N}} \subset \mathcal{P}_2(\mathbb{X})$ converges weakly to μ, written as $\mu_k \rightharpoonup \mu$, if

$$\lim_{k\to\infty} \int_{\mathbb{X}} f(x)\mathrm{d}\mu_k = \int_{\mathbb{X}} f(x)\mathrm{d}\mu \quad \text{for all } f \in C_b(\mathbb{X}),$$

where $C_b(\mathbb{X})$ is the Banach space of continuous, bounded and real-valued functions on $\mathbb{X}$.

Lemma 3.0.1 (Gluing lemma [12, 18])**.** *Let $\mathbb{X}_1$, $\mathbb{X}_2$, and $\mathbb{X}_3$ be three copies of $\mathbb{X}$. Given three probability measures $\mu_i(x_i) \in \mathcal{P}_2(\mathbb{X}_i)$, $i = 1, 2, 3$ and the couplings $\pi_{12} \in \Pi(\mu_1, \mu_2)$, and $\pi_{13} \in \Pi(\mu_1, \mu_3)$, there exists a probability measure $\pi(x_1, x_2, x_3) \in \mathcal{P}_2(\mathbb{X}_1 \times \mathbb{X}_2 \times \mathbb{X}_3)$ such that $(x_1, x_2)_{\#}\pi = \pi_{12}$ and $(x_1, x_3)_{\#}\pi = \pi_{13}$. Furthermore, the measure π is unique if either π_{12} or π_{13} are induced by a transport map.*

Thus, for any two given couplings, which are consistent along the shared coordinate, the gluing lemma states that we can find a multi-coupling on the product space $(\mathbb{X}_1 \times \mathbb{X}_2 \times \mathbb{X}_3)$ whose projections onto each pair of coordinates match the given couplings. We now briefly discuss three subtopics of interest in sequel.

3.1 Gaussian marginals

In case $\mu_i \sim \mathcal{N}(m_i, C_i)$ for $i \in \{0, 1\}$ are Gaussian with mean μ_i and covariance C_i, respectively, the solution to OMT can be given in closed form [61]

$$W_2(\mu_0, \mu_1) = \sqrt{||m_0 - m_1||^2 + \mathrm{tr}(C_0 + C_1 - 2S)} \tag{3.4}$$

where tr(.) stands for trace and S is an optimal (uniquely defined) cross-covariance term which turns out to be

$$S = (C_0 C_1)^{\frac{1}{2}} = C_0^{1/2}(C_0^{1/2} C_1 C_0^{1/2})^{1/2} C_0^{-1/2}. \tag{3.5}$$

The McCann geodesic μ_t for all $0 \leq t \leq 1$ is a Gaussian distribution with mean $m_t = (1-t)m_0 + tm_1$ and covariance

$$C_t = C_0^{-1/2}((1-t)C_0 + t(C_0^{1/2} C_1 C_0^{1/2})^{1/2})^2 C_0^{-1/2}. \tag{3.6}$$

3.2 Discrete measures

Suppose the marginals are discrete probability measures on a finite set $X \subset \mathbb{R}^d$, that is, $\mu_0 = \sum_{x_0 \in X} p_{x_0} \delta_{x_0}$ and $\mu_1 = \sum_{x_1 \in X} q_{x_1} \delta_{x_1}$, where the non-negative weights p_{x_0} and q_{x_1} are such that $\sum_{x_0 \in X} p_{x_0} = \sum_{x_1 \in X} q_{x_1} = 1$. The transport plan is now in the form of a matrix $(\Pi_{x_0,x_1})_{(x_0,x_1) \in X \times X}$ and its entries represent the amount of mass moved from x_0 to x_1. The

Kantorovich problem in discrete setting can be written as the following linear program:

$$\begin{aligned} \min_{\Pi} \quad & \sum_{x_0,x_1 \in X} c(x_0,x_1)\Pi_{x_0,x_1} \\ \text{s.t.} \quad & \sum_{x_1 \in X} \Pi_{x_0,x_1} = p_{x_0}, \ \forall x_0 \in X \\ & \sum_{x_0 \in X} \Pi_{x_0,x_1} = q_{x_1}, \ \forall x_1 \in X \\ & \Pi_{x_0,x_1} \geq 0, \ \forall x_0, x_1 \in X, \end{aligned} \tag{3.7}$$

where $c(x_0, x_1) = ||x_0 - x_1||_2^2$ is the transportation cost. (Throughout, we assume that transportation costs are quadratic.) Although cast as a linear program, this problem suffers from a heavy computational cost in large scale applications. It was pointed out in [62] that there are computation advantages by introducing an entropy regularization term since, in that case, the problem can then be solved efficiently using the Sinkhorn's algorithm.

3.3 Multi-marginal optimal transportation

In multi-marginal optimal transport, a set of marginals are given and a law is sought that is consistent with the given marginals and minimizes a cost. This problem and its applications are surveyed in [47, 63]. The Kantorovich formulation of this problem for given marginals $\{\mu_i\}_{i=1}^N$ and transportation cost $c(x_1, \cdots, x_N)$ is to minimize

$$\int_{\mathbb{X}^N} c(x_1, \cdots, x_N) d\gamma(x_1, \cdots, x_N) \tag{3.8}$$

where the multi-coupling $\gamma \in \mathcal{P}_2(\mathbb{X}^N)$ is such that $x_{i\#}\gamma = \mu_i$. This is a linear optimization problem over a weakly compact and convex set for which the numerical methods to solve it efficiently are well studied in [63, 64].

Chapter 4

Regression analysis of distributional data

4.1 Regression in Wasserstein space using measure-valued curves

We generalize regression problems, thought of in the setting of a Euclidean space, to the space of probability measures. To this end, for a given set $\{\mu_{t_i}\}_{i=1}^N \subset \mathcal{P}_2(\mathbb{X})$ of probability measures that are indexed by timestamps $\{t_i\}_{i=1}^N \subset [0,1]$, we seek suitable interpolating *measure-valued curves*. Notice that when μ_{t_i} is absolutely continuous with respect to the Lebesgue measure, (by a slight abuse of notation) we use μ_{t_i} to denote both the measure and its density function, depending on the context.

4.1.1 Measure-valued curves

We consider primarily two classes of functions (curves) from the time interval $[0,1]$ to the state space $\mathbb{X}$, linear and quadratic polynomials, denoted by $\mathrm{Lin}([0,1],\mathbb{X})$ and $\mathrm{Quad}([0,1],\mathbb{X})$,

respectively. Generically, we use Ω to denote either class. In the sequel, we consider probability laws on linear, quadratic, and possibly other classes of functions, so as to build corresponding classes of measure-valued curves.

For instance, in the case of $\Omega = \text{Lin}([0,1],\mathbb{X})$, a probability law can be expressed as a coupling between the endpoints of line segments, i.e., a probability law π on $\mathbb{X}^2 := \mathbb{X} \times \mathbb{X}$. This is due to the fact that there is a bijective correspondence $(X_{0,1})$ between each element in Ω and $\mathbb{X}^2$ using the endpoints at $t = 0$ and $t = 1$, i.e., x_0 and $x_1 \in \mathbb{X}$, such that for any $\omega = (\omega_t)_{t\in[0,1]} \in \Omega$, we have $X_{0,1}(\omega) := (x_0, x_1)$. We equip Ω with the canonical σ-algebra generated by the projection maps $(X_t)_{t\in[0,1]}$, defined by $X_t(\omega) := \omega_t$. In this study, we consider only the probability measures with finite second moments over $\mathbb{X}^2$, that is, $\mathcal{P}_2(\mathbb{X}^2)$, and accordingly the induced probability measures over Ω. Given any probability measure π on $\mathbb{X}^2$, the one-time marginals can be obtained through $\nu_t := ((1-t)x_0 + tx_1)_{\#}\pi,\ t \in [0,1]$.

An alternative representation of a probability law on $\Omega = \text{Lin}([0,1],\mathbb{X})$ may be given in terms of a coupling between one endpoint, x_0, and a velocity v. In this representation, the one-time marginals are cast as $\nu_t := (x_0 + tv)_{\#}\pi,\ t \in [0,1]$. In the rest of this chapter, we use the first representation to define probability laws on $\text{Lin}([0,1],\mathbb{X})$.

Similar setting can be defined for $\Omega = \text{Quad}([0,1],\mathbb{X})$ where Ω is, clearly, bijective to $\mathbb{X}^3$. Herein, any probability law on Ω can be expressed as a probability measure over $\mathbb{X}^3$, namely, $\pi \in \mathcal{P}_2(\mathbb{X}^3)$. Also, the one-time marginals can be obtained via $\nu_t := (x_0 + tx_1 + t^2x_2)_{\#}\pi,\ t \in [0,1]$. For ease of notation, we use x_0, x_1, and x_2 to denote the initial point, velocity, and acceleration, respectively. Although one can consider other parameterizations of quadratic curves, e.g. through three points lying on each curve with suitable timestamps, we derive the results for the former representation without loss of generality.

In the next subsection, we detail the regression formalism of minimizing in the Wasserstein sense the distance of distributional data from respective marginals of measure-valued linear

and quadratic curves in $\mathcal{P}_2(\mathbb{X})$, namely,

$$\mathcal{G}_{\text{Lin}} := \{(\nu_t)_{t\in[0,1]} \subset \mathcal{P}_2(\mathbb{X}) \mid \nu_t = ((1-t)x_0 + tx_1)_{\#}\pi, \pi \in \mathcal{P}_2(\mathbb{X}^2)\}, \tag{4.1}$$

and

$$\mathcal{G}_{\text{Quad}} := \{(\nu_t)_{t\in[0,1]} \subset \mathcal{P}_2(\mathbb{X}) \mid \nu_t = (x_0 + tx_1 + t^2x_2)_{\#}\pi, \pi \in \mathcal{P}_2(\mathbb{X}^3)\}. \tag{4.2}$$

We point out that any $(\nu_t)_{t\in[0,1]}$ in $\mathcal{G}_{\text{Lin}}$, or $\mathcal{G}_{\text{Quad}}$, is absolutely continuous [1, Theorem 1], which amounts to the fact that the metric derivative [65]

$$|\nu'|(t) := \lim_{s\to t} \frac{W_2(\nu_s, \nu_t)}{|s-t|} \leq m(t)$$

is bounded by some function $m(t) \in L^1(0,1)$ for almost all $t \in (0,1)$.

4.1.2 Regression problems

Regression analysis seeks to model the relationship between variables, which in our case are probability measures. We consider time as the independent variable and, thereby, regression in the space of probability measures amounts to identifying a flow of one-time marginals which may capture possible underlying dynamics.

Thus, given a set of "points" $\{\mu_{t_i}\}_{i=1}^N \subset \mathcal{P}_2(\mathbb{X})$, we pose the regression problem

$$\inf_{\nu\in\mathcal{G}} \sum_{i=1}^{N} \lambda_i W_2^2(\nu_{t_i}, \mu_{t_i}), \tag{4.3}$$

where $\mathcal{G}$ is either $\mathcal{G}_{\text{Lin}}$ or $\mathcal{G}_{\text{Quad}}$, and the "weights" $\lambda_i > 0$ $(i = 1, \cdots, N)$ satisfy $\sum_{i=1}^N \lambda_i = 1$.

Linear measure-valued curves represent linear-in-time flows which advance an initial probability measure at $t = 0$ to another one at $t = 1$, and generate correlations across the time

interval. Conversely, these linear curves are specified by correlation of their end points, and therefore, problem (4.3) becomes one of minimizing over $\pi \in \mathcal{P}_2(\mathbb{X}^2)$ that represents the coupling between the marginals at $t = 0$ and $t = 1$. Specifically, (4.3) can be cast as

$$\inf_{\pi \in \mathcal{P}_2(\mathbb{X}^2)} F_1(\pi) := \sum_{i=1}^{N} \lambda_i W_2^2(((1-t_i)x_0 + t_i x_1)_{\#}\pi, \mu_{t_i}). \tag{4.4}$$

In (4.4) we assume $N \geq 3$ since, trivially, for $N = 2$ any coupling between the two endpoints results in a zero cost.

Remark 4.1.1. *It is important to contrast* (4.4) *with the geodesic regression problem [10, 15–17] that seeks a geodesic in Wasserstein space to likewise approximate the distributional data* μ_{t_i} *(*$i \in \{1, \ldots, N\}$*). To this end, note that a curve* $\nu_t = ((1-t)x_0 + tx_1)_{\#}\pi$ *is a Wasserstein geodesic when* π *is an* optimal *coupling between two marginals (typically, the end-point ones); the space of such optimal couplings is a strict subset of* $\mathcal{P}_2(\mathbb{X}^2)$*. Thus, the formulation* (4.4) *is a relaxation of the geodesic regression in a way that may be seen as analogous to Kantorovich's relaxation of Monge's problem. Our motivation stems from the computational complexity of geodesic regression rooted in the fact that* $F_1(\pi)$ *is not displacement convex (see [1, Section III]). In contrast, in the next section, we will see that* (4.4) *can be recast as a multi-marginal transport problem and solved efficiently using Sinkhorn's algorithm.*

Analogously, we define the regression problem for measure-valued quadratic curves by minimizing (4.3) over $(\nu_t)_{t\in[0,1]} \in \mathcal{G}_{\text{Quad}}$, leading to

$$\inf_{\pi \in \mathcal{P}_2(\mathbb{X}^3)} F_2(\pi) := \sum_{i=1}^{N} \lambda_i W_2^2((x_0 + tx_1 + t^2 x_2)_{\#}\pi, \mu_{t_i}). \tag{4.5}$$

In this case, the hypothesis class consists of flows which are quadratic in time. It may represent distributions of inertial (mass) particles moving in the space according to quadratic functions in time under the influence of a conservative force field. As mentioned earlier, the regression formalism can be generalized to any other hypothesis classes, e.g. higher-order

curves (cubic, quartic), sinusoids with variable amplitudes and frequencies, and so on. In the present work, however, we restrict our attention to linear and quadratic measure-valued curves.

The existence of minimizers is stated next.

Proposition 4.1.1. *Problems* (4.4) *and* (4.5) *have minimizing solutions.*

Proof. The proof follows from Proposition 2.3 in [66]. For completeness, we detail the steps of the proof for (4.4); the proof of (4.5) follows similarly.

Let $\{\pi_n\}_{n=1}^{\infty}$ be a minimizing sequence of (4.4). Since the data $\mu_{t_i} \in \mathcal{P}_2(\mathbb{X})$ ($i \in \{1, \dots, N\}$), the sequence $\left\{\int_{\mathbb{X}^2} \|x\|_2^2 \mathrm{d}\pi_n\right\}_{n=1}^{\infty}$ remains bounded. This implies that $\{\pi_n\}_{n=1}^{\infty}$ is tight. Therefore, Prokhorov's theorem guarantees the existence of a sub-sequence weakly converging to some $\pi^* \in \mathcal{P}_2(\mathbb{X}^2)$. The lower semi-continuity of Wasserstein distance shows that $F_1(\pi)$ in (4.4) is a lower semi-continuous functional. As a result, $F_1(\pi^*) \leq \liminf_{n\to\infty} F_1(\pi_n) = \inf_{\pi} F_1(\pi)$. This proves that (4.4) has a minimizer. □

Our next proposition states that (4.4) and (4.5) behave well with respect to scaling time. It is stated for problem (4.4) and highlights the fact that changing the units of time does not affect the solution.

Proposition 4.1.2. *Suppose for given* $\{\mu_{t_i}\}_{i=1}^N \subset \mathcal{P}_2(\mathbb{X})$, *with* $\{t_i\}_{i=1}^N \subset [0, T]$, $\hat{\pi}^T \in \mathcal{P}_2(\mathbb{X}^2)$ *is a minimizer of*

$$\inf_{\pi\in\mathcal{P}_2(\mathbb{X}^2)} \sum_{i=1}^{N} \lambda_i W_2^2(((T - t_i)x_0 + t_i x_1)_{\#}\pi, \mu_{t_i}). \tag{4.6}$$

Then, $\hat{\pi}^1 := (Tx_0, Tx_1)_{\#}\hat{\pi}^T$ *is a minimizer of* (4.4) *for* $\{\frac{t_i}{T}\}_{i=1}^N \subset [0, 1]$.

Proof. For each term in (4.6), let $\hat{\eta}_i(x_0, x_1, y) \in \Pi(\hat{\pi}^T, \mu_{t_i})$ be such that $((T-t_i)x_0+t_i x_1, y)_{\#}\hat{\eta}_i$

is an optimal coupling between its marginals. Such $\hat{\eta}_i$ exists due to Proposition 7.3.1 in [65]. Using (3.1), we have

$$\begin{aligned} W_2^2(((T-t_i)x_0 + t_i x_1)_{\#}\hat{\pi}^T, \mu_{t_i}) &= \int_{\mathbb{X}^3} \|(T-t_i)x_0 + t_i x_1 - y\|_2^2 \mathrm{d}\hat{\eta}_i(x_0, x_1, y) \\ &= \int_{\mathbb{X}^3} \|(1-\frac{t_i}{T})(Tx_0) + \frac{t_i}{T}(Tx_1) - y\|_2^2 \mathrm{d}\hat{\eta}_i(x_0, x_1, y) \\ &= \int_{\mathbb{X}^3} \|(1-\frac{t_i}{T})x_0 + \frac{t_i}{T}x_1 - y\|_2^2 \mathrm{d}\left\{(Tx_0, Tx_1, y)_{\#}\hat{\eta}_i\right\}. \end{aligned}$$

It follows that $\hat{\pi}^1 = (Tx_0, Tx_1)_{\#}\hat{\pi}^T$. □

The proposition above shows the regression problems behave nicely with respect to time scaling and thus, without loss of generality, we can always assume the timestamps normalized to lie within the interval $[0, 1]$. Analogous steps can be carried out to show that $\hat{\pi}^1 = (x_0, Tx_1, T^2x_2)_{\#}\hat{\pi}^T$ is a minimizer of (4.5) for $\{\frac{t_i}{T}\}_{i=1}^N \subset [0, 1]$ when $\hat{\pi}^T$ is a minimizer for a corresponding problem with timestamps over a window $[0, T]$ with $T > 1$.

4.2 Multi-marginal formulation

In this section, we show that measure-valued regression can be recast as a multi-marginal optimal transportation problem. Numerically, this is extremely beneficial when combined with entropy regularization as described in the next section. First, we provide the result for measure-valued quadratic curves in the following.

Theorem 4.2.1. *Problem (4.5) can be recast as*

$$\begin{aligned} \inf_{\pi} F_2(\pi) = &\inf_{\gamma} \int_{\mathbb{X}^{N+3}} \sum_{i=1}^{N} \lambda_i \|x_0 + t_i x_1 + t_i^2 x_2 - y_i\|_2^2 \mathrm{d}\gamma(x_0, x_1, x_2, y_1, \cdots, y_N) \\ &\text{s.t.} \quad y_{i\#}\gamma = \mu_{t_i}, \forall i = 1, \cdots, N, \end{aligned} \tag{4.7}$$

with $\gamma \in \mathcal{P}_2(\mathbb{X}^{(N+3)})$, $\pi \in \mathcal{P}_2(\mathbb{X}^3)$. Moreover, a minimizer of the right-hand side ($\hat{\gamma}$) exists

and $\hat{\pi} = (x_0, x_1, x_2)_{\#}\hat{\gamma}$ *is a minimizer of left-hand side.*

Proof. First, suppose $\pi \in \mathcal{P}_2(\mathbb{X}^3)$ and $\mu_t \in \mathcal{P}_2(\mathbb{X})$ are such that $\nu_t = (x_0 + tx_1 + t^2x_2)_{\#}\pi,\ t \in [0,1]$ and $\eta_t \in \Pi(\pi, \mu_t)$, namely, a coupling between π and μ_t. Define

$$W_{\eta_t}(\nu_t, \mu_t) := \int_{\mathbb{X}^4} \|x_0 + tx_1 + t^2x_2 - y\|_2^2 \mathrm{d}\eta_t(x_0, x_1, x_2, y).$$

for which we have $W_2^2(\nu_t, \mu_t) \leq W_{\eta_t}^2(\nu_t, \mu_t),\ \forall t \in [0,1]$. We can show the tightness of this inequality for some $\hat{\eta}_t$, namely, $W_2^2(\nu_t, \mu_t) = W_{\hat{\eta}_t}^2(\nu_t, \mu_t)$. To do so, we assume that $\hat{\Lambda}_t$ is an optimal coupling between ν_t and μ_t. Also, we define $\rho_t \in \mathcal{P}_2(\mathbb{X}^4)$ as

$$\rho_t := (x_0 + tx_1 + t^2x_2, x_1, x_2, y)_{\#}\hat{\eta}_t.$$

The existence of $\hat{\eta}_t$ amounts to finding the probability measure ρ_t which fulfils the following properties:

$$(z_1, z_4)_{\#}\rho_t = \hat{\Lambda}_t \quad \text{and} \quad (z_1, z_2, z_3)_{\#}\rho_t = (x_0 + tx_1 + t^2x_2, x_1, x_2)_{\#}\pi$$

where $(z_1, z_4)_{\#}\rho_t$ denotes the projection of ρ_t onto the product space of first and last coordinates, and $(z_1, z_2, z_3)_{\#}\rho_t$ is its projection onto the product space of the first three coordinates. Since the projections of $\hat{\Lambda}_t$ and $(x_0 + tx_1 + t^2x_2, x_1, x_2)_{\#}\pi$ onto their first coordinates are consistent, i.e., equal to ν_t, by the application of Gluing Lemma (Lemma 5.2.1), we conclude the existence of ρ_t. Moreover, as the map $(x_0 + tx_1 + t^2x_2, x_1, x_2, y)$ is invertible, $\hat{\eta}_t$ exists as well.

Using the disintegration theorem [65, Theorem 5.3.1], we can extend this result to a family

of measures $\{\mu_{t_i}\}_{i=1}^N \subset \mathcal{P}_2(\mathbb{X})$ to show that for given $\pi \in \mathcal{P}_2(\mathbb{X}^3)$,

$$\sum_{i=1}^N \lambda_i W_2^2(\nu_{t_i}, \mu_{t_i}) = \min_{\substack{\gamma \in \mathcal{P}_2(\mathbb{X}^{N+3}) \\ y_{i\#}\gamma = \mu_{t_i} \\ (x_0,x_1,x_2)_\#\gamma = \pi}} \sum_{i=1}^N \lambda_i W_\gamma^2(\nu_{t_i}, \mu_{t_i})$$

A minimizer of problem above ($\hat{\gamma}$) can be constructed as

$$\mathrm{d}\hat{\gamma}(x_0, x_1, , x_2, y_1, \ldots, y_N) = \mathrm{d}(\hat{\eta}_{t_1}^{x_0,x_1,x_2} \times \ldots \times \hat{\eta}_{t_N}^{x_0,x_1,x_2})(y_1, \ldots, y_N)\mathrm{d}\pi(x_0, x_1, x_2). \quad (4.8)$$

In (4.8), the disintegration of each measure $\hat{\eta}_{t_i}$ is written as $\mathrm{d}\hat{\eta}_{t_i}(x_0, x_1, x_2, y_i) = \mathrm{d}\hat{\eta}_{t_i}^{x_0,x_1,x_2}(y_i)\mathrm{d}\pi(x_0, x_1, x_2)$.

According to Proposition 4.1.1, the minimizer $\hat{\pi}$ of left-hand side in (4.7) exists. Thereby, using (4.8), we can obtain a minimizer of the multi-marginal formulation in (4.7) ($\hat{\gamma}$). This proves existence of a solution for our multi-marginal formulation and also, $\hat{\pi} = (x_0, x_1, x_2)_\#\hat{\gamma}$. The proof is complete. □

Similarly, a multi-marginal formulation for (4.4) is provided in the following corollary. The proof is skipped as it resembles that of Theorem 4.2.1.

Corollary 4.2.1. *Problem (4.4) can be recast as*

$$\begin{aligned} \inf_\pi F_1(\pi) = &\inf_\gamma \int_{\mathbb{X}^{N+2}} \sum_{i=1}^N \lambda_i \|(1-t_i)x_0 + t_i x_1 - y_i\|_2^2 \mathrm{d}\gamma(x_0, x_1, y_1, \cdots, y_N) \\ &\text{s.t.} \quad y_{i\#}\gamma = \mu_{t_i}, \forall i = 1, \cdots, N, \end{aligned} \quad (4.9)$$

with $\gamma \in \mathcal{P}_2(\mathbb{X}^{(N+2)})$, $\pi \in \mathcal{P}_2(\mathbb{X}^2)$. Moreover, a minimizer of the right-hand side ($\hat{\gamma}$) exists and $\hat{\pi} = (x_0, x_1)_\#\hat{\gamma}$ where $\hat{\pi}$ is a minimizer of left-hand side.

The following proposition shows the consistency of our method with regression in Euclidean space when the target distributions are Dirac measures.

Proposition 4.2.1. *If all the observations are Dirac measures, i.e.,* $\mu_{t_i} = \delta_{v_i},\ i = 1, \cdots, N$ *where* $\{v_i\}_{i=1}^N \subset \mathbb{X}$*, we have*

$$\inf_{\pi \in \mathcal{P}_2(\mathbb{X}^2)} F_1(\pi) = \inf_{x_0, x_1 \in \mathbb{X}} \sum_{i=1}^N \lambda_i \|(1-t_i)x_0 + t_i x_1 - v_i\|_2^2,$$

$$\inf_{\pi \in \mathcal{P}_2(\mathbb{X}^3)} F_2(\pi) = \inf_{x_0, x_1, x_2 \in \mathbb{X}} \sum_{i=1}^N \lambda_i \|x_0 + t_i x_1 + t_i^2 x_2 - v_i\|_2^2.$$

Proof. See [1, Proposition 7] for the proof. □

In the original formulation of multi-marginal optimal transport, constraints are typically given on all marginals. However, in (4.7) and (4.9), constraints are only imposed on a subset of marginals of multi-coupling γ. We show that these problems can be written in an original formalism of multi-marginal transportation. The following proposition provides this result for linear curves; similar argument holds for quadratic curves.

Proposition 4.2.2. *For every* $y = (y_1, \cdots, y_N) \in \mathbb{X}^N$ *define*

$$(\hat{x}_0(y), \hat{x}_1(y)) = \operatorname*{arg\,min}_{(x_0, x_1) \in \mathbb{X}^2} \sum_{i=1}^N \lambda_i \|(1-t_i)x_0 + t_i x_1 - y_i\|_2^2$$

which is a well-defined map from $\mathbb{X}^N$ *to* $\mathbb{X}^2$ *(since the linear regression in Euclidean space has a unique solution in a closed form). Then, we have*

$$\begin{aligned} \inf_{\pi} F_1(\pi) = \inf_{\gamma' \in \mathcal{P}_2(\mathbb{X}^N)} \int_{\mathbb{X}^N} \sum_{i=1}^N \lambda_i \|(1-t_i)\hat{x}_0(y) + t_i \hat{x}_1(y) - y_i\|_2^2 \mathrm{d}\gamma'(y_1, \cdots, y_N) \\ \text{s.t.} \quad y_{i\#}\gamma' = \mu_{t_i}, \forall i = 1, \cdots, N, \end{aligned} \tag{4.10}$$

and also, $\hat{\pi} = (\hat{x}_0(y), \hat{x}_1(y))_{\#}\hat{\gamma}'$ *where* $\hat{\pi}$ *and* $\hat{\gamma}'$ *are minimizers of the left- and right-hand sides, respectively.*

Proof. The proof is straightforward by noticing that for any $\gamma \in \mathcal{P}_2(\mathbb{X}^{N+2})$ which respects all the constraints on the marginals, we have

$$\begin{aligned}
&\int_{\mathbb{X}^{N+2}} \sum_{i=1}^{N} \lambda_i \|(1-t_i)x_0 + t_i x_1 - y_i\|_2^2 \mathrm{d}\gamma(x_0, x_1, y_1, \cdots, y_N) \\
&\geq \int_{\mathbb{X}^{(N+2)}} \left(\inf_{z_0, z_1 \in \mathbb{X}} \sum_{i=1}^{N} \lambda_i \|(1-t_i)z_0 + t_i z_1 - y_i\|_2^2 \right) \mathrm{d}\gamma(x_0, x_1, y_1, \cdots, y_N) \\
&= \int_{\mathbb{X}^{(N+2)}} \sum_{i=1}^{N} \lambda_i \|(1-t_i)\hat{x}_0(y) + t_i \hat{x}_1(y) - y_i\|_2^2 \mathrm{d}\gamma(x_0, x_1, y_1, \cdots, y_N).
\end{aligned}$$

By taking the infimum of both sides of inequality above over $\gamma \in \mathcal{P}_2(\mathbb{X}^{N+2})$ and using the identity $\gamma' = (y_1, \cdots, y_N)_{\#}\gamma$ (projection onto the last N coordinates), we arrive at the result. □

Remark 4.2.1. *One implication of Proposition 4.2.2 is that in case all the distributional data are supported on a discrete set of points, so does $\hat{\gamma}'$. Specifically, if each μ_{t_i} is supported on a finite set $X_i \in \mathbb{X}$ for all $i = 1, \cdots, N$, then* $\mathrm{supp}(\hat{\gamma}')$*, i.e., support of $\hat{\gamma}'$, lies within $X_1 \times \cdots X_N$ (it is straightforward to show this result, see Proposition 7 in [48]). Also, $\hat{\pi} = (\hat{x}_0(y), \hat{x}_1(y))_{\#}\hat{\gamma}'$ is concentrated on a finite set, that is, the projection of* $\mathrm{supp}(\hat{\gamma}')$ *under the map $(\hat{x}_0(y), \hat{x}_1(y))$. In this case, the problem of measure-valued curves admits a solution in which only a finite number of measure-valued curves in $\mathbb{X}$ have non-zero measures. Therefore, for discrete target measures the problem reduces to a finite-dimensional linear programming as formulated in the following section. Figure 4.1 illustrates this concept for three discrete measures as the distributional data at three instants of time. Two possible trajectories for a mass particle at t_0 are shown with dotted lines. The solid lines represent the best fitting lines for each trajectory (resulted from linear regression in Euclidean space). One can observe that comparing to the lower fitting line, the upper one leads to a smaller value for the sum of squared residuals in $\mathbb{X}$, as it passes closer to its three corresponding points. By solving the multi-marginal problem in* (4.10)*, a smaller probability measure (weight) is*

expected to be assigned to the lower fitting line to penalize its higher value for the sum of squared residuals.

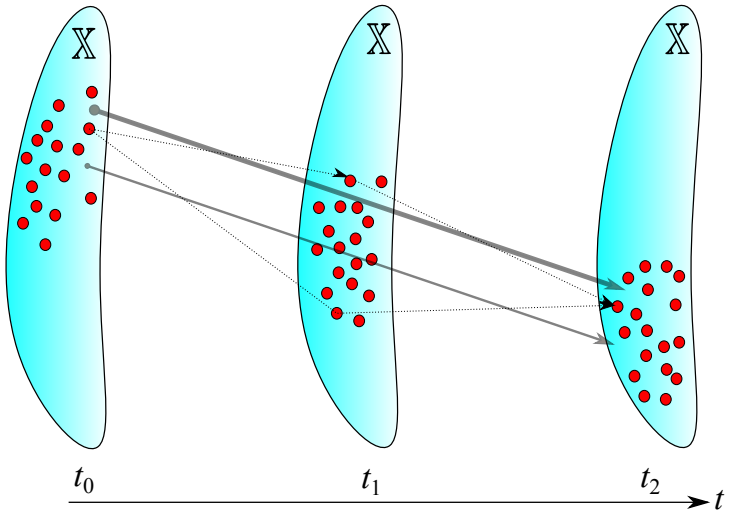


Figure 4.1: Illustration of measure-valued curves for discrete one-time marginals. The dotted lines show two different trajectories for a particle starting from t_0. The solid lines are their corresponding fitting lines resulted from linear regression in $\mathbb{X}$. The sum of squared residuals of the fitting line in the top has a lower value than that of the other one. The solution of multi-marginal problem assigns a higher probability measure (weight) to this fitting line. The thickness of lines is proportional to the likelihood of each line.

4.3 Discretization

In this section, we propose a strategy towards solving the multi-marginal problems introduced in the previous section. First, we express a discretized version of the problem and then invoke the entropy regularization to solve our multi-marginal formulation efficiently. This is beneficial as in many practical situations we only have a set of samples available for each one-time marginal. Thus, we can approximate each distributional data with a sum of Diracs placed at the positions of the available samples.

4.3.1 Discrete multi-marginal formulation

Suppose for a finite set $X \subset \mathbb{X}$, $\mu_{t_i} = \sum_{y \in X} p_y^{t_i} \delta_y$, $i = 1, \cdots, N$ are the given observations, where for each i the non-negatives weights $p_y^{t_i}$ sum up to 1. Without of loss of generality, it is assumed that all μ_{t_i}s are supported on X or a subset of it. We define the multi-marginal problem as seeking a multi-dimensional array $(\Gamma_{x_0,x_1,y_1,\cdots,y_N})_{(x_0,x_1,y_1,\cdots,y_N) \in X^{N+2}}$ with non-negative real elements which solves the following linear programming problem,

$$\begin{aligned} &\min_{\Gamma \geq 0} \sum_{x_0,x_1,y_1,\cdots,y_N \in X} c(x_0, x_1, y_1, \cdots, y_N) \Gamma_{x_0,x_1,y_1,\cdots,y_N} \\ &\text{s.t.} \quad P_{y_j}(\Gamma) = p_{y_j}^{t_j}, \ \forall y_j \in X, \ j = 1, \cdots, N \end{aligned} \tag{4.11}$$

where,

$$c(x_0, x_1, y_1, \cdots, y_N) = \sum_{i=1}^{N} \lambda_i ||(1 - t_i)x_0 + t_i x_1 - y_i||^2$$

is the cost of transport and

$$P_{y_j}(\Gamma) = \sum_{x_0,x_1,y_1,\cdots,y_{j-1},y_{j+1},\cdots,y_N \in X} \Gamma_{x_0,x_1,y_1,\cdots,y_{j-1},y_j,y_{j+1},\cdots,y_N} \tag{4.12}$$

is the projection operator on the marginal of Γ associated with y_j.

Notice that Γ is analogous to the multi-coupling γ in (4.9). Comparing to the definition of π in continuous setting, we have $\Pi_{x_0,x_1} = P_{x_0,x_1}(\Gamma), \ \forall x_0, x_1 \in X$ as the projection of multi-dimensional array Γ onto (x_0, x_1) obtained by summing over all the remaining entries. This leads to a probability measure over the space of linear functions represented by the endpoints in X.

Remark 4.3.1. *Linear programming* (4.11) *is equivalent to* (4.9) *if* X *is chosen rich enough to contain* $\text{supp}(\hat{\pi})$ *as described in Remark 4.2.1. This assumption is not required if, instead*

of (4.9)*, we write the discrete version of* (4.10) *(see Remark 4.2.1). However, as explained in the next subsection, we continue with the discrete formulation in* (4.9) *due to the structure of its transportation cost which entails a lower time and space complexities in order to implement Sinkhorn's algorithm.*

The previous formalism deals with the case of measure-valued lines in discrete setting. A similar formalism for quadratic curves seeks a multi-dimensional array Γ, such that $(\Gamma_{x_0,x_1,x_2,y_1,\cdots,y_N})_{(x_0,x_1,x_2,y_1,\cdots,y_N)\in X^{N+3}}$, with non-negative elements which solves

$$\begin{aligned} &\min_{\Gamma\geq 0} \sum_{x_0,x_1,x_2,\{y_i\}_{i=1}^N\in X} c(x_0,x_1,x_2,y_1,\cdots,y_N)\Gamma_{x_0,x_1,x_2,y_1,\cdots,y_N} \\ &\text{s.t.} \quad P_{y_j}(\Gamma)=p_{y_j}^{t_j},\ \forall y_j\in X,\ j=1,\cdots,N \end{aligned} \tag{4.13}$$

where,

$$c(x_0,x_1,x_2,y_1,\cdots,y_N)=\sum_{i=1}^{N}\lambda_i||x_0+t_ix_1+t^2x_2-y_i||^2,$$

and

$$P_{y_j}(\Gamma)=\sum_{x_0,x_1,x_2,y_1,\cdots,y_{j-1},y_{j+1},\cdots,y_N\in X}\Gamma_{x_0,x_1,x_2,y_1,\cdots,y_{j-1},y_j,y_{j+1},\cdots,y_N}.$$

4.3.2 Entropy regularization

The linear programming problems in (4.11) and (4.13) suffer from a high computational burden. However, the more efficient Sinkhorn iteration can be employed to converge to the optimal solution of entropy-regularized problem as explained in the following.

Given two discrete probability measures $\mu=\sum_{x\in\mathbb{X}}p_x\delta_x$ and $\nu=\sum_{x\in\mathbb{X}}q_x\delta_x$, supported on a finite set $X\subset\mathbb{X}$, the relative entropy (Kullback-Leibler divergence) of μ with respect to

ν [67] is defined as

$$H(\mu|\nu) = \begin{cases} \sum_{x\in\mathbb{X}} p_x \log \frac{p_x}{q_x} & \text{if } \mu \ll \nu \\ +\infty & \text{otherwise,} \end{cases}$$

where $\mu \ll \nu$ indicates that μ is absolutely continuous with respect to ν and $0 \log 0$ is defined to be 0. Also, define

$$H(\mu) := H(\mu|1) = \sum_{x\in\mathbb{X}} p_x \log p_x,$$

which is effectively the negative of entropy of μ.

In the rest of this section, we present the results for measure-valued lines, however, one can readily state analogous results for quadratic curves. The entropy regularized version of our multi-marginal formulation is the convex problem

$$\begin{aligned} \min_{\Gamma\geq 0} \quad & \sum_{x_0,x_1,y_1,\cdots,y_N\in X} c(x_0,x_1,y_1,\cdots,y_N)\Gamma_{x_0,x_1,y_1,\cdots,y_N} + \epsilon H(\Gamma) \\ \text{s.t.} \quad & P_{y_j}(\Gamma) = p_{y_j}^{t_j},\ \forall y_j \in X,\ j = 1,\cdots,N, \end{aligned} \tag{4.14}$$

where $\epsilon > 0$ is a regularization parameter.

There are effective strategies to solve entropy regularized optimal transport problems, for instance, the alternating projection method (iterative Bergman projections [13,68,69]), which is based on projecting sequentially an initial Γ onto the subset corresponding to each marginal constraint.

Sinkhorn's algorithm [62] is another approach which enjoys a slightly better performance in terms of space complexity and parallel computation as discussed in detail in [13]. In this method, the optimal solution $\hat{\Gamma}$ is expressed in terms of the Lagrange dual variables, which

may be computed by Sinkhorn iterations. In the following, we first briefly touch upon this method, then by presenting a similar idea to that used in [70], we explain how to improve the performance of this algorithm in terms of time and space complexities.

It can be shown [70] that, for any $x_0, x_1, y_1, \cdots, y_N \in X$, the minimizer of (4.14) is of the form

$$\hat{\Gamma}_{x_0,x_1,y_1,\cdots,y_N} = \exp(-\frac{c(x_0, x_1, y_1, \cdots, y_N)}{\epsilon}) \times a_{y_1}^{t_1} \times \cdots \times a_{y_N}^{t_N}, \tag{4.15}$$

for suitable values of $a_{y_j}^{t_j},\ j = 1, \cdots, N$. These are dual variables in the dual problem (see e.g. [63]). In Sinkhorn's algorithm, the $a_{y_j}^{t_j}$'s in (4.17) can be found by iteratively updating their values via

$$a_{y_j}^{t_j} \leftarrow a_{y_j}^{t_j} \times p_{y_j}^{t_j} / P_{y_j}(\hat{\Gamma}), \forall j = 1, \cdots, N, y_j \in X. \tag{4.16}$$

It is known that in the scheme above, the sequence converges at least linearly to a minimizer of (4.14) (see e.g. [69, 70]).

The computational drawback of Sinkhorn's algorithm lies in computing the projections $P_{y_j}(\hat{\Gamma})$ in (4.16), as these grow exponentially in the number of snapshots (N). Furthermore, a large amount of memory is required to store the array $\hat{\Gamma}$ at each iteration which leads to a space complexity issue. However, the specific structure of the cost in (4.15) can be exploited to mitigate the aforementioned bottlenecks. Similar ideas have been advanced in [70, 71].

Notice that we can partially decouple the cost as

$$c(x_0, x_1, y_1, \cdots, y_N) = \sum_{i=1}^{N} \lambda_i c_i(x_0, x_1, y_i)$$

where,

$$c_i(x_0, x_1, y_i) = ||(1 - t_i)x_0 + t_i x_1 - y_i||^2.$$

The first implication of this decoupling is that, it is now not needed to store all the elements of $c(x_0, x_1, y_1, \cdots, y_N)$, but only those required to calculate c_is. Moreover, the minimizer in (4.15), can be decoupled as

$$\hat{\Gamma}_{x_0,x_1,y_1,\cdots,y_N} = \prod_{i=1}^{N} a_{y_i}^{t_j} \exp(-\frac{c_i(x_0, x_1, y_i)}{\epsilon}). \tag{4.17}$$

In the following, we explain how to leverage this structure to calculate $P_{y_1}(\hat{\Gamma})$ more efficiently. The same procedure can be utilized to compute other projections, i.e., $P_{y_j}(\hat{\Gamma})$, $j = 2, \cdots, N$. One can easily observe that $P_{y_1}(\hat{\Gamma})$ for fixed $x_0, x_1 \in X$, reads

$$P_{y_1|x_0,x_1}(\hat{\Gamma}) = a_{y_1}^{t_1} \exp(-\frac{c_1(x_0, x_1, y_1)}{\epsilon}) \prod_{i=2}^{N} \left(\sum_{y_i \in X} a_{y_i}^{t_j} \exp(-\frac{c_i(x_0, x_1, y_i)}{\epsilon}) \right), \tag{4.18}$$

for any $y_1 \in X$, and hence,

$$P_{y_1}(\hat{\Gamma}) = \sum_{x_0,x_1 \in X} P_{y_1|x_0,x_1}(\hat{\Gamma}). \tag{4.19}$$

The benefit of this approach is that the term

$$\prod_{i=2}^{N} \left(\sum_{y_i \in X} a_{y_i}^{t_j} \exp(-\frac{c_i(x_0, x_1, y_i)}{\epsilon}) \right)$$

in (4.18) is independent of y_1 and thus it is the same for all $y_1 \in X$. The complexity of computing this term for all $x_0, x_1 \in X$ is $\mathcal{O}((N-1)|X|^3)$, where $|X|$ is the cardinality of the discrete set X. This leads to $\mathcal{O}(N|X|^3)$ as the total computational complexity of each Sinkhorn iteration by using (4.19) to compute the projections. Notice that computing

the projections $P_{y_j}(\hat{\Gamma})$ by summing over all the indices $x_0, x_1, y_1, \cdots, y_{j-1}, y_{j+1}, \cdots, y_N$ as defined in (4.12) scales exponentially in the value of N, i.e., the computational complexity of one Sinkhorn update in (4.16) is $\mathcal{O}(|X|^{N+2})$. Therefore, leveraging the structure of cost in our multi-marginal formulation, decreases the computational complexity of the Sinkhorn iterations substantially.

4.4 Gaussian case

Suppose the data are Gaussian distributions $\mu_{t_i} \sim N(0, C_{y_i})$, $i = 1, \cdots, N$, where the C_{y_i}'s are symmetric and positive definite matrices. The means of distributions are assumed to be zero for simplicity and without loss of generality. This is due to the fact that for Gaussian measures, the means can be treated separately via ordinary regression in Euclidean space and thereby, the means for the optimal curve in $(\mathcal{P}_2(\mathbb{X}), W_2)$ can be computed as a function of t.

In practical settings where for each marginal only a set of samples is available, we can approximate each C_{y_i} with the sample covariance. The following proposition recasts (4.7) as a Semi-Definite Programming (SDP).

Proposition 4.4.1. *Consider $\mu_{t_i} \sim N(0, C_{y_i})$, i.e., Gaussian "points". A minimizing $\hat{\gamma}$ in (4.7) is Gaussian with zero mean and covariance of the form*

$$C_\gamma = \begin{bmatrix} C_{x_0} & S_{x_0x_1} & S_{x_0x_2} & S_{x_0y_1} & \cdots & S_{x_0y_N} \\ S^T_{x_0x_1} & C_{x_1} & S_{x_1x_2} & S_{x_1y_1} & \cdots & S_{x_1y_N} \\ S^T_{x_0x_2} & S^T_{x_1x_2} & C_{x_2} & S_{x_2y_1} & \cdots & S_{x_2y_N} \\ S^T_{x_0y_1} & S^T_{x_1y_1} & S^T_{x_2y_1} & C_{y_1} & \cdots & S_{y_1y_N} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ S^T_{x_0y_N} & S^T_{x_1y_N} & S^T_{x_2y_N} & S^T_{y_1y_N} & \cdots & C_{y_N} \end{bmatrix} \tag{4.20}$$

that solves

$$\min_{C_\gamma \succeq 0} \sum_{i=1}^{N} \lambda_i.\mathrm{tr}(C_{x_0} + t_i^2 C_{x_1} + t_i^4 C_{x_2} + C_{y_i} + 2t_i S_{x_0x_1} + 2t_i^2 S_{x_0x_2} + 2t_i^3 S_{x_1x_2} - 2S_{x_0y_i} - 2t_i S_{x_1y_i} - 2t_i^2 S_{x_2y_i}). \tag{4.21}$$

where $C_\gamma \succeq 0$ *indicates that* C_γ *is positive semi-definite.*

Notice that in C_γ, the sub-matrices C_{y_i}s are given, while the other blocks are unknown.

Proof. As the marginals $\{\mu_{t_i}\}_{i=1}^N$ in (4.7) are Gaussian and the cost function is quadratic in $x_0, x_1, x_2, y_1, \cdots, y_N$, it follows that $\hat{\gamma}$ in (4.7) is also Gaussian as in the cost and constraints only second-order moments are involved. Simple calculation shows the quadratic cost in (4.7) can be written as that in (4.21). □

It should be noted that since $\hat{\pi} = (x_0, x_1, x_2)_\# \hat{\gamma}$, one can express the optimal curve in $\mathcal{G}_{\mathrm{Quad}}$ (defined in (4.2)) as

$$\nu_t \sim N(0, C_{x_0} + t^2 C_{x_1} + t^4 C_{x_2} + t(S_{x_0x_1} + S_{x_0x_1}^T) + t^2(S_{x_0x_2} + S_{x_0x_2}^T) + t^3(S_{x_1x_2} + S_{x_1x_2}^T)), \tag{4.22}$$

for $t \in [0, 1]$, for the optimal solution of (4.21).

Similar results can be derived for multi-marginal formulation of measure-valued lines in (4.9). In particular, a minimizing $\hat{\gamma}$ in (4.9) is Gaussian with zero mean and covariance of the form

$$C_\gamma = \begin{bmatrix} C_{x_0} & S_{x_0x_1} & S_{x_0y_1} & \cdots & S_{x_0y_N} \\ S_{x_0x_1}^T & C_{x_1} & S_{x_1y_1} & \cdots & S_{x_1y_N} \\ S_{x_0y_1}^T & S_{x_1y_1}^T & C_{y_1} & \cdots & S_{y_1y_N} \\ \vdots & \vdots & \vdots & \ddots & \\ S_{x_0y_N}^T & S_{x_1y_N}^T & S_{y_1y_N}^T & \cdots & C_{y_N} \end{bmatrix}, \tag{4.23}$$

that solves

$$\min_{C_\gamma \succeq 0} \sum_{i=1}^{N} \lambda_i . \mathrm{tr}((1-t_i)^2 C_{x_0} + t_i^2 C_{x_1} + C_{y_i} + 2t_i(1-t_i) S_{x_0 x_1} - 2(1-t_i) S_{x_0 y_i} - 2t_i S_{x_1 y_i}). \tag{4.24}$$

To exemplify our regression approach for Gaussian distributional data we consider a one-dimensional Ornstein–Uhlenbeck process modeled by an Itô stochastic differential equation

$$\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t \mathrm{d}t + 2\mathrm{d}\mathbf{W}_t$$

where $(\mathbf{W}_t)_{t \geq 0}$ is a one-dimensional standard Wiener process. Such a process models the dynamics of an over-damped Hookean spring in the presence of thermal fluctuations. Starting from $\mathbf{X}_0 = 0$, the variance of $\mathbf{X}_t$ reads

$$\sigma^2(t) = 2(1 - \exp(-2t)).$$

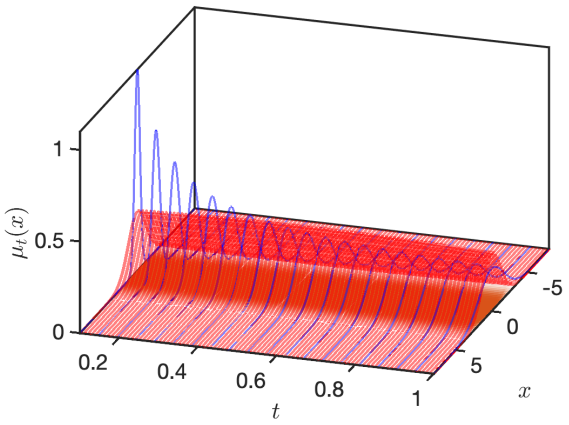
We consider the one-time marginals of this process at 20 different timestamps starting from $t = 0.1$ to $t = 1$ with equal time steps. In practical settings where only a set of samples from each one-time marginal is available, we can approximate the Gaussian distributions using the sample means and variances. The SDPs in (4.21) and (4.24) are solved separately to obtain the optimal multi-couplings $\hat{\gamma}$ and $\hat{\pi}$ in each case. In addition, for the sake of comparison we find the best geodesic which passes as close as possible to these 20 Gaussian marginals. This can be done easily as the marginals are one dimensional, noticing that the geodesic between two Gaussian distributions with standard deviations σ_0 and σ_1 is Gaussian for all $t \in [0, 1]$ with standard deviation $\sigma_t = (1-t)\sigma_0 + t\sigma_1$. Therefore, the geodesic regression in this setting becomes a linear regression in $\mathbb{R}^1$ seeking the values of $\sigma_0, \sigma_1 > 0$. Figure 4.2 illustrates the obtained curves in Wasserstein space for different values of t along with the dataset. Blue

curves are the target marginals. One can notice that the measure-valued quadratic curves capture the variation in the dataset better than measure-valued linear curves. Also, the geodesic regression has the poorest performance among the three. This ensues from the fact that in geodesic regression a curve in Wasserstein space with highest correlated endpoints is sought. However, in the framework of this study, this constraint is relaxed which can also moderate underfitting. In Fig. 4.2, some of the measure-valued linear or quadratic curves are represented in each sub-figure. The intensity of color is proportional to the likelihood of each path. From a fluid mechanical point of view, this can be thought of as a flux for the mass particles. More amount of mass transports through the darker regions.

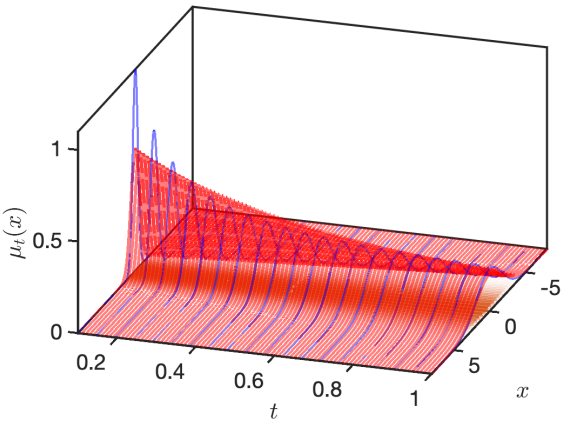
4.5 Gaussian mixtures

Linear combinations of Gaussian measures can model multi-modal densities, which are broadly used to study properties of populations with several subgroups. More generally, the set of all finite Gaussian mixture distributions ($\mathcal{GM}(\mathbb{X})$) is a dense subset of $\mathcal{P}_2(\mathbb{X})$ in the Wasserstein metric [48]. In fact, in principle, we can approximate any measure in $\mathcal{P}_2(\mathbb{X})$ with arbitrary precision with parameters for the Gaussian mixture determined via the Expectation-Maximization algorithm.

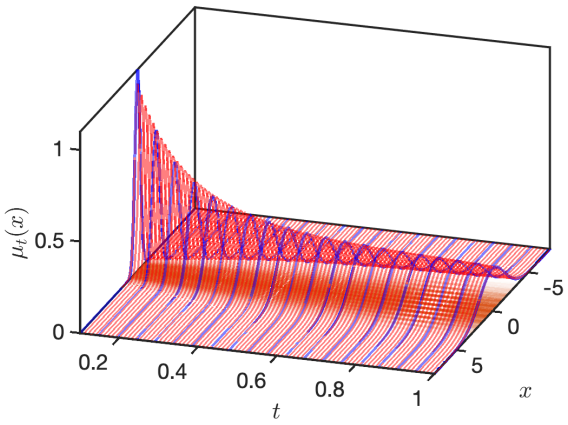
While the displacement interpolation of Gaussian distributions remains Gaussian, for Gaussian mixtures this invariance does not hold. Nevertheless, we may want to retain the Gaussian mixture structure of the interpolation due to their physical or statistical features. In [48,49], a Wasserstein-type distance on Gaussian mixture models is proposed by restricting the set of feasible coupling measures in the optimal transport problem to Gaussian mixture models. This gives rise to a geometry that inherits properties of optimal transport while it preserves the Gaussian mixture structure. Specifically, for positive integers K_0 and K_1,



(a) Geodesic regression



(b) Regression via measure-valued lines



(c) Regression via measure-valued quadratic curves

Figure 4.2: Regression results for one-dimensional Gaussian marginals. Blue curves are the given distributions and red ones are the optimal curves in the Wasserstein space. The intensity of color in linear and quadratic curves is proportional to the likelihood of each path.

consider the following Gaussian mixture models on $\mathbb{X}$,

$$\mu_0 = p^0_{\nu_1}\nu^0_1 + \cdots + p^0_{\nu_{K_0}}\nu^0_{K_0}, \quad \mu_1 = p^1_{\nu_1}\nu^1_1 + \cdots + p^1_{\nu_{K_1}}\nu^1_{K_1},$$

where each ν^i_j is a Gaussian distribution and $p^i = \begin{bmatrix} p^i_{\nu_i} & \cdots & p^i_{\nu_{K_i}} \end{bmatrix}^T$, $i = 0, 1$, are probability vectors. Now define a Wasserstein-type distance between the two Gaussian mixtures μ_0 and

$\mu_1 \in \mathcal{GM}(\mathbb{X})$ by minimizing

$$\int_{\mathbb{X}^2} \|x - y\|_2^2 \, \mathrm{d}\pi(x, y)$$

over $\pi \in \Pi(\mu_0, \mu_1) \bigcap \mathcal{GM}(\mathbb{X}^2)$. The square root of minimum defines a metric on $\mathcal{GM}(\mathbb{X})$ denoted by $W_M(\mu_0, \mu_1)$ [48, 49]. Clearly,

$$W_2(\mu_0, \mu_1) \leq W_M(\mu_0, \mu_1), \quad \forall \mu_0, \mu_1 \in \mathcal{GM}(\mathbb{X}).$$

The problem above has an equivalent discrete formulation. In particular, by viewing the Gaussian mixtures as discrete probability distributions on the Wasserstein space of Gaussian distributions, we can show [48]

$$W_M^2(\mu_0, \mu_1) = \min_{w \in \Pi(p^0, p^1)} \sum_{i,j} w_{ij} W_2^2(\nu_i^0, \nu_j^1), \tag{4.25}$$

where $\Pi(p^0, p^1)$ denotes the space of joint distributions between the probability vectors p^0 and p^1. The space of Gaussian mixtures equipped with this metric is a geodesic space for which one can define the displacement interpolation (see [48, 49] for further details).

This Wasserstein-type distance between the discrete distributions on the Wasserstein space of Gaussian distributions, allows for the notion of measure-valued curves being carried over into the case of Gaussian mixtures. In other words, in the space of Gaussian distributions, the displacement interpolations (Eq. (3.6)) play the role of straight lines in Euclidean space. Therefore, the goal is to find a probability measure over the space of geodesics of Gaussian distributions, for which the one-time marginals approximate a set of Gaussian mixtures indexed with timestamps. To do so, consider the set $X = \{\nu_i\}_{i=1}^K$ which consists of a finite

number of Gaussian distributions. Also, the available data

$$\left\{\mu_{t_i} = \sum_{\nu \in X} p_\nu^{t_i} \nu \right\}_{i=1}^{N},$$

is a family of Gaussian mixtures, each associated with a timestamp $t_i \in [0,1]$. Each μ_{t_i} can also be thought of as a discrete probability measure over the space of Gaussian measures supported on X (or a subset of X). By analogy with the formalism for measure-valued lines (Eq. (4.4)), we minimize

$$\min_{w \in \Omega} \sum_{i=1}^{N} \lambda_i W_M^2(\sum_{j\ell} w_{j\ell}\, g_{t_i}^{\nu_j \nu_\ell} \,,\, \mu_{t_i}), \tag{4.26}$$

where $\Omega = \left\{ w \in \mathbb{R}_+^{K \times K} \mid \sum_{j\ell} w_{j\ell} = 1 \right\}$ and $g_t^{\nu_j \nu_\ell}$ represents the displacement interpolation between ν_j and ν_ℓ. This problem can be recast as a multi-marginal optimal transport problem which enjoys a linear structure by pursuing the same strategy introduced in Section 4.2. In particular, (4.26) is equivalent to seeking a multi-dimensional array $(\Gamma_{\sigma_0,\sigma_1,\nu_1,\cdots,\nu_N})_{(\sigma_0,\sigma_1,\nu_1,\cdots,\nu_N) \in X^{N+2}}$ with non-negative real elements which solves

$$\begin{aligned} &\min_{\Gamma \geq 0} \sum_{\sigma_0,\sigma_1,\nu_1,\cdots,\nu_N \in X} c(\sigma_0,\sigma_1,\nu_1,\cdots,\nu_N)\Gamma_{\sigma_0,\sigma_1,\nu_1,\cdots,\nu_N} \\ &\text{s.t.} \quad P_{\nu_j}(\Gamma) = p_{\nu_j}^{t_j},\ \forall \nu_j \in X,\ j = 1,\cdots,N \end{aligned} \tag{4.27}$$

where

$$c(\sigma_0,\sigma_1,\nu_1,\cdots,\nu_N) = \sum_{i=1}^{N} \lambda_i W_2^2(g_{t_i}^{\nu_j \nu_\ell}, \nu_i),$$

and $P_{\nu_j}(\Gamma)$ is the projection operator on the marginal of Γ associated with ν_j, cf. (4.12). Also, the minimizer of (4.26) ($\hat{w}$) can be obtained by the projection $\hat{w} = P_{\sigma_0,\sigma_1}(\hat{\Gamma})$, where $\hat{\Gamma}$ is the minimizer of (4.27).

The formalism above is a linear programming which can be solved efficiently as one can solve the entropy regularized version of it by leveraging the generalized Sinkhorn algorithm as described in Section 4.3.

We exemplify this approach for Gaussian mixtures with the following toy example. We consider a finite set of probability measures which consists of 4 Gaussian distributions as depicted in Fig. 4.3. The distributional data at 4 instants of time are constructed by choosing some probability vectors over the elements of this set. These target distributions are shown in Fig. 4.4. The linear programming in (4.27) is solved, which results in a curve in $\mathcal{P}_2(\mathbb{X})$ for which the one-time marginals are Gaussian mixtures. The result of regression for this problem is illustrated in Fig. 4.5 at some timestamps. One can observe that the one-time marginals of the obtained curve capture the variation of the distributional data in time.

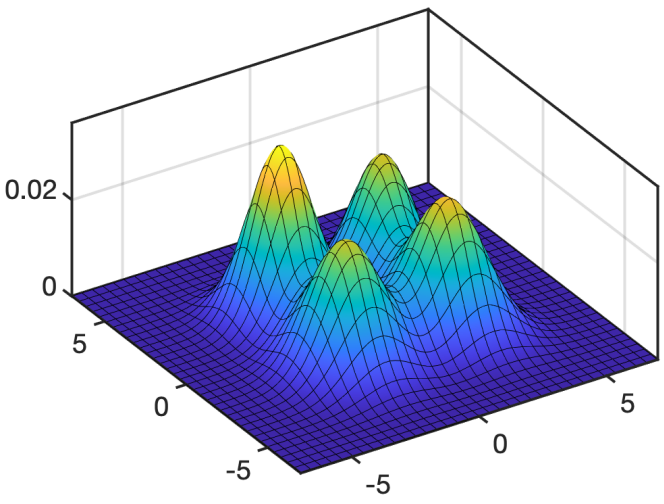
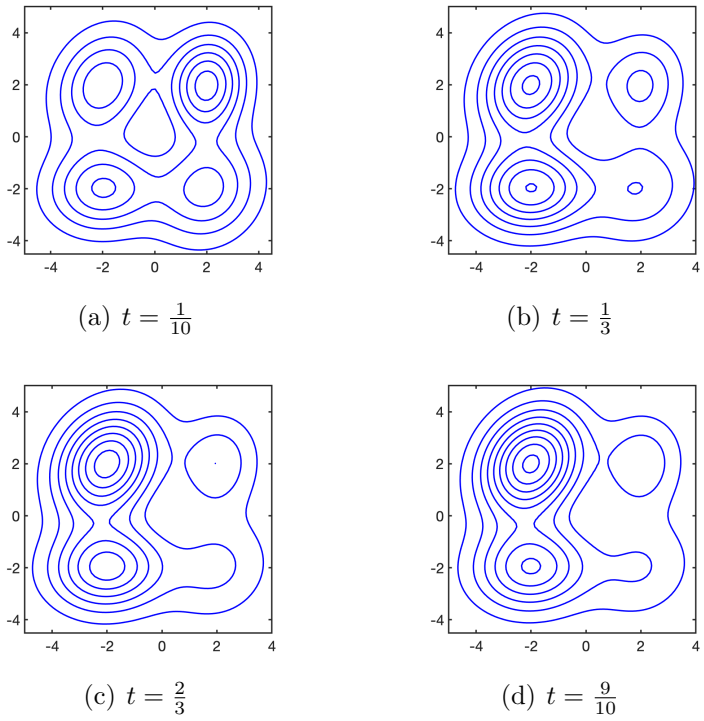


Figure 4.3: Gaussian Basis

4.6 Estimation of invariant measures

As a potential application of the proposed regression, we describe an approach to approximate the Perron-Frobenius operator and stationary distribution (if any exists) associated with a dynamical system using a few available distributional snapshots. Most studies in



(a) $t = \frac{1}{10}$ (b) $t = \frac{1}{3}$

(c) $t = \frac{2}{3}$ (d) $t = \frac{9}{10}$

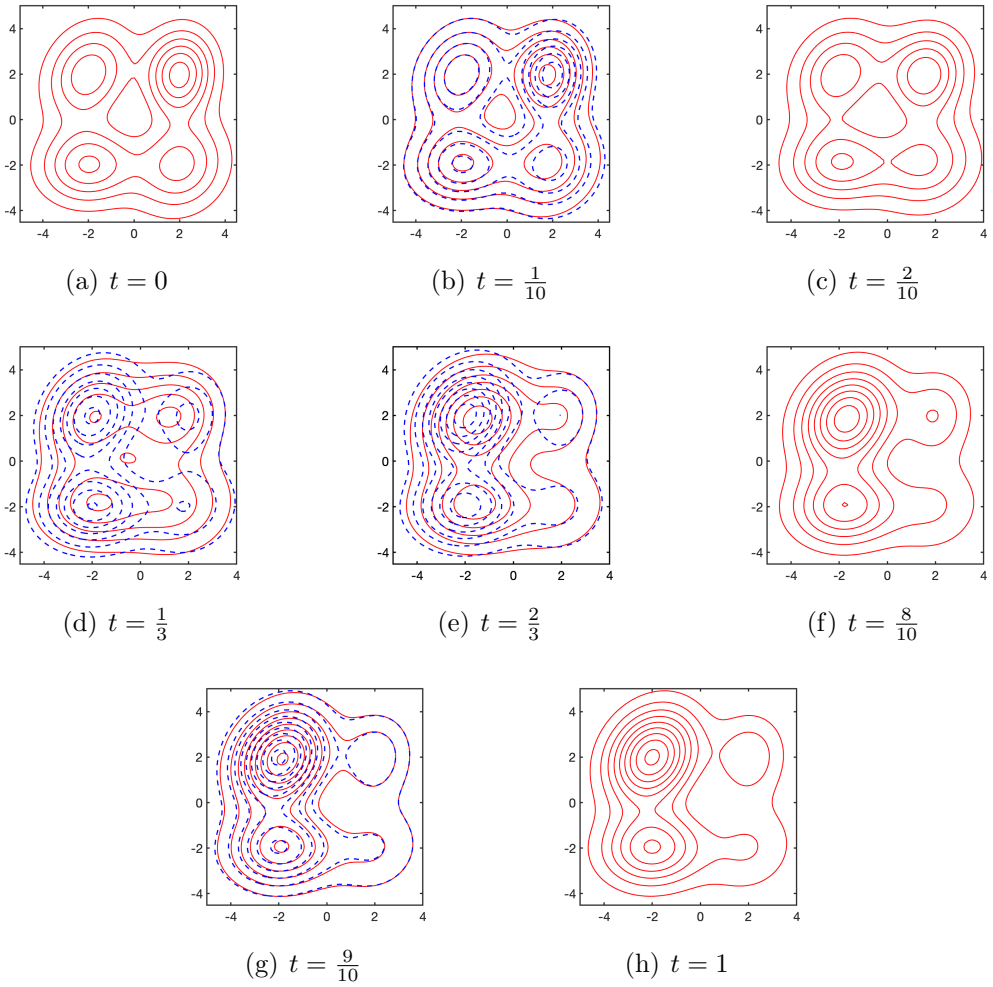
Figure 4.4: Distributional data

the literature present numerical computation of invariant measures for known dynamics, or where the pointwise correspondence between the successive points in time is available (See [45] and references therein). In our approach, we hypothesize no information on the underling dynamics or the trajectories of particles.

A discrete-time dynamical system

$$x_{k+1} = S(x_k)$$

on the measure space $(\mathbb{X}, \mathcal{B}(\mathbb{X}), \lambda)$ is defined by a λ-measurable state transition map $S : \mathbb{X} \to \mathbb{X}$. This map is assumed to be non-singular, which guarantees that the push-forward operator under S preserves the absolute continuity of (probability) measures with respect to λ. In continuous-time setting, the state transition law can be represented by a flow map $x_{t+\tau} = S_\tau(x_t)$ for $\tau \geq 0$, where x_t denotes the state of dynamics at time t. We assume



(a) $t = 0$ (b) $t = \frac{1}{10}$ (c) $t = \frac{2}{10}$

(d) $t = \frac{1}{3}$ (e) $t = \frac{2}{3}$ (f) $t = \frac{8}{10}$

(g) $t = \frac{9}{10}$ (h) $t = 1$

Figure 4.5: The result of measure-valued geodesics regression for Gaussian mixtures.

that the dynamics is time-invariant. The evolution of probability measures under S_τ can be written as $\mu_{t+\tau} = S_{\tau\#}\mu_t$.

Let $L^1(\mathbb{X}) := L^1(\mathbb{X}, \mathcal{B}(\mathbb{X}), \lambda)$ be the space of integrable functions on $\mathbb{X}$, then the Perron-Frobenius operator (PFO), $P_\tau : L^1(\mathbb{X}) \to L^1(\mathbb{X})$, is defined by

$$\int_A P_\tau f \; d\lambda = \int_{S_\tau^{-1}(A)} f \; d\lambda, \quad \forall A \in \Sigma, \tag{4.28}$$

for $f \in L^1(\mathbb{X})$. When f is a density associated with the probability measure μ_f, PFO can be thought of as a push-forward map, that is, $P_\tau \mu_f = S_{\tau\#}\mu_f$. The connection between the

dynamics and PFO can be seen in that the PFO translates the center of a Dirac measure $\delta_x \in L^1(\mathbb{X})$ in compliance with the underlying dynamics, that is, $S_{\tau\#}\delta_x = \delta_{S_\tau(x)}$. It also relates to the Koopman operator which acts on the observable functions through a duality correspondence [72].

It is standard that PFO is a Markov operator, namely, a linear operator which maps probability densities to probability densities. It is also a weak contraction (non-expansive map), in that, $\|P_\tau f\|_{L^1} \leq \|f\|_{L^1}$ for any $f \in L^1(\mathbb{X})$. If $\mu = S_{\tau\#}\mu$, then μ is an invariant measure for S_τ. For many dynamical systems, the PFO drives the densities into an invariant one (measure, in general), which is unique if the map S_τ is ergodic with respect to λ.

For non-deterministic dynamics where $S_\tau(x)$ is a an $\mathbb{X}$-valued random variable on some implicitly given probability space, the Perron-Frobenius operator reads

$$P_\tau f(x) = \int_{\mathbb{X}} K_\tau(y, x)\ f(y)\ \mathrm{d}\lambda(y), \tag{4.29}$$

where the transition density function is denoted by $K_\tau(y, x)\ :\ \mathbb{X}\times\mathbb{X} \to [0, \infty]$. The transition density function exists, if $S_\tau(x)$ does not assign non-zero measures to null sets [73].

The most popular method in the literature to discretize PFO is Ulam's method [39, 40]. In this approach, the state-space ($\mathbb{X}$) is divided into a finite number of disjoint measurable boxes $\{B_1, ..., B_n\}$. The PFO is approximated with a $n \times n$ matrix with elements p_{ij}. To do so, first we can choose a number (q) of test points (samples) $\{x_l^i\}_{l=1}^q$ within each Box B_i randomly. Then, the elements of this matrix can be estimated by

$$p_{ij} = \frac{1}{q}\sum_{l=1}^{q} \mathbf{1}_{B_j}(S(x_l^i))$$

where $\mathbf{1}_{B_j}$ denotes the indicator function for the box B_j.

Ulam's method requires the trajectories of test points to be available, which is not the case in

many practical situations, where the trajectories of test points (i.e., mass particles, agents, or so on) are missing. We can use the method of this study for regression, to estimate the Perron-Frobenius operator, and subsequently invariant measure corresponding to some dynamical system based on the collective behavior of particles. In other words, we postulate no knowledge of the underlying dynamics and assume that only a limited number of one-time marginal distributions is available at different timestamps t_i, $i \in \{1, \ldots, N\}$. In fact, these one-time marginals are the evolution of some initial distribution at $t = 0$ under the action of discretized dynamics $x_{t+\tau} = S_\tau(x_t)$. As mentioned in Proposition 4.1.2, the time can be scaled to lie within the interval $[0, 1]$. These distributions are quantized by suitable partitioning of the domain $\mathbb{X} = \bigcup_{\ell=1}^{n} B_\ell$, which is a compact set, followed by counting the particles in each of the n boxes B_ℓ to obtain μ_{t_i}, with Diracs placed at the center of each interval. The Minimizer of multi-marginal formulation of the regression problem (i.e., Eq. (4.11)), provides a coupling $\hat{\pi}$ between the distributions at two instants of time $t = 0$ and $t = 1$. Also, this can be thought of as a probability measure over the space of linear curves in $\mathbb{X}$, which indicates how much mass is transporting along the lines from $t = 0$ to $t = 1$. Putting these two views together, one can conclude that $\hat{\pi}$ gives a correlation law between the distributions of particles at $t = 0$ and $t = 1$, where the mass particles move at constant speeds from $t = 0$ to $t = 1$. It should be noted that the entropy regularization of cost can be employed to find $\hat{\pi}$ efficiently, as discussed in Section 4.3.

Notice that $\hat{\pi}$ contains the information on the distributions of the particles at $t = 0$ and $t = 1$, namely, $p_{\{t=0\}}$ and $p_{\{t=1\}}$ respectively, as well as the correlation law between the two end-points. Therefore, we can determine a transition probability matrix (of a Markov chain)

$$Q(\ell, \ell') = \pi^*(\ell, \ell')/p_{\{t=0\}}(\ell), \tag{4.30}$$

for $\ell, \ell' \in \{1, \ldots, n\}$. From a measure-theoretic point of view Q can be seen as the dis-

integration of $\hat{\pi}$. This transition probability matrix can be deemed as a finite-dimensional approximation of the Perron-Frobenius operator corresponding to the underlying dynamics, that is, P_τ in either (4.28) or (4.29) for $\tau = 1$. Assuming that the underlying dynamics is time-invariant (or time-homogeneous for non-deterministic dynamics), the invariant distribution of dynamical system can be approximated by the stationary vector of Q.

To exemplify this approach, we apply it to logistic map in order to predict its asymptotic statistical properties for different values of population-growth parameter. The logistic model for population growth is

$$x_{k+1} = T(x_k) = r x_k (1 - x_k), \tag{4.31}$$

where $x_k \in [0, 1]$, $k \in \{0, 1, \ldots\}$, and r is the population-growth parameter, see [74]. The behavior of dynamics changes from regular to chaotic as the parameter r varies from 0 to 4. We visualize the results for two values of r, namely, $r = 3$ and $r = 4$.

For $2 \leq r \leq 3$, starting from any initial point in $(0, 1)$, the population will eventually approach the same value $\frac{r-1}{r}$, so-called "attractor". However, as r approaches 3 the convergence becomes increasingly slow. For $r = 4$, it is known that this system displays highly chaotic behavior; in fact, starting from any initial point $x_0 \in (0, 1)$, the sequence $\{x_k \,|\, k = 1, 2, \ldots\}$ covers densely the interval $[0, 1]$, see [75]. Yet, the dynamical system is statistically stable in that, any initial probability distribution tends towards an invariant measure with density

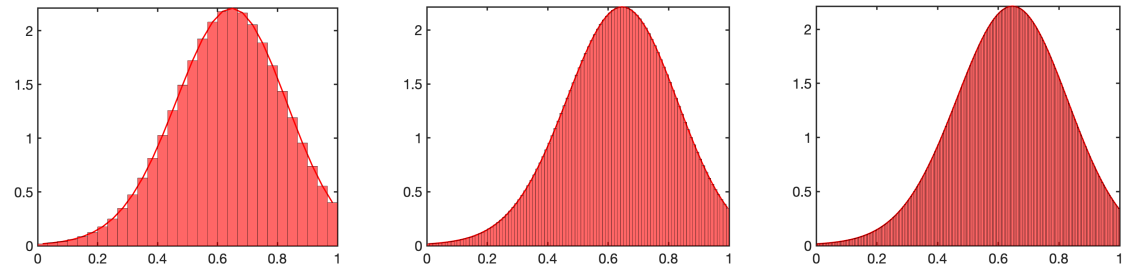
$$f_s(x) = \frac{1}{\pi\sqrt{x(1-x)}}. \tag{4.32}$$

Our aim is to estimate where the mass particles will eventually concentrate by the iterates of logistic map using only a few probability distributions obtained from the evolution of an initial distribution under the action of this map. We do not hypothesize any information on

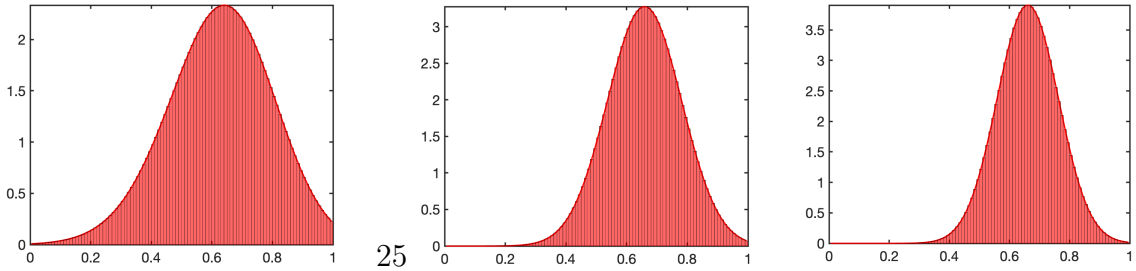
the correlations between each pair of the probability distributions. Namely, the logistic map is only used to construct the distributional data. To do so, the interval $[0, 1]$ is partitioned into n sub-intervals of equal width and the evolution of 1000 points, uniformly selected in $[0, 1]$, is used to construct N distributional data under the successive iterates of logistic map. We index the data with timestamps $t_i = \frac{i-1}{N-1}$, $i = 1, \cdots, N$. The logistic map herein can be though of as the flow map of a dynamical system for the time lag $\tau = \frac{1}{N-1}$. We provide the results for different values of N, n, and the regularization parameter ϵ in Sinkhorn's algorithm, to examine the sensitivity of the results to these parameters.

The transition probability matrix Q in Eq. (4.30) is computed for different values of N, n, and ϵ and accordingly the stationary distribution of Q is obtained. The results are depicted in Fig. 4.6 which show that the stationary distribution is concentrated around the stable fixed point of the logistic map ($x_{k+1} = 3x_k(1 - x_k)$) at $x = \frac{2}{3}$. In particular, the results for three values of n are illustrated in the first row. The second row represents the impact of N on estimated stationary distribution. As the number of distributional data varies from 3 to 9, we observe the stationary distribution is concentrated more densely around the fixed point. Finally, the third row relates to the sensitivity of results to regularization parameter ϵ. Although for smaller values of ϵ the convergence of Sinkhorn iterates to the optimal solution becomes slower, we achieve a better result for the stationary distribution in terms of having a lower variance.

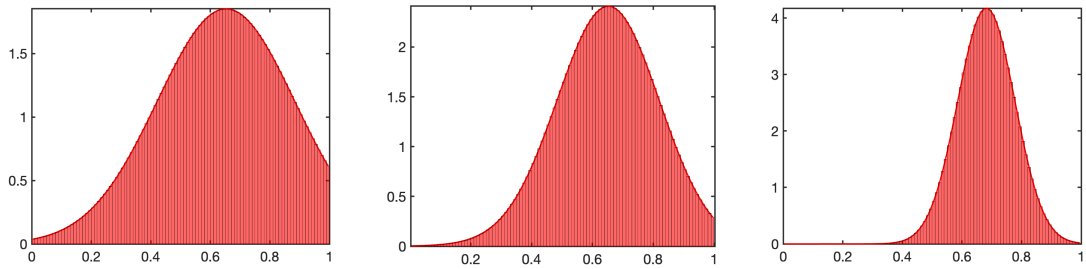
Figure 4.7 depicts the approximated invariant measure for the logistic map where $r = 4$. In this case $[0, 1]$ is partitioned into 50 equi-length sub-intervals and we construct 5 distributional data by the iterates of logistic map starting from a uniform distribution. The blue curve represents the analytic invariant measure given in (4.32).



(a) $N = 5,\ n = 30,\ \epsilon = 0.1$ (b) $N = 5,\ n = 100,\ \epsilon = 0.1$ (c) $N = 5,\ n = 200,\ \epsilon = 0.1$



(d) $N = 3,\ n = 100,\ \epsilon = 0.05$ (e) $N = 6,\ n = 100,\ \epsilon = 0.05$ (f) $N = 9,\ n = 100,\ \epsilon = 0.05$



(g) $N = 6,\ n = 100,\ \epsilon = 0.2$ (h) $N = 6,\ n = 100,\ \epsilon = 0.1$ (i) $N = 6,\ n = 100,\ \epsilon = 0.03$

Figure 4.6: The stationary distribution of the Markov chain (histogram and red fitting curve) for logistic map $x_{k+1} = 3x_k(1 - x_k)$. The figures show the concentration of stationary distribution around the single stable fixed point of logistic map at $x = \frac{2}{3}$ for different values of N, n, and ϵ.

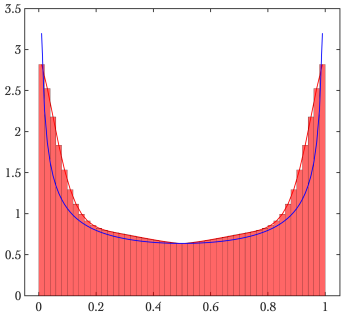


Figure 4.7: The stationary distribution of the Markov chain (histogram and red fitting curve) is compared to the invariant density of the logistic map for $r = 4$ (blue).

Chapter 5

Data-Driven Approximation of the Perron-Frobenius Operator Using the Wasserstein Metric

5.1 Transfer operators

In this section we discuss the Perron-Frobenius operator and Koopman operators. These encode information on the underlying dynamical equations, which are nonlinear, in general. The operators are linear albeit on infinite-dimensional spaces, the space of distributions and observables, respectively. Although our study focuses on approximating the Perron-Frobenius operator, we concisely summarize the duality between the two [29].

5.1.1 Notation

The three-tuple $(\mathbb{X}, \Sigma, \lambda)$ represents a measure space $\mathbb{X} \subset \mathbb{R}^d$ equipped with a sigma-algebra Σ and measure λ. Typically, and unless otherwise stated, $\mathbb{X} = \mathbb{R}^d$, Σ is the Borel algebra, and λ the Lebesgue measure. The Banach space $L^p(\mathbb{X})$ $(1 \leq p \leq \infty)$ is the space of p-Lebesgue

integrable functions endowed with the norm $\|\cdot\|_{L^p}$. We denote by $(\mathcal{P}_2(\mathbb{X}), W_2)$ the Wasserstein space where $\mathcal{P}_2(\mathbb{X})$ is the set of Borel probability measures with finite second moments, and W_2 the Wasserstein distance. The push-forward of a measure ν by the measurable map $S : \mathbb{X} \to \mathbb{X}$ is denoted by $\nu' = S_{\#}\nu \in \mathcal{P}_2(\mathbb{X})$, meaning $\nu'(B) = \nu(S^{-1}(B))$ for every Borel set B. If a measure $\mu_f \in \mathcal{P}_2(\mathbb{X})$ is absolutely continuous with respect to the Lebesgue measure, then we can assign to μ_f, a density $f \in L^1(\mathbb{X})$, that is, a positive function with unit L^1-norm, such that $\mu_f(B) = \int_B f d\lambda$, for every Borel set B. The Dirac measure at point x is denoted by δ_x.

5.1.2 Perron-Frobenius operator

A discrete-time dynamical system

$$x_{k+1} = S(x_k)$$

on $\mathbb{X}$ is defined by a λ-measurable state transition map $S : \mathbb{X} \to \mathbb{X}$. This map is assumed to be non-singular throughout this chapter, which guarantees that the push-forward operator under S preserves the absolute continuity of (probability) measures with respect to λ. The time is assumed to be discrete. In other words, for the time lag τ, the evolution of measures under S can be written as $\mu_{t_k+\tau} = S_{\#}\mu_{t_k}$, $(k = 1, 2, \ldots)$; for convenience we compress the notation by writing $\mu_{t_k} =: \mu_k$.

The Perron-Frobenius operator (PFO), $P : L^1(\mathbb{X}) \to L^1(\mathbb{X})$, is defined by

$$\int_A Pf \; d\lambda = \int_{S^{-1}(A)} f \; d\lambda, \quad \forall A \in \Sigma$$

for $f \in L^1(\mathbb{X})$. When f is a density associated with the probability measure μ_f, PFO can be thought of as a push-forward map, that is, $P\mu_f = S_{\#}\mu_f$. The connection between the dynamics and PFO can be seen in that the PFO translates the center of a Dirac measure

$\delta_x \in L^1(\mathbb{X})$ in compliance with the underlying dynamics, that is, $S_{\#}\delta_x = \delta_{S(x)}$.

It is standard that PFO is a Markov operator, namely, a linear operator which maps probability densities to probability densities. It is also a weak contraction (non-expansive map), in that, $\|Pf\|_{L^1} \leq \|f\|_{L^1}$ for any $f \in L^1(\mathbb{X})$. For many dynamical systems, the PFO drives the densities into an invariant one (measure, in general) which is unique if the map S is ergodic with respect to λ.

5.1.3 Koopman operator

The Koopman operator (KO) with respect to S, $U : L^\infty(\mathbb{X}) \to L^\infty(\mathbb{X})$, is the infinite-dimensional linear operator

$$Uf(x) = f(S(x)), \quad \forall x \in \mathbb{X}, \quad \forall f \in L^\infty(\mathbb{X}),$$

see e.g., [76]. This is a positive operator and a weak contraction, that is, $\|Uf\|_{L^\infty} \leq \|f\|_{L^\infty}$ for any $f \in L^\infty(\mathbb{X})$.

It is straightforward to see that KO is the dual of PFO, namely,

$$\langle Pf, g\rangle_\lambda = \langle f, Ug\rangle_\lambda, \quad \forall f \in L^1(\mathbb{X}), \quad g \in L^\infty(\mathbb{X})$$

where $\langle \cdot, \cdot \rangle_\lambda$ is the duality pairing between $L^1(\mathbb{X})$ and $L^\infty(\mathbb{X})$. To reconstruct the underlying dynamics (S) from KO, we can pick the full-state observable $g(x) = x$, where g is a vector-valued observable and KO acts on it componentwise.

5.1.4 Data-driven approximation of transfer operators

As mentioned earlier, the most popular method in the literature to discretize PFO is the Ulam's method [39, 40]. In this method, the state-space ($\mathbb{X}$) is divided into a finite number

of disjoint measurable boxes $\{B_1, ..., B_n\}$. The PFO is approximated with a $n \times n$ matrix with elements p_{ij}. To do so, first we choose a large number (k) of test points $\{x_l^i\}_{l=1}^k$ within each Box B_i randomly. Then, the elements of this matrix can be estimated by

$$p_{ij} = \frac{1}{k} \sum_{l=1}^{k} \mathbf{1}_{B_j}(S(x_l^i))$$

where $\mathbf{1}_{B_j}$ denotes the indicator function for the box B_j.

Extended dynamic mode decomposition (EDMD) [5], on the other hand, approximates the Koopman operator for an available time series of data, i.e., $\{x_i\}_{i=1}^m$. First, a dictionary of observables $D = \{\phi_i(\cdot)\}_{i=1}^k$ is chosen. We then consider the vector-valued function $\Phi = [\phi_1 \ \phi_2 \ \dots \ \phi_k]^T$. We stack up the values of this function at the snapshots in two matrices as

$$\Phi_{[1,m-1]} = [\Phi(x_1) \ \dots \ \Phi(x_{m-1})],$$
$$\Phi_{[2,m]} = [\Phi(x_2) \ \dots \ \Phi(x_m)].$$

A finite-dimensional approximation of the restriction of the Koopman operator on the span of D can be sought by considering a $k \times k$ matrix K that satisfies

$$\Phi_{[2,m]} = K\Phi_{[1,m-1]}. \tag{5.1}$$

Depending on the values of m and k, the system of equations (5.1), may be over- or under-determined. For example, if it is over-determined, K can be obtained by solving a corresponding least-squares problem.

5.2 Rudiments of Wasserstein space

In this section, we recall the definition and some properties of the Wasserstein distance [59, 60], which are used in this chapter.

Let μ_0 and μ_1 be two probability measures in $\mathcal{P}_2(\mathbb{X})$. In the Monge's formulation of optimal transport, a mapping $T^* : \mathbb{X} \to \mathbb{X}$ is sought such that $T^*_{\#}\mu_0 = \mu_1$ and

$$\int_{\mathbb{X}} \|T^*(x) - x\|_2^2 \, d\mu_0 \leq \int_{\mathbb{X}} \|T(x) - x\|_2^2 \, d\mu_0$$

for any transport map T such that $T_{\#}\mu_0 = \mu_1$. This is the minimization of a quadratic cost over the space of maps $T : \mathbb{X} \to \mathbb{X}$ which "transport" mass $d\mu_0(x)$ at x so as to match the final distribution μ_1. If μ_0 and μ_1 are absolutely continuous, Brenier's characterization states that the optimal transport problem has a unique solution obtained as gradient of a convex function ϕ, that is a monotone map $T^* = \nabla\phi(x)$ [77].

In case a transport map fails to exist, as is the case when μ_0 is a discrete probability measure and μ_1 is absolutely continuous, we consider a relaxation of Monge's problem, known as the Kantorovich's formulation, in which one seeks a joint distribution (referred to as coupling) π on $\mathbb{X} \times \mathbb{X}$, having marginals μ_0 and μ_1 along the two coordinates, namely,

$$W_2^2(\mu_0, \mu_1) := \inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{\mathbb{X}\times\mathbb{X}} \|x - y\|^2 d\pi(x, y)$$

where $\Pi(\mu_0, \mu_1)$ is the space of "couplings" with marginals μ_0 and μ_1. In this, a minimizer always exists, and we use $\Pi^*(\mu_0, \mu_1)$ to denote the space of optimal couplings between the marginals μ_0 and μ_1. In case the optimal transport map for the Monge problem exists, the consistency between the two problems can be realized through the relation $\pi = (x, T^*(x))_{\#}\mu_0$.

The square root of the optimal cost, namely $W_2(\mu_0, \mu_1)$, defines a metric on $\mathcal{P}_2(\mathbb{X})$ referred to as the Wasserstein metric [12, 18]. Moreover, assuming that T^* exists, the constant-speed

geodesic between μ_0 and μ_1 is given by

$$\mu_t = \{(1-t)x + tT^*(x)\}_{\#}\mu_0, \quad 0 \leq t \leq 1,$$

and known as *McCann's displacement interpolation* [78].

In the following, we state an important lemma from measure theory which will be used in the proof of main theorem in this chapter.

Lemma 5.2.1 (Gluing lemma [12, 18])**.** *Let $\mathbb{X}_1$, $\mathbb{X}_2$, and $\mathbb{X}_3$ be three copies of $\mathbb{X}$. Given three probability measures $\mu_i(x_i) \in \mathcal{P}_2(\mathbb{X}_i)$, $i = 1, 2, 3$ and the couplings $\pi_{12} \in \Pi(\mu_1, \mu_2)$, and $\pi_{13} \in \Pi(\mu_1, \mu_3)$, there exists a probability measure $\pi(x_1, x_2, x_3) \in \mathcal{P}_2(\mathbb{X}_1 \times \mathbb{X}_2 \times \mathbb{X}_3)$ such that $(x_1, x_2)_{\#}\pi = \pi_{12}$ and $(x_1, x_3)_{\#}\pi = \pi_{13}$. Furthermore, the measure π is unique if either π_{12} or π_{13} are induced by a transport map.*

That is, the gluing lemma states that for any two given couplings, which are consistent along one coordinate, we can find a measure on the product space $(\mathbb{X}_1 \times \mathbb{X}_2 \times \mathbb{X}_3)$ whose projections onto each pair of coordinates match the given couplings, respectively. With this, we are ready to present the main results in the next section.

5.3 Main results

In this section, we formally define the problem of PFO approximation in the presence of distributional snapshots for a dynamical system. As already noted, it is assumed that there is no information on the correlation between each pair of data points (distributions). We seek system dynamics, $S : \mathbb{X} \to \mathbb{X}$, as a λ-measurable map such that it can serve as a model for the flow encoded in the sequence of data points $\mu_1, \mu_2, \ldots, \mu_m$. This is in the sense that, either $S_{\#}\mu_k = \mu_{k+1}$ over the data set for $k \in \{1, \ldots, m-1\}$ (exact matching), or that the discrepancy between $S_{\#}\mu_k$ and μ_{k+1}, for the successive data points, is small in the average

over the available record of distributions. Below, in Section 5.3.1, we first develop the case where S is a linear map

$$S \ : \ x \mapsto Ax,$$

with $A \in \mathbb{R}^{d\times d}$. Then, in Section 5.3.2, we detail the approach for the case where $S(\cdot) = \sum_{j=1}^{n} \theta_j y_j(\cdot)$ is nonlinear (in general) expressed in terms of a linear combination of specified basis functions y_j, $j \in \{1, \dots, n\}$.

5.3.1 First-order approximation

We first draw an analogy with the EDMD problem by stating the problem to find a matrix that satisfies the condition in Eq. (5.1). Thus, given a sequence of probability measures $\{\mu_i\}_{i=1}^{m}$ in $\mathcal{P}_2(\mathbb{X})$, we seek to find a matrix $A \in M(d)$ (the space of real $d \times d$ matrices) such that

$$[\mu_2 \ \mu_3 \ \dots \ \mu_m] = (Ax)_{\#}[\mu_1 \ \mu_2 \ \dots \ \mu_{m-1}]. \tag{5.2}$$

In (5.2), similar to EDMD, the probability distributions (μ_1, μ_2,...) are stacked in arrays, where one is the shifted version of the other. The push-forward operator acts on "stacked up" measures separately.

Typically, the problem is over-determined, in which case there might not exist a matrix A that satisfies (5.2), we consider the following regression-type formulation.

Problem 1. *Determine a matrix $A \in M(d)$ that minimizes*

$$F(A) = \sum_{i=1}^{m-1} W_2^2(Ax_{\#}\mu_i, \mu_{i+1}). \tag{5.3}$$

If (5.2) has a solution, it trivially coincides with the minimizer of Problem 1 and $F(A) = 0$.

If, on the other hand, all the measures are Dirac, that is, $\mu_i = \delta_{x_i},\ i = 1, \ldots, m$, the problem to satisfy (5.2) reduces to an ordinary DMD problem. This shows the consistency of DMD with our formulation on measures.

Next, we provide a stationarity condition that can be used to obtain the solution to Problem 1.

Theorem 5.3.1. *Consider a sequence of absolutely continuous probability measures* $\{\mu_i\}_{i=1}^m$ *in* $\mathcal{P}_2(\mathbb{X})$. *If a minimizer* $A \in M(d)$ *for* (5.3) *exists and is nonsingular, then there exist unique* $\eta_i(x_i, x_{i+1}) \in \Pi(\mu_i, \mu_{i+1})$ *for each* $i \in \{1, \ldots, m\}$ *such that*

$$(Ax_i, x_{i+1})_{\#}\eta_i \in \Pi^*(Ax_{i\#}\mu_i, \mu_{i+1}),$$

and moreover, A *satisfies*

$$\sum_{i=1}^{m-1} \int_{\mathbb{X}\times\mathbb{X}} (Ax_i - x_{i+1}) x_i^T \, d\eta_i(x_i, x_{i+1}) = 0. \tag{5.4}$$

In the theorem, each probability measure η_i is a coupling between two distributional snapshots μ_i and μ_{i+1} such that the push-forward measure $(Ax_i, x_{i+1})_{\#}\eta_i$ is an optimal coupling between its marginals. In turn, since these marginals are absolutely continuous by virtue of the fact that A is nonsingular, the latter coupling (i.e., $(Ax_i, x_{i+1})_{\#}\eta_i$) is singular and "sits" on the graph of a "Monge map." As explained in the proof of the theorem, application of the Gluing lemma shows that each η_i exists and is unique. At this point, the absolute continuity of the marginals is essential; later on, we will discuss how to relax this assumption so as to include a class of discrete measures as well.

Proof of Theorem 5.3.1: According to the assumption that A is a minimizer of (5.3), the

Fermat's condition

$$\frac{d}{d\epsilon}F(A+\epsilon\delta A)|_{\epsilon=0}=0 \tag{5.5}$$

holds for any tangent direction δA, that is, any matrix in $M(d)$. Without loss of generality, we consider only one of the terms in (5.3) and define

$$G(A)=W_2^2(Ax_{\#}\mu_1,\mu_2).$$

To calculate the directional derivative (Gateaux derivative) of $G(A)$, first we show that for any real ϵ and $\delta A\in M(d)$

$$\begin{aligned}&G(A+\epsilon\delta A)-G(A)\leq \\ &\langle\int_{\mathbb{X}\times\mathbb{X}}2(Ax_1-x_2)x_1^T d\eta_1(x_1,x_2),\epsilon\delta A\rangle_F+O(\epsilon^2)\end{aligned} \tag{5.6}$$

where $\langle\cdot,\cdot\rangle_F$ is the Frobenius inner product and η_1 is as stated in the theorem. To do so, let the measure $\gamma_1(x_1,x_1',x_2)\in\mathcal{P}_2(\mathbb{X}^3)$ be such that $(x_1,x_1')_{\#}\gamma_1=(x_1,Ax_1)_{\#}\mu_1$ and $(x_1',x_2)_{\#}\gamma_1\in\Pi^*(Ax_{1\#}\mu_1,\mu_2)$. Since these two constraints coincide along x_1', by application of the Gluing lemma, we conclude that γ_1 exists. Moreover, as the projection of γ_1 onto (x_1',x_2) is the optimal coupling between two absolutely continuous measures, it is induced by a transport map (Monge map), and thus the choice of γ_1 is unique by once again invoking the Gluing lemma. Then, $\eta_1:=(x_1,x_2)_{\#}\gamma_1$ where its uniqueness immediately results from that of γ_1. Hence,

$$\begin{aligned}&G(A+\epsilon\delta A)-G(A)\leq \\ &\int_{\mathbb{X}_1\times\mathbb{X}_2}(\|(A+\epsilon\delta A)x_1-x_2\|_2^2-\|Ax_1-x_2\|_2^2)d\eta_1(x_1,x_2).\end{aligned}$$

This follows from the fact that $G(A+\epsilon\delta A)$ is the Wasserstein distance (i.e., the minimum

among all the couplings between $(A+\epsilon\delta A)x_{1\#}\mu_1$ and μ_2). Finally, by expanding the integrand above with respect to ϵ, (5.6) is derived.

Without loss of generality we take $\epsilon > 0$. According to (5.6), we can readily conclude that

$$\limsup_{\epsilon\to 0} \frac{G(A+\epsilon\delta A)-G(A)}{\epsilon} \leq$$
$$\langle \int_{\mathbb{X}_1\times\mathbb{X}_2} 2(Ax_1-x_2)x_1^T d\eta_1(x_1,x_2), \delta A\rangle_F.$$

The next step of proof is to show that

$$\liminf_{\epsilon\to 0} \frac{G(A+\epsilon\delta A)-G(A)}{\epsilon} \geq$$
$$\langle \int_{\mathbb{X}_1\times\mathbb{X}_2} 2(Ax_1-x_2)x_1^T d\eta_1(x_1,x_2), \delta A\rangle_F.$$

This last inequality follows from the semi-concavity of the squared Wasserstein distance [65, Proposition 7.3.6].

By combining the "lim inf" and "lim sup" results, it readily follows that

$$\frac{d}{d\epsilon} G(A+\epsilon\delta A)|_{\epsilon=0} = \tag{5.7}$$
$$\langle \int_{\mathbb{X}\times\mathbb{X}} 2(Ax_1-x_2)x_1^T d\eta_1(x_1,x_2), \delta A\rangle_F.$$

Finally, writing the directional derivative for all the terms in (5.3) and using Fermat's condition the proof is complete. □

Remark 5.3.1. In the statement of Theorem 5.3.1 we assume the existence of a minimizer A to Problem 1. We now explain that this assumption holds in many reasonable settings, as for instance, in the case where the probability measures have compact support. To see this, note that $F(A)$ is coercive, i.e., $F(A)\to+\infty$ as $\|A\|_F\to+\infty$ for absolutely continuous μ_i's

with compact support. Further, using the lower semi-continuity of W_2 (see Proposition 7.1.3 and Lemma 5.2.1 in [65]), we conclude the lower semi-continuity of $F(A)$ with respect to the Frobenius norm. These two observations guarantee the existence of a solution to Problem 1. □

Remark 5.3.2. Equation (5.7) shows how to generate a gradient flow, and thereby a steepest descent direction for minimizing $F(A)$. Specifically,

$$\nabla_A F(A) = 2\sum_{i=1}^{m-1} \int_{\mathbb{X}_i \times \mathbb{X}_{i+1}} (Ax_i - x_{i+1}) x_i^T \, d\eta_i(x_i, x_{i+1}), \tag{5.8}$$

allows us to construct a gradient-type numerical optimization to find the minimizer of (5.3). □

Remark 5.3.3. We note in passing that the setting of our approximation Problem 1, can be used to construct pseudo-metrics for various applications. Specifically, an admissible set of transformations $\mathcal{F}$ may be available (e.g., rotations, translations, scalings of images and so on), and that these are natural for the problem at hand, and thought to "incur no cost." Thence, a distance can be defined between distributions as follows

$$W_{\mathcal{F}}^2(\mu_0, \mu_1) = \inf_{S \in \mathcal{F}} W_2^2(S_{\#}\mu_0, \mu_1).$$

Such a construction is relevant in image registration where alignment/scaling may be desired. □

5.3.2 Higher-order approximations

In this subsection, we extend the previous result to non-linear models for the underlying dynamics.

We consider system dynamics, $S : \mathbb{X} \to \mathbb{X}$, a λ-measurable map, to be expressed as a linear

combination of basis functions $y_j : \mathbb{X} \to \mathbb{X}$, with $j \in \{1, \ldots, n\}$, i.e.,

$$S(x; \Theta) = \sum_{j=1}^{n} \theta_j y_j(x).$$

where $\Theta = [\theta_1 \ \ldots \ \theta_n]^T \in \mathbb{R}^n$.

The set of basis functions may be chosen to include polynomials. In such a case, the corresponding-order moments of the distributional snapshots need to exist, so that integrals remain finite.

Extending (5.3) to this new setting, we now consider the problem to minimize

$$F(\Theta) = \sum_{i=1}^{m-1} W_2^2(S(x; \Theta)_{\#}\mu_i, \mu_{i+1}), \tag{5.9}$$

over $\Theta \in \mathbb{R}^n$. We follow a strategy that is similar to that in the proof of Theorem 5.3.1, to derive a first-order optimality condition for Θ in the form

$$\sum_{i=1}^{m-1} \int_{\mathbb{X}\times\mathbb{X}} (Y(x_i))^T (S(x_i; \Theta) - x_{i+1}) d\eta_i(x_i x_{i+1}) = 0. \tag{5.10}$$

Here, $Y(x_i) = [y_1(x_i) \ \ldots \ y_n(x_i)] \in \mathbb{R}^{d\times n}$ and, as before, $\eta_i(x_i, x_{i+1}) \in \Pi(\mu_i, \mu_{i+1})$ is such that

$$(S(x_i; \Theta), x_{i+1})_{\#}\eta_i \in \Pi^*(S(x_i; \Theta)_{\#}\mu_i, \mu_{i+1}).$$

In a similar manner, the absolute continuity of μ_i's guarantees the existence and uniqueness of all the η_i's.

Equation (5.10) extends our formalism to nonlinear dynamics, parametrized by the span of Y, for approximating the PFO. In a way similar to (5.8), we consider the gradient of $F(\Theta)$

in (5.9) with respect to Θ,

$$\nabla_\Theta F = 2\sum_{i=1}^{m-1}\int_{\mathbb{X}\times\mathbb{X}}(Y(x_i))^T(S(x_i;\Theta)-x_{i+1})d\eta_i(x_i,x_{i+1}), \tag{5.11}$$

and employ a gradient-type descent to find the minimizing value for Θ.

5.4 Simulation results

5.4.1 Gaussian distributions

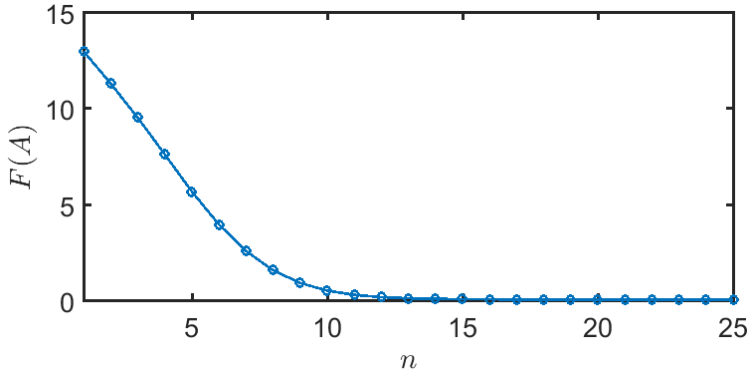


Figure 5.1: Value $F(A_n)$ as a function of iterated steps in (5.15).

We exemplify our framework with numerical results for the case where the distributional snapshots are Gaussian. In this case, the Wasserstein distance between distributions can be written in closed-form.

Consider[1] $\mu_0 = \mathcal{N}(m_0, C_0)$ and $\mu_1 = \mathcal{N}(m_1, C_1)$. The transportation problem admits a solution in closed-form [61, 79], with transportation (Monge) map

$$T^*: x \to C_0^{-1}(C_0C_1)^{1/2} = C_0^{-1/2}(C_0^{1/2}C_1C_0^{1/2})^{1/2}C_0^{-1/2}x,$$

[1] $\mathcal{N}(m, C)$ denotes a Gaussian distribution with mean m and covariance C

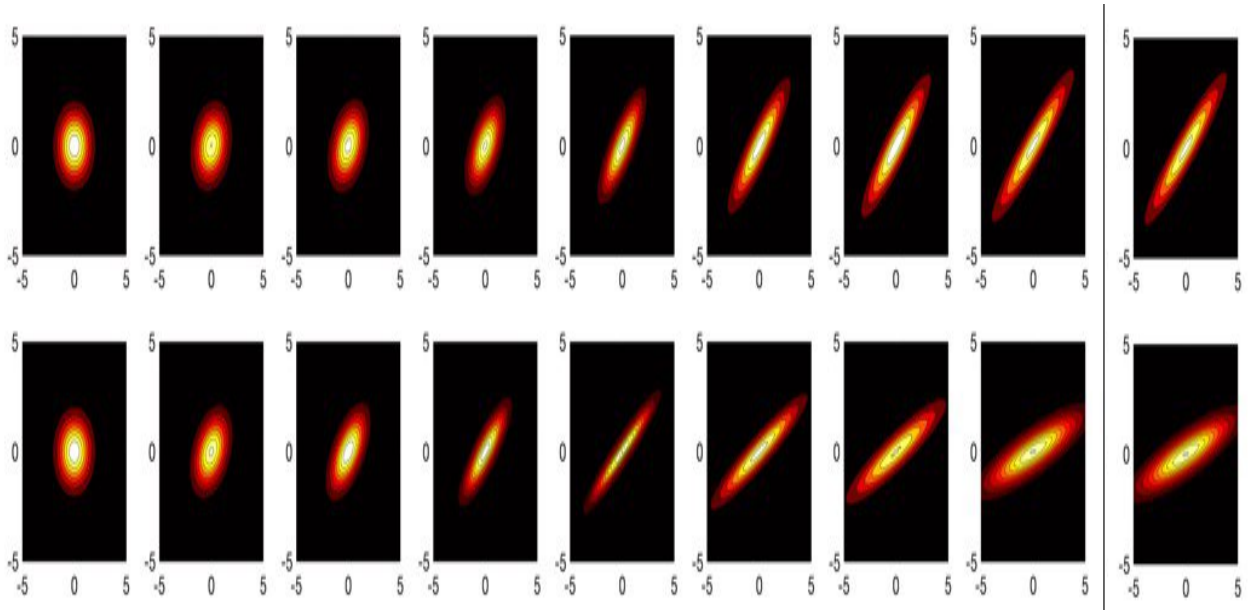


Figure 5.2: The rows exemplify the convergence of $(A_n x)_{\#}\mu_1 \rightarrow \mu_2$ and $(A_n x)_{\#}\left((A_n x)_{\#}\mu_1\right) \rightarrow \mu_3$, respectively, as $n = 1, \ldots, 8$, towards μ_2 and μ_3, which are displayed on the right and separated by a vertical line (with μ_2 on top of μ_3).

and transportation cost $W_2(\mu_0, \mu_1)$ given by

$$\sqrt{\|m_0 - m_1\|^2 + \mathrm{tr}(C_0 + C_1 - 2(C_1^{1/2} C_0 C_1^{1/2})^{1/2})} \tag{5.12}$$

where tr(.) stands for trace.

We begin with a collection $\mu_i = \mathcal{N}(0, C_i),\ i = 1, \ldots, m$ as our distributional snapshots; for simplicity we have assumed zero-means. The cost (5.3) reads

$$F(A) = \sum_{i=1}^{m-1} tr(AC_iA^T + C_{i+1} - 2(C_{i+1}^{1/2} AC_iA^T C_{i+1}^{1/2})^{1/2}). \tag{5.13}$$

The gradient $\nabla_A F(A)$, for the case of Gaussian snapshots, is expressed below directly in terms of the data C_i, $i \in \{1, \ldots, m\}$.

Proposition 5.4.1. *Given Gaussian distributions $\mu_i = \mathcal{N}(0, C_i),\ i = 1, \ldots, m$, and a non-*

singular $A \in M(d)$,

$$\nabla_A F = 2\left\{ A\sum_{i=1}^{m-1} C_i - (\sum_{i=1}^{m-1}(C_{i+1}AC_iA^T)^{1/2})A^{-T} \right\}. \tag{5.14}$$

To determine a minimizer for (5.13), we utilize a first-order iterative algorithm, taking steps proportional to the negative of the gradient in (5.14), namely,

$$A_{n+1} = A_n - \alpha\nabla_A F(A_n), \quad n = 1, 2, \ldots \tag{5.15}$$

for a small learning rate $\alpha > 0$.

As a guiding example, and for the sake of visualization, we consider the two-dimensional state-space $\mathbb{X} = \mathbb{R}^2$, in which probability measures are evolving according to linear non-deterministic dynamics,

$$x_{k+1} = \begin{bmatrix} -\frac{1}{2} & 2 \\ -1 & \frac{3}{2} \end{bmatrix} x_k + \frac{2}{5}\begin{bmatrix} \Delta\omega_k^1 \\ \Delta\omega_k^2 \end{bmatrix}, \quad k = 1, 2, \ldots$$

starting from $\mu_1 = \mathcal{N}(0, I_2)$, with I_2 a 2×2 identity matrix. We take $\Delta\omega_k^1,\ \Delta\omega_k^2 = \mathcal{N}(0, 1)$ to be independent white noise processes.

This dynamical system is an example of a first-order autoregressive process (AR(1)) which can also be thought of as an Euler-ÄìMaruyama approximation of a two-dimensional Ornstein-Uhlenbeck stochastic differential equation where $\Delta\omega_k^1$ and $\Delta\omega_k^2$ are the increments of two independent Wiener processes with unit step size.

We note that A is neither symmetric nor positive definite, which implies that it is not a "Monge map" and, thus, the flow of distributions is not a geodesic path in the Wasserstein metric.

Using the first five iterates ($m = 6$), we employ (5.15) to obtain dynamics solely on the basis of these 5 distributional snapshots. We initialize (5.15) with $\alpha = 0.1$ and experimented with different starting choices for A_1. Specifically, we took A_1 to be the identity matrix I_2, and also, the average $A_1 = \frac{1}{m-1}\sum_{i=1}^{m-1} C_i^{-1}(C_i C_{i+1})^{1/2}$, without any perceptible difference in the convergence to a minimizer. For the first choice, $A_1 = I_2$, the values of $F(A_n)$ in successive iterations is shown in Fig. 5.1.

Our data C_i ($i \in \{1, \ldots, 6\}$) is generated starting from $\mu_1 = \mathcal{N}(0, C_1)$ with $C_1 = I_2$, i.e., the 2×2 identity matrix, and the gradient search for the minimizer is initialized using $A_1 = I_2$ as well. In Fig. 5.2 we display contours of probability distributions. Specifically, on the right hand side, separated by a vertical line, we display the contours for $\mu_2 = \mathcal{N}(0, C_2)$ and $\mu_3 = \mathcal{N}(0, C_3)$, with μ_2 on top of μ_3. Then, horizontally, from left to right, we display contours corresponding to the approximating sequence of distributions. The first row exemplifies the convergence

$$(A_n x)_{\#}\mu_1 \to \mu_2,$$

whereas the second row, exemplifies the convergence

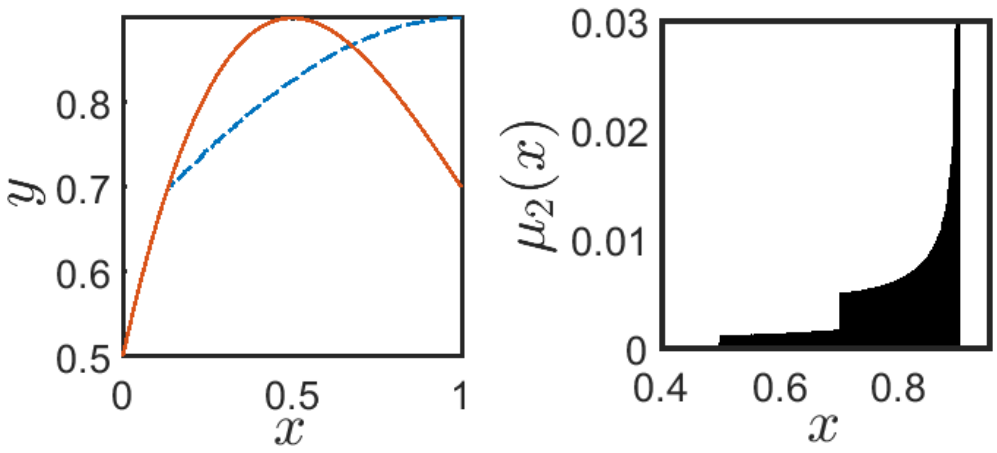
$$(A_n x)_{\#}\left((A_n x)_{\#}\mu_1\right) \to \mu_3,$$

as $n = 1, \ldots, 8$.

5.4.2 Non-linear dynamics

For our second example, to highlight the use of the approach, we consider the C^1 (continuously differentiable) map $S : x \mapsto S(x)$, on $\mathbb{X} = \mathbb{R}$ with

$$S(x) = 0.7 + 0.6(1 - x) - 0.8(1 - x)^3. \tag{5.16}$$



(a) (b)

Figure 5.3: The two maps in (a) transport a uniform distribution on $[0, 1]$ to the same discontinuous density in (b). Monge map (blue) is injective but not in C^1 everywhere. The non-injective map (red) is in C^1.

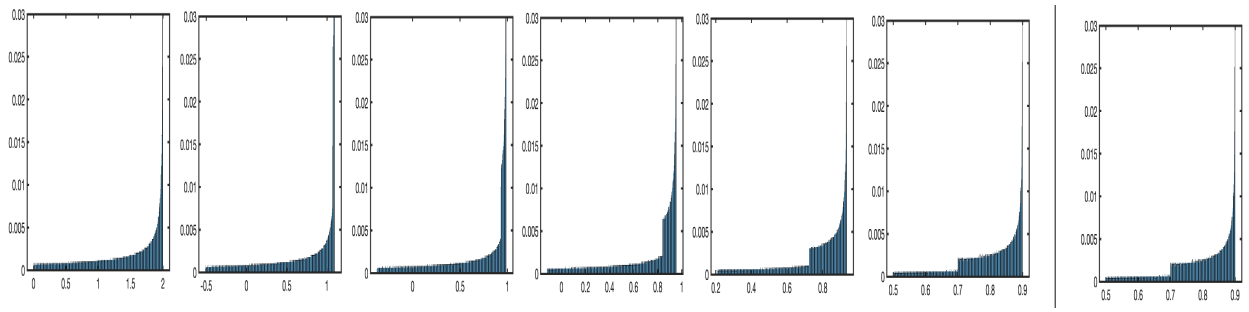


Figure 5.4: The evolution of uniform distribution under $S(x; \Theta)$ at different iterations of the algorithm. On the right-hand side the target density is depicted. In the beginning (left) no jump discontinuity is observed.

The idea for this example has been borrowed from [80]. The map S is depicted in Fig. 5.3(a), in red solid curve, and pushes forward a uniform distribution on $[0, 1]$ to distribution with discontinuous density. This density is shown in Fig. 5.3(b). Due to the fact that the density is discontinuous, the optimal transport (Monge) map has a "corner" (not smooth) and is displayed in Fig. 5.3(a), with a dashed blue curve.

The method outlined in this chapter allows us to seek a transportation map, within a suitably parametrized class of functions, that pushes forward μ_1 (here, this is the uniform distribution

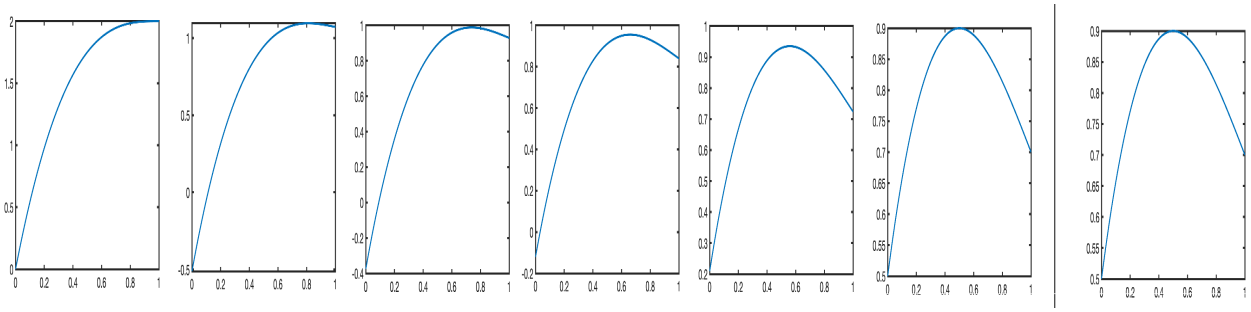


Figure 5.5: The transport map $S(x;\Theta)$ at different iterations of the algorithm. This shows the convergence to the non-injective map.

on $[0,1]$) to μ_2 displayed in Fig. 5.3(b). To this end, we select the representation

$$S(x;\Theta) = \theta_3 + \theta_2(1-x) + \theta_1(1-x)^3,$$

in the basis $Y = \{1, (1-x), (1-x)^3\}$, and seek to determine the parameters θ_k ($k \in \{1,2,3\}$) via a gradient-descent as in (5.11).

The two probability distributions are approximated using 100 sample points (drawn independently). We initialize with $\theta_1 = -2$, $\theta_2 = 0$, and $\theta_3 = 2$. A discrete optimal transport problem is solved to find the joint distributions η_i in (5.11) at each time step. The convergence is depicted in Fig. 5.5, where successive iterants are displayed from left to right below the resulting pushforward distribution. On the right hand side, separated by vertical lines, the target μ_1 is displayed above the cubic map in (5.16).

It is worth observing that, as illustrated in Fig. 5.5, our initialization corresponds to an injective map resulting in no discontinuity in the first pushforward distribution. In successive steps however, as the distributions converge to μ_1 and the maps to $S(x)$ in (5.16), a discontinuity appears tied to the non-injectivity of the maps with updated parameters.

Chapter 6

Conclusion and Future work

6.1 Conclusion

In this dissertation we focused on related concepts and strategies for obtaining dynamical models and approximating the transfer operators from data, namely, the Perron–Frobenius and Koopman operators which provide natural dual settings to investigate the dynamics of complex systems.

Chapter 2 considers the case that only a limited number of data samples are available for modeling an otherwise exceedingly high dimensional process. The dimensionality of the process, which may represent visual or distributional fields, in conjunction with the limited observation record requires careful analysis. It is precisely this regime of "small data," i.e., "few samples," that has been a challenge in traditional signal analysis since its inception [81], and has led to entropic regularization among other methodologies. DMD represents a more recent development that aims to identify suitable linear dynamics that can explain the data.

Historically, DMD has roots and ramifications that relate to theory of the Koopman operator [82–84]. Data that originate from periodic and quasi-periodic attractors of nonlinear

dynamics can also be dealt with in the same framework [4]. Thus the concept of the gap metric, as a tool to quantify how subspaces spanned by data impact modeling assumptions, is expected to be applicable in this more general setting. This chapter summarizes some of the findings in a developing treatise into the topic of extracting dynamics from high dimension distributional fields [85], specifically, the relevance of the gap metric as a tool to provide guidance in selecting appropriate dimensionality for models for such processes.

In Chapter 4 we presented an approach to estimate flow from distributional data. It can be seen as a generalization of Euclidean regression to the Wasserstein space relying on measure-valued curves. It represents a relaxation of geodesic regression in Wasserstein space. The apparently nonlinear primal problem can be recast as a multi-marginal optimal transport, leading to a formulation as a linear program. Entropic regularization and a generalized Sinkhorn algorithm can be effectively employed to solve this multi-marginal problem.

The proposed framework can be used to estimate correlation between given distributional snapshots. Potential applications of the theory are envisioned to aggregate data inference [24], estimating meta-population dynamics [20], power spectra tracking [10], and more generally, system identification [46]. The framework encompasses the case where probability laws are sought for dynamical systems, generating curves to approximate data sets.

In Chapter 5, we proposed an approach to interpolate distributional snapshots by identifying suitable underlying dynamics. It is assumed that no information on statistical dependence between successive pairs of distributions is available. The scheme we propose aims at modeling a Perron-Frobenius operator associated with underlying unknown dynamics. It is based on formulating a regression-type optimization problem in the Wasserstein metric, weighing in distances between successive distributional snapshots. A first-order necessary condition is derived that leads to a gradient-descent algorithm. The method extends to search for nonlinear dynamics assuming a suitable parametrization of the nonlinear state transition map in terms of selected basis functions. Two academic examples are presented to highlight

the approach as applied in two cases, the first specializing to Gaussian distributions and the second dealing with more general distributions (albeit with one-dimensional support for simplicity).

6.2 Future work

The present work follows a long line of contributions within the data-driven modeling field to approximate the transfer operators, see e.g. [29–31]. There is a wide range of possible applications as well as extensions of the theory that lay ahead.

- In Chapter 2 we considered the case that only a limited number of data samples are available for modeling an otherwise exceedingly high dimensional process. It was assumed that a linear mechanism underlies the data. The other potential scenario is the case where the data are observed along the trajectories of a nonlinear mechanism. In this setting, the Koopman operator plays an important role as it is a linear operator acting on the observable functions. There is a wide range of studies focusing on the approximation of Koopman operator from data using a dictionary of functions. The notion of gap metric can be employed in a more general setting to check the underlying assumptions in approximating the Koopman operator.

- In Chapter 4 we presented an approach to estimate flow from distributional data. Future research along these lines, of utilizing higher-order curves and general dynamics, should prove useful in application that may include weather prediction, modeling traffic, besides more traditional ones in computer vision.

- In Chapter 5, we proposed an approach to interpolate distributional snapshots by identifying suitable underlying dynamics. This method can be combined with deep neural networks to estimate the distribution of high-dimensional data. This is a very important problem to approximate the underlying distribution for high-dimensional

data such as images when only a set of samples is the only available data. There is a very rich literature on this topic among which we can mention the generative adversarial nets [86, 87] and normalizing flows [88, 89].

Bibliography

[1] Amirhossein Karimi, Luigia Ripani, and Tryphon T. Georgiou. Statistical learning in Wasserstein space. *IEEE Control Systems Letters*, 5(3):899–904, 2020.

[2] Amirhossein Karimi and Tryphon T. Georgiou. The challenge of small data: Dynamic mode decomposition, redux. *arXiv preprint arXiv:2104.04005*, 2021.

[3] Amirhossein Karimi and Tryphon T. Georgiou. Regression analysis of distributional data through multi-marginal optimal transport. *arXiv preprint arXiv:2106.15031*, 2021.

[4] J Nathan Kutz, Steven L Brunton, Bingni W Brunton, and Joshua L Proctor. *Dynamic mode decomposition: data-driven modeling of complex systems*. SIAM, 2016.

[5] Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data–driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

[6] Joshua L Proctor, Steven L Brunton, and J Nathan Kutz. Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems*, 15(1):142–161, 2016.

[7] Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.

[8] Soledad Le Clainche and José M Vega. Higher order dynamic mode decomposition. *SIAM Journal on Applied Dynamical Systems*, 16(2):882–925, 2017.

[9] Marc Niethammer, Yang Huang, and François-Xavier Vialard. Geodesic regression for image time-series. In *International conference on medical image computing and computer-assisted intervention*, pages 655–662. Springer, 2011.

[10] Xianhua Jiang, Zhi-Quan Luo, and Tryphon T. Georgiou. Geometric methods for spectral analysis. *IEEE Transactions on Signal Processing*, 60(3):1064–1074, 2011.

[11] Yi Hong, Roland Kwitt, Nikhil Singh, Brad Davis, Nuno Vasconcelos, and Marc Niethammer. Geodesic regression on the Grassmannian. In *European Conference on Computer Vision*, pages 632–646. Springer, 2014.

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

[13] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[14] Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. On the relation between optimal transport and schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169(2):671–691, 2016.

[15] Jérémie Bigot, Raúl Gouet, Thierry Klein, Alfredo López, et al. Geodesic PCA in the Wasserstein space by convex PCA. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 53, pages 1–26. Institut Henri Poincaré, 2017.

[16] Elsa Cazelles, Vivien Seguy, Jérémie Bigot, Marco Cuturi, and Nicolas Papadakis. Geodesic PCA versus log-PCA of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing*, 40(2):B429–B456, 2018.

[17] Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013.

[18] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Gradient flows with metric and differentiable structures, and applications to the Wasserstein space. *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl*, 15(3-4):327–343, 2004.

[19] Vivien Seguy and Marco Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems*, pages 3312–3320, 2015.

[20] Jonathan M Nichols, Jeffrey A Spendelow, and James D Nichols. Using optimal transport theory to estimate transition probabilities in metapopulation dynamics. *Ecological Modelling*, 359:311–319, 2017.

[21] Yongxin Chen and Johan Karlsson. State tracking of linear ensembles via optimal mass transport. *IEEE Control Systems Letters*, 2(2):260–265, 2018.

[22] Karthik Elamvazhuthi and Piyush Grover. Optimal transport over nonlinear systems via infinitesimal generators on graphs. *arXiv preprint arXiv:1612.01193*, 2016.

[23] Mathias Hudoba De Badyn, Utku Eren, Behçet Açikmeşe, and Mehran Mesbahi. Optimal mass transport and kernel density estimation for state-dependent networked dynamic systems. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 1225–1230. IEEE, 2018.

[24] Isabel Haasler, Axel Ringh, Yongxin Chen, and Johan Karlsson. Estimating ensemble

flows on a hidden Markov chain. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 1331–1338. IEEE, 2019.

[25] Jean-David Benamou, Thomas O Gallouët, and François-Xavier Vialard. Second-order models for optimal transport and cubic splines on the Wasserstein space. *Foundations of Computational Mathematics*, 19(5):1113–1143, 2019.

[26] Gary Froyland. An analytic framework for identifying finite-time coherent sets in time-dependent dynamical systems. *Physica D: Nonlinear Phenomena*, 250:1–19, 2013.

[27] Péter Koltai, Johannes von Lindheim, Sebastian Neumayer, and Gabriele Steidl. Transfer operators from optimal transport plans for coherent set detection. *arXiv preprint arXiv:2006.16085*, 2020.

[28] Stefan Klus, Feliks Nüske, Péter Koltai, Hao Wu, Ioannis Kevrekidis, Christof Schütte, and Frank Noé. Data-driven model reduction and transfer operator approximation. *Journal of Nonlinear Science*, 28(3):985–1010, 2018.

[29] Stefan Klus, Péter Koltai, and Christof Schütte. On the numerical approximation of the Perron-Frobenius and Koopman operator. *arXiv preprint arXiv:1512.05997*, 2015.

[30] Igor Mezić. On numerical approximations of the Koopman operator. *arXiv preprint arXiv:2009.05883*, 2020.

[31] Stefan Klus, Feliks Nüske, Sebastian Peitz, Jan-Hendrik Niemann, Cecilia Clementi, and Christof Schütte. Data-driven approximation of the Koopman generator: Model reduction, system identification, and control. *Physica D: Nonlinear Phenomena*, 406:132416, 2020.

[32] Tyrus Berry, Dimitrios Giannakis, and John Harlim. Bridging data science and dynamical systems theory. *arXiv preprint arXiv:2002.07928*, 2020.

[33] Ivo F Sbalzarini. Modeling and simulation of biological systems from image data. *Bioessays*, 35(5):482–490, 2013.

[34] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

[35] Or Yair, Mirela Ben-Chen, and Ronen Talmon. Parallel transport on the cone manifold of spd matrices for domain adaptation. *IEEE Transactions on Signal Processing*, 67(7):1797–1811, 2019.

[36] Joel A Rosenfeld, Benjamin Russo, Rushikesh Kamalapurkar, and Taylor T Johnson. The occupation kernel method for nonlinear system identification. *arXiv preprint arXiv:1909.11792*, 2019.

[37] Joel A Rosenfeld, Rushikesh Kamalapurkar, L Gruss, and Taylor T Johnson. Dynamic mode decomposition for continuous time systems with the Liouville operator. *arXiv preprint arXiv:1910.03977*, 2019.

[38] J.P. Cunningham and Z. Ghahramani. Linear dimensionality reduction: Survey, insights and generalizations. *The Journal of Machine Learning Research*, 16(1):2859–2900, 2015.

[39] Tien-Yien Li. Finite approximation for the Frobenius-Perron operator. a solution to Ulam's conjecture. *Journal of Approximation theory*, 17(2):177–186, 1976.

[40] Gary Froyland, Georg A Gottwald, and Andy Hammerlindl. A computational method to extract macroscopic variables and their dynamics in multiscale systems. *SIAM Journal on Applied Dynamical Systems*, 13(4):1816–1846, 2014.

[41] Christopher J Bose and Rua Murray. Dynamical conditions for convergence of a maximum entropy method for Frobenius–Perron operator equations. *Applied mathematics and computation*, 182(1):210–212, 2006.

[42] Jiu Ding. A maximum entropy method for solving Frobenius-Perron operator equations. *Applied mathematics and computation*, 93(2-3):155–168, 1998.

[43] Michael Dellnitz, Andreas Hohmann, Oliver Junge, and Martin Rumpf. Exploring invariant sets and invariant measures. *CHAOS: An Interdisciplinary Journal of Nonlinear Science*, 7(2):221–228, 1997.

[44] Oliver Junge and Ioannis G. Kevrekidis. On the sighting of unicorns: A variational approach to computing invariant sets in dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(6):063102, 2017.

[45] Milan Korda, Didier Henrion, and Igor Mezić. Convex computation of extremal invariant measures of nonlinear dynamical systems and Markov processes. *Journal of Nonlinear Science*, 31(1):1–26, 2021.

[46] Amirhossein Karimi and Tryphon T. Georgiou. Data-driven approximation of the Perron-Frobenius operator using the Wasserstein metric. *arXiv preprint arXiv:2011.00759*, 2020.

[47] Brendan Pass. Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1771–1790, 2015.

[48] Julie Delon and Agnès Desolneux. A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

[49] Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2018.

[50] Mihailo R Jovanović, Peter J Schmid, and Joseph W Nichols. Sparsity-promoting dy-

namic mode decomposition. *Physics of Fluids*, 26(2):024103, 2014.

[51] Caglayan Dicle, Hassan Mansour, Dong Tian, Mouhacine Benosman, and Anthony Vetro. Robust low rank dynamic mode decomposition for compressed domain crowd and traffic flow analysis. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016.

[52] GW Stewart and Ji-Guang Sun. Matrix perturbation theory academic press. *San Diego*, 1990.

[53] Tosio Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

[54] Tryphon T Georgiou. On the computation of the gap metric. *Systems & Control Letters*, 11(4):253–257, 1988.

[55] Tryphon T Georgiou and Malcolm C Smith. Optimal robustness in the gap metric. In *Proceedings of the 28th IEEE Conference on Decision and Control,*, pages 2331–2336. IEEE, 1989.

[56] Kemin Zhou and John Comstock Doyle. *Essentials of robust control*, volume 104. Prentice hall Upper Saddle River, NJ, 1998.

[57] Petre Stoica and Randolph L Moses. *Spectral analysis of signals*. Pearson Prentice Hall, NJ, 2005.

[58] Clancey R Rowley. Github repository.

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

[60] Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.

[61] Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. Wasserstein riemannian geometry of gaussian densities. *Information Geometry*, 1(2):137–179, 2018.

[62] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[63] Luca Nenna. *Numerical methods for multi-marginal optimal transportation*. PhD thesis, 2016.

[64] Jean-David Benamou, Guillaume Carlier, and Luca Nenna. Generalized incompressible flows, multi-marginal transport and Sinkhorn algorithm. *Numerische Mathematik*, 142(1):33–54, 2019.

[65] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

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

[67] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[68] Jean-David Benamou, Guillaume Carlier, and Luca Nenna. A numerical method to solve multi-marginal optimal transport problems with Coulomb cost. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 577–601. Springer, 2016.

[69] Heinz H Bauschke and Adrian S Lewis. Dykstra's algorithm with Bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

[70] Isabel Haasler, Axel Ringh, Yongxin Chen, and Johan Karlsson. Multi-marginal optimal transport and Schrödinger bridges on trees. *arXiv preprint arXiv:2004.06909*, 2020.

[71] Filip Elvander, Isabel Haasler, Andreas Jakobsson, and Johan Karlsson. Multi-marginal optimal transport using partial information with applications in robust localization and sensor fusion. *Signal Processing*, 171:107474, 2020.

[72] Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. Modern Koopman theory for dynamical systems. *arXiv preprint arXiv:2102.12086*, 2021.

[73] Stefan Klus, Feliks Nüske, Péter Koltai, Hao Wu, Ioannis Kevrekidis, Christof Schütte, and Frank Noé. Data-driven model reduction and transfer operator approximation. *Journal of Nonlinear Science*, 28(3):985–1010, 2018.

[74] Andrzej Lasota and Michael C Mackey. *Chaos, fractals, and noise: stochastic aspects of dynamics*. Springer Science & Business Media, 2013.

[75] Jiu Ding and Aihui Zhou. *Statistical properties of deterministic systems*. Springer Science & Business Media, 2010.

[76] Marko Budišić, Ryan Mohr, and Igor Mezić. Applied Koopmanism. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4):047510, 2012.

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

[78] Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

[79] 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.

[80] Caroline Moosmüller, Felix Dietrich, and Ioannis G Kevrekidis. A geometric approach to the transport of discontinuous densities. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):1012–1035, 2020.

[81] John Parker Burg, David G Luenberger, and Daniel L Wenger. Estimation of structured covariance matrices. *Proceedings of the IEEE*, 70(9):963–974, 1982.

[82] Igor Mezic and Andrzej Banaszuk. Comparison of systems with complex behavior: spectral methods. In *Proceedings of the 39th IEEE Conference on Decision and Control (Cat. No. 00CH37187)*, volume 2, pages 1224–1231. IEEE, 2000.

[83] Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1):309–325, 2005.

[84] Clarence W Rowley, IGOR Mezić, Shervin Bagheri, Philipp Schlatter, Dans Henningson, et al. Spectral analysis of nonlinear flows. *Journal of fluid mechanics*, 641(1):115–127, 2009.

[85] Amirhossein Karimi. *Statistical learning in Wasserstein space*. PhD thesis, University of California, Irvine, 2021.

[86] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[87] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[88] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[89] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.