DATA-DRIVEN APPROXIMATION OF THE PERRON-FROBENIUS OPERATOR USING THE WASSERSTEIN METRIC

A PREPRINT

Amirhossein Karimi
Department of Mechanical
and Aerospace Engineering
University of California, Irvine
California, USA
amirhosk@uci.edu

Tryphon T. Georgiou
Department of Mechanical
and Aerospace Engineering
University of California, Irvine
California, USA
tryphon@uci.edu

November 3, 2020

ABSTRACT

This manuscript introduces a regression-type formulation for approximating the Perron-Frobenius Operator by relying on distributional snapshots of data. These snapshots may represent densities of particles. The Wasserstein metric is leveraged to define a suitable functional optimization in the space of distributions. 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. The framework is exemplied with numerical simulations.

***K*eywords** Regression analysis · Perron-Frobenius Operator · Wasserstein space

1 Introduction

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 [15]. 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 [14].

Modeling and approximation of transfer operators often relies on samples of along collections of trajectories, e.g., see [5, 16, 17, 22]. 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. Herein, we are concerned with such problems where dynamics are to be inferred from data on density flows. We advance a viewpoint that leverages the geometry of *Optimal Mass Transport* (OMT) and the *Wasserstein metric* on distributions, to identify underlying dynamics.

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 [27]. Another instance is domain adaptation, which aims at finding a model on a *target data* distribution, by training on a *source data* distribution [10, 34].

A popular and effective method in identifying dynamics using snapshots of data is due to Schmid and Sesterhenn (2008) and is known as DMD (*Dynamic Mode Decomposition*). Their algorithm aimed at modeling time-series measurements of fluid flow data [28]. The connection between DMD and the Koopman operator was pointed out and

discussed in [26]; a reformulation as a least-squares regression problem was proposed in 2014 [29] and an extension, referred to as *extended DMD*, for approximating the eigenvalues and eigenfunctions of the Koopman operator was proposed in [32, 33].

The Liouville operator [25] is another example of a linear operator associated with non-linear dynamics; this is the infinitesimal generator for the Koopman operator [24]. 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 [11].

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 [13, 18]. 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 [6, 12, 14]). 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 [15]. 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.

In this paper, 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. 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) [4, 9, 31].

The paper is organized as follows. Notation and preliminaries on transfer operators are presented in Section 2, and rudiments of Wasserstein geometry needed for the development of the method are explained in Section 3. These tools are then used in Section 4 to derive a first-order necessary condition for two different approximations of Perron-Frobenius operator. Further, a gradient-descent approach in finding sought parameters in a system identification setting is presented. The proposed framework is highlighted via two numerical examples in Section 5.

2 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 [14].

2.1 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.

2.2 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 paper, 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.

2.3 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

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

see e.g., [8]. 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,

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

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 component-wise.

2.4 Data-driven approximation of transfer operators

As mentioned earlier, the most popular method in the literature to discretize PFO is the Ulam's method [13, 18]. 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) [32], 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 $\Phi = [\phi_1 \; \phi_2 \; \ldots \; \phi_k]^T$. We stack up the values of this function at the snapshots in two matrices as

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

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

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

Depending on the values of m and k, the system of equations (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.

3 Rudiments of Wasserstein space

In this section, we recall the definition and some properties of the Wasserstein distance [1, 30], which are used in this paper.

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 $T^* = \nabla\phi(x)$ [7].

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 [2, 31]. Moreover, assuming that T^* exists, the constant-speed geodesic between μ_0 and μ_1 is given by

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

and known as *McCann's displacement interpolation* [21].

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

Lemma 1 (Gluing lemma [2,31]) *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.

4 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 4.1, 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 4.2, 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, $j \in \{1, \ldots, n\}$.

4.1 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. (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

$$[\mu_2 \; \mu_3 \; \ldots \; \mu_m] = (Ax)_{\#}[\mu_1 \; \mu_2 \; \ldots \; \mu_{m-1}]. \tag{2}$$

In (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 (2), we consider the following regression-type formulation.

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

$$F(A) = \sum_{i=1}^{m-1} W_2^2(Ax_{\#}\mu_i, \mu_{i+1}). \tag{3}$$

If (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 (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 2 *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* (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

$$\sum_{i=1}^{m-1} \int_{\mathbb{X}\times\mathbb{X}} (Ax_i - x_{i+1})x_i^T d\eta_i(x_i, x_{i+1}) = 0. \tag{4}$$

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 2: According to the assumption that A is a minimizer of (3), the Fermat's condition

$$\frac{d}{d\epsilon} F(A + \epsilon\delta A)|_{\epsilon=0} = 0 \tag{5}$$

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 (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 $\delta A \in M(d)$

$$\begin{aligned} &G(A + \epsilon\delta A) - G(A) \leq \\ &\langle \int_{\mathbb{X}\times\mathbb{X}} 2(Ax_1 - x_2)x_1^T d\eta_1(x_1, x_2), \epsilon\delta A\rangle_F + O(\epsilon^2) \end{aligned} \tag{6}$$

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, (6) is derived.

Without loss of generality we take $\epsilon > 0$. According to (6), we can readily conclude that

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

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 [3, Proposition 7.3.6].

By combining the "$\liminf$" and "$\limsup$" results, it readily follows that

$$\frac{d}{d\epsilon}G(A+\epsilon\delta A)|_{\epsilon=0} = \tag{7}$$
$$\langle \int_{\mathbb{X}\times\mathbb{X}} 2(Ax_1-x_2)x_1^T d\eta_1(x_1,x_2), \delta A\rangle_F.$$

Finally, writing the directional derivative for all the terms in (3) and using Fermat's condition the proof is complete. □

Remark 1 In the statement of Theorem 2 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 [3]), 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 2 Equation (7) shows how to generate a gradient flow, and thereby a steepest descent direction for minimizing $F(A)$. Specifically,

$$\nabla_A F(A) = 2\sum_{i=1}^{m-1}\int_{\mathbb{X}_i\times\mathbb{X}_{i+1}} (Ax_i-x_{i+1})x_i^T d\eta_i(x_i,x_{i+1}), \tag{8}$$

allows us to construct a gradient-type numerical optimization to find the minimizer of (3). □

Remark 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. □

4.2 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 $j\in\{1,\dots,n\}$, i.e.,

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

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

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

$$F(\Theta) = \sum_{i=1}^{m-1} W_2^2(S(x;\Theta)_{\#}\mu_i,\mu_{i+1}), \tag{9}$$

over $\Theta\in\mathbb{R}^n$. We follow a strategy that is similar to that in the proof of Theorem 2, to derive a first-order optimality condition for Θ in the form

$$\sum_{i=1}^{m-1}\int_{\mathbb{X}\times\mathbb{X}} (Y(x_i))^T(S(x_i;\Theta)-x_{i+1})d\eta_i(x_ix_{i+1}) = 0. \tag{10}$$

Here, $Y(x_i) = [y_1(x_i) \ \dots \ y_n(x_i)] \in \mathbb{R}^{d\times n}$ 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 (10) extends our formalism to nonlinear dynamics, parametrized by the span of Y, for approximating the PFO. In a way similar to (8), we consider the gradient of $F(\Theta)$ in (9) with respect to Θ,

$$\nabla_\Theta F = 2\sum_{i=1}^{m-1}\int_{\mathbb{X}\times\mathbb{X}} (Y(x_i))^T(S(x_i;\Theta) - x_{i+1})d\eta_i(x_i, x_{i+1}), \tag{11}$$

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

5 Simulation results

5.1 Gaussian distributions

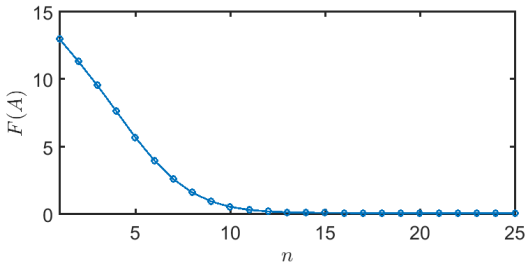


Figure 1: Value $F(A_n)$ as a function of iterated steps in (15).

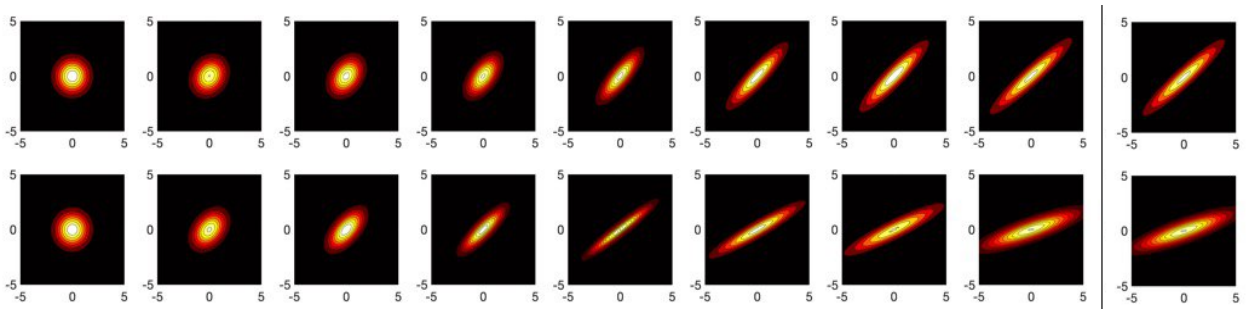


Figure 2: The rows exemplify the convergence of $(A_n x)_{\#}\mu_1 \to \mu_2$ and $(A_n x)_{\#}\left((A_n x)_{\#}\mu_1\right) \to \mu_3$, respectively, as $n = 1, \dots, 8$, towards μ_2 and μ_3, which are displayed on the right and separated by a vertical line (with μ_2 on top of μ_3).

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 [19, 20], with transportation (Monge) map

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

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

$$\sqrt{\|m_0 - m_1\|^2 + \mathrm{tr}(C_0 + C_1 - 2(C_1^{1/2}C_0C_1^{1/2})^{1/2})} \tag{12}$$

where $\mathrm{tr}(.)$ stands for trace.

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

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 (3) reads

$$F(A) = \sum_{i=1}^{m-1} tr(AC_iA^T + C_{i+1} - 2(C_{i+1}^{1/2} AC_iA^T C_{i+1}^{1/2})^{1/2}). \tag{13}$$

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 3 *Given Gaussian distributions $\mu_i = \mathcal{N}(0, C_i),\ i = 1, \ldots, m$, and a non-singular $A \in M(d)$,*

$$\nabla_A F = 2 \left\{ A \sum_{i=1}^{m-1} C_i - (\sum_{i=1}^{m-1} (C_{i+1} AC_iA^T)^{1/2}) A^{-T} \right\}. \tag{14}$$

To determine a minimizer for (13), we utilize a first-order iterative algorithm, taking steps proportional to the negative of the gradient in (14), namely,

$$A_{n+1} = A_n - \alpha \nabla_A F(A_n),\ \ n = 1, 2, \ldots \tag{15}$$

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,

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

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-Maruyama 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 (15) to obtain dynamics solely on the basis of these 5 distributional snapshots. We initialize (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. 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. 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

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

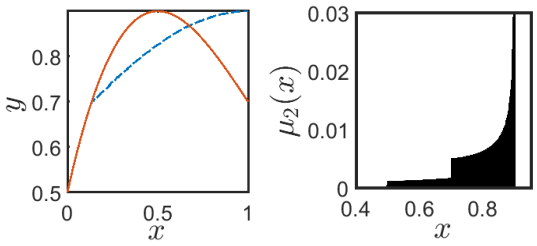
as $n = 1, \ldots, 8$.

5.2 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

$$S(x) = 0.7 + 0.6(1 - x) - 0.8(1 - x)^3. \tag{16}$$

The idea for this example has been borrowed from [23]. The map S is depicted in Fig. 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. 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. 3(a), with a dashed blue curve.



(a) (b)

Figure 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.

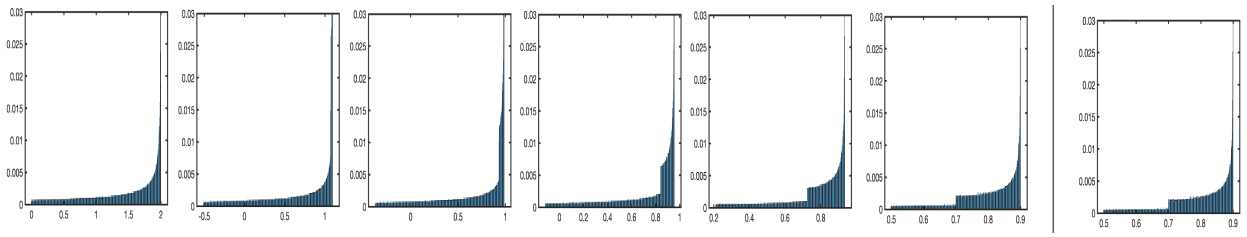


Figure 4: The evolution of uniform distribution under $S(x; \Theta)$ at different iterations of the algorithm. On the right-hand side the target density is depicted. In the beginning (left) no jump discontinuity is observed.

The method outlined in this paper 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 on $[0, 1]$) to μ_2 displayed in Fig. 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 (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 (11) at each time step. The convergence is depicted in Fig. 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 (16).

It is worth observing that, as illustrated in Fig. 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 (16), a discontinuity appears tied to the non-injectivity of the maps with updated parameters.

6 Concluding remarks

We presented 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

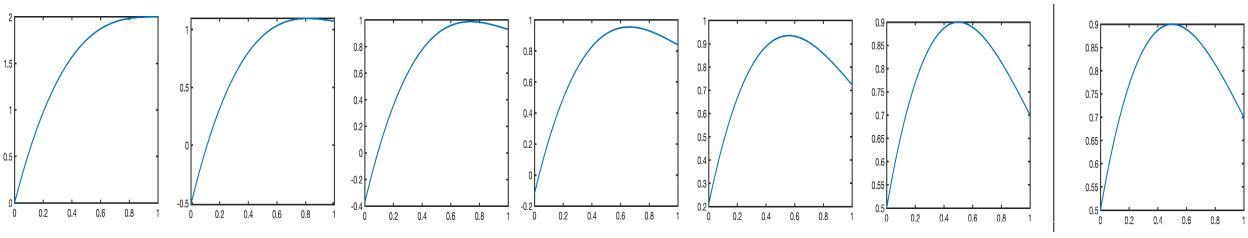


Figure 5: The transport map $S(x; \Theta)$ at different iterations of the algorithm. This shows the convergence to the non-injective map.

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).

References

[1] 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.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Gradient flows with metric and differentiable structures, and applications to the Wasserstein space. *Atti Accad. Naz. Lincei, Mat. Appl*, 15:327–343, 2004.

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

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

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

[6] 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.

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

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

[9] 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.

[10] 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.

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

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

[13] 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.

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

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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.

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

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

[23] 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.

[24] 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.

[25] 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.

[26] Clarence W Rowley, Igor Mezić, Shervin Bagheri, Philipp Schlatter, and Dan S Henningson. Spectral analysis of nonlinear flows. *Journal of fluid mechanics*, 641(1):115–127, 2009.

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

[28] Peter Schmid and Joern Sesterhenn. Dynamic mode decomposition of numerical and experimental data. *APS*, 61:MR–007, 2008.

[29] Jonathan H Tu, Clarence W Rowley, Dirk M Luchtenburg, Steven L Brunton, and J Nathan Kutz. On dynamic mode decomposition: Theory and applications. *arXiv preprint arXiv:1312.0041*, 2013.

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

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

[32] 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.

[33] Matthew O Williams, Clarence W Rowley, and Ioannis G Kevrekidis. A kernel-based approach to data-driven Koopman spectral analysis. *arXiv preprint arXiv:1411.2260*, 2014.

[34] 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.