SIAM J. Applied Dynamical Systems
Vol. 22, No. 1, pp. 269–310

Optimal Transport for Parameter Identification of Chaotic Dynamics via Invariant Measures*

Yunan Yang†, Levon Nurbekyan‡, Elisa Negrini§, Robert Martin¶, and Mirjeta Pasha‖

Abstract. We study an optimal transportation approach for recovering parameters in dynamical systems with a single smoothly varying attractor. We assume that the data are not sufficient for estimating time derivatives of state variables but enough to approximate the long-time behavior of the system through an approximation of its physical measure. Thus, we fit physical measures by taking the Wasserstein distance from optimal transportation as a misfit function between two probability distributions. In particular, we analyze the regularity of the resulting loss function for general transportation costs and derive gradient formulas. Physical measures are approximated as fixed points of suitable PDE-based Perron–Frobenius operators. Test cases discussed in the paper include common low-dimensional dynamical systems.

Key words. dynamical system, parameter identification, optimal transportation, Wasserstein metric, continuity equation, inverse problems

MSC codes. 37M21, 49Q22, 82C31, 34A55, 65N08, 93B30

DOI. 10.1137/21M1421337

1. Introduction.

The problem of parameter identification in dynamical systems is common in many areas of science and engineering, such as signal processing [30], optimal control [34, 56], secure communications [64, 30], as well as biology [63, 36], to mention a few. The main idea of parameter identification for a dynamical system is to identify a mathematical model of the real-world system and adapt its parameters until the simulations obtained with the mathematical model are close to experimental data. The models usually represent time-dependent processes with numerous state variables and many interactions between variables. In many applications, one can derive the form of the mathematical model from some knowl-

* Received by the editors May 24, 2021; accepted for publication (in revised form) by I. Belykh July 19, 2022; published electronically February 16, 2023.

https://doi.org/10.1137/21M1421337

Funding: The first author gratefully acknowledges the support by National Science Foundation through grant number DMS-1913129. The second author was partially supported by AFOSR MURI FA 9550 18-1-0502 grant. The third author acknowledges that results in this paper were obtained in part using a high-performance computing system acquired through NSF MRI grant DMS-1337943 to WPI. The fourth author was partially supported by AFOSR Grants FA9550-20RQCOR098 (PO: Leve) and FA9550-20RQCOR100 (PO: Fahroo).

† Department of Mathematics, Cornell University, Ithaca, NY 14850 USA (yy837@cornell.edu).

‡ Department of Mathematics, UCLA, Los Angeles, CA 90095 USA (lnurbek@math.ucla.edu).

§ Institute for Pure and Applied Mathematics, Los Angeles, CA 90095 USA (enegrini@wpi.edu).

¶ U.S. Army Research Office, Research Triangle Park, Durham, NC 27709 USA (robert.s.martin163.civ@army.mil).

‖ Department of Mathematics, Tufts University, Medford, MA 02155 USA (mirjeta.pasha@tufts.edu).

edge about the process under investigation but, in general, the parameters of such a model must be inferred from empirical observations of time series data. The initial parameter values are usually based on, for instance, some preliminary knowledge of the real-world system. The type of mathematical model and the parameter identification algorithm chosen strongly influence the accuracy of the estimates.

More formally, suppose that we have noisy observations

$$\mathbf{X}^* = (\mathbf{x}^*(t_0) + \eta_0, \mathbf{x}^*(t_1) + \eta_1, \dots, \mathbf{x}^*(t_n) + \eta_n)\,,$$

where $\{t_0, t_1, \dots, t_n\}$ are sampling times, $\mathbf{x}^*$ is the solution of the autonomous dynamical system $\dot{\mathbf{x}} = v(\mathbf{x}, \theta^*)$, and $\{\eta_0, \eta_1, \dots, \eta_n\}$ are measurement errors or uncertainties. The goal is to find θ^* from $\mathbf{X}^*$.

Most common parameter estimation techniques estimate θ by integrating $\dot{\mathbf{x}} = v(\mathbf{x}, \theta)$ and fitting the resulting trajectory $\mathbf{X}(\theta) = (\mathbf{x}(t_0), \mathbf{x}(t_1), \dots, \mathbf{x}(t_n))$ to data $\mathbf{X}^*$ via optimization

$$\inf_{\theta \in \Theta} \|\mathbf{X}(\theta) - \mathbf{X}^*\|^2$$

for a suitably chosen norm $\|\cdot\|$. For a linear map $\theta \mapsto v(\mathbf{x}, \theta)$ and an L^2 norm, the problem above reduces to the least-squares problem that tends to overfit measurement errors [48, 45]. For a nonlinear map $\theta \mapsto v(\mathbf{x}, \theta)$, this approach leads to a so-called single shooting method [55] that uses a single initial condition to produce a trajectory. However, relying only on one trajectory may not result in meaningful approximations of the desired solution for chaotic systems due to their sensitivity to initial data. The multiple shooting algorithm deals with this issue by using multiple trajectories to estimate parameters [7]. For a more complete review we refer to [1] and [52]. Because of their universal approximation properties, neural networks and combinations of the above methods with neural networks have also been used recently for parameter identification of dynamical systems [8, 51, 58, 57].

An alternative approach is to fit the time derivatives of the state. More precisely, assume that $\dot{\mathbf{x}}^*$ is either measured directly or estimated from $\mathbf{X}^*$ yielding

$$\mathbf{V}^* = (\dot{\mathbf{x}}^*(t_0) + \xi_0, \dot{\mathbf{x}}^*(t_1) + \xi_1, \dots, \dot{\mathbf{x}}^*(t_n) + \xi_n)\,,$$

where $\{\xi_0, \xi_1, \dots, \xi_n\}$ are measurement or estimation errors. The parameter estimation is then performed via an optimization problem

$$\inf_{\theta \in \Theta} \|\mathbf{V}^* - v(\mathbf{X}^*, \theta)\|^2 + R(\theta)$$

for a suitably chosen norm $\|\cdot\|$ and a regularization $R(\theta)$, where we denote $v(\mathbf{X}^*, \theta) = (v(\mathbf{x}^*(t_0), \theta), v(\mathbf{x}^*(t_1), \theta), \dots, v(\mathbf{x}^*(t_n), \theta))$ by slightly abusing the notation. Sparse identification of nonlinear dynamics [16] is one such notable method, where one has a linear model $v(\mathbf{x}, \theta) = \sum_i \theta_i \psi_i(\mathbf{x})$ with a suitably chosen dictionary of basis functions $\{\psi_i\}$ and a sparsity enforcing regularization term $R(\theta) = \|\theta\|_1$.

We are interested in parameter estimation problems where trajectories are sensitive to initial conditions and estimation parameters. In particular, we consider the case where the time derivatives $\mathbf{V}^*$ cannot be estimated due to the lack of observational data, slow sampling, discontinuous or inconsistent time trajectories, and noisy measurements [15, 11, 72, 65, 5]. The

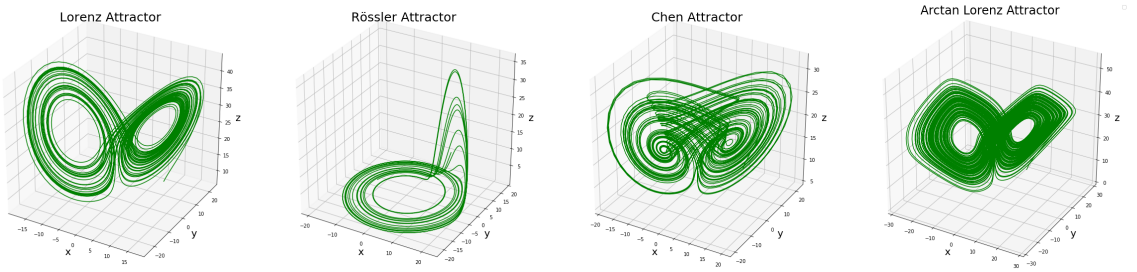


Figure 1. *From left to right: the Lorenz, Rössler, Chen, and Arctan Lorenz attractors.*

methods described above incur many challenges or are inapplicable in such settings. Hence, following [41], we "suppress" the time variable and consider the state-space distribution of data

$$\rho^* = \frac{1}{n+1}\sum_{i=0}^{n} \delta_{\mathbf{x}^*(t_i)}.$$

We say that a dynamical system $\dot{\mathbf{x}} = v(\mathbf{x}, \theta)$ admits a *physical measure* $\rho(\theta)$ [73, Definition 2.3], [53, section 9.3], if for a Lebesgue positive set of initial conditions $\mathbf{x}(0) = x$, one has that

$$\rho(\theta) = \lim_{T\to\infty} \frac{1}{T}\int_0^T \delta_{\mathbf{x}(t)} dt.$$

Therefore, as an alternative, we can fit physical measures instead of trajectories for systems admitting such measures. The convergence of ρ^* to $\rho(\theta)$ highly depends on the data availability and the fractal dimension of the attractor. Here, we assume that the observed trajectory $\mathbf{X}^*$ provides a reasonable estimation of $\rho(\theta)$. This assumption might be too restrictive for systems with high-dimensional attractors. Nevertheless, numerous systems appearing in physics, biology, and other fields admit low-dimensional attractors, for example, observed in [50] and [37, Figures 1(B), S1].

In this work, we focus on dynamical systems with a unique physical measure. More precisely, the parameter estimation problem reduces to the optimization problem

$$\inf_{\theta\in\Theta} f(\theta) := d(\rho_\epsilon(\theta), \rho^*), \tag{1.1}$$

where $\rho_\epsilon(\theta)$ is an approximation of $\rho(\theta)$ with an approximation (regularization) parameter $\epsilon > 0$, and d is a suitable metric in the space of probability measures.

Note that the definition of physical measures reflects their stability with respect to perturbations of initial conditions. Additionally, ρ^* can provide an accurate estimate of $\rho(\theta^*)$ even if we perform slow sampling; that is, when the time derivatives $\mathbf{V}^*$ cannot be estimated (subsection 6.2.4).

The difficulty and efficiency of the parameter estimation problem (1.1) depend significantly on the choice of the approximation method ρ_ϵ and the metric d. The Wasserstein metric from optimal transportation (OT) [71] has recently gained popularity as a metric of choice

in numerous fields such as image processing [42], machine learning [6], large-scale inverse problems [28], and statistical inference [9], only to mention a few. Interested readers may further refer to [60]. The Wasserstein metric is beneficial for several reasons. First, it is well-defined for singular measures and, unlike the Kullback—Leibler divergence, reflects both the local intensity differences and the global geometry mismatches [28]. Additionally, the L^p and total variation norms lead to weakly pronounced minima with small basins of attraction when the supports are disjoint or only partially intersect. Second, recent works in both deterministic and Bayesian inverse problems have demonstrated that the Wasserstein metric is robust to noise [27, 24]. Thanks to the geometric nature of the OT problem, the Wasserstein metric is primarily sensitive to global changes such as translation and dilation and is robust to small local perturbations such as noisy measurements of ρ^*. Since we are primarily interested in problems where the latter occurs, the Wasserstein metric is an adequate choice.

Our first main goal of this work is to study OT distances as the objective function for parameter identification problems in dynamical systems building on insights from [41]. An important element of the method (1.1) is the surrogate model $\rho_\epsilon(\theta)$. In [41], the authors build a histogram from a single long-time trajectory, where ϵ is the bin width. Although effective, one drawback of this approximation method is the inability to differentiate $\rho_\epsilon(\theta)$ with respect to θ. Consequently, it relies on a potentially slow derivative-free optimization method to solve (1.1). Our second main goal is to explore an alternative scheme for the approximation $\rho_\epsilon(\theta)$ that is differentiable in θ, and rigorously study the regularity of $f(\theta)$ in (1.1). One can then devise more efficient gradient-based optimization algorithms to solve (1.1).

In this work, we propose a partial differential equation (PDE)-based approximation method for $\rho(\theta)$. Note that $\rho(\theta)$ is a distributional solution of the stationary continuity PDE

$$-\nabla \cdot (v(\mathbf{x}, \theta)\rho(\mathbf{x})) = 0. \tag{1.2}$$

Hence, we consider a regularized solution $\rho_\epsilon(\theta)$ of (1.2) and turn (1.1) into a PDE-constrained optimization problem. We choose the teleportation regularization from Google's PageRank algorithm [39] because of its simplicity in implementation and other favorable properties such as the uniqueness, absolute continuity, and differentiability (with respect to θ) of $\rho_\epsilon(\theta)$. The numerical method for computing $\rho_\epsilon(\theta)$ is based on its representation as a fixed point of a suitable Perron–Frobenius operator.

Approximating physical measures by PDE and fixed points of Perron–Frobenius operators instead of directly simulating single long-time trajectories is not new [23, 3]. Some of these methods come with rigorous convergence guarantees, especially for uniformly hyperbolic systems [23, Theorem 4.14], and are more computationally efficient because of considering $\rho_\epsilon(\theta)$ that are supported on tight covers of $\mathrm{supp}(\rho)$ [23, section 4]. However, the differentiability of the resulting approximations with respect to the parameters is unclear and warrants separate careful analyses. Here, we do not analyze the convergence of $\rho_\epsilon(\theta)$ to $\rho(\theta)$, but the numerical evidence in subsection 6.2.5 and the discussion in section 3 suggest that this convergence occurs for a suitable class of dynamical systems. Instead, we focus on studying the properties of OT-based distances and the viability of the overall approach at the expense of employing a less accurate yet more straightforward approximation method for the differentiability analysis. Thus, our work serves as a foundation for possibly other OT-based techniques with different but differentiable approximation methods for the physical measures. Formally, we assume

that (1) the dynamical system of interest, $\dot{x} = v(x, \theta)$, where $\theta \in \Theta$, has one unique physical invariant measure, and (2) the distributional solution to (1.2) with the same $v(x, \theta)$ is unique and recovers the physical invariant measure to the dynamical system. We refer to section 3 for more details.

The discussion above leads to our next essential contribution: the regularity analysis of the optimal transport cost with respect to the inference parameter for generic cost functions; see section 4. Although the gradient formula is well known in the literature, its validity analysis seems to be missing except in special cases where the optimal transport cost can be calculated explicitly [61, Lemma 2.4]. In the nonparametric setting, such analysis can be found in [68, Theorem 2.4] for probability measures on finite spaces and [66, Proposition 7.17] for probability measures on $\mathbb{R}^d$. For probability measures modeled by push-forward maps, see [6].

Similarly to related results in the literature, we rely on Kantorovich's formulation of the OT problem and the regularity theory of optimal value functions [12]. Under rather mild conditions, we prove that the transportation cost is directionally differentiable everywhere. In general, the directional derivative is nonlinear and depends on the structure of Kantorovich potentials. To this end, we find a sufficient condition in terms of the geometry of the optimal transport plans that guarantees the linearity of the directional derivative providing a descent direction for the optimal transport cost. To the best of our knowledge, this condition is new in the literature.

The paper is arranged as follows. In section 2, we review challenges of chaotic dynamics, the advantages provided by a PDE perspective (1.2), and a short introduction to optimal transport. In section 3, we describe a regularized forward problem based on the PDE perspective and discuss a numerical scheme that enforces positivity and strict mass conservation. The solution to the forward problem is computed as finding the dominant eigenvector of a Markov matrix. In section 4, we present theoretical regularity analysis for evaluating gradients of optimal transport costs with respect to the model parameters. In section 5, we introduce two different ways to compute gradients for our PDE-constrained optimization problem using the implicit function theorem and the adjoint-state method. Numerical results for the Lorenz, Rössler, and Chen systems are presented in section 6. In section 7, we summarize our results and describe several future research directions.

2. Background.

In this section, we present essential background of dynamical systems and OT theory.

2.1. Dynamical systems.

This section reviews some basic terminology in the field of dynamical systems that will appear throughout the paper.

2.1.1. Chaotic dynamical systems.

A continuous-time dynamical system represents the behavior of a system in which the time-dependent flow of a point in a geometrical state space, $\mathbf{x}$, is governed by a function of that state, $v(\mathbf{x})$, such that

$$\frac{d\mathbf{x}}{dt} = \dot{\mathbf{x}} = v(\mathbf{x}), \tag{2.1}$$

in the form of an ordinary differential equation (ODE).

While linear first-order dynamical systems, $\dot{\mathbf{x}} = A\mathbf{x}$, admit only stable, unstable, and periodic solutions, the more general class of nonlinear dynamical systems can exhibit a range

of more complex long-time behaviors due to locally bounded regions of instability. It is this local region of instability that enables the emergence of chaotic behavior.

While a formal definition of chaos remains elusive, it is generally characterized by sensitive dependence to initial conditions, and expansivity; see [25] for more details. In particular, it is this sensitive dependence on initial conditions that results in the apparent randomness characteristic of chaotic systems. This randomness results from a combination of local instability causing exponential divergence of nearby trajectories and state-space mixing that occurs when this exponential divergence is restabilized such that a nontrivial attractor forms. This combination makes long-time predictions impossible despite the purely causal nature of the governing system. It is also this sensitivity that makes the classical trajectory-based parameter inference problem challenging when the observed dynamics are obscured by noise, slow sampling, and other corruption, as described in section 1.

2.1.2. From trajectory samples to the physical measure. We shift from the trajectory-based to distribution-based perspective to remedy the aforementioned stability and data availability issues. Mathematically, statistical properties of (2.1) can be characterized by the occupation measure $\rho_{x,T}$ defined as

$$\rho_{x,T}(B) = \frac{1}{T}\int_0^T \mathbb{1}_B(\mathbf{x}(s))ds = \frac{\int_0^T \mathbb{1}_B(\mathbf{x}(s))ds}{\int_0^T \mathbb{1}_{\mathbb{R}^d}(\mathbf{x}(s))ds}, \tag{2.2}$$

where $T > 0$, $\mathbb{1}$ is the indicator function, B is any Borel measurable set, and $\mathbf{x}(\cdot)$ is the time-dependent trajectory starting at x. System (2.1) has robust statistical properties if there exist a set of positive Lebesgue measure U and an invariant probability measure ρ such that $\rho_{x,T}$ converges weakly to ρ for all initial conditions $x \in U$. Such ρ are called *physical* [73, Definition 2.3], [53, section 9.3]. *Sinai–Ruelle–Bowen measures* [23, 73, 53] are archetypal examples of physical measures.

In general, the existence and properties of such measures are rather intricate and require careful analysis. For a more detailed account of these topics, we refer to [73] for general systems, and [70, 69] for the Lorenz system. Furthermore, in some cases, one can recover ρ as the zero-noise limit of stationary measures of the corresponding stochastic dynamical systems [18, 43, 47, 23].

As we will show in section 3, direct simulation of ρ for parameter identification faces the difficulty of not having access to the gradients of the loss function. Consequently, one has to rely on gradient-free space-search methods. Motivated by these challenges, we take a PDE perspective on ρ and formulate the parameter inference problem as a PDE-constrained optimization.

2.2. Optimal transportation. In this subsection, we give a brief overview of the topic of OT, first brought up by Monge in 1781.

We first introduce the original Monge's problem. Let $\Omega \subset \mathbb{R}^d$ be an arbitrary domain, and $\mu, \nu \in \mathscr{P}(\Omega)$ arbitrary probability measures supported in Ω. A transport map $T : \Omega \to \Omega$ is mass preserving if for any measurable set $B \subseteq \Omega$

$$\mu(T^{-1}(B)) = \nu(B).$$

If this condition is satisfied, ν is said to be the push-forward of μ by T, and we write $\nu = T_\sharp \mu$. In the case μ, ν are absolutely continuous, that is, $d\mu(x) = f(x)dx$ and $d\nu(y) = g(y)dy$, we have that T is a mass-preserving map if

$$f(x) = g(T(x)) \cdot |\det(\nabla T(x))|, \quad x \in \Omega.$$

The transport cost function $c(x, y)$ maps pairs $(x, y) \in \Omega \times \Omega$ to $\mathbb{R} \cup \{+\infty\}$, which denotes the cost of transporting one unit mass from location x to y. The most common choice of $c(x, y)$ is $|x - y|^p$, $p \in \mathbb{N}$, where $|x - y|$ denotes the Euclidean distance between vectors x and y. Given a mass-preserving map T, the total transport cost is

$$\int_\Omega c(x, T(x)) f(x)\, dx.$$

While there are many maps T that can perform the relocation, we are interested in finding the optimal map that minimizes the total cost. So far, we have informally defined the optimal transport problem, which induces the so-called Wasserstein distance defined below, associated with cost function $c(x, y) = |x - y|^p$.

Definition 2.1 (the Wasserstein distance). *We denote by $\mathscr{P}_p(\Omega)$ the set of probability measures with finite moments of order p. For all $p \in [1, \infty)$,*

$$W_p(\mu, \nu) = \left(\inf_{T_{\mu,\nu} \in \mathcal{M}} \int_\Omega |x - T_{\mu,\nu}(x)|^p \, d\mu(x) \right)^{\frac{1}{p}}, \quad \mu, \nu \in \mathscr{P}_p(\Omega), \tag{2.3}$$

where $\mathcal{M}$ is the set of all maps that push-forward μ into ν.

The definition (2.3) is the original static formulation of the optimal transport problem with a specific cost function. In mid-20th century, Kantorovich relaxed the constraints, turning it into a linear programming problem, and also formulated the dual problem [66]. Instead of searching for a map T, a transport plan π is considered, which is a measure supported in the product space $\Omega \times \Omega$. The Kantorovich problem is to find an optimal transport plan as follows:

$$\mathcal{T}_c(\mu, \nu) = \inf_\pi \left\{ \int_{\Omega \times \Omega} c(x, y) d\pi \mid \pi \geq 0 \text{ and } \pi \in \Pi(\mu, \nu) \right\}, \tag{2.4}$$

where $\Pi(\mu, \nu) = \{\pi \in \mathscr{P}(\Omega \times \Omega) \mid (P_1)_\sharp \pi = \mu, (P_2)_\sharp \pi = \nu\}$. Here, $\mathscr{P}(\Omega \times \Omega)$ stands for the set of all the probability measures on $\Omega \times \Omega$, functions $P_1(x, y) = x$ and $P_2(x, y) = y$ denote projections over the two coordinates, and $(P_1)_\sharp \pi$ and $(P_2)_\sharp \pi$ are two measures obtained by pushing forward π with these two projections.

Since every transport map determines a transport plan of the same cost, Kantorovich's problem is weaker than the original Monge's problem. If the cost function $c(x, y)$ is of the form $|x - y|^p$ and μ and ν are absolutely continuous with respect to the Lebesgue measure, solutions to the Kantorovich and Monge problems coincide under certain conditions. When $p > 1$, the strict convexity of $|x - y|^p$ guarantees that there is a unique solution to Kantorovich's problem (2.4) which is also the unique solution to Monge's problem (2.3).

3. The forward model. While matching shadow state-space density in [41] provided a potential route to resolve issues related to the chaotic divergence of state-space trajectories and data availability, the direct estimation of state-space density from trajectory data still retained two major challenges. One significant issue was the inability to efficiently calculate a gradient of the Wasserstein metric with respect to the parameters, forcing the reliance on evolutionary or other gradient-free optimization methods. Another major issue was related to the time required to converge to the density estimate asymptotically, as particularly highlighted in [41, Figure 7], where the self-Wasserstein metric is observed to oscillate as it converges with more ODE time steps. This slow convergence is related to the long and intermittent switching times between lobes of the butterfly attractor. While the invariant measure of the Lorenz system is known to exist [69], the long measurement times with respect to the switching times complicate the parameter inference problem. The problem is exacerbated in more expensive and complicated dynamics such as the thruster model [41].

To address these challenges, we instead directly solve for the solution of the stationary continuity equation (1.2). This choice not only removes the issue of slow convergence with respect to the slowest system processes but also provides a forward model that can be differentiated for building the required gradients needed to tackle the parameter inference problem directly. This alternative forward model follows the approach described in [10] in converting from the trajectory samples to the probability measure for the Bayesian estimation problem, as detailed in subsection 3.1, but then recasts this forward Perron–Frobenius operator as a Markov process for determining the steady-state solution as described in subsection 3.3.

Our approach is close in spirit to other cell-based or grid-based frameworks that introduce a suitable Perron–Frobenius operator and compute its fixed points [23, section 4]. Some of these methods, such as the software package GAIO developed by Dellnitz and Junge [21, 22], represent the attractors via a hierarchy of covers by cells: cells that do not intersect the support of the invariant measure are ignored so that the data structures and computational requirements for this method are smaller than the ones required for our grid-based approach. In some cases, such as uniformly hyperbolic systems, these methods come with convergence guarantees [23, Theorem 4.14]. Many other subdivision methods have been successfully applied to the numerical analysis of complex dynamical behavior; see, for instance, [20, 26, 67]. A more comprehensive list of examples can be found in [19, 38].

We regularize our Perron–Frobenius operator via teleportation regularization from Google's PageRank method [39], which ensures the uniqueness and regularity of the fixed point. This step is similar to stochastic perturbation techniques for approximating physical measures [18, 43, 47, 23]. Intuitively, teleportation amounts to stopping the dynamics at a random time and restarting it from a randomly chosen initial point. The regularization parameter ϵ controls the restarting frequency: the smaller the ϵ, the rarer the restart. This regularization is somewhat similar to "snapshot attractors" described in [62] where attractors are estimated by following the dynamics from randomly chosen initial conditions for a fixed time. Here, we do not analyze the convergence of $\rho_\epsilon(\theta)$ to the physical invariant measure, but the numerical evidence in section 6.2.5 suggest that this convergence does take place for the tested examples. Intuitively, if we restart the dynamics from the basin of attraction and do so very rarely, we should approximate the physical measure. Additionally, general results in [47] hint at a convergence result similar to [23, Theorem 4.14] for uniformly hyperbolic attractors. Analyzing

the convergence of our model and the differentiability of other forward models described here is an exciting future research direction that we plan to pursue. For additional methods based on Markov partitions and chains we refer to [13, 35, 32]. Formally, we assume that (1) the dynamical system of interest, $\dot{x} = v(x, \theta)$, where $\theta \in \Theta$, has one unique physical invariant measure, and (2) the distributional solution to (1.2) with the same $v(x, \theta)$ is unique and recovers the physical invariant measure to the dynamical system.

3.1. From linear advection to stationary eigenvectors. In converting the dynamical system from the trajectory samples to the probability measure, the governing equation is converted from a nonlinear ODE for the system state "point," $\mathbf{x}$, to a linear PDE (1.2) for the state-space density $\rho(\mathbf{x})$.

Note that a causal dynamical system includes no diffusion. It then corresponds to (1.2), a linear advection of probability density in state space. Subsection 3.2 describes a particular simple low-order discretization of this linear advection problem. While adding physical diffusion is a relatively simple modification of the numerical method, the more significant issue with this approach relates to excess diffusion. Although the zero diffusion case can be relaxed for stochastic dynamical systems where $D_{ij} \neq 0$, the upwinding scheme required to stabilize the advection introduces an artificial diffusion, which is the predominant numerical error as described in [10]. This numerical diffusion is expected to dominate physical diffusion for the moderate spatial resolution that is tractable for the forward model unless the dynamics of the system are highly stochastic. As this numerical diffusion is irreducible at finite computational cost, the addition of finite diffusion to the ODE model is explored in subsection 6.2 when attempting to understand the class of problems for which inference with respect to the binned direct ODE solution is viable.

3.2. Finite volume discretization. A finite volume discretization of the resulting continuity equation defined on the domain Ω, as described in [10], is then obtained. The finite volume discretization combined with a zero-flux boundary condition, $v = 0$ on the boundaries $\partial\Omega$, enforces strict mass conservation whenever the discrete integration by parts formulation is used [31]. Only the first-order operator split upwind discretization is used in this work to enforce positivity of the probability density, as will be shown to be a consequence of the form of the discrete operator.

We first discretize (1.2) on a d-dimensional uniform mesh in space and time with no added diffusion, which gives us the following equation for the explicit time evolution of the probability density,

$$\frac{\rho^{(l+1)}(x_i) - \rho^{(l)}(x_i)}{\Delta t} = -\sum_{i_d=1}^{d} \frac{F^{(l)}_{(i_d)}(x_i + \Delta x_{(i_d)}/2) - F^{(l)}_{(i_d)}(x_i - \Delta x_{(i_d)}/2)}{\Delta x_{(i_d)}}.$$

Here, the point x_i refers to the ith cell center vector and $\Delta x_{(i_d)}$ refers to the mesh spacing in the i_dth direction, $i_d = 1, \ldots, d$. The upwind i_d-direction flux at the lth time step, $F^{(l)}_{(i_d)}$, is then approximated using the face center velocity assuming uniform density within the cell centered at x_i as follows:

$$F^{(l)}_{(i_d)}\left(x_i - \frac{\Delta x_{(i_d)}}{2}\right) = v^+_{(i_d-\frac{1}{2})}\rho^{(l)}(x_i - \Delta x_{(i_d)}) + v^-_{(i_d-\frac{1}{2})}\rho^{(l)}(x_i),$$

where the upwind velocities $v^+_{(i_d)} = \max(v_{(i_d)}, 0)$ and $v^-_{(i_d)} = \min(v_{(i_d)}, 0)$ refer to the i_dth component of the velocity vector split between positive and negative values, and

$$v^+_{(i_d-\frac{1}{2})} := v^+_{(i_d)}\left(x_i - \frac{\Delta x_{(i_d)}}{2}\right), \quad v^-_{(i_d-\frac{1}{2})} := v^-_{(i_d)}\left(x_i - \frac{\Delta x_{(i_d)}}{2}\right).$$

Inserting these fluxes into the discrete equation yields the following expression for the future time density, $\rho^{(l+1)}$:

$$\rho_0^{(l+1)} = \rho_0^{(l)} + \Delta t \sum_{i_d=1}^{d} \frac{\left(v^+_{(i_d-\frac{1}{2})}\rho^{(l)}_- + v^-_{(i_d-\frac{1}{2})}\rho^{(l)}_0\right) - \left(v^+_{(i_d+\frac{1}{2})}\rho^{(l)}_0 + v^-_{(i_d+\frac{1}{2})}\rho^{(l)}_+\right)}{\Delta x_{(i_d)}},$$

where $\rho^{(l)}_0 = \rho^{(l)}(x_i)$, $\rho^{(l)}_- = \rho^{(l)}(x_i - \Delta x_{(i_d)})$, and $\rho^{(l)}_+ = \rho^{(l)}(x_i + \Delta x_{(i_d)})$. The equation above can be rewritten in matrix-vector format:

$$\rho^{(l+1)} = \rho^{(l)} + K_{mat}\rho^{(l)} = (I + K_{mat})\rho^{(l)}.$$

For steady-state distributions, $\rho^{(l+1)} = \rho^{(l)} = \rho^{eq}$. This corresponds to finding a nonzero solution $\rho^{(eq)}$ to the following linear system

$$K_{mat}\rho^{(eq)} = \left[\sum_{i_d=1}^{d} \frac{\Delta t}{\Delta x_{(i_d)}} K_{(i_d)}\right]\rho^{(eq)} = 0,$$

where for $i_d = 1, \ldots, d$ we have

$$K_{(i_d)} = \begin{bmatrix}
\ddots & & & & & \\
 & -v^-_{(i_d-\frac{3}{2})} & & & & \\
\ddots & \vdots & & & & \\
 & v^-_{(i_d-\frac{3}{2})} - v^+_{(i_d-\frac{1}{2})} & -v^-_{(i_d-\frac{1}{2})} & & & \\
\ddots & \vdots & \vdots & & & \\
 & +v^+_{(i_d-\frac{1}{2})} & v^-_{(i_d-\frac{1}{2})} - v^+_{(i_d+\frac{1}{2})} & -v^-_{(i_d+\frac{1}{2})} & & \\
 & & \vdots & \vdots & \ddots & \\
 & & +v^+_{(i_d+\frac{1}{2})} & v^-_{(i_d+\frac{1}{2})} - v^+_{(i_d+\frac{3}{2})} & & \\
 & & & \vdots & \ddots & \\
 & & & +v^+_{(i_d+\frac{3}{2})} & & \\
 & & & & \ddots &
\end{bmatrix}. \tag{3.1}$$

We remark that each $K_{(i_d)}$, $i_d = 1, \ldots, d$, is a tridiagonal matrix, while the offsets for the three diagonals vary for different $i_d's$. For example, consider the case that $\Omega \subseteq \mathbb{R}^3$ is a cuboid,

discretized with grid size n_x, n_y, n_z in the x, y, z dimension, respectively. Then, $K_{(1)}$ is nonzero at the first lower diagonal, the main diagonal, and the first upper diagonal; $K_{(2)}$ is nonzero at the n_xth lower diagonal, the main diagonal, and the n_xth upper diagonal; $K_{(3)}$ is nonzero at the $(n_x \times n_y)$th lower diagonal, the main diagonal, and the $(n_x \times n_y)$th upper diagonal.

We highlight that the solution $\rho^{(l)}$ at any lth time step satisfies the mass conservation property. That is,

$$\rho^{(l)} \cdot \mathbf{1} = \rho^{(l+1)} \cdot \mathbf{1} = \rho^{(eq)} \cdot \mathbf{1}, \quad \text{where } \mathbf{1} = [1, 1, \ldots, 1]^\top.$$

It is a direct consequence of the fact that columns of K_{mat} sum to zero. Note also that the off-diagonal terms are all positive or zero while the diagonal terms are all negative or zero by construction. One can construct a *column-stochastic matrix* M,

$$M = I + cK_{mat}.$$

M can be positive definite if we ensure that c is small enough.

Since the main focus of this paper is parameter identification, the velocity field v is parameter dependent. Thus, we will highlight the dependency on the parameter θ by using notation $v(\theta)$, $K_{mat}(\theta)$, $K_{(i_d)}(\theta)$, and $\rho^{(eq)}(\theta)$ hereafter.

The upper bound on c unsurprisingly also depends on θ. Nevertheless, if we assume that v depends continuously on θ and we operate in a bounded domain Ω, we can choose c small enough to serve all $\theta' s$ of interest. For instance, we can choose

$$0 < c < \min_{i_d} \frac{\Delta x_{(i_d)}}{2\Delta t \max\limits_{x\in\Omega,\theta\in\Theta} |v_{(i_d)}(x,\theta)|}. \tag{3.2}$$

3.3. Finding the stationary distribution of a Markov chain. From the previous section, we learned that $\rho(\theta)$ is the solution of

$$M(\theta)\rho = \rho, \quad \rho \cdot \mathbf{1} = 1, \tag{3.3}$$

where,

$$M(\theta) = I + cK_{mat}(\theta), \quad K_{mat}(\theta) = \sum_{i_d=1}^{d} \frac{\Delta t}{\Delta x_{(i_d)}} K_{(i_d)}(\theta)$$

with $K_{(i_d)}(\theta)$ given in (3.1), and c is chosen to satisfy (3.2). While the matrix, M, was built from a finite volume causal flow model, it was noted that this flux also approximates a discrete cell-to-cell transition probability for a point randomly sampled from the volume of one cell to its neighbor cells, which mirrors the propagator of a Markov chain as described in [46].

A priori we have that the off-diagonal entries of $M(\theta) = I + cK_{mat}(\theta)$ are nonnegative. Additionally, we know that $M(\theta)$ is column stochastic. Thus, by Gershgorin's theorem [40] we have that the spectral radius of M is not greater than one. On the other hand, we know $\mathbf{1} = [1, 1, \ldots, 1]^\top$ is an eigenvector for $M^\top$ which is a row-stochastic matrix, and so $\lambda = 1$ is an

eigenvalue for both M and $M^\top$. The spectral radius of M has to be equal to 1. Furthermore, by a limiting argument, we can show that the eigenspace of M corresponding to the eigenvalue $\lambda = 1$ contains vectors with nonnegative entries.

However, the dimension of this eigenspace may be bigger than one, which complicates our analysis. Thus, we regularize M via the so-called teleportation trick, which is well known from Google's PageRank method [39]. That is, given a small positive constant ϵ, we consider

$$M_\epsilon(\theta) = (1-\epsilon)M + \epsilon n^{-1}\mathbf{1}\,\mathbf{1}^\top = (1-\epsilon)(I + cK_{mat}(\theta)) + \frac{\epsilon}{n}\mathbf{1}\,\mathbf{1}^\top. \tag{3.4}$$

Note that the off-diagonal entries of M_ϵ are at least $\frac{\epsilon}{n} > 0$. The regularization also connects all cells, achieving similar regularizing effects by having a diffusion term. Moreover, M_ϵ is still column stochastic. Based on the following Perron–Frobenius theorem, the spectral radius of M_ϵ must be 1.

Theorem 3.1 (Perron-Frobenius theorem [54]). *If all entries of a Markov matrix A are positive, then A has a unique equilibrium; there is only one eigenvalue equal to 1. All other eigenvalues are strictly smaller than 1.*

Consequently, the eigenspace $\{\rho : M_\epsilon(\theta)\rho = \rho\}$ is one dimensional and has a generator with all positive entries. Hence, the equation

$$M_\epsilon(\theta)\rho = \rho, \quad \rho \cdot \mathbf{1} = 1, \quad \rho > 0, \tag{3.5}$$

has a unique solution that converges to a solution of (3.3) as $\epsilon \to 0$. We can analyze the error between ρ_0 and ρ_ϵ, where

$$M\rho_0 = \rho_0, \quad M_\epsilon\rho_\epsilon = \rho_\epsilon, \quad \rho_0 \cdot \mathbf{1} = \rho_\epsilon \cdot \mathbf{1} = 1.$$

The error analysis traces back to the classical root-finding problem. We define $\Delta\rho_\epsilon = \rho_\epsilon - \rho_0$. Using the forward error analysis, we obtain that

$$(M-I)\Delta\rho_\epsilon = (M-I)\rho_\epsilon = \epsilon\left(M - n^{-1}\mathbf{1}\,\mathbf{1}^\top\right)\rho_\epsilon, \quad \Delta\rho_\epsilon \cdot \mathbf{1} = 0.$$

Solving for $\Delta\rho_\epsilon$ from the linear system above can improve the current "root" ρ_ϵ, which is precisely the principle behind Newton's method. Using backward error analysis, starting from $M_\epsilon\rho_\epsilon = \rho_\epsilon$, we obtain that

$$\left((1-\epsilon)M + \epsilon n^{-1}\mathbf{1}\,\mathbf{1}^\top - I\right)(\rho_0 + \Delta\rho_\epsilon) = 0.$$

Up to the first-order terms, we have

$$(M-I)\Delta\rho_\epsilon = \epsilon\left(M - n^{-1}\mathbf{1}\,\mathbf{1}^\top\right)\rho_0, \quad \Delta\rho_\epsilon \cdot \mathbf{1} = 0.$$

The above equation implies that $\|\Delta\rho_\epsilon\|$ is $\mathcal{O}(\epsilon)$, showing the convergence $\rho_\epsilon \to \rho_0$ as we decrease ϵ. This is further verified by our numerical examples in subsection 6.2.5.

Numerically, the problem (3.5) can be solved by mature tools from numerical linear algebra such as the power method and the Richardson iteration [39]. We present one direct solve method in subsection B.1 using the sparsity of K_{mat}.

4. Optimal transport for parameter inference. Here, we discuss gradient evaluation of optimal transport-based costs with respect to the inference parameters. Assume that $\Omega \subset \mathbb{R}^d$ is a compact set, and $c : \Omega^2 \to \mathbb{R}$ is a continuous cost function. The main goal of this section is to discuss the differentiability of the objective function

$$f(\theta) = \mathcal{T}_c(\rho(\cdot, \theta), \rho^*), \quad \theta \in \Theta,$$

where $\{\rho(\cdot, \theta)\}_{\theta\in\Theta}$ is a family of parameter-dependent probability measures on Ω, and $\mathcal{T}_c$ is the optimal transport cost defined in (2.4). Throughout the paper, we assume that $\Omega \subset \mathbb{R}^d$ is compact, $\rho^* \in \mathscr{P}(\Omega)$ is an arbitrary probability measure, and

A1. $\Theta \subset \mathbb{R}^m$ is an open set, and $\{\rho(\cdot, \theta)\}_{\theta\in\Theta} \subset \mathscr{P}(\Omega)$ is a family of absolutely continuous probability measures.

A2. For a.e. $x \in \Omega$ the mapping $\theta \mapsto \rho(x, \theta)$ is differentiable, and $|\nabla_\theta \rho(x, \theta)| \le \eta(x)$, $\theta \in \Theta$, for some $\eta \in L^1(\Omega)$. Note that by slightly abusing the notation, we use the same notation for probability measures and their densities.

A3. $c : \Omega^2 \to \mathbb{R}$ is continuous and nonnegative.

Occasionally, we need the following hypothesis.

A4. For a.e. $x \in \Omega$ the mapping $\theta \mapsto \rho(x, \theta)$ is locally semiconvex, and $\nabla^2_\theta \rho(x, \theta) \ge -h(x)$, $\theta \in \Theta$, for some $h \in L^1(\Omega)$.

Proofs for results of this section can be found in Appendix A.

4.1. Preliminaries. First, we recall preliminary results from the OT theory that can be found in [71, 4, 66]. A key tool in OT is the Kantorovich duality [71, Theorem 1.3] that states

$$\mathcal{T}_c(\mu, \nu) = \sup_{(\phi,\psi)\in\Phi_c(\mu,\nu)} \int_\Omega \phi(x) d\mu(x) + \int_\Omega \psi(y) d\nu(y), \quad \mu, \nu \in \mathscr{P}(\Omega), \tag{4.1}$$

where $\Phi_c(\mu, \nu) \subset C(\Omega) \times C(\Omega)$ is the set of pairs (ϕ, ψ) such that $\phi(x) + \psi(y) \le c(x, y)$ for all $(x, y) \in \Omega^2$. The maximizing pairs (ϕ, ψ) in (4.1) are called Kantorovich potentials. The c-transform of a function $x \mapsto \phi(x)$ is defined as

$$\phi^c(y) = \inf_{x\in\Omega} \{c(x, y) - \phi(x)\}.$$

Similarly, the c-transform of a function $y \mapsto \psi(y)$ is defined as

$$\psi^c(x) = \inf_{y\in\Omega} \{c(x, y) - \psi(y)\}.$$

A function $x \mapsto \phi(x)$ (resp., $y \mapsto \psi(y)$) is called c-concave if there exists a function ψ (resp., ϕ) such that $\phi = \psi^c$ (resp., $\psi = \phi^c$).

Since Ω is compact and c is continuous, we obtain that c is bounded. Thus, the set $\Phi_c(\mu, \nu)$ in (4.1) can be further restricted to uniformly bounded pairs of conjugate c-concave functions, that is, pairs of $(\phi, \phi^c) \in \Phi_c(\mu, \nu)$, where $\phi = \phi^{cc}$, and $0 \le \phi \le \|c\|_\infty$, $-\|c\|_\infty \le \phi^c \le 0$ [12, Remarks 1.12–13]. We denote this set by K_c.

Since the modulus of continuity of $y \mapsto c(x, y) - \phi(x)$ (resp., $x \mapsto c(x, y) - \phi^c(y)$) is bounded by that of c for all x (resp., y), K_c is uniformly equicontinuous, uniformly bounded,

and, consequently, precompact in $C(\Omega) \times C(\Omega)$ by the Arzelà–Ascoli theorem [66, section 1.2]. Additionally, since the c-transform is continuous under the uniform convergence, K_c is compact in $C(\Omega) \times C(\Omega)$, and the existence of Kantorovich potentials in K_c is guaranteed [66, Proposition 1.11].

4.2. The differentibility of the transport cost in the parameter space. Here, we heavily rely on the Kantorovich duality (4.1) and the regularity theory of optimal value functions [12, Chapter 4]. Recall that f is directionally differentiable at $\theta_0 \in \Theta$ if

$$\lim_{t \to 0+} \frac{f(\theta_0 + t\Delta\theta) - f(\theta_0)}{t} = f'(\theta_0, \Delta\theta)$$

for all $\Delta\theta \in \mathbb{R}^m$ [12, section 2.2]. Furthermore, if $\Delta\theta \mapsto f'(\theta_0, \Delta\theta)$ is linear, we say that f is Gâteaux differentiable at θ_0 and denote by $\nabla f(\theta_0)$ the generator of this linear map.

Next, denote by $\mathcal{S}(\theta) \subset K_c$ the set of Kantorovoch potentials for the OT from $\rho(\cdot, \theta)$ to ρ^*.

Proposition 4.1. *Assume that* A1–A3 *hold.*

(i) f is everywhere directionally differentiable, and

$$f'(\theta_0, \Delta\theta) = \sup_{(\phi, \phi^c) \in \mathcal{S}(\theta_0)} \int_\Omega \phi(x) \nabla_\theta \rho(x, \theta_0) dx \cdot \Delta\theta \tag{4.2}$$

for all $\theta_0 \in \Theta$, and $\Delta\theta \in \mathbb{R}^m$.

(ii) f is Gâteaux differentiable at $\theta_0 \in \Theta$ if and only if

$$\int_\Omega \phi_1(x) \nabla_\theta \rho(x, \theta_0) dx = \int_\Omega \phi_2(x) \nabla_\theta \rho(x, \theta_0) dx \tag{4.3}$$

for all $(\phi_1, \phi_1^c), (\phi_2, \phi_2^c) \in \mathcal{S}(\theta_0)$. In this case, we have that

$$\nabla f(\theta_0) = \int_\Omega \phi(x) \nabla_\theta \rho(x, \theta_0) dx \tag{4.4}$$

for an arbitrary pair of Kantorovich potentials $(\phi, \psi) \in \Phi_c(\rho(\cdot, \theta_0), \rho^*)$.

The proof is in subsection A.1.

Proposition 4.1 asserts that f is directionally differentiable at all points and that its directional derivative is a one-homogeneous closed convex function. Since we are interested in descent directions of f, we focus on cases when the directional derivative is a linear function and thus provides a descent direction in the form of the negative gradient. In what follows, we prove that f is generically differentiable even without (4.3). Furthermore, we find sufficient structural conditions on the optimal transport plans between $\rho(\cdot, \theta_0)$ and ρ^* to guarantee (4.3).

Theorem 4.2. *Assume that* A1–A3 *hold. Then f is locally Lipschitz continuous, and* (4.4) *holds a.e.. Additionally, if* A4 *holds, then f is locally semiconvex, and* (4.4) *holds up to a set of Hausdorff dimension $d - 1$.*

The proof can be found in subsection A.2.

There is a natural degree of freedom for Kantorovich potentials given by the addition of constants; that is, (ϕ, ϕ^c) is a pair of Kantorovich potentials if and only if $(\phi+\lambda, \phi^c-\lambda)$ is also a pair of Kantorovich potentials for an arbitrary constant λ. As a corollary of Proposition 4.1, we obtain that the Gâteaux differentiability of f is guaranteed if the addition of constants is the only degree of freedom for Kantorovich potentials.

Corollary 4.3. *Assume that* A1–A3 *hold, and* $\theta_0 \in \Theta$ *is such that* $\phi_2 - \phi_1$ *is constant* $\rho(\cdot, \theta_0)$ *a.e. for all pairs of Kantorovich potentials* (ϕ_1, ψ_1), (ϕ_2, ψ_2). *Then* f *is Gâteaux differentiable at* θ_0, *and* (4.4) *holds.*

In general, Kantorovich potentials are not unique up to constants. In what follows, we provide a sufficient condition for such uniqueness. Essentially, the OT should not amount to transportation between disjoint parts of $\mathrm{supp}(\rho(\cdot, \theta_0))$ and $\mathrm{supp}(\rho^*)$.

More formally, assume that $\rho, \rho^* \in \mathscr{P}(\Omega)$ are such that $\mathrm{int}(\mathrm{supp}(\rho)) \neq \emptyset$. Furthermore, denote by $\Gamma_0(\rho, \rho^*)$ the set of optimal transport plans; that is, minimizers in (2.4). We have that

$$\mathrm{int}(\mathrm{supp}(\rho)) = \cup_k O_k, \tag{4.5}$$

where O_k are disjoint open and connected sets. Next, denote by

$$E_k = \mathrm{cl}\left(\{y : (x, y) \in \mathrm{supp}(\pi) \text{ for some } x \in \mathrm{cl}(O_k),\ \pi \in \Gamma_0(\rho, \rho^*)\}\right). \tag{4.6}$$

In other words, E_k is the set where the mass from $\mathrm{cl}(O_k)$ is transported to.

Definition 4.4. *We say that* $\mathrm{cl}(O_k)$ *and* $\mathrm{cl}(O_l)$ *are linked in the OT from* ρ *to* ρ^* *with a transport cost* c, *if there exist* $\{i_j\}_{j=1}^m$ *such that* $k = i_1, l = i_m$, *and* $E_{i_j} \cap E_{i_{j+1}} \neq \emptyset$, $1 \leq j \leq m$.

Theorem 4.5. *Assume that* $c \in C^1(\Omega^2)$, $\rho, \rho^* \in \mathscr{P}(\Omega)$, *and*

$$\mathrm{supp}(\rho) = \mathrm{cl}(\mathrm{int}(\mathrm{supp}(\rho))). \tag{4.7}$$

Furthermore, suppose that $\{O_k\}$ *and* $\{E_k\}$ *are defined as in* (4.5) *and* (4.6), *respectively. Assume that all* $\{\mathrm{cl}(O_k)\}$ *are mutually linked. Then* $\phi_2 - \phi_1$ *is constant* ρ*-a.e. for all pairs of Kantorovich potentials* (ϕ_1, ψ_1), (ϕ_2, ψ_2).

The proof is presented in subsection A.3. Theorem 4.5 and Corollary 4.3 yield the following corollary.

Corollary 4.6. *Assume that* A1–A3 *hold, and* $\rho = \rho(\cdot, \theta_0)$ *satisfies the hypotheses in Theorem* 4.5. *Then* f *is Gâteaux differentiable at* θ_0.

In particular, if $\rho(\cdot, \theta_0)$ *is supported on a closure of an open connected set, then* f *is Gâteaux differentiable at* θ_0.

The following proposition illustrates the sharpness of Corollary 4.6. Incidentally, the same example illustrates that a smooth dependence on θ with respect to the flat L^2 metric does not guarantee smooth dependence on θ with respect to the Wasserstein metric.

Proposition 4.7. *Assume that* $\Omega = [0, 4]$ *and* $c(x, y) = |x - y|^p$ *for some* $p > 1$ *(so that* $\mathcal{T}_c = W_p^p$*). Consider*

$$\rho(x,\theta) = (0.5+\theta)\,\chi_{[0,1]}(x) + (0.5-\theta)\,\chi_{[2,3]}(x),\ |\theta| < 0.5,$$
$$\rho^*(y) = 0.5\chi_{[1,2]}(y) + 0.5\chi_{[3,4]}(y),$$

where χ_A *is the characteristic function of set* $A \subset \mathbb{R}$. *Then we have that*

1. $\{\rho(\cdot,\theta)\}$ *satisfies* A1–A3.
2. $\{\rho(\cdot,\theta)\}$ *is not absolutely continuous in* $\mathscr{P}_p(\Omega)$.
3. $\rho \mapsto W_p^p(\rho,\rho^*)$ *is not Gâteaux differentiable at* $\rho(\cdot,\theta)$ *for all* $|\theta| < 0.5$.
4. $[0,1]$ *and* $[2,3]$ *are linked in the OT from* $\rho(\cdot,\theta)$ *to* ρ^* *for all* $|\theta| < 0.5$ *except* $\theta = 0$.
5. $\theta \mapsto W_p^p(\rho(\cdot,\theta),\rho^*)$ *is differentiable for all* $|\theta| < 0.5$ *except* $\theta = 0$.

The proof can be found in subsection A.4.

4.3. Qualitative error analysis for the gradient. In this subsection, we prove that almost-optimal solutions of Kantorovich's dual problem would provide accurate approximations of ∇f.

Proposition 4.8. *Assume that* A1–A3 *hold, and* f *is Gâteaux differentiable at* $\theta_0 \in \Theta$. *For every* $\epsilon > 0$ *there exists a* $\delta > 0$ *such that for all* $(\phi,\psi) \in \Phi_c(\rho(\cdot,\theta_0),\rho^*)$ *satisfying* $I(\phi,\psi,\theta_0) > f(\theta_0) - \delta$ *one has that*

$$\left|\nabla_\theta f(\theta_0) - \int_\Omega \phi^{cc}(x)\nabla_\theta \rho(x,\theta_0)dx\right| < \epsilon.$$

The proof is presented in subsection A.5.

Remark 4.9. Proposition 4.8 asserts that one needs to calculate c-transforms of suboptimal ϕ for accurate gradients. This can be done very efficiently for costs of the form $c(x,y) = \sum_{i=1}^d h_i(x_i - y_i)$, where h_i are even and strictly convex functions [44, section 4.1]. For OT algorithms that produce c-concave iterates, such as in [44], no further considerations are necessary.

5. Gradient calculation. Our parameter-dependent synthetic data obtained through the forward model is given by a finite volume approximation

$$\rho(x,\theta) = \sum_{i=1}^n \rho_i(\theta)\frac{\chi_{C_i}(x)}{|C_i|}, \tag{5.1}$$

where $n = n_x n_y n_z$ is the total grid size, each C_i is the finite volume cell, the parameter $\theta \in \Theta \subset \mathbb{R}^m$, and $\rho(\theta) = (\rho_i(\theta))_{i=1}^n$ is the solution to (3.5) for some fixed $c, \epsilon > 0$. Furthermore, after discretization, our reference data are given by

$$\rho^*(y) = \sum_{i=1}^n \rho_i^* \frac{\chi_{C_i}(y)}{|C_i|}.$$

By slightly abusing the notation we denote $\rho^* = (\rho_i^*)_{i=1}^n$. Our goal is to solve

$$\min_\theta f(\theta) = \mathcal{T}_c(\rho(\cdot,\theta),\rho^*) \tag{5.2}$$

by gradient-based algorithms, where $\mathcal{T}_c$ is the optimal transport cost defined in (2.4). To apply Corollary 4.6, which will guarantee the differentiability of f, we need to verify A2 for (5.1) and that the connected components of $\mathrm{supp}\rho(\cdot,\theta)$ are linked according to Definition 4.4. Since in all our experiments in section 6, $\mathrm{supp}\rho(\cdot,\theta) = \cup_{i:\rho_i(\theta)>0} C_i$ is connected, the latter condition is satisfied. Therefore, we just need to verify A2, which is equivalent to the differentiability of $\theta \mapsto \rho(\theta)$. This verification is part of subsection 5.1.

Once all assumptions are verified, we have that

$$\nabla_\theta \rho(x,\theta) = \sum_{i=1}^{n} \nabla_\theta \rho_i(\theta) \frac{\chi_{C_i}(x)}{|C_i|}.$$

Therefore,

$$\nabla f(\theta) = \sum_{i=1}^{n} \nabla_\theta \rho_i(\theta)\phi_i(\theta), \quad \text{where } \phi_i(\theta) = \frac{\int_{C_i} \phi(x,\theta)dx}{|C_i|}. \tag{5.3}$$

Here, $\phi(\cdot,\theta)$ is a Kantorovich potential for an OT from $\rho(\cdot,\theta)$ to ρ^*. Kantorovich potentials can be calculated by one of many available OT solvers such as [44, 33]. Hence, we focus on calculating $\nabla_\theta \rho_i(\theta)$.

5.1. Gradient descent via implicit function theorem. First, we verify A2; that is, the differentiability of $\theta \mapsto \rho(\theta)$.

Lemma 5.1. *Assume that $\theta \mapsto A(\theta)$, $\theta \in \Theta$, is a C^1 matrix valued function such that $A(\theta)$ is column stochastic with strictly positive entries for all $\theta \in \Theta$. Then the system of equations*

$$A(\theta)\rho = \rho, \quad \rho \cdot \mathbf{1} = 1, \tag{5.4}$$

has a unique solution $\rho = \rho(\theta)$ for all $\theta \in \Theta$. Moreover, $\theta \mapsto \rho(\theta)$ is continuously differentiable with $\zeta_k(\theta) = \partial_{\theta_k}\rho(\theta)$ being the unique solution of

$$(A(\theta) - I)\zeta_k = -\partial_{\theta_k} A(\theta)\rho(\theta), \quad \zeta_k \cdot \mathbf{1} = 0, \tag{5.5}$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$.

Proof. The existence and uniqueness of $\rho(\theta)$ is a consequence of the Perron–Frobenius theorem as explained in subsection 3.3. Denote by $B(\theta)$ the matrix obtained from $A(\theta) - I$ by adding an $(n+1)$th row vector $\mathbf{1}^\top$. Then we have that $\ker(B(\theta)) = \{\mathbf{0}\}$, and so $\mathrm{rank}(B(\theta)) = n$, and n rows of $B(\theta)$ are linearly independent. Moreover, since $\ker(A(\theta) - I) = \mathrm{span}\{\rho(\theta)\}$, we have that $\mathrm{rank}(A(\theta) - I) = n - 1$. Thus, the first n rows of $B(\theta)$ are linearly dependent, and any list of n independent rows must contain the last row $\mathbf{1}^\top$. Since $\theta \mapsto A(\theta)$ is continuous, linearly independent vectors stay so in a neighborhood of each θ. Hence, we fix θ and without loss of generality assume that the rows of $B(\theta)$ from 2 to $n+1$ are linearly independent in a neighborhood of θ.

Denote

$$F(\theta,\rho) = \widetilde{B(\theta)}\rho - e_n,$$

where $\widetilde{B(\theta)}$ is the matrix obtained from $B(\theta)$ by dropping the first row and e_n is the nth standard basis vector. Then we have that $\rho(\theta)$ is the unique solution of $F(\theta,\rho)=0$, and $D_\rho F(\theta,\rho)=\widetilde{B(\theta)}$ is non-degenerate. Thus, the Implicit function theorem applies and we obtain that $\theta \mapsto \rho(\theta)$ is continuously differentiable. Therefore, we can differentiate (5.4) and obtain (5.5). Moreover, $\ker(B(\theta))=\{\mathbf{0}\}$ yields that the solution of (5.5) is unique. ∎

Applying Lemma 5.1 to $A(\theta)=M_\epsilon(\theta)$ we obtain that the solution of (3.5) is differentiable and (5.3) holds. Thus, we can devise a gradient descent algorithm as follows:

$$\begin{cases} M_\epsilon(\theta^l)\rho^l=\rho^l, \quad \rho^l\cdot\mathbf{1}=1, \\ (M_\epsilon(\theta^l)-I)\zeta_k^l=-\partial_{\theta_k}M_\epsilon(\theta^l)\rho^l, \quad \zeta_k^l\cdot\mathbf{1}=0, \quad 1\le k\le m, \\ (\phi^l,\psi^l)\in \underset{\phi_i+\psi_j\le c(x_i,x_j)}{\operatorname{argmax}} [\phi\cdot\rho^l+\psi\cdot\rho^*], \\ \theta_k^{l+1}=\theta_k^l-\tau^l\ \phi^l\cdot\zeta_k^l, \quad 1\le k\le m, \end{cases} \tag{5.6}$$

where $\tau^l>0$ is a proper step size for the gradient descent algorithm.

5.2. Gradient descent via adjoint method. Here we discuss an alternative approach to calculate the gradient (5.3) via the adjoint-state method.

Lemma 5.2. *Assume that $\theta\mapsto A(\theta)$ satisfies the hypotheses in Lemma 5.1, $\rho(\theta)$ is the solution of (5.4), and $\phi\in\mathbb{R}^n$ is an arbitrary vector. Then the linear system*

$$(A(\theta)^\top-I)\lambda=-\phi+\phi\cdot\rho(\theta)\ \mathbf{1} \tag{5.7}$$

is consistent with a one-dimensional solution set. Moreover, for any solution λ one has that

$$\partial_{\theta_k}(\phi\cdot\rho(\theta))=\lambda\cdot\partial_{\theta_k}A(\theta)\rho(\theta).$$

Proof. Since $\operatorname{im}(A(\theta)^\top-I)=\ker(A(\theta)-I)^\perp$, we have to show that

$$-\phi+\phi\cdot\rho(\theta)\ \mathbf{1}\in\ker(A(\theta)-I)^\perp=\operatorname{span}\{\rho(\theta)\}^\perp.$$

A simple calculation yields the result

$$(-\phi+\phi\cdot\rho(\theta)\ \mathbf{1})\cdot\rho(\theta)=-\phi\cdot\rho(\theta)+\phi\cdot\rho(\theta)\ \mathbf{1}\cdot\rho(\theta)=0.$$

Furthermore, since $\ker(A(\theta)^\top-I)=\operatorname{span}\{\mathbf{1}\}$, the solution set of (5.7) is a one-dimensional coset of $\operatorname{span}\{\mathbf{1}\}$.

Finally, assume that λ is an arbitrary solution of (5.7). Then applying (5.5) we obtain that

$$\begin{aligned}\partial_{\theta_k}(\phi\cdot\rho(\theta)) = \phi\cdot\zeta_k &= (\phi\cdot\rho(\theta)\ \mathbf{1} - (A(\theta)^\top - I)\lambda)\cdot\zeta_k \\ &= \phi\cdot\rho(\theta)\ \mathbf{1}\cdot\zeta_k - \lambda\cdot(A(\theta)-I)\zeta_k = \lambda\cdot\partial_{\theta_k}A(\theta)\rho(\theta).\end{aligned}$$

Applying Lemma 5.2 to $A(\theta) = M_\epsilon(\theta)$, we obtain an alternative, but equivalent, gradient descent algorithm:

$$\text{(5.8)}\qquad \begin{cases} M_\epsilon(\theta^l)\rho^l = \rho^l, \quad \rho^l\cdot\mathbf{1} = 1, \\ (\phi^l,\psi^l) \in \underset{\phi_i+\psi_j\le c(x_i,x_j)}{\operatorname{argmax}} [\phi\cdot\rho^l + \psi\cdot\rho^*], \\ (M_\epsilon(\theta^l)^\top - I)\lambda^l = -\phi^l + \phi^l\cdot\rho^l\ \mathbf{1}, \quad \lambda^l\cdot\mathbf{1} = 0, \\ \theta_k^{l+1} = \theta_k^l - \tau^l\ \lambda^l\cdot\partial_{\theta_k}M_\epsilon(\theta^l)\rho^l, \quad 1\le k\le m. \end{cases}$$

Here, $\tau^l > 0$ is a chosen step size to guarantee enough decrease in the objective function. Note that we add a condition $\lambda^l\cdot\mathbf{1}$ to ensure the uniqueness of λ^l.

We present a numerical scheme for efficiently solving systems of equations (5.6) and (5.8) in subsection B.1.

5.3. The gradient of $M_\epsilon(\theta)$. For both algorithms (5.6) and (5.8) we need to evaluate $\partial_{\theta_i}M_\epsilon(\theta)$. Denote by $H(x) = \frac{dx^+}{dx}$ the Heaviside function. We then have

$$\partial_{\theta_i}v^+ = H(v)\partial_{\theta_i}v, \quad \partial_{\theta_i}v^- = (1-H(v))\partial_{\theta_i}v.$$

We can also consider smoothed versions of H such as

$$H_k(x) = \frac{d}{dx}k\log(1+e^{\frac{x}{k}}) = \frac{e^{\frac{x}{k}}}{1+e^{\frac{x}{k}}}.$$

It is not hard to show that H_k is smooth and $\lim_{k\to 0^+}H_k(x) = H(x)$. Based on (3.4), we derive that

$$\partial_{\theta_i}M_\epsilon = (1-\epsilon)c\cdot\partial_{\theta_i}K_{mat} = (1-\epsilon)c\cdot\sum_{i_d=1}^{d}\frac{\Delta t}{\Delta x_{(i_d)}}\partial_{\theta_i}K_{(i_d)}(\theta),$$

where each matrix $\partial_{\theta_i}K_{(i_d)}(\theta)$ has three nonzero diagonals for each pair of (i,i_d) where $1\le i\le m$, $1\le i_d\le d$, while the offsets of the diagonals depend on i_d, as we have discussed earlier regarding (3.1). We emphasize that $\partial_{\theta_i}K_{(i_d)}(\theta)$ shares the same tridiagonal structure with $K_{(i_d)}(\theta)$ for each i_d as illustrated below;

$$\partial_{\theta_i} K_{(i_d)}(\theta) = \begin{bmatrix} \ddots & & & & \\ & \ddots & & & \\ \ddots & & -\left(1 - H_k(v_{(i_d-\frac{1}{2})})\right)\partial_{\theta_i} v_{(i_d-\frac{1}{2})} & & \\ & \ddots & \vdots & \ddots & \\ \ddots & & \left(1 - H_k(v_{(i_d-\frac{1}{2})})\right)\partial_{\theta_i} v_{(i_d-\frac{1}{2})} - H_k(v_{(i_d+\frac{1}{2})})\partial_{\theta_i} v_{(i_d+\frac{1}{2})} & & \ddots \\ & \ddots & \vdots & \ddots & \\ & & H_k(v_{(i_d+\frac{1}{2})})\partial_{\theta_i} v_{(i_d+\frac{1}{2})} & & \ddots \\ & & & \ddots & \\ & & & & \ddots \end{bmatrix}.$$

One can also compute $\partial_{\theta_i} K_{(i_d)}(\theta)$ through automatic differentiation; see subsection B.2 for details of implementation and performance comparison.

6. Numerical results. In this section, we show several numerical results on dynamical system parameter identification following the methodology described in the earlier sections. The forward problem is to solve for the steady state of the corresponding PDE (1.2) rather than the ODE system (2.1). The objective function that compares the observed and the synthetic invariant measures is the quadratic Wasserstein metric (W_2) from OT. The optimization algorithm implemented for all inversion tests is the gradient descent method with backtracking line search to control the step size [59].

6.1. Chaotic system examples. We test our proposed method on three classic chaotic systems: the Lorenz, Rössler, and Chen systems. These models are widely used benchmarks that illustrate typical features of dynamical systems with instabilities and nonlinearities that give rise to deterministic chaos. We also perform an inversion test on a modified arctan Lorenz system in which the unknown parameters are nonlinear with respect to the flow velocity in terms of monomial basis. The true parameters are selected such that the dynamical systems exhibit chaotic behaviors; see the illustration of time trajectories in Figure 1.

6.1.1. Lorenz system. Consider the following Lorenz system

$$\begin{cases} \dot{x} &= \sigma(y - x), \\ \dot{y} &= x(\rho - z) - y, \\ \dot{z} &= xy - \beta z. \end{cases} \tag{6.1}$$

The equations form a simplified mathematical model for atmospheric convection, where x, y, z denote variables proportional to convective intensity, horizontal, and vertical temperature differences. The parameters σ, β, ρ are proportional to the Prandtl number, Rayleigh number, and a geometric factor. The true parameter values that we will try to infer are $\sigma = 10$, $\beta = 8/3$, $\rho = 28$. These are well-known parameter values for which the Lorenz system shows a chaotic behavior.

6.1.2. Rössler system. Consider the following Rössler System:

$$\begin{cases} \dot{x} &= -y - z, \\ \dot{y} &= x + ay, \\ \dot{z} &= b + z(x - c). \end{cases} \tag{6.2}$$

Here x, y, z denote variables, while a, b, c are the parameters we want to infer. The system exhibits continuous-time chaos and is described by the above three coupled ODEs. The true parameters that we try to infer are $a = 0.1,\ b = 0.1,\ c = 14$.

6.1.3. Chen system. Consider the following Chen system [17]:

$$\begin{cases} \dot{x} &= a(y - x), \\ \dot{y} &= (c - a)x - xz + cy, \\ \dot{z} &= xy - bz. \end{cases} \tag{6.3}$$

Again, x, y, z are variables and a, b, c are parameters we will infer. The system has a double-scroll chaotic attractor. The true parameters that we will infer are $a = 40,\ b = 3,\ c = 28$.

6.1.4. Arctan Lorenz system. The parameters in the earlier examples are all coefficients of the monomial basis. Here, we modify the right-hand side of the Lorenz system (6.1) to create a new dynamical system such that the particle flow velocity is nonlinear with respect to the monomial basis:

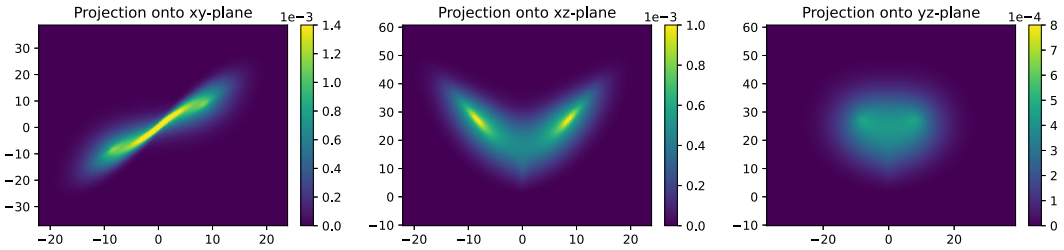
$$\begin{cases} \dot{x} &= 50 \arctan\left(\sigma(y - x)/50\right), \\ \dot{y} &= 50 \arctan\left(x(\rho - z)/50 - y/50\right), \\ \dot{z} &= 50 \arctan\left((xy - \beta z)/50\right). \end{cases} \tag{6.4}$$

Again, x, y, z are variables, and σ, ρ, β are parameters we want to infer. The reference values are set to be $(10, 28, 8/3)$, the same as the original Lorenz system.

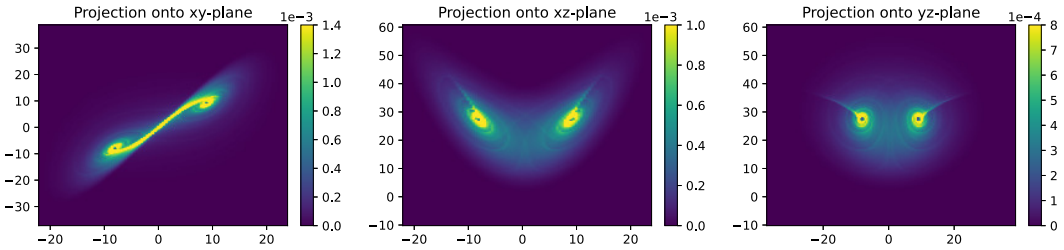
6.2. The invariant measures. Here, we follow the numerical scheme described in subsection 3.3 and approximate the invariant measure through the regularized PDE surrogate model, represented by the corresponding probability density function (PDF), for the three dynamical systems at the given sets of parameters.

We compare PDFs obtained through the steady-state solution to (1.2) with the histogram accumulated from long-time trajectories from direct numerical simulation (DNS). That is, we solve systems (6.1)–(6.3) forward in time using the explicit Euler scheme with time step Δt from $t = 0$ to its final time $t = T$. We then compute the physical invariant measure following (2.2). Moreover, we use time trajectories that are enforced with either the intrinsic or the extrinsic noises.

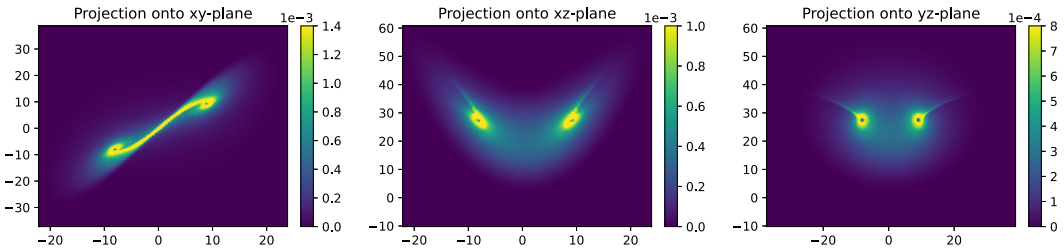
6.2.1. Numerical illustrations. Comparisons for the Lorenz system (6.1) are displayed in Figure 2. The three plots in the top row show the x-y, x-z, and y-z projections of the dominant eigenvector of the Markov matrix M_ϵ. The grid size for the finite volume discretization of (1.2) is $93 \times 153 \times 143$. The teleportation parameter is $\epsilon = 10^{-6}$. In the second row, we



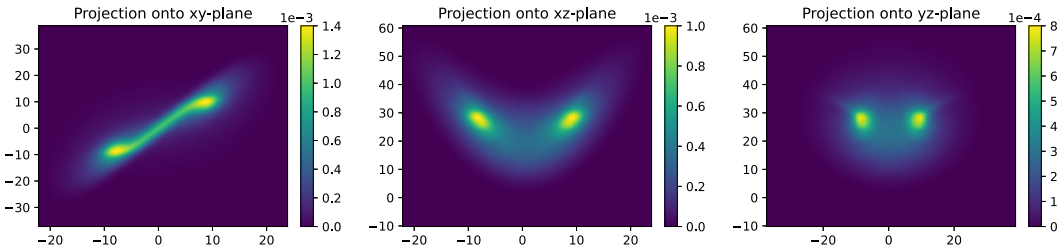
(a) Steady-state solution to (1.2)



(b) Histogram accumulated from noise-free Lorenz system time trajectory

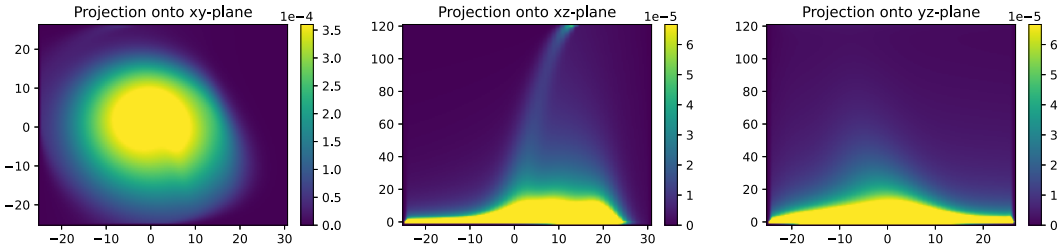


(c) Histogram accumulated from Lorenz system time trajectory with *intrinsic* noise

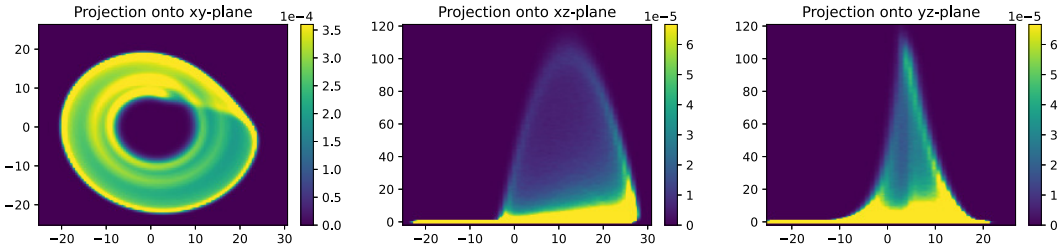


(d) Histogram accumulated from Lorenz system time trajectory with *extrinsic* noise

Figure 2. *Lorenz system. Top row: the steady state on the grid size* $93 \times 153 \times 143$ *by solving* (1.2). *The teleportation parameter is* $\epsilon = 10^{-6}$. *Second row: projections of physical invariant measure from noise-free time trajectory for* $T = 2 \times 10^6$. *Third row: projections of physical invariant measure from time trajectory with intrinsic noise* $\omega \sim \mathcal{N}(0, \mathbf{I})$. *Last row: projections of physical invariant measure from time trajectory with extrinsic noise* $\gamma \sim \mathcal{N}(0, \mathbf{I})$.



(a) Steady-state solution to (1.2)



(b) Histogram accumulated from Rössler system time trajectory with intrinsic noise

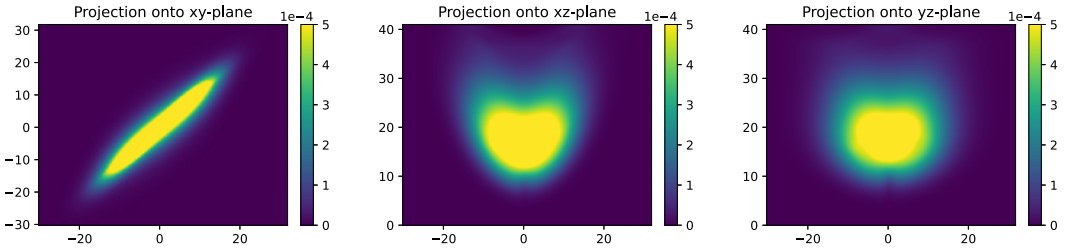
Figure 3. *Rössler system. Top row: the steady-state solution to* (1.2) *on the grid size* $94 \times 87 \times 106$. *The teleportation parameter is* $\epsilon = 10^{-6}$. *Bottom row: the histogram accumulated from Rössler system time trajectory for total time* $T = 1 \times 10^6$ *with intrinsic noise* $\omega \sim \mathcal{N}(0, 0.2\mathbf{I})$.

see the corresponding three projections of the physical invariant measure from *noise-free* time trajectory for total time $T = 2\times10^6$. The third row and the bottom row show three projections of the physical invariant measure from time trajectories of the same total time T but with *intrinsic* noise $\omega \sim \mathcal{N}(0, \mathbf{I})$ (the noise occurs on the right-hand side of the dynamical system as $\dot{\mathbf{x}} = v(\mathbf{x})+\omega$) and *extrinsic* noise $\gamma \sim \mathcal{N}(0, \mathbf{I})$ (the observation of the time trajectory suffers from noise as $\mathbf{x}_\gamma = \mathbf{x} + \gamma$), respectively. The bin size for all three histograms is a cube of volume 0.5^3.

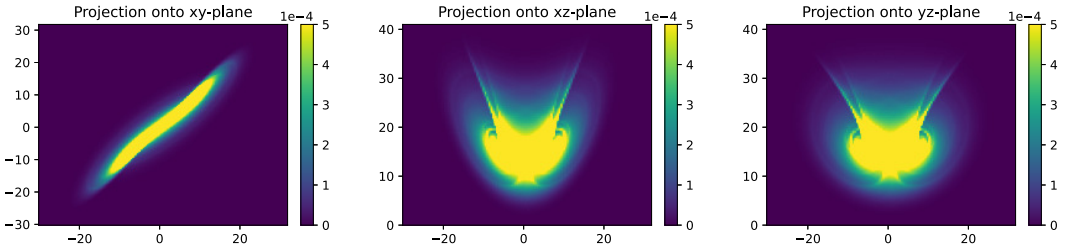
Similar plots for the Rössler system (6.2) are presented in Figure 3. Top row shows the steady-state solution to (1.2) computed on a grid size is $94 \times 87 \times 106$. The teleportation parameter is $\epsilon = 10^{-6}$. For the bottom row, the Rössler system time trajectory runs for a total time $T = 1 \times 10^6$ with an intrinsic noise $\omega \sim \mathcal{N}(0, 0.2\mathbf{I})$. The bin size for the histogram is a cube of volume 0.6^3.

Figure 4 shows the comparisons for the Chen system (6.3). The first row displays the three projections of the steady-state solution to (1.2) on a $104 \times 104 \times 69$ grid. The teleportation parameter is $\epsilon = 10^{-6}$. The bottom row shows the projections of the physical invariant measure accumulated from time trajectory with intrinsic noise for a total time $T = 5 \times 10^5$. The bin size for the histogram is a cube of volume 0.5^3. The intrinsic noise $\omega \sim \mathcal{N}(0, 0.2\mathbf{I})$.

6.2.2. The effect of noise. It is important to understand the fundamental limitations and challenges of converging the low-order solver for (1.2), particularly the role that the addition of the extrinsic and intrinsic noises play here as an approximation of the diffusive errors expected in the PDE solver.



(a) Steady-state solution to (1.2)



(b) Histogram accumulated from Chen system time trajectory with intrinsic noise

Figure 4. *Chen system. Top row: the steady-state solution to* (1.2) *on the grid size* $125 \times 125 \times 83$. *The teleportation parameter is* $\epsilon = 10^{-6}$. *Bottom row: the histogram accumulated from Chen system time trajectory with* $T = 5 \times 10^5$ *and intrinsic noise* $\omega \sim \mathcal{N}(0, 0.2\mathbf{I})$.

After the ODE is solved, the extrinsic noise applied to the trajectory corresponds to an effective Gaussian blur of the DNS results. In the limit of long-time DNS simulation, the true density is the result of taking every point on the invariant measure, represented by a delta function in state space based on the DNS solution, and then replacing it with a Gaussian ball of equal integral mass with width defined by the standard deviation of the noise. This process is equivalent to the Gaussian blur common in image processing.

The intrinsic noise case is more complicated. Since the three examples we have all admit nontrivial basins of attraction, the shape of the distribution depends on both the magnitude of the noise injected into the system and the dissipation rate in directions orthogonal to the attractor. Fluctuations off the attractor place the system in states subject to additional dissipation as the dynamics drive the solution back towards the attractor. The resulting trajectories are biased random walks that balance the diffusion of the noise with contraction in the stable state-space directions. While the extrinsic noise corresponds to a spatially uniform low pass filter, the blurring resulting from the intrinsic noise depends on the local stability and shape of the attractor in state space.

6.2.3. The effect of mesh size and numerical diffusion. While of a form dominated by diffusion, numerical errors of the PDE solver have a dependence on the flow velocity $\propto v^2 \Delta t$, as described in [10]. This is the well-known numerical diffusion that motivates running computational fluid dynamics solvers with a Courant–Freidrich–Lewy (CFL) condition number as close to 1 as possible for low-order methods to minimize the numerical diffusivity. While

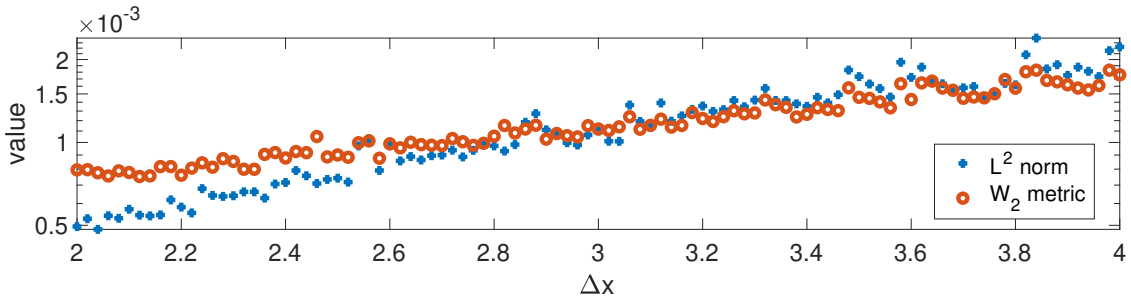


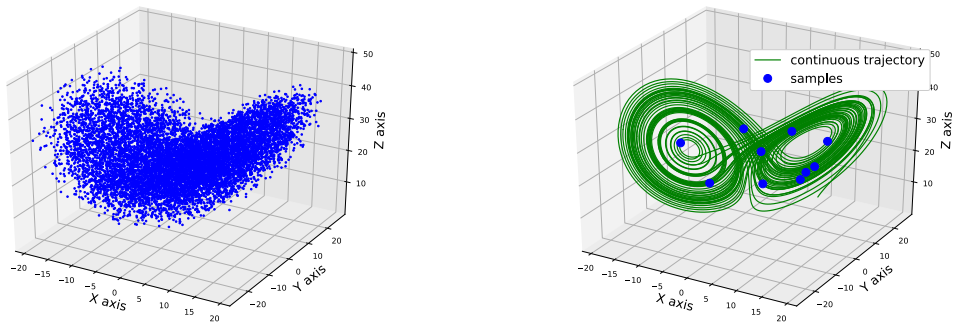
Figure 5. *The W_2 metric and the L^2 difference between the PDF accumulated from DNS with bin volume $(\Delta x)^3$ and the PDF solved as the steady-state solution to (1.2) with spatial spacing Δx. The PDFs are for the Lorenz system at the true parameters.*

in this work, we seek a steady-state solution, the time step of the forward operator has effectively been selected to comply with this CFL restriction in the act of ensuring that the forward operator is at least positive semidefinite in (3.2). Substituting the CFL restriction, $\Delta t = \Delta x / v_{max}$, into the expression for the numerical diffusivity, it can be seen that numerical diffusion in the PDE solver is effectively $\propto v^2 \Delta x / v_{max}$, which is bounded by $v_{max} \Delta x$, suggesting first-order convergence with Δx if v_{max} is bounded. More detailed numerical analysis for the convergence and numerical errors can be found in [49]. The linear convergence is also seen in Figure 5, where we compare the differences between the PDF accumulated from the Lorenz system DNS with again $T = 2 \times 10^6$ and the steady-state solution to (1.2), both evaluated at the true parameters for the Lorenz system. The histogram bin size changes as we use different Δx's in the finite volume discretization.

We remark that all the inversion tests in this paper use $\Delta x = 3$. It is for demonstration only and thus far from being optimal. The size of the Markov matrix M grows $\propto \Delta x^{-3}$ as Δx decreases, making it very expensive to compute the steady state at a fine mesh. Mesh-refinement strategies could help provide better parameter estimates while saving computational costs of the forward solve. This, along with more efficient numerical implementations, will be left to future work.

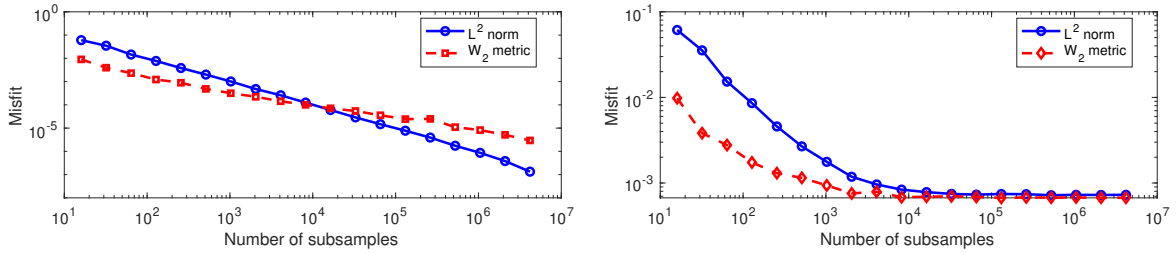
6.2.4. The effect of random samples. One main advantage of the proposed framework is that we allow the trajectory data to be "slowly" sampled, in which case we do not have access to the state-space velocity or velocity estimates, i.e., the $\dot{\mathbf{x}}$. In Figure 6(a), we illustrate the total samples of the trajectory that will be used in the parameter inference, while Figure 6(b) displays the relationship of the first 10 samples in the time series with the continuous trajectory in the corresponding time window. One can observe that our random samples of state-space positions are "sparse" and could not accurately estimate the state-space velocity. Later in subsection 6.3.3, we use the reference measure constructed from such slowly sampled and completely randomized state measurements to perform parameter identification.

In Figure 7, we numerically investigate the relationship between the amount of state-space position samples and the approximation error for the invariant measure. In Figure 7(a), we set the reference density to be the histogram accumulated from 10^8 samples and compare it with the histogram accumulated from much fewer samples. We observe the classical Monte



(a) 10^4 samples of the trajectory (b) Zoom-in view of the first 10 samples

Figure 6. *Left:* 10^4 *random samples of the Lorenz trajectory; Right: illustration of the first* 10 *samples of Figure* 6(a) *compared with the continuous trajectory.*



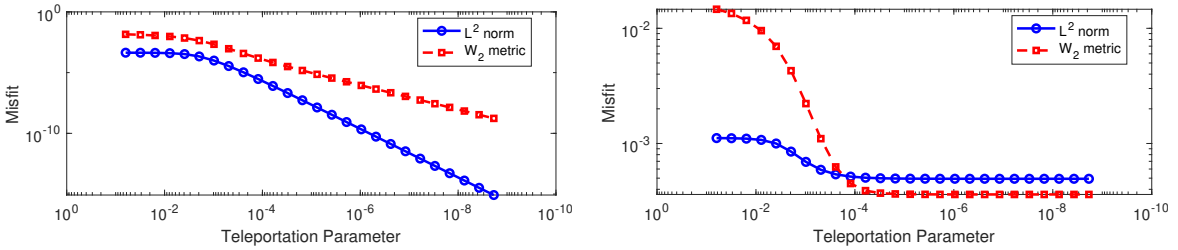
(a) DNS vs. DNS (b) DNS vs. PDE steady state

Figure 7. *Left: the misfit between the density accumulated from subsampled data and the one from the entire trajectory. Right: the misfit between the density accumulated from subsampled data and the computed steady-state solution to the continuity equation.*

Carlo error, $\mathcal{O}(1/\sqrt{N})$, where N is the number of samples. In Figure 7(b), we change the reference density to the steady-state solution to the continuity equation (see (1.2)). The error plateaus for large N since the modeling error, mainly due to the numerical diffusion discussed in subsection 6.2.3, becomes the dominant factor of the mismatch when N is large enough. It also indicates that we do not need too many trajectory samples to perform parameter identification.

6.2.5. The effect of the teleportation parameter. To obtain the steady-state solution, we used the so-called teleportation trick to regularize the Markov matrix; see subsection 3.2 for details. Here, we numerically investigate the impact of the teleportation parameter ϵ on the obtained steady-state solution.

In Figure 8(a), we use the steady-state density in which the teleportation parameter $\epsilon = 0$ as the reference data. We then compare it with those generated with a nonzero ϵ in terms of the L^2 norm and W_2 metric. The misfit monotonically decreases to zero as $\epsilon \to 0$. When the reference density is replaced by the histogram accumulated from trajectory samples, the misfit again plateaued when ϵ becomes small since the modeling error, mainly the numerical



(a) PDE steady state vs. PDE steady state (b) PDE steady state vs. DNS

Figure 8. *The L^2 norm and W_2 metric when the steady-state solution of various teleportation parameters is compared with the steady state without teleportation (left), and with a fixed invariant measure obtained from the trajectory samples (right).*

diffusion from the finite volume solver, becomes the dominant factor of their difference. As discussed in subsection 6.2.3, the error from numerical diffusion could be effectively reduced as the mesh is refined, i.e., $\Delta x \to 0$.

6.3. Parameter inference. One main goal of this work is to perform parameter identification using the invariant measure, a macroscopic statistical quantity, as the data, rather than inferring the parameter directly from time trajectories. All steady-state distributions in this section are solved on a mesh with spacing $\Delta x = 3$.

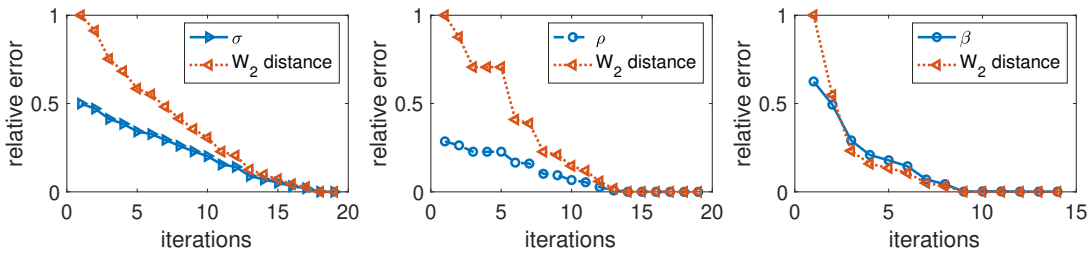
6.3.1. Single parameter inference. We first focus on the single-parameter reconstruction by assuming that the other parameters in the dynamical systems are accurately known. Figure 9(a) shows the single-parameter inversions of the Lorenz system where the ones for Rössler and Chen systems can be found in subsection C.1. All experiments use the squared W_2 metric as the objective function; see (5.2). One can see that both the objective function that measures the data mismatch and the relative error of the reconstructed parameters decay to zero rapidly.

We remark that in these tests, the target invariant measure (our reference data) is simulated as the steady-state solution to (1.2) at the true parameters, using the same PDE solver that produces the synthetic data. Later, to mimic the realistic scenarios, we will show numerical inversion tests where the reference data directly come from time trajectories and thus contain both noise and model discrepancy.

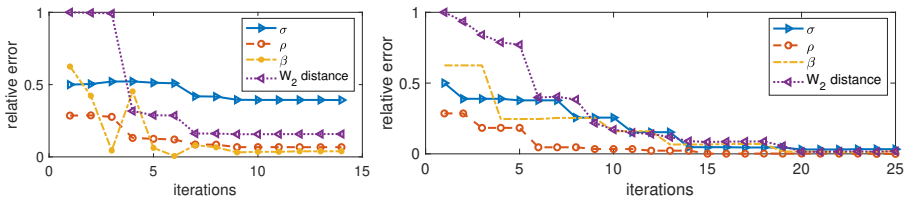
6.3.2. Multiparameter inference via coordinate gradient descent. For numerical tests we consider here, all dynamical systems have three parameters, while our observation is the invariant measure $\rho(\theta_1, \theta_2, \theta_3)$. Under certain assumptions for the continuous dependency on the parameters, the first-order variation gives

$$\delta\rho = \rho_{\theta_1}\delta\theta_1 + \rho_{\theta_2}\delta\theta_2 + \rho_{\theta_3}\delta\theta_3,$$

which highlights the issue of multiparameter inversion. In the forward problem, a small perturbation in each parameter causes a corresponding perturbation in the data ρ, but in the inverse problem, the observed misfit in ρ could be contributed from any of the parameters, causing nonzero and possibly wrong gradient updates.



(a) Single-parameter inversion



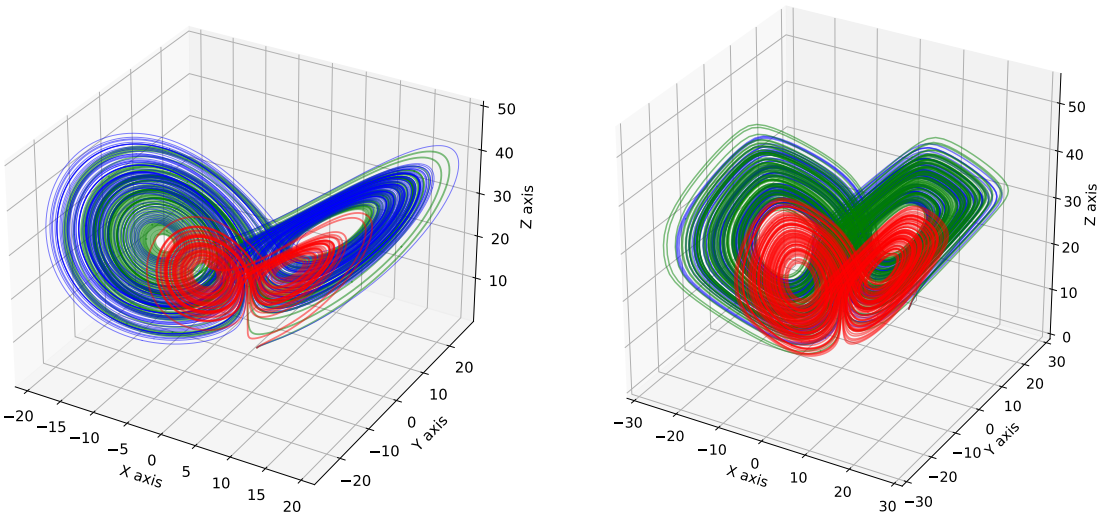
(b) Multi-parameter inversion

Figure 9. *Top row: Lorenz system single-parameter inference starting with $\sigma = 5$ (left), $\rho = 20$ (middle), $\beta = 1$ (right), respectively. Bottom row: multiparameter inference by updating three parameters simultaneously (bottom left) and using coordinate gradient descent (bottom right) with initial guess $(\sigma, \rho, \beta) = (5, 20, 1)$. The reference PDF is generated through the same numerical solver producing the synthetic PDF.*

Numerical strategies exist to reduce the interparameter trade-off. One may mitigate the interparameter dependency either from the formulation of the optimization problem or through the optimization algorithm. Here, we separate the parameters in the optimization algorithm by using the coordinate gradient descent by only updating one parameter at one iteration.

Figure 9(b) shows the Lorenz system multiparameter inversion. We remark again that the reference data in these tests are produced by the same PDE solver that produces the synthetic data and thus contains no modeling discrepancy. The left plot in Figure 9(b) shows the convergence history of simultaneously updating all three parameters, but the iterates get stuck at an incorrect set of values with no feasible descent direction. On the other hand, the right plot shows the convergence result using coordinate gradient descent. The gradient descent algorithm quickly converges to the true value $(\sigma, \rho, \beta) = (10, 28, 8/3)$ starting from $(5, 20, 1)$. The different convergence behaviors of the two plots in Figure 9(b) demonstrate that the reconstruction process is affected by the interparameter interaction.

6.3.3. Parameter inference for chaotic systems with noise. In this work, we formulate an inverse problem into a nonlinear regression problem, usually subject to at least three sources of error: model discrepancy, data noise, and optimization error. As discussed earlier, the almost perfect reconstructions in the previous section are achieved under the so-called "inverse crime" regime and thus are immune to the first two types of errors. Here, we set up tests to avoid the "inverse crime" regime. We first solve the dynamical system forward in time with a fixed time step Δt from $t = 0$ to $T = 2 \times 10^6$, achieving the DNS solution. We then randomly subsample 10^4 state-space positions; see Figure 6 for their illustrations. The reference data, i.e., the target estimated invariant measure, is obtained from the histogram



(a) Lorenz with extrinsic noise

(b) Arctan Lorenz with intrinsic noise

Figure 10. *Comparison among the dynamics produced by the initial parameter (red); true parameter (green); reconstructed parameters (blue) for two examples.*

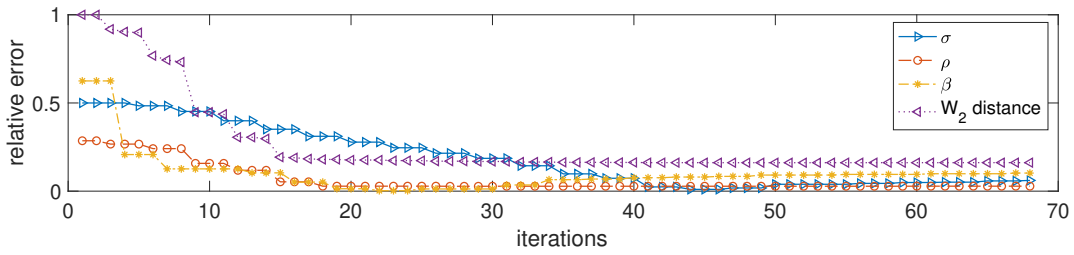


Figure 11. *Lorenz system (6.1): Multiparameter inference using coordinate gradient descent with initial guess* $(\sigma, \rho, \beta) = (5, 20, 1)$. *The reference PDF is the histogram from the time trajectory with extrinsic noise.*

that results from binning the subsampled data into cubic boxes in $\mathbb{R}^3$. Moreover, we also use time trajectories incorporating intrinsic and extrinsic noises. Starting from the initial guess $(5, 20, 1)$, the multiparameter inversion for the Lorenz system (6.1) with extrinsic noise converges to $(\sigma, \rho, \beta) = (10.63, 28.82, 3.04)$, and the test with the intrinsic noise converges to $(10.50, 28.41, 2.89)$. For the Arctan Lorenz system (6.4), the reconstruction converges to $(11.37, 27.64, 2.35)$ starting from $(5, 20, 2)$, where the reference data are polluted by the intrinsic noise. We demonstrate the reconstructed dynamics in Figure 10. The plot for the convergence history of the Lorenz example is shown in Figure 11. More numerical results can be found in Appendix C.

Earlier in subsection 6.2.3, we have analyzed the numerical error between the synthetic steady-state solution using the first-order finite volume method. It is shown both in Figure 5 and by numerical analysis that the error grows linearly with Δx. It is also a good characterization of the model discrepancy and could be utilized to design specific stopping criteria

to avoid parameter overfitting. For example, Figure 5 could serve as the baseline: whenever the objective function (W_2 metric in our case) is minimized to a value smaller than the model discrepancy, one should execute early stopping: terminate the iterative parameter reconstruction to avoid overfitting the noise. In machine learning, early stopping is designed to monitor the generalization error of one model and stop training when the generalization error begins to degrade, which is quite similar to the situation we encounter here.

7. Conclusion. In this paper, we propose a data-driven approach for parameter estimation of chaotic dynamical systems. There are two significant contributions. First, we shift from an ODE forward model to the related PDE forward model through a focus upon physical measures. Instead of using time trajectories as the inference data, we treat statistics accumulated from the direct numerical simulation as the observable, whose analog in the PDE setting is the steady-state solution to (1.2). As a result, the original parameter identification problem is translated into a data-fitting, PDE-constrained optimization problem. We then use an upwind scheme based on the finite volume method to discretize and solve the forward problem. Second, we use the quadratic Wasserstein metric from OT to measure the difference between the synthetic and the reference datasets. We provide a rigorous analysis of the differentiability of Wasserstein-based parameter estimation and then derive two ways of calculating the Wasserstein gradient following a discretize-then-optimize approach. In particular, the adjoint approach is efficient as the computational cost of gradient evaluation is independent of the size of the unknown parameters, making the method scalable for large-scale parameterization of the velocity fields. Finally, we show several numerical results to demonstrate the promise of this new approach for chaotic dynamical system parameter identification.

For this method, sufficient data are required for the histogram estimate to converge to the reference distribution. As in any nonparametric density estimate, the amount of data is dependent on the coarseness of the approximation and level of stochastic error tolerated. In this work, knowledge of the full state is also presumed. The approximated invariant measure from the time trajectories as our reference data might be a singular probability measure with highly complex support that has fractional fractal dimension. Thus, we use the regularized forward PDE model as a surrogate in solving this inverse problem. We approximate the steady-state solution to the PDE model with first-order accuracy based on the finite volume upwind discretization. Due to the sparsity of the Markov matrix and a coarse grid, we can evaluate the gradient of the resulting PDE-constrained optimization problem quite efficiently in terms of both memory and computational complexity. The Wasserstein metric from OT is our objective function, which can compare measures with singular and compact support and handle the fractional fractal dimension of the reference invariant measure.

Appendix A. Proofs from section 4.

A.1. Proof of Proposition 4.1.

Proof. We fix $\theta_0 \in \Theta$ and first prove that (i) implies (ii). Note that (4.3) follows immediately from (4.2). Furthermore, assume that $(\phi, \psi) \in \Phi_c(\rho(\cdot, \theta_0), \rho^*)$ is an arbitrary pair of Kantorovich potentials. Note that (ϕ, ψ) are not necessarily in $\mathcal{S}(\theta_0)$. Since $\int_\Omega \nabla_\theta \rho(x, \theta_0) dx = 0$, we can add an arbitrary constant to ϕ and assume that $\sup \phi = \|c\|_\infty$. In that case, we obtain that $(\phi^{cc}, \phi^c) \in \mathcal{S}(\theta_0)$, and

$$\phi^{cc}(x) = \phi(x),\ x \in \operatorname{supp}(\rho(\cdot, \theta_0)), \quad \text{and} \quad \phi^c(y) = \psi(y),\ y \in \operatorname{supp}(\rho^*).$$

Next, we have that $\operatorname{supp}(\nabla_\theta \rho(\cdot, \theta_0)) \subset \operatorname{supp}(\rho(\cdot, \theta_0))$. Therefore, we have that

$$\int_\Omega \phi(x) \nabla_\theta \rho(x, \theta_0) dx = \int_\Omega \phi^{cc}(x) \nabla_\theta \rho(x, \theta_0) dx,$$

and (4.4) follows from (4.2) and (4.3).

Next, we prove (i). We apply [12, Proposition 4.12] with $U = \Theta$, $X = C(\Omega) \times C(\Omega)$, $\Phi = C = K_c$, and an objective function given by

$$I(\phi, \psi, \theta) = \int_\Omega \phi(x) \rho(x, \theta) dx + \int_\Omega \psi(y) \rho^*(y) dy.$$

For $\theta_1, \theta_2 \in \Theta$ such that $[\theta_1, \theta_2] \subset \Theta$, we have that

$$|I(\phi_2, \psi_2, \theta_2) - I(\phi_1, \psi_1, \theta_1)| \le \|\phi_2 - \phi_1\|_\infty + \|\psi_2 - \psi_1\|_\infty + \|\phi_1\|_\infty \|\eta\|_1 |\theta_2 - \theta_1|, \tag{A.1}$$

and so I is continuous. Since K_c is compact, the supcompactness condition holds. Furthermore, A2 and the dominated convergence theorem yield the directional differentiability of $I(\phi, \psi, \cdot)$ with

$$I'(\phi, \psi, \theta_0, \Delta\theta) = \int_\Omega \phi(x) \nabla_\theta \rho(x, \theta_0) dx \cdot \Delta\theta.$$

Finally, assume that $t_n \to 0+$, $(\phi_n, \psi_n) \in K_c$, $\Delta\theta \in \mathbb{R}^m$, and $(\phi_n, \psi_n) \to (\phi, \psi) \in K_c$. Then by the dominated convergence theorem we have that

$$\begin{aligned} &\lim_{n\to\infty} \frac{I(\phi_n, \psi_n, \theta_0 + t_n \Delta\theta) - I(\phi_n, \psi_n, \theta_0 \theta)}{t_n} \\ &= \lim_{n\to\infty} \int_\Omega \phi_n(x) \frac{\rho(x, \theta_0 + t_n \Delta\theta) - \rho(x, \theta_0)}{t_n} dx = I'(\phi, \psi, \theta_0, \Delta\theta). \end{aligned}$$

Thus, all conditions in [12, Proposition 4.12] are satisfied and (4.2) follows. ∎

A.2. Proof of Theorem 4.2.

Proof. Assume that A1–A3 hold. Then (A.1) yields that $\theta \mapsto I(\phi, \psi, \theta)$ is locally Lipschitz for all $(\phi, \psi) \in C(\Omega) \times C(\Omega)$. Invoking (4.1), we conclude that f is locally Lipschitz and a.e. differentiable by Rademacher's theorem [29, section 3.1].

Next, assume that A4 also holds and denote $C_0 = \|c\|_\infty \|h\|_1$. For arbitrary $(\phi, \phi^c) \in K_c$ we have that

$$\begin{aligned} I(\phi, \phi^c, \theta) + \frac{C_0 |\theta|^2}{2} &= \int_\Omega \phi(x) \left(\rho(x, \theta) + \frac{h(x)|\theta|^2}{2} \right) dx + \int_\Omega \phi^c(y) \rho^*(y) dy \\ &\quad + \left(\|c\|_\infty \|h\|_1 - \int_\Omega \phi(x) h(x) dx \right) \frac{|\theta|^2}{2}. \end{aligned}$$

Since $0 \le \phi \le \|c\|_\infty$, and $\theta \mapsto \rho(x,\theta) + \frac{h(x)|\theta|^2}{2}$ is convex for a.e. x, we obtain that $\theta \mapsto I(\phi,\phi^c,\theta) + \frac{C_0|\theta|^2}{2}$ is convex. Invoking Kantorovich duality again, we obtain that

$$f(\theta) + \frac{C_0|\theta|^2}{2} = \sup_{(\phi,\phi^c)\in K_c} I(\phi,\phi^c,\theta) + \frac{C_0|\theta|^2}{2}$$

is convex. Thus, by a theorem of Anderson and Klee [2] f is differentiable up to a set of Hausdorff dimension $d-1$. ∎

A.3. Proof of Theorem 4.5.

Proof. Fix an arbitrary pair of Kantorovich potentials (ϕ_1,ψ_1), (ϕ_2,ψ_2). Note that (4.7) guarantees that $\mathrm{int}(\mathrm{supp}(\rho)) \neq \emptyset$, and $\{O_k\}$, $\{E_k\}$ are well defined.

First, we prove that $\phi_2 - \phi_1$ is constant on $\mathrm{cl}(O_k)$ for all k. Fix an optimal plan $\pi_0 \in \Gamma_0(\rho,\rho^*)$. For all $x \in \mathrm{supp}(\rho)$ there exists $y \in \Omega$ such that $(x,y) \in \mathrm{supp}(\pi_0)$. Therefore $\phi_i(x)+\psi_i(y) = c(x,y)$, and so $\phi_i(x) = \psi_i^c(x)$ for $x \in \mathrm{supp}(\rho)$. Furthermore, since $c \in C^1(\Omega^2)$ is locally Lipschitz continuous, ϕ_i are locally Lipschitz continuous in O_k. Thus, by Rademacher's theorem we have that ϕ_i are a.e. differentiable in O_k, and by [66, Proposition 1.15] we obtain that $\nabla\phi_2 = \nabla\phi_1$ a.e. in O_k. Since O_k are connected and ϕ_i are continuous, we obtain that $\phi_2 - \phi_1 = \lambda_k$ in $\mathrm{cl}(O_k)$ for some constants λ_k.

Next, we show that $\lambda_k = \lambda_l$ for all k,l. We start with a claim that

$$\psi_i(y) = \inf_{x\in\mathrm{cl}(O_k)} \{c(x,y) - \phi_i(x)\}, \quad y \in E_k. \tag{A.2}$$

Indeed, we have that $y = \lim_{n\to\infty} y_n$, where y_n are such that $(x_n,y_n) \in \mathrm{supp}(\pi_n)$ for some $\pi_n \in \Gamma_0(\rho,\rho^*)$, and $x_n \in \mathrm{cl}(O_k)$. Therefore, for all n we have that $\phi_i(x_n)+\psi_i(y_n) = c(x_n,y_n)$, and so

$$\psi_i(y_n) = \inf_{x\in\mathrm{cl}(O_k)} \{c(x,y_n) - \phi_i(x)\}.$$

Since both ψ_i and $y \mapsto \inf_{x\in\mathrm{cl}(O_k)}\{c(x,y) - \phi_i(x)\}$ are continuous, we deduce (A.2). Next, $\phi_2 - \phi_1 = \lambda_k$ in $\mathrm{cl}(O_k)$, and (A.2) yields that $\psi_2 - \psi_1 = -\lambda_k$ in E_k.

Now fix arbitrary k,l. Since $\mathrm{cl}(O_k), \mathrm{cl}(O_l)$ are linked, there exist $\{i_j\}_{j=1}^m$ such that $k = i_1, l = i_m$, and $E_{i_j} \cap E_{i_{j+1}} \neq \emptyset$, $1 \le j \le m$. Since $\psi_2 - \psi_1 = -\lambda_{i_j}$ in E_{i_j}, and $\psi_2 - \psi_1 = -\lambda_{i_{j+1}}$ in $E_{i_{j+1}}$, we obtain that $\lambda_{i_j} = \lambda_{i_{j+1}}$ for all j. Thus, $\lambda_k = \lambda_l$, and, consequently, $\phi_2 - \phi_1 = \lambda$ in $\mathrm{int}(\mathrm{supp}(\rho)) = \cup_k O_k$. Finally, (4.7) and the continuity of ϕ_i yield that $\phi_2 - \phi_1 = \lambda$ in $\mathrm{supp}(\rho)$. ∎

A.4. Proof of Proposition 4.7.

Proof. The proof is based on the following points.

1. We have that $|\partial_\theta \rho(x,\theta)| = |\chi_{[0,1]}(x) - \chi_{[2,3]}(x)| \le 1$, for all $x \in \Omega$.
2. Assume that $-0.5 < \theta_1 < \theta_2 < 0.5$. In $\mathbb{R}$, OT maps are precisely the order-preserving ones [71, section 2.2]. The total mass of $[0,1]$ with respect to $\rho(\cdot,\theta_1)$ and $\rho(\cdot,\theta_2)$ is $0.5+\theta_1$ and $0.5+\theta_2$, respectively. Since $0.5+\theta_1 < 0.5+\theta_2$, all of the mass of $\rho(\cdot,\theta_1)$ from $[0,1]$ has to be transported to $[0,1]$ with a linear transport map $T(x) = \frac{0.5+\theta_1}{0.5+\theta_2}x$.

Meanwhile, the excess mass of $\rho(\cdot,\theta_2)$ in $[0,1]$, supported on $\left[\frac{0.5+\theta_1}{0.5+\theta_2},1\right]$, has to be transported from $[2,3]$, and therefore has to travel a distance ≥ 1. Since the excess mass of $\rho(\cdot,\theta_2)$ left in $[0,1]$ is equal to $0.5+\theta_2-(0.5+\theta_1)=\theta_2-\theta_1$, we obtain that the transport cost is at least $(\theta_2-\theta_1)\cdot 1^p$. Thus,

$$W_p(\rho(\cdot,\theta_1),\rho(\cdot,\theta_2)) \geq |\theta_2-\theta_1|^{\frac{1}{p}} \quad \forall \theta_1,\theta_2 \in (-0.5,0.5),$$

which means that $\theta \mapsto \rho(\cdot,\theta)$ is not absolutely continuous with respect to W_p metric.

3. Fix an arbitrary $|\theta|<0.5$. We only use the fact that $\operatorname{supp}(\rho(\cdot,\theta)) \subsetneq [0,4]$. Assume by contradiction that $\rho \mapsto W_p^p(\rho,\rho^*)$ is Gâteaux differentiable at $\rho(\cdot,\theta)$ in the sense of [66, Definition 7.12]; that is, there exists a measurable function g such that

$$\frac{d}{d\epsilon} W_p^p(\rho(\cdot,\theta)+\epsilon(\widetilde{\rho}-\rho(\cdot,\theta)),\rho^*)|_{\epsilon=0+} = \int_0^4 g(x)(\widetilde{\rho}(x)-\rho(x,\theta))dx$$

for all $\widetilde{\rho} \in \mathscr{P}([0,4]) \cap L^\infty([0,4])$. Let $\phi \in C([0,4])$ be an arbitrary Kantorovich potential. From [66, Proposition 7.17] we have that ϕ is in the subdifferential of $\rho \mapsto W_p^p(\rho,\rho^*)$ at $\rho(\cdot,\theta)$; that is,

$$W_p^p(\rho,\rho^*) \geq W_p^p(\rho(\cdot,\theta),\rho^*) + \int_0^4 \phi(x)(\rho(x)-\rho(x,\theta))dx \quad \forall \rho \in \mathscr{P}([0,4]).$$

Hence, we have that

$$\frac{d}{d\epsilon} W_p^p(\rho(\cdot,\theta)+\epsilon(\widetilde{\rho}-\rho(\cdot,\theta)),\rho^*)|_{\epsilon=0+} \geq \int_0^4 \phi(x)(\widetilde{\rho}(x)-\rho(x,\theta))dx.$$

Combining this inequality with the preceding equality, we obtain

$$\int_0^4 (\phi(x)-g(x))\rho(x,\theta)dx \geq \int_0^4 (\phi(x)-g(x))\widetilde{\rho}(x)dx$$

for all $\widetilde{\rho} \in \mathscr{P}(\Omega) \cap L^\infty(\Omega)$ and Kantorovich potentials ϕ. Fix an arbitrary potential ϕ_0 and take $\widetilde{\rho}(x)=\chi_{(1,2)}(x)$. Furthermore, for every $\lambda \in \mathbb{R}$ consider

$$\phi_\lambda(x) = \phi_0(x) + \lambda(x-1)(2-x)\chi_{(1,2)}(x).$$

Note that ϕ_λ is continuous and $\phi_\lambda = \phi_0$ in $\operatorname{supp}\rho(\cdot,\theta)$. Thus, if (ϕ_0,ψ_0) is a pair of Kantorovich potentials, then (ϕ_λ,ψ_0) is also a pair of Kantorovich potentials. Plugging $\phi=\phi_\lambda$ into the inequality above we obtain

$$\begin{aligned}\int_0^4 (\phi_0(x)-g(x))\rho(x,\theta)dx &\geq \int_1^2 (\phi_0(x)-g(x))dx + \lambda\int_1^2 (x-1)(2-x)dx \\ &= \int_1^2 (\phi_0(x)-g(x))dx + \frac{\lambda}{6}\end{aligned}$$

for all $\lambda \in \mathbb{R}$, which is a contradiction.

4. For this and the following item, we need an explicit characterization of the OT map, T_θ, from $\rho(\cdot,\theta)$ to ρ^*. For $\theta = 0$, we have that ρ^* is a translation of $\rho(\cdot,0)$. Thus, we have that

$$T_0(x) = x + 1, \quad W_p^p(\rho(\cdot,0),\rho^*) = 1.$$

Next, for $\theta > 0$ we have that $\rho([0,1],\theta) = 0.5 + \theta > 0.5 = \rho^*([1,2])$. Therefore,

$$T_\theta(x) = \begin{cases} 1 + \dfrac{0.5+\theta}{0.5}x, & x \in \left[0, \dfrac{0.5}{0.5+\theta}\right], \\ 3 + \dfrac{0.5+\theta}{0.5}\left(x - \dfrac{0.5}{0.5+\theta}\right), & x \in \left[\dfrac{0.5}{0.5+\theta}, 1\right], \\ 3 + \dfrac{\theta}{0.5} + \dfrac{0.5-\theta}{0.5}(x-2), & x \in [2,3]. \end{cases} \tag{A.3}$$

For $\theta < 0$ we have that $\rho([0,1],\theta) = 0.5 + \theta < 0.5 = \rho^*([1,2])$. Therefore,

$$T_\theta(x) = \begin{cases} 1 + \dfrac{0.5+\theta}{0.5}x, & x \in [0,1], \\ 1 + \dfrac{0.5+\theta}{0.5} + \dfrac{0.5-\theta}{0.5}(x-2), & x \in \left[2, 2 - \dfrac{\theta}{0.5-\theta}\right], \\ 3 + \dfrac{0.5-\theta}{0.5}\left(x - 2 + \dfrac{\theta}{0.5-\theta}\right), & x \in \left[2 - \dfrac{\theta}{0.5-\theta}, 3\right]. \end{cases} \tag{A.4}$$

For all θ, the connected components of $\operatorname{int}(\operatorname{supp}(\rho(\cdot,\theta)))$ are

$$O_1 = (0,1), \quad O_2 = (2,3).$$

Furthermore, using the definition (4.6) and invoking (A.3), (A.4) we obtain

$$E_1 = \begin{cases} [1,2], \ \theta = 0, \\ [1,2] \cup \left[3, 3 + \dfrac{\theta}{0.5}\right], \ \theta > 0, \\ \left[1, 1 + \dfrac{0.5+\theta}{0.5}\right], \ \theta < 0, \end{cases} \quad E_2 = \begin{cases} [3,4], \ \theta = 0, \\ \left[3 + \dfrac{\theta}{0.5}, 4\right], \ \theta > 0, \\ \left[1 + \dfrac{0.5+\theta}{0.5}, 2\right] \cup [3,4], \ \theta < 0. \end{cases}$$

Thus, we have that

$$E_1 \cap E_2 = \begin{cases} \emptyset, & \theta = 0, \\ \left\{3 + \dfrac{\theta}{0.5}\right\}, & \theta > 0, \\ \left\{1 + \dfrac{0.5+\theta}{0.5}\right\}, & \theta < 0, \end{cases}$$

which means that $\operatorname{cl}(O_1), \operatorname{cl}(O_2)$ are linked for all $|\theta| < 0.5$ except $\theta = 0$.

5. The differentiability of $\theta \mapsto W_p^p(\rho(\cdot,\theta),\rho^*)$ at $\theta \neq 0$ follows from Corollary 4.6, and Item 4 above. Recall that $W_p^p(\rho(\cdot,0),\rho^*) = 1$. Next, assume that $\theta > 0$. From (A.3),

$$W_p^p(\rho(\cdot,\theta),\rho^*) = \sum_{k=1}^{3} \int_{I_k} |T_\theta(x) - x|^p \rho(x,\theta) dx, \tag{A.5}$$

$$\text{where} \quad I_1 = \left[0, \frac{0.5}{0.5+\theta}\right], \quad I_2 = \left[\frac{0.5}{0.5+\theta}, 1\right], \quad I_3 = [2,3].$$

For $x \in I_1 \cup I_3$ we use the elementary inequality

$$|T_\theta(x) - x|^p \geq 1 + p(T_\theta(x) - x - 1).$$

For $x \in I_2$, we have that

$$|T_\theta(x) - x|^p \geq 2^p.$$

Plugging these inequalities into (A.5) and using (A.3) for evaluating elementary integrals, we obtain

$$W_p^p(\rho(\cdot,\theta),\rho^*) \geq 1 + (2^p + p - 1)\theta - p\theta^2, \quad 0 < \theta < 0.5,$$

and so

$$\liminf_{\theta \to 0+} \frac{W_p^p(\rho(\cdot,\theta),\rho^*) - W_p^p(\rho(\cdot,0),\rho^*)}{\theta} \geq 2^p + p - 1.$$

For $\theta < 0$, we have that

$$W_p^p(\rho(\cdot,\theta),\rho^*) = \sum_{k=1}^{3} \int_{J_k} |T_\theta(x) - x|^p \rho(x,\theta) dx, \tag{A.6}$$

where $J_1 = [0,1]$, $J_2 = [2, 2 - \frac{\theta}{0.5-\theta}]$, and $J_3 = [2 - \frac{\theta}{0.5-\theta}, 3]$. Furthermore,

$$\begin{aligned} |T_\theta(x) - x|^p &\geq 1 + p(T_\theta(x) - x - 1), \quad x \in J_1 \cup J_3, \\ |T_\theta(x) - x|^p &\geq 0, \quad x \in J_2. \end{aligned}$$

Plugging these inequalities into (A.6), we obtain

$$W_p^p(\rho(\cdot,\theta),\rho^*) \geq 1 + (p+1)\theta + p\theta^2, \quad -0.5 < \theta < 0,$$

and so

$$\limsup_{\theta \to 0-} \frac{W_p^p(\rho(\cdot,\theta),\rho^*) - W_p^p(\rho(\cdot,0),\rho^*)}{\theta} \leq p + 1.$$

Since $2^p + p - 1 > p + 1$ for $p > 1$, we obtain that $\theta \mapsto W_p^p(\rho(\cdot,\theta),\rho^*)$ is not differentiable at $\theta = 0$. ∎

A.5. Proof of Proposition 4.8.

Proof. Assume by contradiction that there exist $(\phi_n, \psi_n) \in \Phi_c(\rho(\cdot, \theta_0), \rho^*)$ and $\epsilon_0 > 0$ such that $I(\phi_n, \psi_n) > f(\theta_0) - \frac{1}{n}$ and

$$\left| \nabla_\theta f(\theta_0) - \int_\Omega \phi_n^{cc}(x) \nabla_\theta \rho(x, \theta_0) dx \right| \geq \epsilon_0. \tag{A.7}$$

Note that by adding a suitable constant to ϕ_n, we can assume that $\sup \phi_n = \|c\|_\infty$. Thus, $(\phi_n^{cc}, \phi_n^c) \in K_c$ and

$$f(\theta_0) \geq I(\phi_n^{cc}, \phi_n^c, \theta_0) \geq I(\phi_n, \psi_n, \theta_0) > f(\theta_0) - \frac{1}{n}.$$

Since K_c is compact, we have that $(\phi_n^{cc}, \phi_n^c) \to (\phi, \phi^c) \in K_c$ at least through a subsequence. Thus,

$$I(\phi, \phi^c, \theta_0) = \lim_{n \to \infty} I(\phi_n^{cc}, \phi_n^c, \theta_0) = f(\theta_0),$$

and so $(\phi, \phi^c) \in \mathcal{S}(\theta_0)$. Hence, from Proposition 4.1 we have that

$$\begin{aligned} &\left| \nabla_\theta f(\theta_0) - \int_\Omega \phi_n^{cc}(x) \nabla_\theta \rho(x, \theta_0) dx \right| \\ &= \left| \int_\Omega \phi(x) \nabla_\theta \rho(x, \theta_0) dx - \int_\Omega \phi_n^{cc}(x) \nabla_\theta \rho(x, \theta_0) dx \right| \leq \|\phi - \phi_n^{cc}\|_\infty \|\eta\|_1, \end{aligned}$$

which contradicts (A.7) and finishes the proof. ∎

Appendix B. Numerical schemes for computing the gradient.

B.1. Numerical scheme for (5.6) and (5.8). We remark that the first equations in both (5.6) and (5.8) are the same, which corresponds to solving the forward problem (3.5) given the current iterate of the unknown parameter θ^l. There are at least three ways to solve the linear system: (1) the power method, (2) Richardson iteration, and (3) the sparse linear solve. We refer the readers to [39] for more details about the first two approaches and explain (3) in more detail.

In (3.5), we are interested in finding the solution ρ to the linear system

$$M_\epsilon \rho = (1 - \epsilon) M \rho + \frac{\epsilon}{n} \mathbf{1} \mathbf{1}^\top \rho = \rho, \tag{B.1}$$

where $\mathbf{1} = [1, 1, \ldots, 1]^\top$, M is defined in (3.3), and ϵ is our teleportation (regularization) parameter. Thus, we can rewrite the linear system as

$$((1 - \epsilon) M - I) \rho = -\frac{\epsilon}{n} \mathbf{1} \mathbf{1}^\top \rho.$$

Since the biggest eigenvalue of $(1 - \epsilon)M$ is $1 - \epsilon < 1$, the matrix on the left-hand side is invertible, and the solution is unique. We have

$$\rho^* = \frac{\rho}{\mathbf{1}^\top \rho} = -((1 - \epsilon) M - I)^{-1} \frac{\epsilon}{n} \mathbf{1}, \tag{B.2}$$

where ρ^* is our solution that we seek as $\mathbf{1}^\top \rho^* = 1$.

Regarding the second equation of (5.6), we consider the general linear system as below to solve for ζ given the right-hand side y where

$$(M_\epsilon - I)\zeta = y. \tag{B.3}$$

Based on Lemma 5.1, we know the right-hand side of (5.6), which we denote as y, satisfies $y \cdot \mathbf{1} = 0$, and $M_\epsilon - I$ has a one-dimensional null space with generator ρ^*. We seek a unique solution ζ^*, where $\mathbf{1}^\top \zeta^* = 0$. Note that (B.3) is equivalent to

$$((1-\epsilon)M - I)\zeta = y - \frac{\epsilon}{n}\mathbf{1}\mathbf{1}^\top \zeta.$$

Since $\mathbf{1}^\top \zeta^* = 0$, we obtain that ζ^* must satisfy $((1-\epsilon)M - I)\zeta^* = y$. As above, $(1-\epsilon)M - I$ is invertible, and this system has a unique solution. Therefore,

$$\zeta^* = ((1-\epsilon)M - I)^{-1} y. \tag{B.4}$$

Regarding the third equation of (5.8), we consider the general linear system as below to solve for ζ given the right-hand side b, where

$$(M_\epsilon^\top - I)\zeta = b. \tag{B.5}$$

Based on Lemma 5.2, we know the right-hand side of (5.8), which we denote as b, satisfies $b \cdot \rho^* = 0$, and $M_\epsilon^\top - I$ has a one-dimensional null space with generator $\mathbf{1}$. We seek a unique solution ζ^* where $\mathbf{1}^\top \zeta^* = 0$. Note that (B.5) is equivalent to

$$((1-\epsilon)M^\top - I)\zeta = b - \frac{\epsilon}{n}\mathbf{1}\mathbf{1}^\top \zeta.$$

Since $\mathbf{1}^\top \zeta^* = 0$, we obtain that ζ^* must satisfy $((1-\epsilon)M^\top - I)\zeta = b$. As above, $(1-\epsilon)M^\top - I$ is invertible, and this system has a unique solution. Therefore

$$\zeta^* = ((1-\epsilon)M^\top - I)^{-1} b. \tag{B.6}$$

Note that both the matrix $(1-\epsilon)M - I$ and its transpose are sparse. Therefore, it is relatively efficient to solve the linear systems that are essential for gradient calculation. The other components in (5.6) and (5.8) are rather straightforward once we solve (B.1) and (B.3) (or (B.5)).

B.2. Automatic differentiation. In the previous sections, we explained how to directly compute $\nabla_\theta K_{mat}(\theta)$ (or equivalently $\nabla_\theta M_\epsilon(\theta)$), which is necessary to calculate $\nabla_\theta \rho(\theta)$. However, if the numerical scheme for the forward problem changes, the structure of $K_{mat}(\theta)$ changes, and consequently, one has to rederive the explicit form of $\nabla_\theta K_{mat}(\theta)$. Such situations occur when using a higher-order finite volume method or switching to other standard numerical schemes such as the discontinuous Galerkin method. In order to make our code more flexible, we also implemented an automatic differentiation version using the Python library JAX [14].

We compute the full Jacobian matrices of $K_{mat}(\theta)$ using the `jacfwd` function. It uses forward-mode automatic differentiation, the most efficient choice when working with "tall"

matrices like those in this paper. The method of automatic differentiation is extremely valuable when working with real-world data. In many realistic situations, such as weather forecast, we do not have access to the underlying dynamical system, and thus we cannot compute $\nabla_\theta K_{mat}(\theta)$ directly. In future work, we plan on using neural networks to approximate the dynamical system from data. Given the large number of parameters and the complex functional form of a deep neural network, it would be impossible to derive $\nabla_\theta K_{mat}(\theta)$ explicitly, making the automatic differentiation approach necessary.

Appendix C. More numerical results.

C.1. Single parameter inversion for the Rössler and Chen systems. Figures 12 and 13 show the single-parameter inversion for the Rössler and Chen systems, respectively. The reference data is produced by the same PDE solver as the synthetic data but evaluated at the true set of parameters.

C.2. Convergence history of parameter inference with noisy time trajectories. Figures 11 and 14 are the inversion results where the Lorenz time trajectory is polluted by extrinsic and intrinsic noises, respectively. The properties of the time trajectories that are affected by the intrinsic and extrinsic noises are the same as the ones in Figure 2. As one can see from all the single-parameter and multiparameter inversions, it gets more challenging to achieve reconstruction with high accuracy than the previous noise-free cases. In particular, the overfitting phenomenon occurs, which can be directly observed for β in the single-parameter

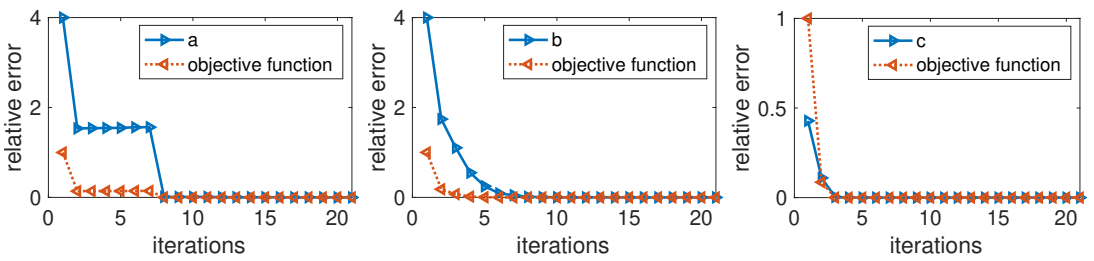


Figure 12. *Rössler system single-parameter inference starting with $a = 0.5$ (left), $b = 0.5$ (middle), $c = 10$ (right), respectively. The reference PDF is generated by the truth $(a, b, c) = (0.1, 0.1, 14)$ through the same numerical solver for the synthetic data.*

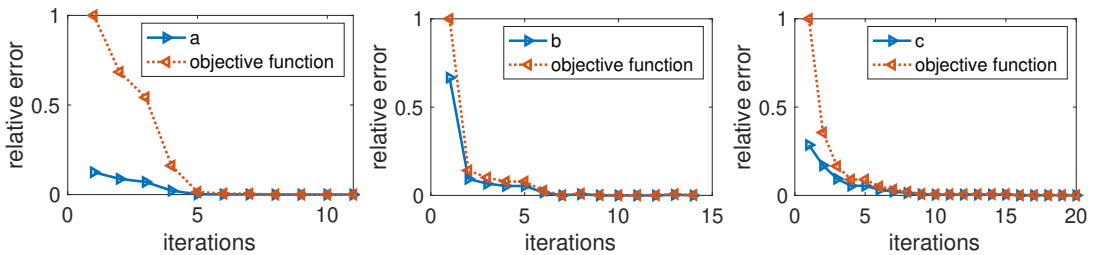
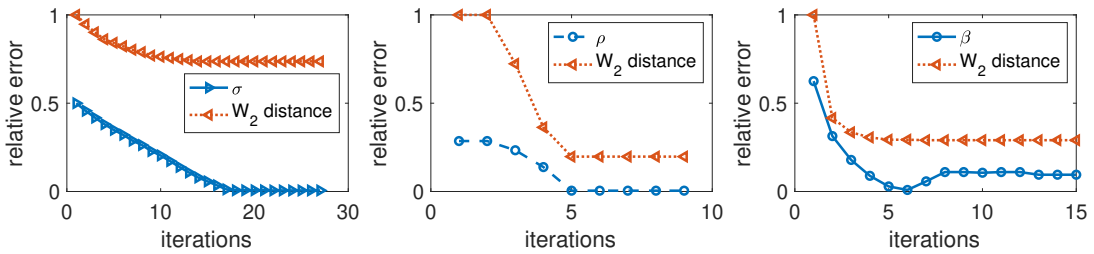
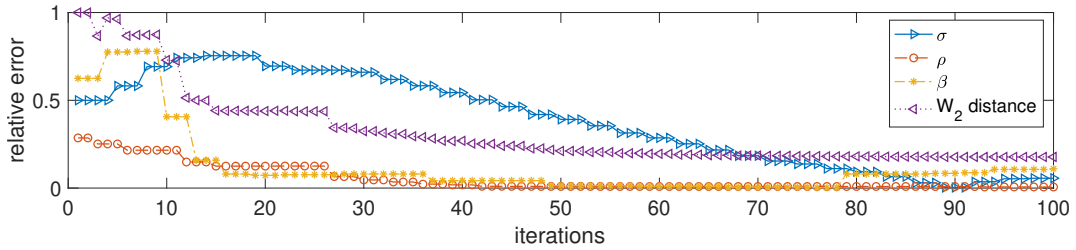


Figure 13. *Chen system single-parameter inference starting with $a = 45$ (left), $b = 5$ (middle), $c = 20$ (right), respectively. The reference PDF is generated by the truth $(a, b, c) = (40, 3, 28)$ through the same numerical solver for the synthetic data.*



(a) Single-parameter inversion



(b) Multi-parameter inversion

Figure 14. *Top row: Lorenz system single-parameter inference starting with* $\sigma = 5$ *(left),* $\rho = 20$ *(middle),* $\beta = 1$ *(right), respectively. Bottom row: Multiparameter inference using coordinate gradient descent with initial guess* $(\sigma, \rho, \beta) = (5, 20, 1)$*. The reference PDF is the histogram from the time trajectory with intrinsic noise.*

inversion (the top right plot in both figures) and the three-parameter joint inversion (the bottom plots). As the number of iterations increases, the reconstructed β first reaches the actual value but immediately deviates away as the objective function keeps being minimized to fit the noise.

Acknowledgments. We thank Prof. Adam Oberman for initiating our collaboration. We also thank Prof. Alex Townsend for his constructive suggestions. This work was in part completed during the long program on High Dimensional Hamilton-Jacobi PDEs held in the Institute for Pure and Applied Mathematics (IPAM) at UCLA, March 9-June 12, 2020. The authors thank the program's organizers, IPAM scientific committee, and staff for the hospitality and stimulating research environment. The authors are also grateful to the peer referees for their time, comments, and constructive suggestions during the review process.

REFERENCES

[1] L. A. Aguirre and C. Letellier, *Modeling nonlinear dynamics and chaos: A review*, Math. Probl. Eng., 2009 (2009) 238960.

[2] R. D. Anderson and V. L. Klee, Jr, *Convex functions and upper semi-continuous collections*, Duke Math. J., 19 (1952), pp. 349-357.

[3] A. Allawala and J. Marston, *Statistics of the stochastically forced Lorenz attractor by the Fokker-Planck equation and cumulant expansions*, Phys. Rev. E (3), 94 (2016), 052218.

[4] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Lect. Math. ETH Zürich, 2nd ed., Birkhäuser, Basel, 2008.

[5] S. J. Araki, J. W. Koo, R. S. Martin, and B. Dankongkakul, *A grid-based nonlinear approach to noise reduction and deconvolution for coupled systems*, Phys. D, 417 (2021), 132819.

[6] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, Proc. Mach. Learn. Res. (PMLR), 70 (2017), pp. 214–223.
[7] E. Baake, M. Baake, H. Bock, and K. Briggs, *Fitting ordinary differential equations to chaotic data*, Phys. Rev. A (3), 45 (1992), p. 5524.
[8] R. Bakker, J. C. Schouten, C. L. Giles, F. Takens, and C. M. Van Den Bleek, *Learning chaotic attractors by neural networks*, Neural Comput., 12 (2000), pp. 2355–2383.
[9] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert, *On parameter estimation with the Wasserstein distance*, Inf. Inference, 8 (2019), pp. 657–676.
[10] T. R. Bewley and A. S. Sharma, *Efficient grid-based Bayesian estimation of nonlinear low-dimensional systems with sparse non-Gaussian PDFs*, Automatica J. IFAC, 48 (2012), pp. 1286–1290.
[11] B. P. Bezruchko and D. A. Smirnov, *Extracting Knowledge From Time Series: An Introduction to Nonlinear Empirical Modeling*, Springer, Heidelberg, 2010.
[12] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer Ser. Oper. Res., Springer, New York, 2000.
[13] R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, Lecture Notes in Math. 470, Springer, Berlin, 1975.
[14] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, *JAX: Composable Transformations of Python+NumPy Programs*, 2018, http://github.com/google/jax.
[15] J. Brocker, U. Parlitz, and M. Ogorzalek, *Nonlinear noise reduction*, Proc IEEE, 90 (2002), pp. 898–918.
[16] S. L. Brunton, J. L. Proctor, and J. N. Kutz, *Discovering governing equations from data by sparse identification of nonlinear dynamical systems*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. 3932–3937.
[17] G. Chen and T. Ueta, *Yet another chaotic attractor*, Internat. J. Bifur. Chaos, 9 (1999), pp. 1465–1466.
[18] W. Cowieson and L.-S. Young, *SRB measures as zero-noise limits*, Ergodic Theory Dynam. Systems, 25 (2005), pp. 1115–1138, https://doi.org/10.1017/S0143385704000604.
[19] M. Dellnitz, G. Froyland, and O. Junge, *The algorithms behind GAIO—set oriented numerical methods for dynamical systems*, in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems, Springer, Berlin, 2001, pp. 145–174.
[20] M. Dellnitz and O. Junge, *Almost invariant sets in Chua's circuit*, Internat. J. Bifur. Chaos, 7 (1997), pp. 2475–2485.
[21] M. Dellnitz and O. Junge, *An adaptive subdivision technique for the approximation of attractors and invariant measures*, Comput. Vis. Sci., 1 (1998), pp. 63–68.
[22] M. Dellnitz and O. Junge, *On the approximation of complicated dynamical behavior*, SIAM J. Numer. Anal., 36 (1999), pp. 491–515.
[23] M. Dellnitz and O. Junge, *Set oriented numerical methods for dynamical systems*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., Elsevier, Amsterdam, 2002, pp. 221–264.
[24] M. M. Dunlop and Y. Yang, *Stability of Gibbs posteriors from the Wasserstein loss for Bayesian full waveform inversion*, SIAM/ASA J. Uncertain. Quantif., 9 (2021), pp. 1499–1526.
[25] S. Effah-Poku, W. Obeng-Denteh, and I. Dontwi, *A study of chaos in dynamical systems*, J. Math., 2018, (2018), 1808953.
[26] M. Eidenschink, *Exploring Global Dynamics: A Numerical Algorithm Based on the Conley Index Theory*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 1995.
[27] B. Engquist, K. Ren, and Y. Yang, *The quadratic Wasserstein metric for inverse data matching*, Inverse Problems, 36 (2020), 055001.
[28] B. Engquist and Y. Yang, *Optimal transport based seismic inversion: Beyond cycle skipping*, Comm. Pure Appl. Math., 2020.
[29] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.
[30] J. C. Feng, *Reconstruction of Chaotic Signals With Applications to Chaos-Based Communications*, World Scientific, Singapore, 2008.
[31] D. C. D. R. Fernández, P. D. Boom, and D. W. Zingg, *A generalized framework for nodal first derivative summation-by-parts operators*, J. Comput. Phys., 266 (2014), pp. 214–239.

[32] B. Fiedler, *Handbook of Dynamical Systems*, Elsevier, Amsterdam, 2002.
[33] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. H. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, *POT: Python optimal transport*, J. Mach. Learn. Res., 22 (2021), pp. 1–8.
[34] A. L. Fradkov and R. J. Evans, *Control of chaos: Methods and applications in engineering*, Annu. Rev. Control, 29 (2005), pp. 33–56.
[35] G. Froyland, *Extracting dynamical behavior via Markov models*, in Nonlinear Dynamics and Statistics, Birkhäuser Boston, Boston, 2001, pp. 281–321.
[36] A. Gábor and J. R. Banga, *Robust and efficient parameter estimation in dynamic models of biological systems*, BMC Syst. Biol., 9 (2015), pp. 1–25.
[37] W. Gilpin, *Chaos as an interpretable benchmark for forecasting and data-driven modelling*, in Thirty-Fifth Conference on Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2021, https://openreview.net/forum?id=enYjtbjYJrf.
[38] D. Givon, R. Kupferman, and A. Stuart, *Extracting macroscopic dynamics: Model problems and algorithms*, Nonlinearity, 17 (2004), pp. R55–R127.
[39] D. F. Gleich, *Pagerank beyond the web*, SIAM Rev., 57 (2015), pp. 321–363.
[40] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, MD, 2013.
[41] C. Greve, K. Hara, R. Martin, D. Eckhardt, and J. Koo, *A data-driven approach to model calibration for nonlinear dynamical systems*, J. Appl. Phys., 125 (2019), 244901.
[42] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent, *Optimal mass transport for registration and warping*, Int. J. Comput. Vision, 60 (2004), pp. 225–240.
[43] W. Huang, M. Ji, Z. Liu, and Y. Yi, *Concentration and limit behaviors of stationary measures*, Phys. D, 369 (2018), pp. 1–17.
[44] M. Jacobs and F. Léger, *A fast approach to optimal transport: The back-and-forth method*, Numer. Math., 146 (2020), pp. 513–544.
[45] L. Jaeger and H. Kantz, *Unbiased reconstruction of the dynamics underlying a noisy chaotic time series*, Chaos, 6 (1996), pp. 440–450.
[46] E. Kaiser, B. R. Noack, L. Cordier, A. Spohn, M. Segond, M. Abel, G. Daviller, J. Östh, S. Krajnović, and R. K. Niven, *Cluster-based reduced-order modelling of a mixing layer*, J. Fluid Mech., 754 (2014), pp. 365–414.
[47] Y. Kifer, *General random perturbations of hyperbolic and expanding transformations*, J. Anal. Math., 47 (1986), pp. 111–150.
[48] E. J. Kostelich, *Problems in estimating dynamics from data*, Phys. D, 58 (1992), pp. 138–152.
[49] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge Texts Appl. Math. 31, Cambridge University Press, Cambridge, 2002.
[50] J.-C. Loiseau, *Data-driven modeling of the chaotic thermal convection in an annular thermosyphon*, Theor. Comput. Fluid Dyn., 34 (2020), pp. 339–365.
[51] L. Lu, X. Meng, Z. Mao, and G. E. Karniadakis, *DeepXDE: A deep learning library for solving differential equations*, SIAM Rev., 63 (2021), pp. 208–228.
[52] K. McGoff, S. Mukherjee, and N. Pillai, *Statistical inference for dynamical systems: A review*, Statist. Surv., 9 (2015), pp. 209–252.
[53] A. Medio and M. Lines, *Nonlinear Dynamics: A Primer*, Cambridge University Press, Cambridge, 2001.
[54] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, Philadelphia, SIAM, 2000.
[55] C. Michalik, R. Hannemann, and W. Marquardt, *Incremental single shooting—a robust method for the estimation of parameters in dynamical systems*, Comput. Chem. Eng., 33 (2009), pp. 1298–1305.
[56] M. Nakagawa, *Chaos and Fractals in Engineering*, World Scientific, Singapore, 1999.
[57] E. Negrini, G. Citti, and L. Capogna, *A Neural Network Ensemble Approach to System Identification*, preprint, arXiv:2110.08382, 2021.
[58] E. Negrini, G. Citti, and L. Capogna, *System identification through Lipschitz regularized deep neural networks*, J. Comput. Phys., 444 (2021), 110549.

[59] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2006.
[60] G. Peyré and M. Cuturi, *Computational Optimal Transport: With Applications to Data Science*, Foundations and trends in machine learning, Now, Hanover, MA, 2019.
[61] T. Rippl, A. Munk, and A. Sturm, *Limit laws of the empirical Wasserstein distance: Gaussian distributions*, J. Multivariate Anal., 151 (2016), pp. 90–109.
[62] Y. Robin, P. Yiou, and P. Naveau, *Detecting changes in forced climate attractors with Wasserstein distance*, Nonlinear Process Geophys., 24 (2017), pp. 393–405.
[63] M. Rodriguez-Fernandez, J. A. Egea, and J. R. Banga, *Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems*, BMC Bioinform., 7 (2006), pp. 1–18.
[64] H. Ruan, T. Zhai, and E. E. Yaz, *A chaotic secure communication scheme with extended Kalman filter based parameter estimation*, in Proceedings of 2003 IEEE Conference on Control Applications, Vol. 1, IEEE, Piscataway, NJ, 2003, pp. 404–408
[65] S. H. Rudy, J. N. Kutz, and S. L. Brunton, *Deep learning of dynamics and signal-noise decomposition with time-stepping constraints*, J. Comput. Phys., 396 (2019), pp. 483–506.
[66] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, Progr. Nonlinear Differential Equations Appl. 87, Springer, Cham, Switzerland, 2015.
[67] S. Siegmund and P. Taraba, *Approximation of box dimension of attractors using the subdivision algorithm*, Dyn. Syst., 21 (2006), pp. 1–24.
[68] M. Sommerfeld and A. Munk, *Inference for empirical Wasserstein distances on finite spaces*, J. R. Stat. Soc. Ser. B Stat. Methodol., 80 (2018), pp. 219–238.
[69] W. Tucker, *The Lorenz attractor exists,* C. R. Acad. Sci. Ser I Math., 328 (1999), pp. 1197–1202.
[70] W. Tucker, *A rigorous ODE solver and Smale's 14th problem*, Found. Comput. Math., 2 (2002), pp. 53–117, https://doi.org/10.1007/s002080010018.
[71] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, American Mathematical Society, Providence, RI, 2003.
[72] H. Ye and G. Sugihara, *Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality*, Science, 353 (2016), pp. 922–925.
[73] L.-S. Young, *What are SRB measures, and which dynamical systems have them?*, J. Stat. Phys., 108 (2002), pp. 733–754.