Probabilistic Model Validation for Uncertain Nonlinear Systems ⋆

Abhishek Halder, Raktim Bhattacharya

Department of Aerospace Engineering, Texas A&M University, College Station, TX 77843-3141, United States

Abstract

This paper presents a probabilistic model validation methodology for nonlinear systems in time-domain. The proposed formulation is simple, intuitive, and accounts both deterministic and stochastic nonlinear systems with parametric and nonparametric uncertainties. Instead of hard invalidation methods available in the literature, a relaxed notion of validation in probability is introduced. To guarantee provably correct inference, algorithm for constructing probabilistically robust validation certificate is given along with computational complexities. Several examples are worked out to illustrate its use.

Key words: Model validation, uncertainty propagation, optimal transport, Wasserstein distance.

1 Introduction

A model serves as a mathematical abstraction of the physical system, providing a framework for system analysis and controller synthesis. Since such mathematical representations are based on assumptions specific to the process being modeled, it's important to quantify the reliability to which the model is consistent with the physical observations. Model quality assessment is imperative for applications where the model needs to be used for prediction (e.g. weather forecasting, stock market) or safety-critical control design (e.g. aerospace, nuclear, systems biology) purposes.

Here it is important to realize that a model can only be validated against experimental observations, not against another model. Thus a *model validation problem* can be stated as: *given a candidate model and experimentally observed measurements of the physical system, how well does the model replicate the experimental measurements?* It has been argued in the literature [3–6] that the term 'model validation' is a misnomer since it would take infinite number of experimental observations to do so. Hence the term 'model invalidation' or 'falsification' [7] is preferred. In this paper, instead of hard invalidation, we will consider the validation/invalidation problem in a probabilistically relaxed sense.

1.1 Related literature

Broadly speaking, there have been three distinct frameworks in which the model validation problem has been attempted till now. **One** is a discrete formulation in *temporal logic framework* [8] which has been extended to account probabilistic models [8,9]. **Second** is the $\mathcal{H}_\infty$ *control framework* where time-domain [5,10,11], frequency domain [4,12] and mixed domain [13] model validation methods have been studied extensively assuming structured norm-bounded uncertainty in linear dynamics setting. The **third** framework involves deductive inference based on barrier certificates [6] which was shown to encompass a large class of nonlinear models including differential-algebraic equations [14], dynamic uncertainties described by integral quadratic constraints [15], stochastic [16] and hybrid dynamics [17].

In statistical setting, model validation has been addressed from system identification perspective [18,19] where the main theme is to validate an identified nominal model through correlation analysis of the residuals. A polynomial chaos framework has also been proposed [20] for model validation. Gevers *et. al.* [21] have connected the robust control framework with prediction error based identification for frequency-domain validation of linear systems. In another vein, using Bayesian conditioning, Lee and Poolla [22] showed that for *parametric* uncertainty models, the statistical validation

⋆ The material in this paper was partially presented in IEEE Conference on Decision and Control (CDC) 2011 [1] and 2012 [2]. Corresponding author: A. Halder. Tel. +979-583-6070, Fax +979-845-6051.

Email addresses: ahalder@tamu.edu (Abhishek Halder), raktim@tamu.edu (Raktim Bhattacharya).

problem may be reduced to the computation of relative weighted volumes of convex sets. However, for *nonparametric* models: "the situation is significantly more complicated" [22] and to the best of our knowledge, has not been addressed in the literature. Recently, in the spirit of weak stochastic realization problem [23], Ugrinovskii [24] investigated the conditions for which the output of a stochastic nonlinear system can be realized through perturbation of a nominal stochastic *linear* system.

In practice, one often encounters the situation where a model is either proposed from physics-based reasoning or a reduced order model is derived for computational convenience. In either case, the model can be linear or nonlinear, continuous or discrete-time, and in general, it's not possible to make any a-priori assumption about the noise. Given the experimental data and such a candidate model for the physical process, our task is to answer: "to what extent, the proposed model is valid?" In addition to quantify such degree of validation, one must also be able to demonstrate that the answer is *provably correct* in the face of uncertainty. This brings forth the notion of *probabilistically robust model validation*. In this paper, we will show how to construct such a *robust validation certificate*, guaranteeing the performance of probabilistic model validation algorithm.

1.2 Contributions of this paper

With respect to the literature, the contributions of this paper are as follows.

(1) Instead of interval-valued structured uncertainty (as in $\mathcal{H}_\infty$ control framework) or moment based uncertainty (as in parametric statistics framework), this paper deals with model validation in the sense of nonparametric statistics. Uncertainties in the model are quantified in terms of the probability density functions (PDFs) of the associated random variables. We argue that such a formulation offers several advantages. *Firstly*, we show that model uncertainties in the parameters, initial states and input disturbance, can be propagated accurately by spatio-temporally evolving the joint state and output PDFs. Since experimental data usually come in the form of histograms, it's a more natural quantification of uncertainty than specifying sets [6] to which the trajectories are contained at each instant of time. However, if needed, such sets can be recovered from the supports of the instantaneous PDFs. *Secondly*, as we'll see in Section 5, instead of simply invalidating a model, our methodology allows to estimate the probability that a proposed model is valid or invalid. This can help to decide which specific aspects of the model need further refinement. Hard invalidation methods don't cater such constructive information. *Thirdly*, the framework can handle both discrete-time and continuous-time nonlinear models which need not be polynomial. Previous work like [6] dealt with semialgebraic nonlinearities and relied on sum of squares (SOS) decomposition [25] for computational tractability. From an implementation point of view, the approach presented in this paper doesn't suffer from such conservatism.

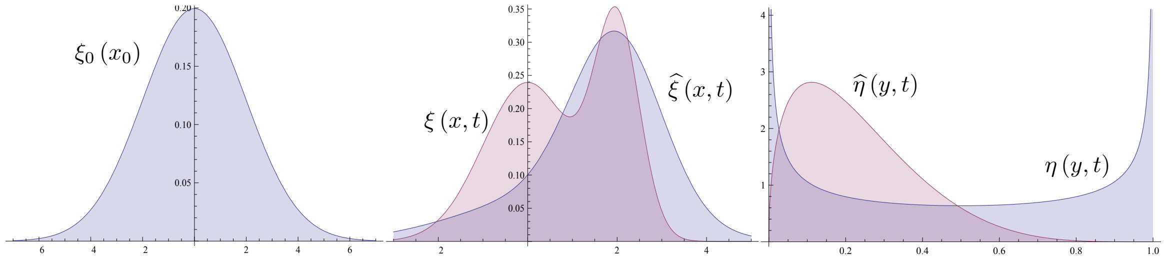
(2) Due to the uncertainties in initial conditions, parameters, and process noise, one needs to compare output ensembles instead of comparing individual output realizations. This requires a metric to quantify closeness between the experimental data and the model in the sense of distribution. We propose *Wasserstein distance* to compare the output PDFs and argue why commonly used information-theoretic notions like *Kullback-Leibler divergence* may not be appropriate for this purpose.

(3) We show that the uncertainty propagation through continuous or discrete-time dynamics can be done via numerically efficient meshless algorithms, even when the model is high-dimensional and strongly nonlinear. Moreover, we outline how to compute the Wasserstein distance in such settings. Further, bringing together ideas from analysis of randomized algorithms, we give sample-complexity bounds for robust validation inference.

The paper is organized as follows. In Section 2, we describe the problem setup. Then we expound on the three steps of our validation framework, viz. uncertainty propagation, distributional comparison and construction of validation certificates in Section 3, 4 and 5, respectively. We provide numerical examples in Section 6, to illustrate the ideas presented in this paper. The concept of worst-case initial uncertainty related to model discrimination, is addressed in Section 7. Section 8 presents some results for discrete-time linear Gaussian systems, followed by conclusions in Section 9.

Notation

We use the superscript $^\top$ to denote matrix transpose, $\otimes$ to denote Kronecker product, and the symbol $\wedge$ to denote minimum of two real numbers. The notation ${}_rF_s\left(a_1,\ldots,a_r;b_1,\ldots,b_s;x\right)$ stands for generalized hypergeometric function. The symbols $\mathcal{N}\left(.,.\right)$, $\mathcal{U}\left(.\right)$, and $\mathcal{A}\left(.\right)$ are used for normal, uniform and arcsine distributions, respectively. We use the notation $\xi_0\left(.\right)$ to denote the joint PDF over initial states and parameters. $\xi\left(.,t\right)$ and $\widehat{\xi}\left(.,t\right)$ denote joint PDFs over instantaneous states and parameters, for the true and model dynamics, respectively. Similarly, $\eta\left(.,t\right)$ and $\widehat{\eta}\left(.,t\right)$, respectively denote joint PDFs over output spaces y and $\widehat{y}$ at time t, for the true and model dynamics. The symbol $\widetilde{x}$ is used to denote the extended state vector obtained by augmenting the state (x) and parameter (p) vectors. We use χ to denote indicator function and $\#$ to denote cardinality. Unless stated otherwise, $\delta\left(.\right)$ stands for Dirac delta. The symbol I_ℓ denotes the ℓ-by-ℓ identity



(a) (b) (c)

Fig. 1. The proposed model validation framework compares experimentally observed output PDF $\eta(y,t)$ with the model-predicted output PDF $\widehat{\eta}(\widehat{y},t)$, the comparison being made with respect to some suitable metric at each instant of measurement availability. The state dynamics evolves the initial joint PDF $\xi_0(x_0)$ (Fig. 1(a)) to instantaneous joint state PDFs $\xi(x,t)$ and $\widehat{\xi}(\widehat{x},t)$ (Fig. 1(b)). The associated output PDFs $\eta(y,t)$ and $\widehat{\eta}(y,t)$ may share the same support ($[0,1]$ as shown in Fig. 1(c)), but have different shapes. Hence, instead of matching output supports, we propose matching output PDFs at all times, for validating a model.

matrix, ∇_x denotes gradient operator with respect to vector x, vec$(\cdot)$ stands for the vectorization operator, and $\|\cdot\|_F$ denotes the Frobenius norm. tr$(\cdot)$ and det$(\cdot)$ stand for trace and determinant of a matrix. The abbreviations *a.s.* and *i.p.* refer to convergence in *almost sure* and *in probability* sense. The shorthand ∂_α means partial derivative with respect to variable α, supp$(\cdot)$ denotes support of a function, and erf$(\cdot)$ stands for error function.

2 Problem Setup

2.1 Intuitive idea

The proposed framework is based on the evolution of densities in output space, instead of evolution of individual trajectories, as in the Lyapunov framework. Intuitively, characteristics of the input to output mapping is revealed by the growth or depletion of trajectory concentrations in the output-space. Growth in concentration, or increased density, defines regions in where the trajectories accumulate. This corresponds to regions with slow time scale dynamics or time invariance. Similarly, depletion of concentration in a set implies fast-scale dynamics or unstable manifold. We refer the readers to [26] for an introduction to analysis of dynamical systems using trajectory densities. This idea of comparing dynamical systems based on density functions, have been presented before by Sun and Mehta [27] in the context of filtering, and by Georgiou [28] in the context of matching power spectral densities.

2.1.1 Proposed approach

Given the experimental measurements of the physical system in the form of a time-varying distribution (such as histograms), we propose to compare the *shape* or *concentration profile* of this measured output density, with that predicted by the model. At every instant of time, if the model-predicted density matches with the experimental one "*reasonably well*" (to be made precise later in the paper), we conclude that the model is validated with high *confidence* (to be computed for guaranteeing quality of inference).

2.1.2 Why compare densities instead of trajectories

The rationale behind comparing the distributional shapes for model validation comes from the fact that the presence of uncertainties mask the difference between individual output realizations. Uncertainties in initial conditions, parameters and noise result different realizations of the trajectory or integral curve of the dynamical system. Regions of high (low) concentration of trajectories correspond to regions of high (low) probability. Thus a model validation procedure should naturally aim to compare concentrations of the trajectories between the measurements and model-predictions, instead of comparing individual realizations of them, which would be meaningful only in the absence of uncertainties.

2.1.3 Why compare densities instead of moments or sets

Density based model validation provides natural advantages over moment based or set containment methods for the following reasons. Moment based methods can be erroneous for nonlinear non-Gaussian systems, as two different trajectory densities may provide the same correlation information. This can be circumvented by including higher order moments, but it is not computationally tractable for high dimensional systems. Set containment arguments can also be erroneous as it is possible that at a given time, two systems have trajectory densities with identical supports but different concentrations (Fig. 1 (c)).

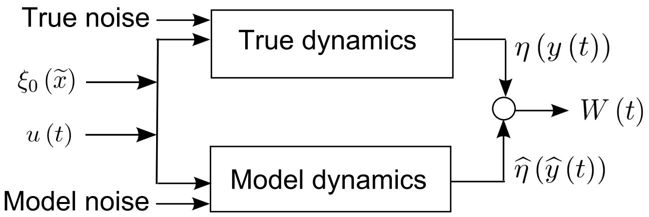

True noise
True dynamics
$\eta\left(y\left(t\right)\right)$
$\xi_0\left(\widetilde{x}\right)$
$W\left(t\right)$
$u\left(t\right)$
$\widehat{\eta}\left(\widehat{y}\left(t\right)\right)$
Model dynamics
Model noise

Fig. 2. Block diagram for the proposed model validation formulation.

A proposed model is validated, if the "distance" between its predicted density and the measured density, remains below a user-specified tolerance level, which need not be fixed over time. For example, take-off and landing are critical operational segments during the flight of a commercial aircraft, and it's unacceptable to have a controller that does not guarantee the robust performance for these critical time-segments with very high probability. This motivates the computation of probability of validation as part of the model validation oracle.

2.2 Methodology

In this section, we formalize the ideas presented above. Fig. 1 and 2 show the outline of the model validation framework proposed here. In this formulation, the systems under comparison are excited with a *known* input signal $u\left(t\right)$, and an initial PDF $\xi_0\left(\widetilde{x}_0\right)$, supported over the extended state space $\widetilde{x} := \left\{x,\ p\right\}^\top$, where the states $x \in \mathbb{R}^{n_s}$, and the parameters $p \in \mathbb{R}^{n_p}$. Given the PDF $\eta\left(y\left(t\right)\right)$ supported over the true output space $y \in \mathbb{R}^{n_o}$, and a candidate model, we compute and then compare the model predicted output PDF $\widehat{\eta}\left(\widehat{y}\left(t\right)\right)$, with $\eta\left(y\left(t\right)\right)$ at each instances of measurement availability $\left\{t_j\right\}_{j=1}^{\tau}$. Thus, one can think of three distinct steps of such a model validation framework. These are:

(1) evolving $\xi_0\left(\widetilde{x}_0\right)$ using the proposed model, to compute $\widehat{\eta}\left(\widehat{y}\left(t\right)\right)$,
(2) measuring an appropriate notion of distance, denoted as $W(t)$ in Fig. 2, between $\eta\left(y\left(t\right)\right)$ and $\widehat{\eta}\left(\widehat{y}\left(t\right)\right)$ at $\left\{t_j\right\}_{j=1}^{\tau}$,
(3) probabilistic quantification of provably correct inference in this framework and providing sample complexity bounds for the same.

Now we will elicit each of these steps.

3 Uncertainty Propagation

3.1 Continuous-time models

3.1.1 Uncertainty propagation for deterministic flow

Consider the continuous-time nonlinear model with state dynamics given by the ODE $\dot{\widehat{x}} = \widehat{f}\left(\widehat{x},\widehat{p}\right)$, where $\widehat{x}\left(t\right) \in \widehat{\mathcal{X}} \subseteq \mathbb{R}^{\widehat{n}_s}$ is the state vector, $\widehat{p} \in \widehat{\mathcal{P}} \subseteq \mathbb{R}^{\widehat{n}_p}$ is the parameter vector, the dynamics $\widehat{f}\left(.,\widehat{p}\right) : \widehat{\mathcal{X}} \mapsto \mathbb{R}^{\widehat{n}_s}$ $\forall\, \widehat{p} \in \widehat{\mathcal{P}}$, and is at least locally Lipschitz . It can be put in an extended state space form

$$\dot{\widehat{\widetilde{x}}} = \widehat{\widetilde{f}}\left(\widehat{\widetilde{x}}\right),\ \widehat{\widetilde{x}} \in \widehat{\mathcal{X}} \times \widehat{\mathcal{P}} \subseteq \mathbb{R}^{\widehat{n}_s + \widehat{n}_p},\ \widehat{\widetilde{f}} = \begin{Bmatrix} \widehat{f}_{\widehat{n}_s \times 1} \\ \mathbf{0}_{\widehat{n}_p \times 1} \end{Bmatrix}. \quad (1)$$

The output equation can be written as

$$\widehat{y} = \widehat{h}\left(\widehat{\widetilde{x}}\right), \quad \widehat{h} : \widehat{\mathcal{X}} \times \widehat{\mathcal{P}} \mapsto \widehat{\mathcal{Y}}, \quad (2)$$

where $\widehat{y}\left(t\right) \in \widehat{\mathcal{Y}} \subseteq \mathbb{R}^{n_o}$ is the output vector. If uncertainties in the initial conditions $\left(x_0 := x\left(0\right)\right)$ and parameters $\left(\widehat{p}\right)$ are specified by the initial joint PDF $\xi_0\left(\widetilde{x}\right)$, then the evolution of uncertainties subject to the dynamics (1), can be described by evolving the joint PDF $\widehat{\xi}\left(\widehat{\widetilde{x}}, t\right)$ over the extended state space. Such spatio-temporal evolution of $\widehat{\xi}\left(\widehat{\widetilde{x}}, t\right)$ is governed by the *stochastic Liouville equation* (SLE) given by

$$\frac{\partial \widehat{\xi}}{\partial t} = \mathscr{L}_{\text{SLE}}\widehat{\xi} = D_1 \widehat{\xi} = -\nabla.\left(\widehat{\xi}\widehat{f}\right) = -\sum_{i=1}^{\widehat{n}_s} \frac{\partial}{\partial \widehat{x}_i}\left(\widehat{\xi}\widehat{f}_i\right), (3)$$

which is a quasi-linear partial differential equation (PDE), first order in both space and time. Notice that, the spatial operator $\mathscr{L}_{\text{SLE}}$ is a drift operator D_1 that describes the *advection* of the PDF in extended state space. The output PDF $\widehat{\eta}\left(\widehat{y}, t\right)$ can be computed from the state PDF as

$$\widehat{\eta}\left(\widehat{y}, t\right) = \sum_{j=1}^{\nu} \frac{\widehat{\xi}\left(\widehat{\widetilde{x}}_j^{\star}\right)}{\left|\det\left(\mathcal{J}\left(\widehat{\widetilde{x}}_j^{\star}\right)\right)\right|}, \quad (4)$$

where $\widehat{\widetilde{x}}_j^{\star}$ is the j^{th} root of the inverse transformation of (2) with $j = 1, 2, \ldots, \nu$, and $\mathcal{J}$ is the Jacobian of this inverse transformation.

3.1.2 Uncertainty propagation for stochastic flow

Consider the continuous-time nonlinear model with state dynamics given by the Itô SDE

$$d\widehat{\widetilde{x}} = \widehat{\widetilde{f}}\left(\widehat{\widetilde{x}}\right)\ dt + \widehat{g}\left(\widehat{\widetilde{x}}\right)\ d\beta, \quad (5)$$

where $\beta\left(t\right) \in \mathbb{R}^{\omega}$ is the ω-dimensional Wiener process at time t, and the noise coupling $\widehat{g} : \widehat{\mathcal{X}} \times \widehat{\mathcal{P}} \mapsto \mathbb{R}^{\left(\widehat{n}_s + \widehat{n}_p\right) \times \omega}$. For the Wiener process $\beta\left(t\right)$, at all times

$$\mathbb{E}\left[d\beta_i\right] = 0,\ \mathbb{E}\left[d\beta_i d\beta_j\right] = Q_{ij} = \alpha_i\, \delta_{ij}\ \forall\, i, j = 1, \ldots, \omega (6)$$

where $\mathbb{E}\left[.\right]$ stands for the expectation operator and δ_{ij} is the Kronecker delta. Thus $Q \in \mathbb{R}^{\omega \times \omega}$ with $\alpha_i > 0 \; \forall \; i = 1, 2, \ldots, \omega$, being the noise strength. The output map is still assumed to be given by (2). In such a setting, the evolution of the state PDF $\widehat{\xi}\left(\widehat{\widetilde{x}}, t\right)$ subject to (5), is governed by the *Fokker-Planck equation* (FPE), also known as *forward Kolmogorov equation*

$$\frac{\partial \widehat{\xi}}{\partial t} = \mathscr{L}_{\text{FPE}} \widehat{\xi} = \left(D_1 \; + D_2\right) \widehat{\xi}$$
$$= -\sum_{i=1}^{\widehat{n}_s} \frac{\partial}{\partial \widehat{x}_i} \left(\widehat{\xi} \widehat{f}_i\right) + \sum_{i=1}^{\widehat{n}_s} \sum_{j=1}^{\widehat{n}_s} \frac{\partial^2}{\partial \widehat{x}_i \partial \widehat{x}_j} \left(\left(\widehat{g} Q \widehat{g}^{\top}\right)_{ij} \widehat{\xi}\right), \quad (7)$$

which is a homogeneous parabolic PDE, second order in space and first order in time. In this case, the spatial operator $\mathscr{L}_{\text{FPE}}$ can be written as a sum of a *drift operator* (D_1) and a *diffusion operator* (D_2). The diffusion term accounts for the smearing of the PDF due to process noise. Once the state PDF is computed through (7), the output PDF can again be obtained from (4).

3.2 Discrete-time models

3.2.1 Uncertainty propagation for deterministic maps

Let $\widehat{\mathcal{X}} \times \widehat{\mathcal{P}} \subseteq \mathbb{R}^{\widehat{n}_s + \widehat{n}_p}$ be a compact set and let $\mathcal{B}\left(\widehat{\mathcal{X}} \times \widehat{\mathcal{P}}\right)$ be the Borel-σ algebra defined on it. Consider the discrete-time nonlinear system with state dynamics given by the vector recurrence relation

$$\widehat{\widetilde{x}}_{k+1} = \widehat{\mathcal{T}}\left(\widehat{\widetilde{x}}_k\right), \; \widehat{\mathcal{T}} : \widehat{\mathcal{X}} \times \widehat{\mathcal{P}} \mapsto \widehat{\mathcal{X}} \times \widehat{\mathcal{P}}, \quad (8)$$

where $\widehat{\mathcal{T}}$ is a measurable nonsingular transformation and the time index k takes values from the ordered index set of non-negative integers $\{0, 1, 2, \ldots\}$. Then the evolution of the joint PDF $\widehat{\xi}\left(\widehat{\widetilde{x}}_k\right)$ is dictated by the *Perron-Frobenius operator* $\widehat{\mathscr{P}}$, given by

$$\int_B \widehat{\mathscr{P}} \widehat{\xi}\left(\widehat{\widetilde{x}}_k\right) \; \mu\left(d\widehat{\widetilde{x}}_k\right) = \int_{\widehat{\mathcal{T}}^{-1}(B)} \widehat{\xi}\left(\widehat{\widetilde{x}}_k\right) \; \mu\left(d\widehat{\widetilde{x}}_k\right) \quad (9)$$

for $B \in \mathcal{B}$. Properties of Perron-Frobenius operator can be found in Chap. 3 of [26]. Further, assuming the output dynamics as $\widehat{y}_k = \widehat{h}\left(\widehat{\widetilde{x}}_k\right)$, one can derive $\widehat{\eta}\left(\widehat{y}_k\right)$ from $\widehat{\xi}\left(\widehat{\widetilde{x}}_k\right)$ using the discrete analogue of (4).

3.2.2 Uncertainty propagation for stochastic maps

In this case, we consider the nonlinear state space representation given by the stochastic maps of general form

$$\widehat{\widetilde{x}}_{k+1} = \widehat{\mathcal{T}}\left(\widehat{\widetilde{x}}_k, \zeta_k\right), \qquad \widehat{\widetilde{y}}_k = \widehat{h}\left(\widehat{\widetilde{x}}_k, \zeta_k\right), \quad (10)$$

where $\zeta_k \in \mathbb{R}^{\omega}$ is the i.i.d. sample drawn from a known distribution for the noise (stochastic perturbations). Here, the dynamics $\widehat{\mathcal{T}}$ is not required to be a nonsingular transformation (Chap. 10, [26]). Since $\widehat{\mathcal{T}}$ defines a Markov Chain on $\widehat{\mathcal{X}} \times \widehat{\mathcal{P}}$, it can be shown that [26,29] evolution of the joint PDFs follow

$$\widehat{\xi}_{k+1} := \widehat{\xi}_{\widehat{\widetilde{x}}_{k+1}}\left(\widehat{\widetilde{x}}\right) = \int_{\widehat{\mathcal{X}} \times \widehat{\mathcal{P}}} \mathcal{K}_{\widehat{\mathcal{T}}}\left(\widehat{\widetilde{x}} | z\right) \xi_{\widehat{\widetilde{x}}_k}\left(z\right) dz,$$
$$\widehat{\eta}_k := \widehat{\eta}_{\widehat{y}_k}\left(\widehat{y}\right) = \int_{\widehat{\mathcal{X}} \times \widehat{\mathcal{P}}} \mathcal{K}_{\widehat{h}}\left(\widehat{y} | z\right) \widehat{\xi}_{\widehat{\widetilde{x}}_k}\left(z\right) dz, \quad (11)$$

where $\mathcal{K}_{\mathcal{T}}\left(\widehat{\widetilde{x}} | z\right)$ and $\mathcal{K}_h\left(\widehat{y} | z\right)$ are the *stochastic kernels* for maps $\widehat{\mathcal{T}}$ and $\widehat{h}$ respectively. (11) can be seen as a special case of the Chapman-Kolmogorov equation [30].

3.3 Computational aspects

For deterministic flow, the Liouville PDE (3) can be solved in exact arithmetic [31] via method-of-characteristics (MOC). Since the characteristic curves for (3) are the trajectories in the extended state space, $\widehat{\xi}\left(\widehat{\widetilde{x}}, t\right)$ and hence $\widehat{\eta}\left(\widehat{y}, t\right)$ can be computed directly along these characteristics. Unlike Monte-Carlo, this is an "on-the-fly" computation and does not involve any approximation, and hence offers a superior performance [31,32] than Monte-Carlo in high dimensions. For deterministic maps, cell-to-cell mapping [33] achieves a finite dimensional approximation of the Perron-Frobenius operator.

For stochastic flow, solving Fokker-Planck PDE (7) is numerically challenging [34] but has seen some recent success [35] in moderate (4 to 5) dimensions. For high dimensional stochastic flows, an extension of the MOC approach has been proposed [36]. For stochastic maps, discretizations for stochastic kernels (11) and (12), can be done through cell-to-cell mapping [33] resulting a random transition probability matrix [37].

4 Distributional Comparison

Once the observed and model-predicted output PDFs $\eta\left(y, t\right)$ and $\widehat{\eta}\left(\widehat{y}, t\right)$, are obtained, we need a metric to compare the *shapes* of these two PDFs at times $\{t_j\}_{j=1}^{\tau}$, when the measurement PDF $\eta\left(y, t_j\right)$ is available. We argue that the suitable metric for this purpose is *Wasserstein distance.*

4.1 Choice of metric

Distances on the space of probability distributions [38], can be broadly categorized into two classes, viz. Csisźar's ϕ-divergence [39] and integral probability metrics [40].

The first includes well-known distances like Kullback-Leibler (KL) divergence, Hellinger distance, χ^2 divergence etc. while the latter includes Wasserstein distance, Dudley metric, maximum mean discrepancy. Total variation distance belongs to both of these classes.

The choice of a suitable metric depends on application. Following the intuitions of Section 2.1, we list the *axiomatic requirements*, that a model validation metric must satisfy:

R.1 The notion of "distance" must measure the *shape difference* between two instantaneous output PDFs. This is because a good model must emulate similar concentration of trajectories as observed in the measurement space, i.e. the respective joint PDFs $\eta(y,t)$ and $\widehat{\eta}(\widehat{y},t)$, over the time-varying output supports, must match at times whenever measurements are available. In particular, the distance must be function of *shape difference but not of shape*, i.e. same amount of shape difference must return same magnitude of distance, irrespective of the individual shapes being compared.

R.2 For meaningful validation inference, the choice of distance must be a metric.

R.3 For a given model-data pair, the supports of $\eta(y,t)$ and $\widehat{\eta}(\widehat{y},t)$ may not match at $t = t_j$, $j = 1, \ldots, \tau$. The distance must be well defined and computable under such circumstances.

R.4 The computation of the distance need not require $\eta(y,t)$ and $\widehat{\eta}(\widehat{y},t)$ to be represented by the same number of samples. For the purpose of model validation, this offers practical advantages since experimental data are often expensive to gather. However, model based simulation can harness the computational resources and hence, simulation sample size is often larger than that of experimental data.

R.5 The distance must be asymptotically consistent with respect to finite sample representations of the PDFs under comparison. Namely, in the infinite sample limit, the empirical estimate of the distance must converge to the actual instantaneous value of the distance. For practical computation, this rate-of-convergence is required to be fast with respect to the number of samples.

Next, we introduce the Wasserstein distance on the manifold of PDFs, which will be shown to fulfil the axiomatic requirements listed above.

Definition 1 ***(Wasserstein distance)*** *Let the ℓ_p norm between two random output vectors $y \in \mathcal{Y} \subseteq \mathbb{R}^{n_o}$, and $\widehat{y} \in \widehat{\mathcal{Y}} \subseteq \mathbb{R}^{n_o}$, be denoted as $\parallel y - \widehat{y} \parallel_p$. Then, the Wasserstein distance of order q, between two PDFs $\eta(y)$ and $\widehat{\eta}(\widehat{y})$, is defined as*

$$ {}_pW_q(\eta,\widehat{\eta}) := \left[\inf_{\rho\in\mathcal{M}_2(\eta,\widehat{\eta})} \int_{\mathcal{Y}\times\widehat{\mathcal{Y}}} \|y-\widehat{y}\|_p^q \; \rho(y,\widehat{y}) \; dy d\widehat{y}\right]^{\frac{1}{q}} \quad (12) $$

where $\mathcal{M}_2(\eta,\widehat{\eta})$ is the set of all joint PDFs supported on $\mathcal{Y}\times\widehat{\mathcal{Y}}$, having finite second moments, with first marginal as η and second marginal as $\widehat{\eta}$.

Remark 1 ***(Generalizations)*** *In general, the sets $\mathcal{Y}$ and $\widehat{\mathcal{Y}}$ can be subsets of any complete, separable metric (Polish) space, equipped with a p^{th} order distance metric. Further, (12) does not require the distributions under comparison to be absolutely continuous. It remains well defined between output measures μ and $\widehat{\mu}$, even when the corresponding PDFs η and $\widehat{\eta}$ don't exist.*

Remark 2 ***(Choice of** $p = q = 2$**)*** *We take Euclidean metric ($p = 2$) as the inter-sample distance between random vectors y and $\widehat{y}$. Further, we set $q = 2$ since it guarantees uniqueness [41] in (12), and has the interpretation of minimum effort needed to morph a density shape to other. Also, Jordan, Kinderlehrer and Otto [42] have rigorously demonstrated that uncertainty propagation in a dynamical system can be seen as a gradient flux of free energy with respect to the Wasserstein distance of order $q = 2$.*

The interpretation of ${}_2W_2$ as mass preserving optimal transport between two given shapes, makes it a strong candidate for model validation purpose. Further, it is known [43] that on the set $\mathcal{M}_2$, ${}_2W_2$ defines a metric. Thus, Wasserstein distance meets **R.1** and **R.2**. Also, **R.3** and **R.4** are satisfied since Definition 1 does not require the supports or cardinality of the sample representations of the PDFs to be the same. This will be illustrated further in Section 4.2, when we describe the computation of ${}_2W_2$ between two scattered point clouds with probability weights. For **R.5**, convergence of sample Wasserstein estimate to its true deterministic value, will be discussed in Section 4.3.1 (Theorem 2).

4.1.1 Limitations of pointwise distances

Commonly used information-theoretic distances like Kullback-Leibler divergence $D_{KL}(\eta \parallel \widehat{\eta}) \triangleq \mathbb{E}[\log(\eta/\widehat{\eta})]$, its symmetrized version $D_{KL}^{\text{symm}} \triangleq D_{KL}(\eta \parallel \widehat{\eta}) + D_{KL}(\widehat{\eta} \parallel \eta)$, are not metrics. On the other hand, Hellinger distance $H(\eta,\widehat{\eta}) \triangleq \frac{1}{\sqrt{2}} \parallel \sqrt{\eta} - \sqrt{\widehat{\eta}} \parallel_{L_2(\mathbb{R}^{n_o})}$, and the square-root of Jensen-Shannon divergence $JSD(\eta,\widehat{\eta}) \triangleq \frac{1}{2}\left[D_{KL}\left(\eta \parallel \frac{1}{2}(\eta+\widehat{\eta})\right) + D_{KL}\left(\widehat{\eta} \parallel \frac{1}{2}(\eta+\widehat{\eta})\right)\right]$ are metrics. However, being pointwise definitions, all of them fail to satisfy **R.3** and **R.4**, resulting computational difficulties for model validation. As for **R.5**, $D_{KL}(\eta \parallel \widehat{\eta})$ is known to be asymptotically consistent, but the rate-of-convergence can be arbitrarily slow

[44,45]. Besides these computational problems, we emphasize here that the information theoretic distances may not discriminate shapes in a geometric sense, as desired in **R.1**. We provide two counterexamples below to illustrate this point. The first counterexample highlights that two PDFs with same randomness need not have similar shapes. The second counterexample demonstrates that D_{KL} may depend on the shapes under comparison.

Counterexample 1 *(**Randomness $\neq$ shape**) Consider the two parametric family of beta densities $\eta_b\left(x;\alpha,\beta\right) \triangleq \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$, $\alpha,\beta > 0$, $x \in [0,1]$, where $B\left(\alpha,\beta\right) \triangleq \int_0^1 t^{\alpha-1}\left(1-t\right)^{\beta-1}\ dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, is the complete beta function, and $\Gamma\left(z\right)$ denotes the gamma function. The differential entropy for beta family can be computed as [46]*

$$\begin{aligned} H_b\left(\alpha,\beta\right) &= -\int_0^1 \eta_b\left(x;\alpha,\beta\right)\log\eta_b\left(x;\alpha,\beta\right)\ dx \\ &= \log B\left(\alpha,\beta\right) - \left(\alpha-1\right)\left(\Psi\left(\alpha\right)-\Psi\left(\alpha+\beta\right)\right) \\ &\quad -\left(\beta-1\right)\left(\Psi\left(\beta\right)-\Psi\left(\alpha+\beta\right)\right), \end{aligned} \tag{13}$$

*where $\Psi\left(z\right) \triangleq \frac{d}{dz}\log\Gamma\left(z\right)$, is the digamma function. Since (13) remains invariant under $\left(\alpha,\beta\right) \mapsto \left(\beta,\alpha\right)$, $\alpha \neq \beta$, $\eta_b\left(x;\alpha,\beta\right)$ and $\eta_b\left(x;\beta,\alpha\right)$ have same entropy, but one is skewed to right and the other to left, as shown in Fig. 3. Fig. 4 shows the isentropic contours of beta PDFs in $\left(\alpha,\beta\right)$ space. Any pair of **distinct** points chosen on these contours, results two beta PDFs with non-identical shapes, as revealed by Fig. 5 and Appendix A.*

Counterexample 2 *($D_{KL} \neq$ **shape difference**) Consider two ν-dimensional homoscedastic Gaussian PDFs $\mathcal{N}\left(m_1,\Sigma_1\right)$ and $\mathcal{N}\left(m_2,\Sigma_2\right)$, such that $\Sigma_1 = \Sigma_2$. Since the only **difference** between the two PDFs is the location of their means, a shape-discriminating distance is expected to be a function of $\parallel m_1 - m_2 \parallel_2$, and should not depend on the covariance matrix i.e. **shapes of the individual PDFs**.*

In this situation, ${}_2W_2 = \|m_1 - m_2\|_2$ [47] and $D_{KL} = \dfrac{1}{2}\left(m_2-m_1\right)^{\top}\Sigma_2^{-1}\left(m_2-m_1\right)$ [48]. If we introduce $m := m_2 - m_1$, then $\dfrac{D_{KL}}{{}_2W_2} = \dfrac{\|m\|_2}{2}r$, where $r := \dfrac{m^{\top}\Sigma_2^{-1}m}{m^{\top}m}$ is the Rayleigh quotient corresponding to the positive semi-definite precision matrix Σ_2^{-1}. It's known (Chap. 7, [49]) that if we denote $\mathcal{K} := \{\lambda : \lambda = \displaystyle\sum_{i=1}^{\nu}\alpha_i\lambda_i, \sum_{i=1}^{\nu}\alpha_i = 1, \alpha_i \geqslant 0, \forall i = 1,2,\ldots,\nu\}$ as the convex hull of the eigenvalues of the precision matrix Σ_2^{-1}, then $r\left(m\right) \in \mathcal{K}$. In particular,

$$r_{min} = \lambda_{min}\left(\Sigma_2^{-1}\right) = \frac{1}{\lambda_{min}\left(\Sigma_2\right)} > 0,$$
$$r_{max} = \lambda_{max}\left(\Sigma_2^{-1}\right) = \frac{1}{\lambda_{max}\left(\Sigma_2\right)} > 0,$$

and these extrema are attained when $m := m_2 - m_1$ respectively coincides with the minimum and maximum eigenvector of Σ_2^{-1}. Thus the spectrum of Σ_2^{-1} governs the magnitude of the ratio $\dfrac{D_{KL}}{{}_2W_2}$, even when $\|m\|_2$ is kept fixed. In particular, the ratio assumes unity iff $r = \dfrac{2}{\|m\|_2} \Rightarrow \Sigma_2^{-1} = \dfrac{2}{\|m\|_2}I_{\nu} \Rightarrow \Sigma_1 = \Sigma_2 = \dfrac{\|m\|_2}{2}I_{\nu}$.

Further discussions on the inadequacy of D_{KL} for capturing shape characteristics and the utility of Wasserstein distance for the same, can be found in [50,51].

4.1.2 Wasserstein gap between dynamical systems

Proposition 1 *(**Single output systems**)[52] At time $t > 0$, let $F\left(y,t\right)$ and $\widehat{F}\left(\widehat{y},t\right)$ be the cumulative distribution functions (CDFs) corresponding to the univariate PDFs $\eta\left(y,t\right)$ and $\widehat{\eta}\left(\widehat{y},t\right)$, respectively. Then*

$${}_2W_2\left(t\right) = \sqrt{\int_0^1\left(F^{-1}\left(\varsigma,t\right)-\widehat{F}^{-1}\left(\varsigma,t\right)\right)^2 d\varsigma}, \tag{14}$$

$$\rho^{\star}\left(y,\widehat{y},t\right) = \min\left(F\left(y,t\right),\widehat{F}\left(\widehat{y},t\right)\right), \tag{15}$$

where $\rho^{\star}$ is the optimizer in (12).

Proposition 2 *(**Linear Gaussian systems**) Consider stable, observable LTI system pairs in continuous and discrete time:*

$$dx_i(t) = A_i x_i(t)dt + B_i d\beta_i(t), \quad y_i(t) = C_i x_i(t), \tag{16}$$
$$x_i(k+1) = A_i x_i(k) + B_i\vartheta_i(k), \quad y_i(k) = C_i x_i(k), \tag{17}$$

where $i = 1,2$. $\beta_i(t)$ are Wiener processes with auto-covariances $Q_i\left(t_1 \wedge t_2\right)$, $t_1,t_2 > 0$, and $\vartheta_i\left(k\right)$ are Gaussian white noises with covariances $Q_i\left(k\right)$. If the initial PDF $\xi_0 = \mathcal{N}\left(\mu_0,\Sigma_0\right)$, then the Wasserstein distance between output PDFs $\eta_i = \mathcal{N}\left(\mu_{y_i},\Sigma_{y_i}\right)$, is given by [47]

$${}_2W_2 = \sqrt{\parallel \mu_{y_1} - \mu_{y_2} \parallel_2^2 + \text{tr}\left(\Sigma_{y_1} + \Sigma_{y_2} - 2\left[\sqrt{\Sigma_{y_1}}\Sigma_{y_2}\sqrt{\Sigma_{y_1}}\right]^{\frac{1}{2}}\right)} \tag{18}$$

where $\mu_{y_i} = C_i\mu_{x_i}$, $\Sigma_{y_i} = C_i\Sigma_{x_i}C_i^{\top}$. For the continuous-time case,

$$\dot{\mu}_{x_i}(t) = A_i\mu_{x_i}(t), \tag{19}$$
$$\dot{\Sigma}_{x_i}(t) = A_i\Sigma_{x_i}(t) + \Sigma_{x_i}(t)A_i^{\top} + B_iQ_iB_i^{\top}, \tag{20}$$

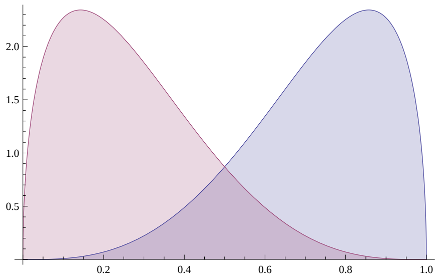


Fig. 3. The two beta densities $\eta_b\left(x;\alpha,\beta\right)$ (left-skewed) and $\eta_b\left(x;\beta,\alpha\right)$ (right-skewed) with $\alpha=4$, $\beta=\frac{3}{2}$, have same entropy/randomness, but have different shapes.

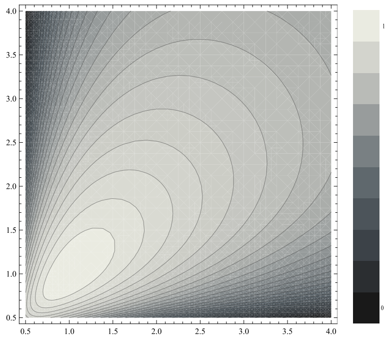


Fig. 4. Isentropic contours of beta family in (α,β) space. The symmetry of the contours about $\alpha=\beta$ line implies $H_b(\alpha,\beta)=H_b(\beta,\alpha)$. This plot also shows that uniform distribution ($\alpha=\beta=1$) is of maximum entropy.

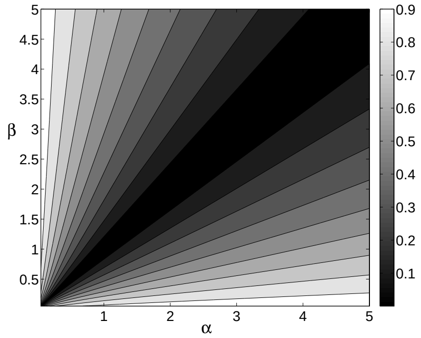


Fig. 5. Iso-Wasserstein contours of ${}_2W_2\left(\eta_b\left(x;\alpha,\beta\right),\eta_b\left(x;\beta,\alpha\right)\right)$ in (α,β) space. Since ${}_2W_2$ is a metric, it has symmetry about $\alpha=\beta$ line, and vanishes only along this line. The computation of ${}_2W_2$ is detailed in Appendix A.

and for the discrete-time case,

$$\mu_{x_i}\left(k+1\right)=A_i\mu_{x_i}\left(k\right), \tag{21}$$
$$\Sigma_{x_i}\left(k+1\right)=A_i\Sigma_{x_i}\left(k\right)A_i^{\top}+B_iQ_iB_i^{\top}, \tag{22}$$

to be solved with $\mu_{x_i}(0)=\mu_0$, and $\Sigma_{x_i}(0)=\Sigma_0$. Deterministic results are recovered from above by setting the diffusion matrix $B_i=0$.

Remark 3 ***(Asymptotic Wasserstein distance)*** *In Table 1, we have listed asymptotic Wasserstein distances between different pairs of stable dynamical systems. The asymptotic ${}_2W_2$ between two deterministic linear systems (**first row**) is zero since the origin being unique equilibria for both systems, Dirac delta is the stationary density for both. For a pair of deterministic affine systems (**second row**), asymptotic ${}_2W_2$ is simply the ℓ_2 norm between their respective fixed points. This holds true even for a pair of nonlinear systems, each having a **unique** globally asymptotically stable equilibrium. For the stochastic linear case (**third row**), $\Sigma_{y\infty}=C\Sigma_{x\infty}C^{\top}$, and $\widehat{\Sigma}_{\widehat{y}\infty}=\widehat{C}\widehat{\Sigma}_{\widehat{x}\infty}\widehat{C}^{\top}$; where $\Sigma_{x\infty},\widehat{\Sigma}_{\widehat{x}\infty}$ respectively solve $A\Sigma_{x\infty}+\Sigma_{x\infty}A^{\top}+BQB^{\top}=0$, and $\widehat{A}\widehat{\Sigma}_{\widehat{x}\infty}+\widehat{\Sigma}_{\widehat{x}\infty}\widehat{A}^{\top}+\widehat{B}\widehat{Q}\widehat{B}^{\top}=0$. Q and $\widehat{Q}$ are process noise covariances associated with Wiener processes $\beta(t)$ and $\widehat{\beta}(t)$. For the **fourth** and **fifth row**, the set of stable equilibria for the true and model nonlinear system, are given by $\{y_i^{\star}\}_{i=1}^{n^{\star}}$ and $\{\widehat{y}_i^{\star}\}_{i=1}^{\widehat{n}^{\star}}$, respectively. Further, we assume that the nonlinear systems have no invariant sets other than these stable equilibria. In such cases, the stationary densities are convex sum of Dirac delta densities, located at these equilibria. The weights for this convex sum, denoted as $m_i^{\star}$ and $\widehat{m}_i^{\star}$, depend on the initial PDF ξ_0. In particular, if we denote $\mathcal{R}_i$ as the **region-of-attraction** of the i^{th} equilibrium, then (see Appendix B)*

$$m_i^{\star}=\int_{\text{supp}(\xi_0)\cap\mathcal{R}_i}\xi_0\left(x_0\right)\,dx_0\;\in\;[0,1]\,. \tag{23}$$

*To further illustrate this idea, a numerical example corresponding to the **fourth row** in Table 1, will be provided in Section 6.*

4.2 Computing multivariate ${}_2W_2$

Computing Wasserstein distance from (12) calls for solving *Monge-Kantorovich optimal transportation plan* [53]. In this formulation, the difference in shape between two statistical distributions is quantified by the minimum amount of work required to convert a shape to the other. The ensuing optimization, often known as *Hitchcock-Koopmans problem* [54–56], can be cast as a linear program (LP), as described next.

Consider a complete, weighted, directed bipartite graph $K_{m,n}\left(U\cup V,E\right)$ with $\#\left(U\right)=m$ and $\#\left(V\right)=n$. If $u_i\in U, i=1,\ldots,m$, and $v_j\in V, j=1,\ldots,n$, then the edge weight $c_{ij}:=\parallel u_i-v_j\parallel_{\ell_2}^2$ denotes the cost of transporting unit mass from vertex u_i to v_j. Then, according to (12), computing ${}_2W_2^2$ translates to

$$\text{minimize}\;\sum_{i=1}^{m}\sum_{j=1}^{n}c_{ij}\,\varphi_{ij} \tag{24}$$

subject to the constraints

$$\sum_{j=1}^{n}\varphi_{ij}=\alpha_i,\qquad\forall\;u_i\in U, \tag{C1}$$

$$\sum_{i=1}^{m}\varphi_{ij}=\beta_j,\qquad\forall\;v_j\in V, \tag{C2}$$

$$\varphi_{ij}\geqslant 0,\qquad\forall\;\left(u_i,v_j\right)\in U\times V. \tag{C3}$$

The objective of the LP is to come up with an optimal mass transportation policy $\varphi_{ij}:=\varphi\left(u_i\to v_j\right)$ associated with cost c_{ij}. Clearly, in addition to constraints

Table 1
For various stable dynamical system pairs, we list asymptotic Wasserstein distance, defined as ${}_2W_2\left(\eta_\infty, \widehat{\eta}_\infty\right)$, where η_∞ and $\widehat{\eta}_\infty$ are the stationary PDFs of the true and model dynamics, respectively.

Systems	Dynamics	Stationary PDFs	Asymptotic ${}_2W_2$
Deterministic linear pair	$\dot{x}(t) = Ax(t),\ y(t) = Cx(t),$ $\dot{\widehat{x}}(t) = \widehat{A}\widehat{x}(t),\ \widehat{y}(t) = \widehat{C}\widehat{x}(t)$	$\eta_\infty = \delta\left(y\right)$ $\widehat{\eta}_\infty = \delta\left(\widehat{y}\right)$	0
Deterministic affine pair	$\dot{x}(t) = Ax(t) + b,\ y(t) = Cx(t) + d,$ $\dot{\widehat{x}}(t) = \widehat{A}\widehat{x}(t) + \widehat{b},\ \widehat{y}(t) = \widehat{C}\widehat{x}(t) + \widehat{d}$	$\eta_\infty = \delta\left(y + CA^{-1}b - d\right)$ $\widehat{\eta}_\infty = \delta\left(\widehat{y} + \widehat{C}\widehat{A}^{-1}\widehat{b} - \widehat{d}\right)$	$\left\Vert \left(d - \widehat{d}\right) - \left(CA^{-1}b - \widehat{C}\widehat{A}^{-1}\widehat{b}\right) \right\Vert_2$
Stochastic linear pair	$dx(t) = Ax(t)dt + Bd\beta(t),\ y(t) = Cx(t),$ $d\widehat{x}(t) = \widehat{A}\widehat{x}(t)dt + \widehat{B}d\widehat{\beta}(t),\ \widehat{y}(t) = \widehat{C}\widehat{x}(t)$	$\eta_\infty = \mathcal{N}\left(0, \Sigma_{y\infty}\right)$ $\widehat{\eta}_\infty = \mathcal{N}\left(0, \widehat{\Sigma}_{\widehat{y}\infty}\right)$	$\left(\text{tr}\left(\Sigma_{y\infty} + \widehat{\Sigma}_{\widehat{y}\infty} - 2\left[\Sigma_{y\infty}^{\frac{1}{2}}\widehat{\Sigma}_{\widehat{y}\infty}\Sigma_{y\infty}^{\frac{1}{2}}\right]^{\frac{1}{2}}\right)\right)^{\frac{1}{2}}$
Deterministic nonlinear and deterministic linear	$\dot{x}(t) = f\left(x(t)\right),\ y(t) = h\left(x(t)\right),$ $\dot{\widehat{x}}(t) = \widehat{A}\widehat{x}(t),\ \widehat{y}(t) = \widehat{C}\widehat{x}(t)$	$\eta_\infty = \sum_{i=1}^{n^\star} m_i^\star \delta\left(y - y_i^\star\right)$ $\widehat{\eta}_\infty = \delta\left(\widehat{y}\right)$	$\left(\sum_{i=1}^{n^\star} \Vert y_i^\star \Vert_2^2\, m_i^\star \delta\left(y - y_i^\star\right)\right)^{\frac{1}{2}}$
Deterministic nonlinear pair	$\dot{x}(t) = f\left(x(t)\right),\ y(t) = h\left(x(t)\right),$ $\dot{\widehat{x}}(t) = \widehat{f}\left(\widehat{x}(t)\right),\ \widehat{y}(t) = \widehat{h}\left(\widehat{x}(t)\right)$	$\eta_\infty = \sum_{i=1}^{n^\star} m_i^\star \delta\left(y - y_i^\star\right)$ $\widehat{\eta}_\infty = \sum_{i=1}^{\widehat{n}^\star} \widehat{m}_i^\star \delta\left(\widehat{y} - \widehat{y}_i^\star\right)$	Monge-Kantorovich optimal transport LP (24), (C1)–(C3)

(C1)–(C3), (24) must respect the necessary feasibility condition

$$\sum_{i=1}^{m} \alpha_i = \sum_{j=1}^{n} \beta_j \tag{C0}$$

denoting the conservation of mass. In our context of measuring the shape difference between two PDFs, we treat the joint probability mass function (PMF) vectors α_i and β_j to be the marginals of some unknown joint PMF φ_{ij} supported over the product space $U \times V$. Since determining joint PMF with given marginals is not unique, (24) strives to find that particular joint PMF which minimizes the total cost for transporting the probability mass while respecting the normality condition.

4.3 Computational complexity for ${}_2W_2$

4.3.1 Sample complexity

For a desired accuracy of Wasserstein distance computation, we want to specify the bounds for number of samples $m = n$, for a given initial PDF. Since the finite sample estimate of Wasserstein distance is a random variable, we need to answer how large should n be, in order to guarantee that the empirical estimate of Wasserstein distance obtained by solving the LP (24), (C1)–(C3) with $m = n$, is close to the true deterministic value of (12) in probability. In other words, given $\epsilon, \delta \in (0, 1)$, we want to estimate a lower bound of $m = n$ as a function of ϵ and δ, such that

$$\begin{aligned}\mathbb{P}\left(\left|{}_2W_2\left(\eta_m^j\left(y\right), \widehat{\eta}_n^j\left(\widehat{y}\right)\right) \right.\right. &- \left.\left. {}_2W_2\left(\eta^j\left(y\right), \widehat{\eta}^j\left(\widehat{y}\right)\right)\right| < \epsilon\right) \\ &> 1 - \delta, \qquad \forall j = 1, 2, \ldots, \tau.\end{aligned}$$

Similar consistency and sample complexity results are available in the literature (see Corollary 9(i) and Corollary 12(i) in [57]) for Wasserstein distance of order $q = 1$. From Hölder's inequality, $W_{q_2} > W_{q_1}$ for $q_2 > q_1$, and hence that sample complexity bound, in general, does not hold for $q = 2$.To proceed, we need the following results.

Lemma 1 *(Appendix C) If X, Y, Z are non-negative random variables such that Y and Z are independent, and $X \leqslant Y + Z$, then for $\epsilon > 0$, we have*

$$\mathbb{P}\left(X > \epsilon\right) \leqslant \mathbb{P}\left(Y + Z > \epsilon\right) \leqslant \mathbb{P}\left(Y > \frac{\epsilon}{2}\right) + \mathbb{P}\left(Z > \frac{\epsilon}{2}\right).$$

Definition 2 ***(Transportation cost inequality)****[58] A probability measure μ is said to satisfy the L_p-transportation cost inequality (TCI) of order q, if there exists some constant $C > 0$ such that for any probability measure ν, ${}_pW_q\left(\mu, \nu\right) \leqslant \sqrt{2CD_{KL}\left(\nu \parallel \mu\right)}$. In short, we write $\mu \in T_q\left(C\right)$. In particular, for $\mu \sim \mathcal{N}\left(m_{\kappa\times 1}, \Sigma_{\kappa\times\kappa}\right)$, we have [59] $\mu \in T_2\left(\lambda_{max}\left(\Sigma\right)\right)$.*

Theorem 1 ***(Rate-of-convergence of empirical measure in Wasserstein metric)****(Thm. 5.3, [60]) For a probability measure $\rho \in T_q\left(\mathscr{C}\right)$, $1 \leqslant q \leqslant 2$, and its n-sample estimate ρ_n, we have*

$$\mathbb{P}\left({}_pW_q\left(\rho, \rho_n\right) > \theta\right) \leqslant K_\theta \,\exp\left(-\frac{n\theta^2}{8\mathscr{C}}\right), \quad \theta > 0, \tag{25}$$

and $\log K_\theta := \dfrac{1}{\mathscr{C}} \inf\limits_{\mu} \#\left(\text{supp}\,\mu\right)\left(\text{diam}\left(\text{supp}\,\mu\right)\right)^2$. *The* *optimization takes place over all probability measures μ of finite support, such that ${}_pW_q(\rho,\mu) \leqslant \theta/4$.*

We now make few notational simplifications. In this subsection, we denote $\eta^j(y)$ and $\widehat{\eta}^j(y)$ by η and $\widehat{\eta}$, and their finite sample representations by η_m and $\widehat{\eta}_n$, respectively. Then we have the following result.

Theorem 2 ***(Rate-of-convergence of empirical Wasserstein estimate)*** *(Appendix D) For true densities η and $\widehat{\eta}$, let corresponding empirical densities be η_m and $\widehat{\eta}_n$, evaluated at respective uniform sampling of cardinality m and n. Let $\mathscr{C}_1$, $\mathscr{C}_2$, be the TCI constants for η and $\widehat{\eta}$, respectively and fix $\epsilon > 0$. Then*

$$\mathbb{P}\left(\left|\,{}_2W_2(\eta_m,\widehat{\eta}_n) - {}_2W_2(\eta,\widehat{\eta})\right| > \epsilon\right)$$
$$\leqslant K_1 \exp\left(-\frac{m\epsilon^2}{32\mathscr{C}_1}\right) + K_2 \exp\left(-\frac{n\epsilon^2}{32\mathscr{C}_2}\right). \qquad (26)$$

Remark 4 *At a fixed time, K_1, K_2, $\mathscr{C}_1$ and $\mathscr{C}_2$ are constants in a given model validation problem, i.e. for a given pair of experimental data and proposed model. However, values of these constants depend on true and model dynamics. In particular, the TCI constants $\mathscr{C}_1$ and $\mathscr{C}_2$ depend on the dynamics via respective PDF evolution operators. The constants K_1 and K_2 depend on η and $\widehat{\eta}$, which in turn depend on the dynamics. For pedagogical purpose, we next illustrate the simplifying case $K_1 = K_2 = K$, $\mathscr{C}_1 = \mathscr{C}_2 = \mathscr{C}$.*

Corollary 1 ***(Sample complexity for empirical Wasserstein estimate)*** *For desired accuracy $\epsilon \in (0,1)$, and confidence $1-\delta$, $\delta \in (0,1)$, the sample complexity $m = n = N_{wass}$, for finite sample Wasserstein computation is given by*

$$N_{wass} = \left(\frac{32\mathscr{C}}{\epsilon^2}\right) \log\left(\frac{2K}{\delta}\right). \qquad (27)$$

4.3.2 Runtime complexity

The LP formulation (24), (C1)–(C3), requires solving for mn unknowns subject to $(m+n+mn)$ constraints. For $m = n$, it can be shown that [61,62] the runtime complexity for solving the LP is $O\left(n_o\, n^{2.5} \log \nu\right)$. Notice that the output dimension n_o enters only through the cost c_{ij} in (24) and hence affects the computational time linearly.

In actual simulations, we found the runtime of the LP (24) to be sensitive on how the constraints were implemented. Suppose, we put (24) in standard form

$$\text{minimize } \widetilde{c}^\top \widetilde{\varphi}, \qquad \text{subject to } A\widetilde{\varphi} = b, \quad \widetilde{\varphi} \geqslant 0, \qquad (28)$$

where $\widetilde{c}_{mn\times 1} := \text{vec}(c)$, $\widetilde{\varphi}_{mn\times 1} := \text{vec}(\varphi)$, $b_{(m+n)\times 1} := [\alpha_{m\times 1}, \beta_{n\times 1}]^\top$. If we let $e_n := [\underbrace{1,1,\ldots,1}_{n\,\text{times}}]^\top$, then the implementation $A_{(m+n)\times mn} = \begin{bmatrix} e_n^\top \otimes I_m \\ I_n \otimes e_m^\top \end{bmatrix}$ was found to achieve fast offline construction of the constraint matrix.

4.3.3 Storage complexity

For $m = n$, the constraint matrix A in (28), is a binary matrix of size $2n \times n^2$, whose each row has n ones. Consequently, there are total $2n^2$ ones in the constraint matrix and the remaining $2n^2(n-1)$ elements are zero. Hence at any fixed time, the sparse representation of the constraint matrix needs # non-zero elements ×3 = $6n^2$ storage. The PMF vectors are, in general, fully populated. In addition, we need to store the model and true sample coordinates, each of them being a n_o-tuple. Hence at any fixed time, constructing cost matrix requires storing $2n_o n$ values. Thus total storage complexity at any given snapshot, is $2n(3n + n_o + 1) = O(n^2)$, assuming $n > n_o$. However, if the sparsity of constraint matrix is not exploited by the solver, then storage complexity rises to $2n\left(n^2 + n_o + 1\right) = O\left(n^3\right)$. For example, if we take $n = 1000$ samples and use double precision arithmetic, then solving the LP at each time requires either megabytes or gigabytes of storage, depending on whether or not sparse representation is utilized by the solver [1]. For $m \neq n$, it is easy to verify that the sparse storage complexity is $(6mn + (m+n)\, n_o + m + n)$, and the non-sparse storage complexity is $(m+n)(mn + n_0 + 1)$.

5 Construction of Validation Certificates

5.1 Probabilistically robust model validation

Often in practice, the exact initial density is not known to facilitate our model validation framework; instead a class of densities may be known. For example, it may be known that the initial density is symmetric unimodal but its exact shape (e.g. normal, semi-circular etc.) may not be known. Even when the distribution-type is known (e.g. normal), it is often difficult to pinpoint the parameter values describing the initial density function. To account such scenarios, consider a random variable $\Delta : \Omega \to E$, that induces a probability triplet $(\Omega, \mathcal{F}, \mathbb{P})$ on the space of initial densities. Here $E \subset \Omega$ and $\#(E) = 1$. The random variable Δ picks up an initial density from the collection of admissible initial densities $\Omega := \{\xi_0^{(1)}(\widetilde{x}), \xi_0^{(2)}(\widetilde{x}), \ldots\}$ according to the law of Δ. For example, if we know $\xi_0 \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ with a given joint distribution over the $\left(\mu_0, \sigma_0^2\right)$ space, then in our model

[1] We used MOSEK (available at `www.mosek.com`) as the LP solver.

Algorithm 1 Construct PRVC

```
Require: $\epsilon, \delta \in (0,1)$, $T$, $\nu$, law of $\Delta$, experimental data $\{\eta_k(y)\}_{k=1}^{\tau}$, model, tolerance vector $\{\gamma_k\}_{k=1}^{\tau}$
$N \leftarrow N_{\mathrm{ch}}(\epsilon, \delta)$    ▷ Using lemma 2
Draw random functions $\xi_0^{(1)}(\widetilde{x}), \xi_0^{(2)}(\widetilde{x}), \ldots, \xi_0^{(N)}(\widetilde{x})$ according to the law of $\Delta$
for $k = 1$ to $\tau$ do    ▷ Index for time step
    for $i = 1$ to $N$ do    ▷ Index for initial density
        for $j = 1$ to $\nu$ do    ▷ Samples drawn from $\xi_0^{(i)}(\widetilde{x})$
            Propagate states using dynamics
            Propagate measurements
        end for
        Propagate $\widehat{\xi}_k^{(i)}\left(\widehat{\widetilde{x}}\right)$    ▷ Use (3), (7), (9) or (11)
        Compute $\widehat{\eta}_k^{(i)}(\widehat{y})$
        Compute ${}_2W_2\left(\eta_k^{(i)}(y), \widehat{\eta}_k^{(i)}(\widehat{y})\right)$    ▷ Distributional comparison by solving LP (24) subject to (C0)–(C3)
        sum $\leftarrow 0$    ▷ Initialize
        if ${}_2W_2\left(\eta_k^{(i)}(y), \widehat{\eta}_k^{(i)}(\widehat{y})\right) \leqslant \gamma_k$ then
            sum $\leftarrow$ sum + 1
        end if
    end for
    $\widehat{p}_N(\gamma_k) \leftarrow \dfrac{\text{sum}}{N}$    ▷ Construct PRVC vector
end for
```

validation framework, one sample from this space will return one distance measure between the instantaneous output PDFs. How many such $\left(\mu_0, \sigma_0^2\right)$ samples are necessary to guarantee the robustness of the model validation oracle? The Chernoff bound provides such an estimate for finite sample complexity.

At time step t_k, let the *validation probability* be $p(\gamma_k) := \mathbb{P}\left({}_2W_2\left(\eta_k(y), \widehat{\eta}_k(\widehat{y})\right) \leqslant \gamma_k\right)$. Here $\gamma_k \in \mathbb{R}^+$ is the prescribed instantaneous tolerance level. If the model validation is performed by drawing N samples from Ω, then the *empirical validation probability* is $\widehat{p}_N(\gamma_k) := \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \chi_{V_k^{(i)}}$ where $V_k^{(i)} := \{\widehat{\eta}_k^{(i)}(\widehat{y}) : {}_2W_2\left(\eta_k^{(i)}(y), \widehat{\eta}_k^{(i)}(\widehat{y})\right) \leqslant \gamma_k\}$. Consider $\epsilon, \delta \in (0,1)$ as the desired accuracy and confidence, respectively.

Lemma 2 ***(Chernoff bound)****[63] For any $\epsilon, \delta \in (0,1)$, if $N \geqslant N_{\mathrm{ch}} := \dfrac{1}{2\epsilon^2}\log\dfrac{2}{\delta}$, then $\mathbb{P}\left(|p(\gamma_k) - \widehat{p}_N(\gamma_k)| < \epsilon\right) > 1 - \delta$.*

The above lemma allows us to construct *probabilistically robust validation certificate* (PRVC) $\widehat{p}_N(\gamma_k)$ through the algorithm below. The PRVC vector, with ϵ accuracy, returns the probability that the model is valid at time t_k, in the sense that the instantaneous output PDFs are no distant than the required tolerance level γ_k. Lemma 2 lets the user control the accuracy ϵ and the confidence δ, with which the preceding statement can be made. Thus the framework enables us to compute a provably correct validation certificate on the face of uncertainty with finite sample complexity.

Algorithm 2 Construct PWVC

```
Require: $\epsilon, \delta \in (0,1)$, $\tau$, $\nu$, law of $\Delta$, experimental data $\{\eta_k(y)\}_{k=1}^{\tau}$, model
$N \leftarrow N_{\mathrm{wc}}(\epsilon, \delta)$    ▷ Using lemma 3
Draw $N$ random functions $\xi_0^{(1)}(\widetilde{x}), \xi_0^{(2)}(\widetilde{x}), \ldots, \xi_0^{(N)}(\widetilde{x})$ according to the law of $\Delta$    ▷ Use MCMC
for $k = 1$ to $\tau$ do    ▷ Index for time step
    for $i = 1$ to $N$ do    ▷ Index for initial density
        for $j = 1$ to $\nu$ do    ▷ Index for samples in the extended state space, drawn from $\xi_0^{(i)}(\widetilde{x})$
            Propagate states using dynamics
            Propagate measurements
        end for
        Propagate $\widehat{\xi}_k^{(i)}\left(\widehat{\widetilde{x}}\right)$    ▷ Use (3), (7), (9) or (11)
        Compute $\widehat{\eta}_k^{(i)}(\widehat{y})$    ▷ Algebraic transformation
        Compute ${}_2W_2\left(\eta_k^{(i)}(y), \widehat{\eta}_k^{(i)}(\widehat{y})\right)$    ▷ Distributional comparison by solving LP
        $\widehat{\gamma}_k^N \leftarrow \max\limits_{i=1,\ldots,N} {}_2W_2\left(\eta_k^{(i)}(y), \widehat{\eta}_k^{(i)}(\widehat{y})\right)$    ▷ Empirically estimate worst-case performance
    end for
end for
```

5.2 *Probabilistically worst-case model validation*

Following [64–66], one can also define a probabilistic notion of the worst-case model validation performance as $\gamma_k^{\mathrm{wc}} := \sup\limits_{\Delta} {}_2W_2\left(\eta_k(y), \widehat{\eta}_k(\widehat{y})\right)$, and its empirical estimate $\widehat{\gamma}_k^N := \max\limits_{i=1,\ldots,N} {}_2W_2\left(\eta_k^{(i)}(y), \widehat{\eta}_k^{(i)}(\widehat{y})\right)$. The sample complexity for probabilistically worst-case model validation is given by the lemma below.

Lemma 3 ***(Worst-case bound)*** *(p. 128, [63]) For any $\epsilon, \delta \in (0,1)$, if $N \geqslant N_{wc} := \dfrac{\log\frac{1}{\delta}}{\log\frac{1}{1-\epsilon}}$, then $\mathbb{P}\left(\mathbb{P}\left({}_2W_2\left(\eta_k(y), \widehat{\eta}_k(\widehat{y})\right) \leqslant \widehat{\gamma}_k^N\right) \geqslant 1 - \epsilon\right) > 1 - \delta$.*

Notice that in general, there is no guarantee that the empirical estimate $\widehat{\gamma}_k^N$ is close to the true worst-case performance γ_k^{wc}. Also, the performance bound is obtained *a posteriori* while the robust validation framework accounted for *a priori* tolerance levels. The corresponding *probabilistically worst-case validation certificate* (PWVC) $\widehat{\gamma}_k^N$ can be computed from the following algorithm. In summary, the algorithm, with high probability $(1-\epsilon)$, only ensures that the output PDFs are at most $\widehat{\gamma}_k^N$ far. The preceding statement can be made with probability at least $1 - \delta$.

6 Illustrative Examples

Example 1 *Continuous-time deterministic dynamics*

Consider the following nonlinear dynamical system

$$\ddot{x} = -ax - b\sin 2x - c\dot{x}, \quad a = 0.1, b = 0.5, c = 1. \quad (29)$$

The system has five fixed points $P_0 = (0,0)$, $P_1^{\pm} = (\pm 1.7495, 0)$, $P_2^{\pm} = (\pm 2.8396, 0)$, which can be solved by noting the abscissa values of the points of intersection of two curves $f(x) = b\sin 2x$ and $g(x) = -ax$, as shown in Fig 6. From linear analysis, it is easy to verify that P_0 and $P_2^{\pm}$ are stable foci while $P_1^{\pm}$ are saddles (Fig. 7). To illustrate our model validation framework, let's assume that 'true data' is generated by the dynamics (29). However, this true dynamics is unknown to the modeler, whose proposed model is a linearization of (29) about the origin. We emphasize here that the purpose of (29) is only to create the synthetic data and to demonstrate the proof-of-concept. In a realistic model validation, the data arrives from experimental measurements, not from another model. For simplicity, we take the outputs same as states for both true and model dynamics.

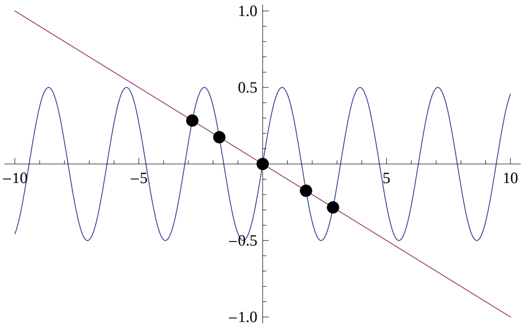


Fig. 6. Points of intersection of the curve $f(x) = b\sin 2x$ and the line $g(x) = -ax$.

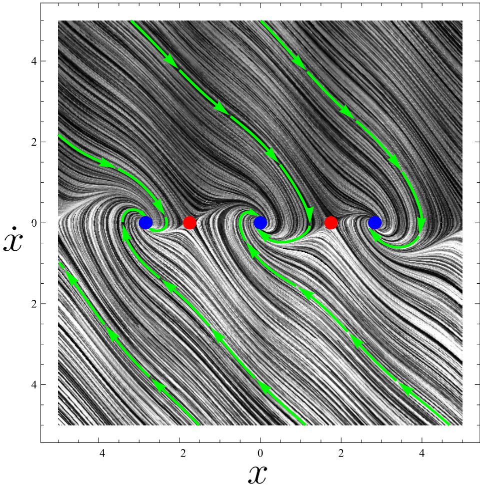


Fig. 7. Phase portrait of the vector field (29) with three stable and two saddle fixed points.

Starting from the bivariate uniform distribution $\mathcal{U}([-\pi,\pi] \times [-\pi,\pi]) =: \xi_0$, we evolve the respective joint PDFs $\xi = \eta$ and $\widehat{\xi} = \widehat{\eta}$, through true and model dynamics via MOC implementation of Liouville equation [31]. The distributional shape discrepancy is captured via the Wasserstein gap $({}_2W_2(\eta, \widehat{\eta}))$ between these instantaneous joint PDFs, as shown in Fig. 8 (*solid line*), computed by solving the LP (24), (C1)–(C3). As the individual joint PDFs converge toward their respective stationary densities, the slope of the Wasserstein time-history decreases progressively. Fig. 9 shows the Wasserstein gap trajectories when ξ_0 is taken to be $\mathcal{N}(0, \sigma_0^2 I_2)$, instead of uniform. In this case, we observe that larger initial dispersion causes larger Wasserstein gap. Suppose the user-specified tolerance level $\{\gamma_j\}_{j=1}^{40}$ is 0.8 for first 10 instances and 0.6 for next 30 instances of measurement availability, as shown by the shaded area in Fig. 9. Given the set of admissible initial densities $\{\xi_0^{(1)}, \ldots, \xi_0^{(9)}\}$ with $\xi_0^{(i)} := \mathcal{N}(0, \sigma_{0i}^2 I_2)$, $i = 1, \ldots, 9$, we can compute the PRVC vector, shown as the dashed line in Fig. 9, to be

$$\left[\underbrace{1, \ldots, 1}_{\text{3 times}}, 0.89, \underbrace{0.78, \ldots, 0.78}_{\text{5 times}}, 0.67, \underbrace{0.56, \ldots, 0.56}_{\text{30 times}}\right]^{\top}.$$

Example 2 *Continuous-time stochastic dynamics*

Here we assume the true data to be generated by (29) with additive white noise having autocorrelation $Q\delta(t_1 - t_2)$, $t_1, t_2 \geqslant 0$. Letting $x_1 = x$ and $x_2 = \dot{x}$, the associated Itô SDE can be written in state-space form similar to (5)

$$\begin{Bmatrix} dx_1 \\ dx_2 \end{Bmatrix} = \begin{Bmatrix} x_2 \\ -ax_1 - b\sin 2x_1 - cx_2 \end{Bmatrix} dt + \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} d\beta, (30)$$

where $\beta(t)$ is a Wiener process with autocorrelation $Q(t_1 \wedge t_2)$. The stationary Fokker-Planck equation for (30) can be solved in closed form (Appendix E)

$$\eta_\infty(x_1, x_2) \propto \exp\left(-\frac{c}{2Q}\left(ax_1^2 + x_2^2 - b\cos 2x_1\right)\right), \quad (31)$$

and one can verify that peaks of (31) appear at the fixed points of the nonlinear drift.

Let the proposed model be the linearization of (30) about the origin. It is well-known [68] that the stationary density of a linear SDE of the form $d\widehat{x} = A\widehat{x}\, dt + B\, d\beta$, is given by

$$\widehat{\eta}_\infty(\widehat{x}) = \mathcal{N}(\mathbf{0}, \Sigma_\infty) = \frac{\exp\left(-\frac{1}{2}\ \widehat{x}^\top \Sigma_\infty^{-1} \widehat{x}\right)}{\sqrt{(2\pi)^2 \det(\Sigma_\infty)}}, \quad (32)$$

provided A is Hurwitz and (A, B) is a controllable pair. The steady-state covariance matrix Σ_∞ solves $A\Sigma_\infty + \Sigma_\infty A^\top + BQB^\top = 0$. For the linearized version of (30),

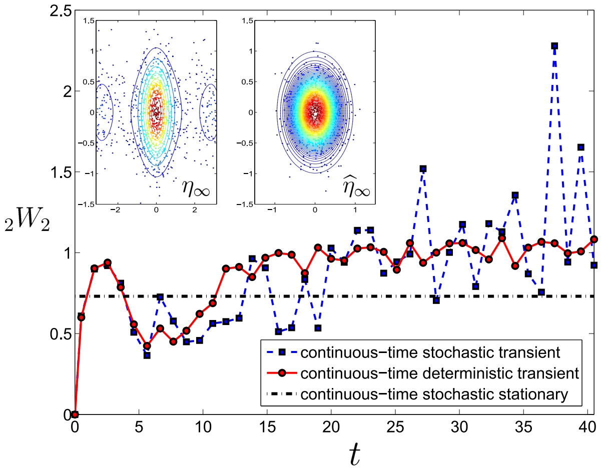


Fig. 8. Starting with $\xi_0 = \mathcal{U}\left([-\pi,\pi]\times[-\pi,\pi]\right)$, the *solid line* shows time history of ${}_2W_2$ measured between the joint state PDFs for (29) and its linearization about the origin. The *dashed line* shows the same between (30) and its linearization about the origin. The *dash-dotted line* shows the stationary ${}_2W_2$ between known η_∞ and $\widehat{\eta}_\infty$ (*contours in the inset plot*), given by (31) and (32) respectively, and is computed by solving the optimal transport LP between their MCMC samples (*scattered points in the inset plot*). All computations were done with 1000 Halton samples [67] from ξ_0 and 50 eigenfunctions in noise KL expansion.

$A = \begin{bmatrix} 0 & 1 \\ (-a-2b) & -c \end{bmatrix}$ and $B = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix}$ satisfy the aforementioned conditions and the stationary density is obtained from (32).

Taking the initial density same as in Example 3.1, we propagated the joint PDFs for (30) and the linear SDE using the KLPF method described in [36]. The *dashed line* in Fig. 8 shows the Wasserstein trajectory for this case. The dash-dotted line in Fig. 8 shows the asymptotic Wasserstein gap between the respective stationary densities (31) and (32). Due to randomized sampling, all stochastic computations are in probabilistically approximate sense [69].

Example 3 *Discrete-time deterministic dynamics*

Let the true data be generated by the Chebyshev map [70] $\mathcal{T} : [-1,1] \mapsto [-1,1]$, given by

$$x_{k+1} = \mathcal{T}\left(x_k\right) = \cos\left(2\cos^{-1} x_k\right). \tag{33}$$

If we let $\xi_k := \xi\left(x_k\right)$, then the PF operator $\mathscr{P} : \xi_k \mapsto \xi_{k+1}$, for (33) can be computed [71] as

$$\mathscr{P}\xi_k = \frac{1}{2\sqrt{2x_k+2}}\left[\xi\left(\sqrt{\frac{x_k+1}{2}}\right) + \xi\left(-\sqrt{\frac{x_k+1}{2}}\right)\right] \tag{34}$$

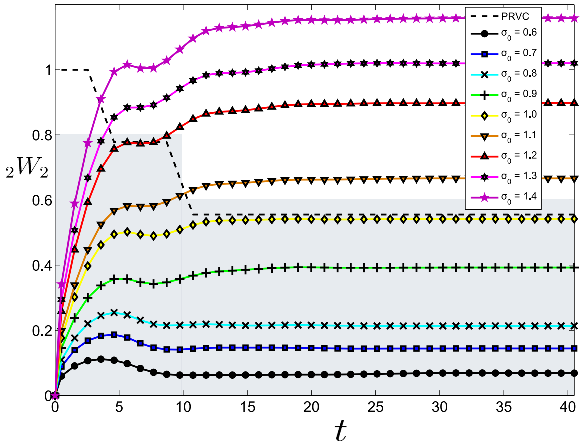


Fig. 9. Starting with $\xi_0 = \mathcal{N}\left(0, \sigma_0^2 I_2\right)$, transient Wasserstein time histories, measured between the joint state PDFs for (29) and its linearization about the origin. In this case, increasing σ_0 increases ${}_2W_2$ *at all times*. Further, notice that ${}_2W_2$ trajectories with larger σ_0, converges to higher asymptotic values. This is due to the fact that the stationary density of (29) is of the form $\eta_\infty(y) = \sum_{i=1}^{5} m_i^\star \delta\left(y - y_i^\star\right)$, and hence depends on ξ_0, as explained in Remark 3 and fourth row of Table 1. The *shaded area* shows user-specified tolerance level $\{\gamma_j\}_{j=1}^{40}$, from which PRVC (*dashed line*) can be computed. In this case, PWVC is simply the ${}_2W_2$ trajectory corresponding to $\sigma_0 = 1.4$.

with stationary PDF $\xi_\infty\left(x\right) = \dfrac{1}{\pi\sqrt{1-x^2}}$, and CDF $F_\infty\left(x\right) = \dfrac{2}{\pi}\sin^{-1}\left(\sqrt{\dfrac{x+1}{2}}\right)$. Notice that for small x_k, (33) behaves like a quadratic transformation. Suppose the following logistic map $\widehat{\mathcal{T}} : [0,1] \mapsto [0,1]$, is proposed to model the data generated by (33):

$$x_{k+1} = \widehat{\mathcal{T}}\left(\widehat{x}_k\right) = 4\widehat{x}_k\left(1-\widehat{x}_k\right). \tag{35}$$

The PF operator $\widehat{\mathscr{P}} : \widehat{\xi}_k \mapsto \widehat{\xi}_{k+1}$, for (35) is given by [26]

$$\widehat{\mathscr{P}}\widehat{\xi}_k = \frac{1}{4\sqrt{1-\widehat{x}_k}}\left[\widehat{\xi}\left(\frac{1+\sqrt{1-\widehat{x}_k}}{2}\right) + \widehat{\xi}\left(\frac{1-\sqrt{1-\widehat{x}_k}}{2}\right)\right] \tag{36}$$

with stationary PDF $\widehat{\xi}_\infty\left(\widehat{x}\right) = \dfrac{1}{\pi\sqrt{\widehat{x}\left(1-\widehat{x}\right)}}$, and CDF $\widehat{F}_\infty\left(\widehat{x}\right) = \dfrac{2}{\pi}\sin^{-1}\left(\sqrt{\widehat{x}}\right)$. Taking the outputs identical to states, the asymptotic Wasserstein distance between

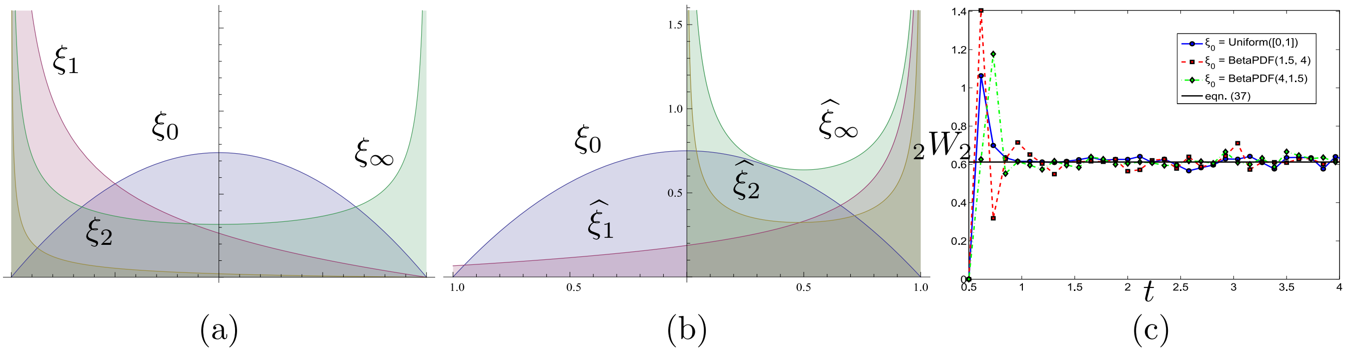


Fig. 10. Starting with $\xi_0(x) = \frac{3}{4}\left(1-x^2\right)$, evolution of PDFs for (a) *true* PF operator (34), and (b) *model* PF operator (36). (c) Wasserstein time histories between PF operators (34) and (36) for various initial PDFs.

(33) and (35), becomes

$$
\begin{aligned}
&{}_2W_2\left(\xi_\infty(x), \widehat{\xi}_\infty(\widehat{x})\right) = \sqrt{\int_0^1 \left(F_\infty^{-1}(\varsigma) - \widehat{F}_\infty^{-1}(\varsigma)\right)^2 d\varsigma} \\
&= \sqrt{\int_0^1 \left(2\sin^2\left(\frac{\pi\varsigma}{2}\right) - 1 - \sin^2\left(\frac{\pi\varsigma}{2}\right)\right)^2 d\varsigma} \\
&= \sqrt{\int_0^1 \left(\frac{1}{2} + \frac{\cos(\pi\varsigma)}{2}\right)^2 d\varsigma} \approx 0.6124.
\end{aligned} \tag{37}
$$

Given an initial density ξ_0, the transient PDFs $\xi(x,t)$ and $\widehat{\xi}(x,t)$ can be computed from (34) and (36) (Fig. 10(a) and (b)). Fig. 10(c) shows the transient Wasserstein time-histories ${}_2W_2\left(\xi(x,t), \widehat{\xi}(x,t)\right)$ for various initial PDFs, which converge to its asymptotic value obtained analytically in (37).

Example 4 *Discrete-time stochastic dynamics*

Consider the true data being generated from the logistic map with multiplicative stochastic perturbation:

$$
x_{k+1} = \mathcal{T}(x_k, \zeta_k) = \zeta_k \mathcal{S}(x_k) = \zeta_k x_k (1 - x_k), \tag{38}
$$

where $\mathcal{S} : [0,1] \mapsto [0,1]$, and $\{\zeta_k\}_0^\infty$ are i.i.d random variables on $[0,4]$, drawn from noise density $\phi(.)$. This map has found applications in population dynamics and size-dependent branching processes [72,73]. The PF operator for (38) is given by (p. 330-331, [26])

$$
\mathscr{P}\xi_k = \int_0^\infty \xi(y)\, \mathcal{K}_{\text{mul}}(x_k, y)\; dy, \tag{39}
$$

with the *multiplicative* stochastic kernel $\mathcal{K}_{\text{mul}}(x_k, y) := \dfrac{1}{\mathcal{S}(y)}\,\phi\left(\dfrac{x_k}{\mathcal{S}(y)}\right)$. In particular, $\zeta_k \sim \mathcal{N}(0,1)$ results $\mathscr{P}\xi_k = \displaystyle\int_0^\infty \xi(y)\, \frac{1}{\sqrt{2\pi}\, y(1-y)}\; e^{-\frac{1}{2}\frac{x^2}{y^2(1-y)^2}}\, dy$. The asymptotic behavior of (38) is known [72] to depend on the noise density $\phi(.)$. Specifically, $\mathbb{E}[\log\zeta_0] < 0, = 0$, and > 0 implies $x_k \xrightarrow{\text{a.s.}} 0$, $x_k \xrightarrow{\text{i.p.}} 0$, and existence of stationary density ξ_∞ on $(0,1)$ $\forall x_0 \neq 0$, respectively. For example, if $\zeta_k \sim \mathcal{N}(0,1)$, then $\displaystyle\int_0^4 \log\zeta \frac{1}{\sqrt{2\pi}} e^{-\frac{\zeta^2}{2}}\, d\zeta = \text{erf}\left(2\sqrt{2}\right)\log(2) - 2\sqrt{\frac{2}{\pi}}\; {}_2F_2\left(\frac{1}{2}, \frac{1}{2}; \frac{3}{2}, \frac{3}{2}; -8\right) \approx -0.32 < 0$, and hence $x_k \xrightarrow{\text{a.s.}} 0$.

Let the proposed model be

$$
\widehat{x}_{k+1} = \widehat{\mathcal{T}}\left(\widehat{x}_k, \widehat{\zeta}_k\right) = \widehat{\mathcal{S}}(x_k) + \widehat{\zeta}_k = \widehat{x}_k + \widehat{\zeta}_k, \tag{40}
$$

where $\widehat{\mathcal{S}} : \mathbb{R} \mapsto \mathbb{R}$, and $\widehat{\zeta}_k \sim \mathcal{N}(0,1)$. The PF operator for a map with additive noise is of the form

$$
\widehat{\mathscr{P}}\widehat{\xi}_k = \int_{-\infty}^\infty \widehat{\xi}(y)\, \mathcal{K}_{\text{add}}(\widehat{x}_k, y)\; dy, \tag{41}
$$

with the *additive* stochastic kernel $\mathcal{K}_{\text{add}}(\widehat{x}_k, y) := \phi\left(\widehat{x}_k - \widehat{\mathcal{S}}(y)\right)$. Consequently, the PF operator for (40) is $\widehat{\mathscr{P}}\widehat{\xi}_k = \displaystyle\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\widehat{x}_k - y)^2}{2}\right) \widehat{\xi}(y)\; dy$. It can be verified (p. 325, [26]) that the successive iterate $\widehat{\mathscr{P}}^k\widehat{\xi}$ converges uniformly to zero as $k \to \infty$, and hence there is no non-trivial stationary density. Given an initial density, the Wasserstein distance can be computed between (39) and (41). This example demonstrates that (in)validating a stochastic map has sensitive dependence on noise density.

Example 5 *Comparison with barrier certificate based model falsification*

Consider the nonlinear model validation problem stated as Example 4 in [6], where the model is $\dot{x} = -px^3$, with parameter $p \in \mathcal{P} = [0.5, 2]$. The measurement data are interval-valued sets $\mathcal{X}_0 = [0.85, 0.95]$ at $t = 0$, and $\mathcal{X}_T = [0.55, 0.65]$ at $t = T = 4$. A barrier certificate of the form $B(x,t) = B_1(x) + tB_2(x)$ was found in [6] through sum-of-squares (SOS) optimization [74] where $B_1(x) = 8.35x + 10.40x^2 - 21.50x^3 + 9.86x^4$, and $B_2(x) = -1.78 + 6.58x - 4.12x^2 - 1.19x^3 + 1.54x^4$. The model was thereby

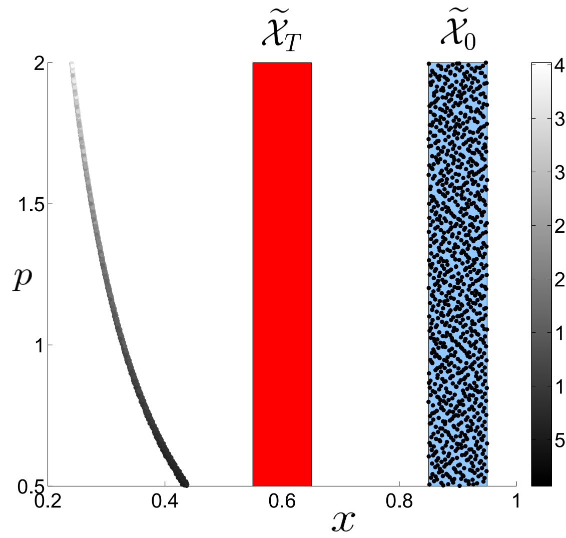


Fig. 11. This plot illustrates how Prajna's barrier certificate-based invalidation result can be recovered in our probabilistic model validation framework. To show $\widetilde{\mathcal{X}}_T$ is not reachable from the set $\widetilde{\mathcal{X}}_0$ in time T = 4, we sample $\widetilde{\mathcal{X}}_0$ uniformly and propagate that uniform ensemble subject to the proposed model dynamics till T = 4. The samples are gray-scale color coded (white = high probability, black = low probability) according to the value of the joint PDF at that location. Here, the model is invalidated since the pair of joint PDFs at initial and final time, does not satisfy the Liouville transport PDE corresponding to the model dynamics, as proved in Theorem 3.

invalidated by the existence of such certificate, i.e. the model $\dot{x} = -px^3$, with parameter $p \in \mathcal{P}$ was shown to be inconsistent with measurements $\{\mathcal{X}_0, \mathcal{X}_T, T\}$.

To tackle this problem in our model validation framework, consider the spatio-temporal evolution of the joint PDF $\xi\left(x, p, t\right)$ over the extended state space $\widetilde{x} = [x \;\; p]^\top$, with initial support $\widetilde{\mathcal{X}}_0 := \mathcal{X}_0 \times \mathcal{P}$, under the action of the extended vector field $\widetilde{f}\left(x, p\right) = \left[-px^3 \quad 0\right]^\top$. Our objective then, is to prove that for $T = 4$, the PDF $\xi_T\left(x_T, p, T\right) = \mathcal{U}\left(x_T, p\right) = 1/\text{vol}\left(\widetilde{\mathcal{X}}_T\right)$ is not finite-time reachable from $\xi_0\left(x_0, p\right) = \mathcal{U}\left(x_0, p\right) = 1/\text{vol}\left(\widetilde{\mathcal{X}}_0\right)$, subject to the proposed model dynamics on the extended state space.

Theorem 3 *The two-point boundary value problem*

$$\frac{\partial \xi}{\partial t} + \nabla_{\widetilde{x}} \cdot \left(\widetilde{f}\left(x, p\right) \xi\right) = \frac{\partial \xi}{\partial t} + \nabla_{x} \cdot \left(-px^3 \xi\right) = 0,$$
$$\xi\left(x(0), p, 0\right) = \xi_0\left(x_0, p\right) = \mathcal{U}\left(x_0, p\right) = 1/\text{vol}\left(\widetilde{\mathcal{X}}_0\right),$$
$$\xi\left(x(T), p, T\right) = \xi_T\left(x_T, p, T\right) = \mathcal{U}\left(x_T, p\right) = 1/\text{vol}\left(\widetilde{\mathcal{X}}_T\right),$$

has no solution for $\xi\left(x, p, t\right)$, *such that* $\int_{\widetilde{\mathcal{X}}(t)} \xi\left(x, p, t\right) dx dp = 1, \forall t \in (0, T)$.

Proof. MOC ODE [31] corresponding to the Liouville PDE $\frac{\partial \xi}{\partial t} + \nabla_{\widetilde{x}} \cdot \left(\widetilde{f}\left(x, p\right) \xi\right) = 0$, yields a solution of the form

$$\xi\left(x, p, t\right) = \xi_0\left(x_0, p\right) \exp\left(-\int_0^t \nabla_{\widetilde{x}} \cdot \left(\widetilde{f}\left(x\left(\tau\right), p\right)\right) d\tau\right) \quad (42)$$

For the model dynamics $\dot{x} = -px^3$, we have $\nabla_{\widetilde{x}} \cdot \left(\widetilde{f}\left(x\left(\tau\right), p\right)\right) = -3p\left(x\left(\tau\right)\right)^2$ and $\dfrac{1}{x^2} = \dfrac{1}{x_0^2} + 2pt$. Consequently (42) results

$$\begin{aligned} \xi\left(x, p, t\right) &= \xi_0\left(x_0, p\right) \left(1 + 2x_0^2 pt\right)^{3/2} \\ &= \frac{1}{\left(1 - 2x^2 pt\right)^{3/2}} \xi_0\left(\pm \frac{x}{\sqrt{1 - 2x^2 pt}}, p\right). \end{aligned} \quad (43)$$

In particular, for $\xi_0\left(x_0, p\right) = 1/\text{vol}\left(\widetilde{\mathcal{X}}_0\right)$, $\xi_T\left(x_T, p, T\right) = 1/\text{vol}\left(\widetilde{\mathcal{X}}_T\right)$, and $T = 4$, (43) requires us to satisfy

$$\left(1 - 8x_T^2 p\right) = \left(\frac{\text{vol}\left(\widetilde{X}_T\right)}{\text{vol}\left(\widetilde{X}_0\right)}\right)^{2/3} > 0 \Rightarrow 1 > 8x_T^2 p. \quad (44)$$

Since $8x_T^2 p$ is an increasing function in both $x_T \in \mathcal{X}_T$ and $p \in \mathcal{P}$, we need at least $1 > 8\left(x_T\right)^2_{\min} p_{\min} = 8 \times (0.55)^2 \times 0.5 = 1.21$, which is incorrect. Thus the PDF $\xi_T\left(x_T, p, T\right) \sim \mathcal{U}\left(x_T, p\right)$ is not finite-time reachable from $\xi_0\left(x_0, p\right) \sim \mathcal{U}\left(x_0, p\right)$ for $T = 4$, via the proposed model dynamics. Hence our measure-theoretic formulation recovers Prajna's invalidation result [6] as a special case. ■

Remark 5 ***(Relaxation of set-based invalidation)*** *Instead of binary (in)validation oracle, we can now measure the "degree of validation" by computing the Wasserstein distance* ${}_2W_2\left(\frac{1}{\left(1-2x_T^2 pT\right)^{3/2}} \frac{1}{vol\left(\widetilde{\mathcal{X}}_0\right)}, \frac{1}{vol\left(\widetilde{\mathcal{X}}_T\right)}\right)$ *between the model predicted and experimentally measured joint PDFs. More importantly, it dispenses off the conservatism in barrier certificate based model validation by showing that the goodness of a model depends on the measures over the same pair of supports* $\widetilde{\mathcal{X}}_0$ *and* $\widetilde{\mathcal{X}}_T$*, than on the supports themselves. Indeed, given a joint PDF* $\xi\left(x_T, p, T\right)$ *supported over* $\widetilde{\mathcal{X}}_T$ *at* $T = 4$*, from (43) we can explicitly compute the initial PDF* $\xi_0\left(x_0, p\right)$ *supported over* $\widetilde{\mathcal{X}}_0$ *that, under the proposed model dynamics, yields the prescribed PDF, i.e.*

$$\xi_0\left(x_0, p\right) = \frac{1}{\left(1 + 8x_0^2 p\right)^{3/2}} \xi\left(\pm \frac{x_0}{\sqrt{1 + 8x_0^2 p}}, p, 4\right). \quad (45)$$

In other words, if the measurements find the initial density given by (45) and final density $\xi\left(x_T, p, T\right)$ at $T = 4$, then the Wasserstein distance at $T = 4$ will be zero, thereby perfectly validating the model. This reinstates the importance of considering the ***reachability of densities*** *over sets than* ***reachability of sets****, for model validation.*

Remark 6 ***(Connections with Rantzer's density function-based invalidation)*** *Similar to barrier certificates, Rantzer's density functions [75] can provide deductive invalidation guarantees (cf. Theorem 1 in [76]) by constructing a scalar function via convex program. Various applications of these two approaches for temporal verification problems have been reported [77]. It is interesting to note that the main idea of Rantzer's density function stems from an integral form of Liouville equation, given by (cf. Lemma A.1 in [75])*

$$\int_{\mathcal{X}_T} \xi \, dx \; - \int_{\mathcal{X}_0} \xi \, dx = \int_0^T \int_{\phi_t(\mathcal{X}_0)} \nabla_x \cdot (\xi f) \, dx \, dt, \quad (46)$$

where the initial set $\mathcal{X}_0$ gets mapped to the set $\mathcal{X}_T$ at time $t = T$, under the action of the flow $\phi_t(\cdot)$ associated with the nonlinear dynamics $\dot{x} = f(x)$. The convex relaxation proposed for invalidation/safety verification (Theorem 1 in [76]), strives to construct an artificial "density" $\xi = \xi_r(x, t)$ satisfying three conditions, viz. (i) $\xi_r(x, 0) > 0, \; \forall \, x \in \mathcal{X}_0$, (ii) $\xi_r(x, T) \leqslant 0, \; \forall \, x \in \mathcal{X}_T$, and (iii) $\nabla_x \cdot (\xi_r f) \geqslant 0, \; \forall x \in \phi_t(\mathcal{X}_0), t \in (0, T)$. From (46), such a construction results a "sign-based invalidation", and is only ***sufficient*** *unless a Slater-like condition [78] is satisfied. On the other hand, the "validation in probability" framework proposed in this paper, relies on Liouville PDE-based exact arithmetic computation of ξ, and is a direct simulation-based non-deductive formulation. In this approach, model invalidation equals violation of (46), not just the sign-mismatch of its left-hand and right-hand side, and hence is* ***necessary and sufficient****. As shown in this subsection, for Liouville-integrable nonlinear vector fields (not necessarily semi-algebraic), our framework can recover the deductive falsification inference while bypassing the* ***additional conservatism*** *due to SOS-based computation.*

7 Effect of Initial Uncertainty

The inference for probabilistic model validation depends on the initial PDF $\xi_0(x_0)$. To account robust inference in the presence of initial PDF uncertainty, the notion of PRVC was introduced in Section 5. However, for many applications, it is desirable to characterize the sensitivity of the gap on the choice of initial PDF. We motivate this issue from two different perspectives.

(i) In predictive modeling applications like systems biology, an important problem is of *model discrimination* [79,80], where one looks for an initial PDF that *maximizes* the gap between two models, which seem to exhibit comparable performance. This idea is similar to optimal input design for system identification.

(ii) In general, ${}_2W_2(t) \in [0, \sup \parallel y(t) - \widehat{y}(t) \parallel_2]$, where the supremum is taken over all inter-sample distances between the measured and model-predicted outputs. Thus, ${}_2W_2$ is un-normalized and its absolute magnitude may be difficult to interpret when validating a single model against experimental data. Hence, given a set of admissible initial PDFs, it is important to quantify "*worst-case*" ${}_2W_2(t)$, defined as $\sup\limits_{\xi_0} {}_2W_2(t)$, which could be used for normalization.

The main result of this section is that the initial PDF that maximizes Wasserstein distance, depends on the model and true dynamics. In particular, we show that for a linear dynamics pair, the gap is oblivious beyond the first two moments of ξ_0. We restrict ourselves to scalar dynamics for this analysis.

7.1 Tools for analysis

Definition 3 ***(Quantile function)*** *Consider the probability space $(\Omega_y, \mathscr{F}, \mathbb{P})$ for the output random variable Y. Let $y := Y(\omega_y)$, for $\omega_y \in \Omega_y$. The quantile function $Q_y : \Omega_y \mapsto [0, 1]$, is defined as the generalized inverse of the CDF for Y, i.e.*

$$Q_y(\varsigma) := \inf\left(y \in \Omega_y : \varsigma \leq \mathbb{P}(Y \leq y)\right). \quad (47)$$

Here $\varsigma \in [0, 1]$ denotes probability mass.

Proposition 3 ***(Quantile transport PDE)****[81] Consider the scalar SDE $dx(t) = f(x) \, dt + g(x) \, d\beta$, where β is the standard Wiener process. Then the* ***quantile Fokker-Planck equation*** *(QFPE), given by*

$$\partial_t Q = f(Q, t) - \frac{1}{2} \partial_Q \left(g(Q, t)\right)^2 + \frac{1}{2} \left(g(Q, t)\right)^2 \frac{\partial_{\varsigma\varsigma} Q}{\left(\partial_\varsigma Q\right)^2},$$

describes the transport of quantile function $Q(\varsigma, t)$ for the process $x(t)$.

Proposition 4 ***(Quantile transformation rule)****[82] For an algebraic map $y = h(x)$, we have*

$$Q_y(\varsigma) = \begin{cases} h \circ Q_x(\varsigma) & \textit{if } h(\cdot) \textit{ is non-decreasing}, \\ h \circ Q_x(1 - \varsigma) & \textit{if } h(\cdot) \textit{ is non-increasing}. \end{cases} \quad (48)$$

Next, we work out some specific results by imposing structural assumptions on the true and model dynamics.

7.2 Deterministic linear systems

Let the dynamics of the two systems be

$$\dot{x_i} = a_i x, \quad y_i = c_i x, \ a_i < 0, c_i > 0, \qquad i = 1, 2. \tag{49}$$

Theorem 4 *For any initial density $\xi_0(x_0)$, the Wasserstein gap between the systems in (49), is given by*

$$_2W_2(t) = \sqrt{m_{20}}\left|c_1 e^{a_1 t} - c_2 e^{a_2 t}\right|, \tag{50}$$

where $m_{20} = \mu_0^2 + \sigma_0^2$, is the second raw moment of $\xi_0(x_0)$, while μ_0 and σ_0 are its mean and standard deviation, respectively.

Proof. For (49), $Q_{y_i} = c_i Q_{x_i}$, and the QFPE reduces to a linear PDE $\partial_t Q_{x_i} = a_i Q_{x_i}$, yielding $Q_{x_i}(\varsigma, t) = Q_0(\varsigma) e^{a_i t}$, where Q_0 is the initial quantile function corresponding to ξ_0. Thus, we have

$$\begin{aligned}(_2W_2(t))^2 &= \int_0^1 (Q_{y_1}(\varsigma,t) - Q_{y_2}(\varsigma,t))^2 \, d\varsigma \\ &= \left(c_1 e^{a_1 t} - c_2 e^{a_2 t}\right)^2 \int_0^1 (Q_0(\varsigma))^2 \, d\varsigma. \end{aligned}\tag{51}$$

Since the quantile function maps probability to the sample space, hence $x_0 = Q_0(\varsigma)$, and $d\varsigma = \xi_0(x_0)\,dx_0$. Consequently, we can rewrite (51) as

$$(_2W_2(t))^2 = \left(c_1 e^{a_1 t} - c_2 e^{a_2 t}\right)^2 \underbrace{\int_{-\infty}^{\infty} x_0^2 \, \xi_0(x_0) \, dx_0}_{m_{20}}.$$

Taking square root to both sides, we obtain the result. It's straightforward to check that $m_{20} = \mu_0^2 + \sigma_0^2$, relating the central moments with m_{20}. ■

Remark 7 *(**$_2W_2$ has limited dependence on ξ_0**) The above result shows that the Wasserstein gap between scalar linear systems, depends on the initial density up to mean and variance. Any other aspect (skewness, kurtosis etc.) of ξ_0, even when it's non-Gaussian, has no effect on $_2W_2(t)$. The next example demonstrates that our result: "the initial PDF with maximum second raw moment, maximizes Wasserstein distance" (Fig. 12), may be counterintuitive in some situations.*

Example 6 *(**Uniform initial PDF may not maximize $_2W_2$**) For (49), let the set of admissible initial PDFs be $S_0 := \{\xi_0 : \text{supp}(\xi_0) = [a,b], \xi_0(x_0) = \frac{1}{(b-a)^{\alpha+\beta-1}B(\alpha,\beta)}(x_0 - a)^{\alpha-1}(b - x_0)^{\beta-1}, \alpha > 0, \beta > 0\}$, i.e. the set of all scaled beta PDFs supported on $[a,b]$. One can readily compute that $\mu_0 = \frac{\alpha b + \beta a}{\alpha+\beta}$, and $\sigma_0^2 = \frac{\alpha\beta(b-a)^2}{(\alpha+\beta)^2(\alpha+\beta+1)}$. For $\alpha = \beta = 1$, $\xi_0 = \mathcal{U}([a,b])$, and for $\alpha = \beta = \frac{1}{2}$, $\xi_0 = \mathcal{A}([a,b])$. Thus, we have*

$$m_{20}(\mathcal{U}[a,b]) = \frac{1}{3}\left(a^2 + b^2 + ab\right), \tag{52}$$

$$m_{20}(\mathcal{A}[a,b]) = \frac{1}{8}\left(3a^2 + 3b^2 + 2ab\right), \tag{53}$$

and hence $m_{20}(\mathcal{A}[a,b]) > m_{20}(\mathcal{U}[a,b])$, $\forall\, b > a$. From Theorem 4, $_2W_2(t)$ trajectory for uniform initial PDF, stays below the same for arcsine initial PDF, as shown in Fig. 13.

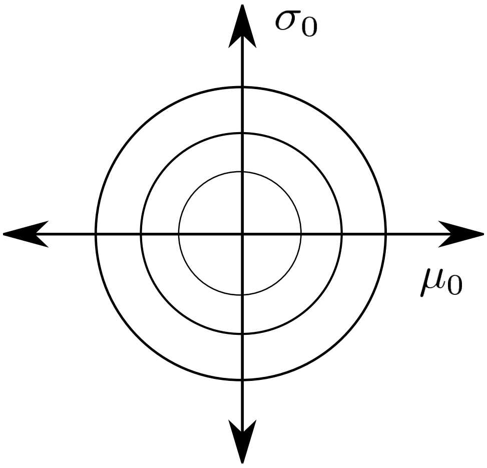


Fig. 12. The results of Section 4 can be summarized through a graphical algorithm illustrated above. For scalar linear systems, given a set of admissible initial PDFs over state space, we construct concentric circles centered at origin, over the two dimensional (μ_0, σ_0) subspace of the (infinite-dimensional) moment space. From (50), ξ_0 corresponding to the circle with largest radius, maximizes $_2W_2(t)$, $\forall t > 0$. For affine systems, (54) implies a similar construction in (μ_0, σ_0) subspace, with circles centered at $\left(-\frac{q(t)}{p(t)}, 0\right)$. The direction of this translation along μ_0 axis, depends on parameters (a_i, b_i, c_i, d_i), $i = 1, 2$, of the systems under comparison.

Remark 8 *(**Discrete-time linear systems**) Consider the true and model maps $x_i^{(k+1)} = a_i x_i^{(k)}$, $y_i^{(k)} = c_i x_i^{(k)}$, $i = 1, 2$, where $k \in \mathbb{N} \cup \{0\}$, denotes the discrete time index. From linear recursion, one can obtain a result similar to (50): $W(k) = \sqrt{m_{20}}\left|c_1 a_1^k - c_2 a_2^k\right|$.*

Remark 9 *(**Linear Gaussian systems**) For the linear Gaussian case, one can verify (50) without resorting to the QFPE. To see this, notice that if $\xi_0(x_0) = \mathcal{N}\left(\mu_0, \sigma_0^2\right)$, then the state PDFs evolve as $\xi_{x_i}(x_i, t) = \mathcal{N}\left(\mu_{x_i}(t), \sigma_{x_i}^2(t)\right)$, where $\mu_{x_i}(t)$ and $\sigma_{x_i}^2(t)$ satisfy their respective state and Lyapunov equations, which, in the scalar case, can be solved in closed form. Since $\eta_{y_i}(y_i, t) = \mathcal{N}\left(c_i\mu_{x_i}(t), c_i^2\sigma_{x_i}^2(t)\right)$, and $_2W_2$ between two Gaussian PDFs is known [47] to be $\sqrt{\left(\mu_{y_1} - \mu_{y_2}\right)^2 + \left(\sigma_{y_1} - \sigma_{y_2}\right)^2}$, the result follows.*

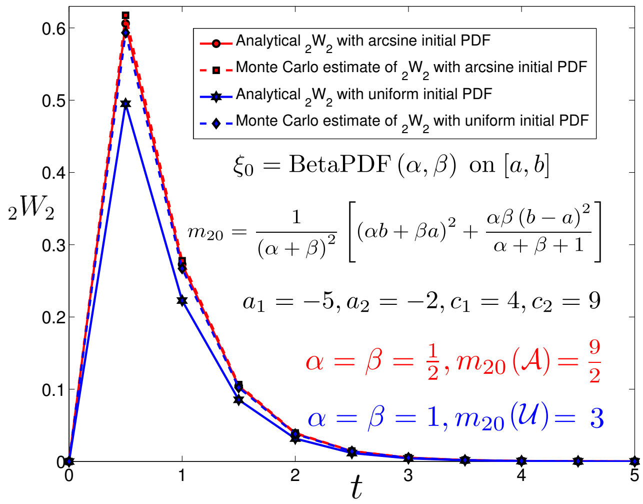


Fig. 13. Wasserstein time histories between linear system pair (49) with ξ_0 as $\mathcal{A}([a,b])$ and $\mathcal{U}([a,b])$, respectively. Here $a=-3$, $b=3$, and we set sampling interval $\Delta t_k = 0.5$. We observe that the Wasserstein gap for $\xi_0 = \mathcal{A}([a,b])$ remains above the same for $\xi_0 = \mathcal{U}([a,b])$, as predicted by Theorem 4. The *solid lines* are direct computation from (50), while the *dashed lines* are Monte Carlo estimates of ${}_2W_2$ using (15).

Remark 10 *(**Affine dynamics**) Instead of (49), if the dynamics are given by $\dot{x}_i = a_i x + b_i$, $y_i = c_i x + d_i$, $i = 1,2$, then by variable substitution, one can derive that $Q_{x_i}(\varsigma,t) = Q_0(\varsigma)e^{a_i t} + \dfrac{b_i}{a_i}\left(e^{a_i t}-1\right)$. Hence, we get*

$$ {}_2W_2(t) = \sqrt{(p(t))^2 m_{20} + 2p(t)\,q(t)\,m_{10} + (q(t))^2}, \tag{54}$$

where $m_{10} = \mu_0$, $p(t) := (c_1 e^{a_1 t} - c_2 e^{a_2 t})$, and $q(t) := \dfrac{b_1 c_1}{a_1}\left(e^{a_1 t}-1\right) - \dfrac{b_2 c_2}{a_2}\left(e^{a_2 t}-1\right) + (d_1 - d_2)$.

7.3 Stochastic linear systems

Consider two stochastic dynamical systems with linear drift and constant diffusion coefficients, given by

$$dx_i = a_i x\, dt + b_i\, d\beta, \quad y_i = c_i x, \qquad i = 1,2, \tag{55}$$

where β is the standard Wiener process.

Theorem 5 *For any initial density $\xi_0(x_0)$, the Wasserstein gap ${}_2W_2(t)$ between the systems in (55), is given by*

$$ {}_2W_2(t) = \sqrt{(p(t))^2 m_{20} + 2p(t)\,r(t)\,s(F_0) + (r(t))^2} \tag{56}$$

where $r(t) := \dfrac{|b_1|c_1}{\sqrt{2a_1}}\sqrt{e^{2a_1 t}-1} - \dfrac{|b_2|c_2}{\sqrt{2a_2}}\sqrt{e^{2a_2 t}-1}$, and $s(F_0) := \sqrt{2}\,\mathbb{E}\left[x_0\, \text{erf}^{-1}(2F_0(x_0)-1)\right]$, F_0 being the CDF of x_0.

Proof. For systems (55), quantile functions for the states evolve as (p. 102, [81])

$$Q_{x_i}(\varsigma,t) = Q_0(\varsigma)e^{a_i t} + |b_i| Q_N(\varsigma)\sqrt{\frac{e^{2a_i t}-1}{2a_i}}, \tag{57}$$

where $Q_N(\varsigma) := \sqrt{2}\,\text{erf}^{-1}(2\varsigma - 1)$, is the standard normal quantile. Thus, the Wasserstein distance becomes

$$\begin{aligned}
\left({}_2W_2(t)\right)^2 &= \int_0^1 \left(c_1 Q_{x_1}(\varsigma,t) - c_2 Q_{x_2}(\varsigma,t)\right)^2 d\varsigma \\
&= (p(t))^2 \int_0^1 (Q_0(\varsigma))^2\, d\varsigma \\
&\quad + 2p(t)\,r(t)\int_0^1 Q_0(\varsigma)\,Q_N(\varsigma)\, d\varsigma \\
&\quad + (r(t))^2 \int_0^1 (Q_N(\varsigma))^2\, d\varsigma.
\end{aligned} \tag{58}$$

Notice that the first and third integrals are m_{20} and 1, respectively. Since $\varsigma = F_0(x_0)$, the second integral becomes

$$\begin{aligned}
&\int_{-\infty}^{\infty} x_0\, F_N^{-1}\circ F_0(x_0)\ \rho_0(x_0)\, dx_0 \\
&= \sqrt{2}\,\mathbb{E}\left[x_0\, \text{erf}^{-1}(2F_0(x_0)-1)\right] = s(F_0).
\end{aligned} \tag{59}$$

This completes the proof. ∎

Remark 11 *(**Gaussian case**) Consider the special case when $\xi_0(x_0) = \mathcal{N}(\mu_0,\sigma_0^2)$. Then $Q_0(\varsigma) = \mu_0 + \sigma_0 Q_N(\varsigma)$, and hence the second integral equals σ_0. Thus, if the initial density is normal, then*

$$ {}_2W_2(t) = \sqrt{(p(t))^2 m_{20} + 2p(t)\,r(t)\,\sigma_0 + (r(t))^2}, \tag{60}$$

a function of μ_0 and σ_0, which can be verified otherwise by solving the mean and variance propagation equations.

8 Upper Bounds for ${}_2W_2$ for Discrete-time Linear Gaussian Systems

The objective of this Section is to derive an upper bound of Wasserstein gap for discrete-time linear systems with $\xi_0(x_0) = \mathcal{N}(0,P_0)$, in terms of the spectrums of the systems under comparison, and initial covariance. This is done by relating ${}_2W_2(k)$ with $D_{KL}(k)$, thus providing an *offline* estimate of the Wasserstein gap. We only provide LTI result below; extension for the LTV case is reported in [2].

Theorem 6 *Consider two discrete-time stable LTI systems $x_{k+1} = Ax_k$, and $\widehat{x}_{k+1} = \widehat{A}\widehat{x}_k$, $k \in \mathbb{N}\cup\{0\}$. Let* *the initial PDF* $\xi_0(x_0) = \mathcal{N}(0, P_0)$. *Then,* ${}_2W_2(k) \leqslant \sqrt{2}\,(tr(P_0))^{1/2}\,\|\widehat{A}^{-k}\|_F\,\Omega_{LTI}(k)$, *where*

$$\Omega_{LTI}(k) \triangleq \left(\|A^k\|_F^2\|\widehat{A}^{-k}\|_F^2\,(tr(P_0))^2 - \log\left(\prod_{i=1}^{n_s}\frac{\vartheta_i^{2k}}{\widehat{\vartheta}_i^{2k}}\right) - n_s\right)^{\frac{1}{2}}$$

where the spectrum for A *is* $\{\vartheta_i\}_{i=1}^{n_s}$, *and for* $\widehat{A}$ *is* $\{\widehat{\vartheta}_i\}_{i=1}^{n_s}$.

Proof We know that

$$\xi_k = \mathcal{N}\left(0, A^k P_0 A^{k^\top}\right) = \mathcal{N}(0, P_k),$$
$$\widehat{\xi}_k = \mathcal{N}\left(0, \widehat{A}^k P_0 \widehat{A}^{k^\top}\right) = \mathcal{N}\left(0, \widehat{P}_k\right). \tag{61}$$

Therefore,

$$D_{KL}\left(\xi_k\|\widehat{\xi}_k\right) = D_{KL}\left(P_k\|\widehat{P}_k\right)$$
$$= \mathrm{tr}\left(\widehat{P}_k^{-1}P_k - I\right) - \log\det\left(\widehat{P}_k^{-1}P_k\right). \tag{62}$$

Now if we assume that the spectrum for P_0 is $\{\rho_i\}_{i=1}^{n_s}$, then from (62), $\det(P_k) = \prod_{i=1}^{n_s}\left(\rho_i\vartheta_i^{2k}\right) \Rightarrow \log\det\left(\widehat{P}_k^{-1}P_k\right) = \log\prod_{i=1}^{n_s}\frac{\vartheta_i^{2k}}{\widehat{\vartheta}_i^{2k}}$. Thus, $D_{KL}\left(P_k\|\widehat{P}_k\right) = \mathrm{tr}\left(\widehat{P}_k^{-1}P_k\right) - \log\prod_{i=1}^{n_s}\frac{\vartheta_i^{2k}}{\widehat{\vartheta}_i^{2k}} - n_s$.

Now, observe that $\mathrm{tr}\left(\widehat{P}_k^{-1}P_k\right) \leqslant \mathrm{tr}\left(\widehat{P}_k^{-1}\right)\mathrm{tr}(P_k)$, since covariance matrices are symmetric positive semi-definite. However, $\mathrm{tr}(P_k) = \mathrm{tr}\left(A^kP_0A^{k^\top}\right) = \mathrm{tr}\left(A^{k^\top}A^kP_0\right) \leqslant \mathrm{tr}\left(A^{k^\top}A^k\right)\mathrm{tr}(P_0) = \|A^k\|_F^2\,\mathrm{tr}(P_0)$; where we have used the fact that trace of a matrix product is invariant under cyclic permutation of the matrices. Likewise, $\mathrm{tr}\left(\widehat{P}_k^{-1}\right) \leqslant \|\widehat{A}^{-k}\|_F^2\,\mathrm{tr}(P_0)$. Combining these results, we get

$$D_{KL}\left(P_k\|\widehat{P}_k\right) \leqslant \underbrace{\|A^k\|_F^2\|\widehat{A}^{-k}\|_F^2\,(\mathrm{tr}(P_0))^2 - \log\prod_{i=1}^{n_s}\frac{\vartheta_i^{2k}}{\widehat{\vartheta}_i^{2k}} - n_s}_{(\Omega_{\mathrm{LTI}}(k))^2}.$$

Now to relate D_{KL} with ${}_2W_2$, we invoke the TCI for Gaussian case [59], which states ${}_2W_2(k) \leqslant \sqrt{2\lambda_{\max}\left(\widehat{P}_k^{-1}\right)D_{KL}(k)}$. But $\lambda_{\max}\left(\widehat{P}_k^{-1}\right) \leqslant \mathrm{tr}\left(\widehat{P}_k^{-1}\right) \leqslant \|\widehat{A}^{-k}\|_F^2\mathrm{tr}(P_0)$. These two, coupled with TCI, results

$${}_2W_2(k) \leqslant \sqrt{2}\,(\mathrm{tr}(P_0))^{1/2}\,\|\widehat{A}^{-k}\|_F\,\Omega_{\mathrm{LTI}}(k). \tag{63}$$

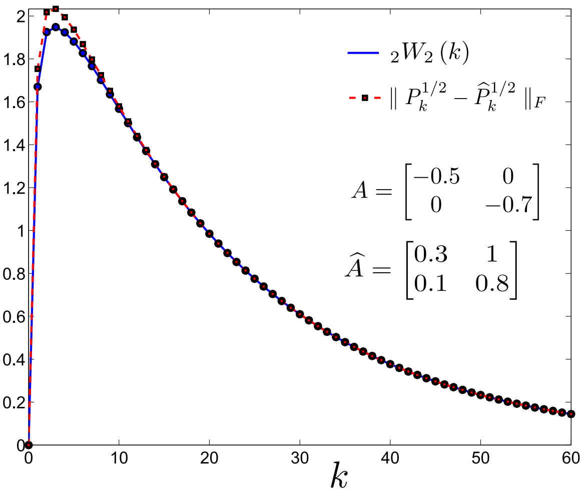


Fig. 14. Starting with $\mathcal{N}(0, P_0)$, time histories for ${}_2W_2(k)$ and its upper bound (64) for two discrete-time LTI systems, with A and $\widehat{A}$ as shown. Since the systems are stable, both ${}_2W_2(k)$ and $\| P_k^{1/2} - \widehat{P}_k^{1/2} \|_F$ asymptotically approach zero.

Remark 12 ***(A sharper upper bound)*** *Instead of relating* ${}_2W_2$ *with the spectrum of the LTI systems, one can obtain a sharper bound (see Appendix F for proof):*

$${}_2W_2(k) \leq \| P_k^{1/2} - \widehat{P}_k^{1/2} \|_F, \tag{64}$$

where $P_k = A^kP_0A^{k^\top}$, $\widehat{P}_k = \widehat{A}^kP_0\widehat{A}^{k^\top}$*; the equality holds when* P_k *and* $\widehat{P}_k$ *commute, resulting an interesting Lie bracket condition on system matrices:* $\left[A^kP_0A^{k^\top}, \widehat{A}^kP_0\widehat{A}^{k^\top}\right] = 0$. *For two Schur-Cohn stable matrices* A *and* $\widehat{A}$, *Fig. 14 illustrates (64) with* $P_0 = \begin{bmatrix}1 & 0\\ 0 & 3\end{bmatrix}$.

9 Conclusions

We have presented a probabilistic model validation framework for nonlinear systems. The notion of soft validation allows us to quantify the degree of mismatch of a proposed model with respect to experimental measurements, thereby guiding for model refinement. A key contribution of this paper is to introduce transport-theoretic Wasserstein distance as a validation metric to measure the difference between distributional shapes over model-predicted and experimentally observed output spaces. The framework presented here applies to any deterministic or stochastic nonlinearity, not necessarily semialgebraic type. In addition to providing computational guarantees for probabilistic inference, we also recover existing nonlinear invalidation results in the literature. Novel results are given for discriminating linear models.

Acknowledgements

This research was supported by NSF award #1016299 with Helen Gill as the program manager. The authors would like to thank P. Khargonekar at University of Florida, and S. Chakravorty at Texas A&M University, for insightful discussions.

References

[1] A. Halder, and R. Bhattacharya, "Model Validation: A Probabilistic Formulation". *IEEE Conference on Decision and Control*, Orlando, Florida, 2011.

[2] A. Halder, and R. Bhattacharya, "Further Results on Probabilistic Model Validation in Wasserstein Metric". *IEEE Conference on Decision and Control*, Maui, Hawaii, 2012.

[3] K. Popper, *Conjectures and Refutations: The Growth of Scientific Knowledge.* Routledge, Second Ed., 2002.

[4] R.S. Smith, and J.C. Doyle, "Model Validation: A Connection Between Robust Control and Identification". *IEEE Transactions on Automatic Control*, Vol. 37, No. 7, pp. 942–952, 1992.

[5] K. Poolla, P. Khargonekar, A. Tikku, J. Krause, and K. Nagpal, "A Time-domain Approach to Model Validation". *IEEE Transactions on Automatic Control*, Vol. 39, No. 5, pp. 951–959, 1994.

[6] S. Prajna, "Barrier Certificates for Nonlinear Model Validation". *Automatica*, Vol. 42, No. 1, pp. 117–126, 2006.

[7] P.B. Brugarolas, and M.G. Safonov, "A Data Driven Approach to Learning Dynamical Systems". *IEEE Conference on Decision and Control*, Las Vegas, Nevada, 2002.

[8] C. Baier, and J.P. Katoen, *Principles of Model Checking.* The MIT Press, First ed., 2008.

[9] F. Ciesinski, and M. Größer, "On Probabilistic Computation Tree Logic". *Validation of Stochastic Systems*, Springer, Eds. Baier, C., Haverkort, B.R., Hermanns, H., Katoen, J.P., and Siegle, M., Lecture Notes in Computer Science 2925, pp. 147–188, 2004.

[10] R. Smith, G.E. Dullerud. "Continuous-time Control Model Validation using Finite Experimental Data". *IEEE Transactions on Automatic Control*, Vol. 41, No. 8, pp. 1094–1105, 1996.

[11] J. Chen, and S. Wang, "Validation of Linear Fractional Uncertain Models: Solutions via Matrix Inequalities". *IEEE Transactions on Automatic Control*, Vol. 41, No. 6, pp. 844–849, 1996.

[12] B. Wahlberg, and L. Ljung, "Hard Frequency-domain Model Error Bounds from Least-squares Like Identification Techniques". *IEEE Transactions on Automatic Control*, Vol. 37, No. 7, pp. 900–912, 1992.

[13] D. Xu, Z. Ren, G. Gu, and J. Chen, "LFT Uncertain Model Validation with Time and Frequency Domain Measurements". *IEEE Transactions on Automatic Control*, Vol. 44, No. 7, pp. 1435–1441, 1999.

[14] S.L. Campbell, *Singular Systems of Differential Equations.* Pitman, First ed., 1980.

[15] A. Megretski, and A. Rantzer, "System Analysis via Integral Quadratic Constraints". *IEEE Transactions on Automatic Control*, Vol. 42, No. 6, pp. 819–830, 1997.

[16] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications.* Springer, Sixth ed., 2003.

[17] A.J. van der Schaft, and H. Schumacher, *An Introduction to Hybrid Dynamical Systems.* Springer, LNCS 251, First ed., 1999.

[18] L. Ljung, and L. Guo, "The Role of Model Validation for Assessing the Size of the Unmodeled Dynamics". *IEEE Transactions on Automatic Control*, Vol. 42, No. 9, pp. 1230–1239, 1997.

[19] L. Ljung, *System Identification: Theory for the User.* Printice-Hall Inc., Second ed., 1999.

[20] R.G. Ghanem, A. Doostan, and J. Red-Horse, "A Probabilistic Construction of Model Validation". *Computer Methods in Applied Mechanics and Engineering*, Vol. 197, No. 29-32, pp. 2585–2595, 2008.

[21] M. Gevers, X. Bombois, B. Codrons, G. Scorletti, and B.D.O. Anderson, "Model Validation for Control and Controller Validation in A Prediction Error Identification Framework–Part I: Theory". *Automatica*, Vol. 39, No. 3, pp. 403–415, 2003.

[22] L.H. Lee, and K. Poolla, "On Statistical Model Validation". *Journal of Dynamic Systems, Mesurement, and Control*, Vol. 118, No. 2, pp. 226–236, 1996.

[23] J. van Schuppen, "Stochastic Realization Problems". *Three Decades of Mathematical System Theory: A Collection of Surveys at the Occasion of the 50th Birthday of Jan C. Willems*, Lecture Notes in Control and Information Sciences, Springer, Vol. 135, pp. 480–523, 1989.

[24] V.A. Ugrinovskii, "Risk-sensitivity Conditions for Stochastic Uncertain Model Validation". *Automatica*, Vol. 45, No. 11, pp. 2651–2658, 2009.

[25] P.A. Parrilo, *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*, PhD thesis, California Institute of Technology, Pasadena, CA, 2000.

[26] A. Lasota, and M. Mackey, *Chaos, Fractals and Noise: Stochastic Aspects of Dynamics.* Applied Mathematical Sciences, Vol. 97, Springer-Verlag, NY, Second ed., 1994.

[27] Y. Sun, and P.G. Mehta, "The Kullback-Leibler Rate Pseudo-Metric for Comparing Dynamical Systems". *IEEE Transactions on Automatic Control*, Vol. 55, No. 7, pp. 1585–1598, 2010.

[28] T.T. Georgiou, "Distances and Riemannian Metrics for Spectral Density Functions". *IEEE Transactions on Signal Processing*, Vol. 55, No. 8, pp. 3995–4003, 2007.

[29] S. Meyn, and R.L. Tweedie, *Markov Chains and Stochastic Stability.* Cambridge University Press, Second ed., 2009.

[30] A. Papoulis, *Random Variables and Stochastic Processes.* McGraw-Hill, NY, Second ed., 1984.

[31] A. Halder, and R. Bhattacharya, "Dispersion Analysis in Hypersonic Flight During Planetary Entry Using Stochastic Liouville Equation". *Journal of Guidance, Control, and Dynamics*, Vol. 34, No. 2, 2011.

[32] A. Halder, and R. Bhattacharya, "Beyond Monte Carlo: A Computational Framework for Uncertainty Propagation in Planetary Entry, Descent and Landing". *AIAA Guidance, Navigation and Control Conference*, Toronto, 2010.

[33] C.S. Hsu, *Cell-to-Cell Mapping: A Method of Global Analysis for Nonlinear Systems*, Applied Mathematical Sciences, Vol. 64, Springer-Verlag, NY; 1987.

[34] H. Risken, *The Fokker-Planck Equation: Methods of Solution and Applications*, Springer-Verlag, NY; 1989.

[35] M. Kumar, S. Chakravorty, and J.L. Junkins, "A Semianalytic Meshless Approach to the Transient Fokker-Planck Equation". *Probabilistic Engineering Mechanics*, Vol. 25, No. 3, pp. 323–331, 2010.

[36] P. Dutta, A. Halder, and R. Bhattacharya, "Uncertainty Quantification for Stochastic Nonlinear Systems using Perron-Frobenius Operator and Karhunen-Loève Expansion". *IEEE Multi-Conference on Systems and Control*, Dubrovnik, Croatia, 2012.

[37] A. Edelman, and B.D. Sutton, "From Random Matrices to Stochastic Operators", *Journal of Statistical Physics*, Vol. 127, No. 6, 2007, pp. 1121–1165.

[38] A.L. Gibbs, and F.E. Su, "On Choosing and Bounding Probability Metrics". *International Statistical Review*, Vol. 70, No. 3, pp. 419–435, 2002.

[39] I. Csiszár, "Information-type Measures of Difference of Probability Distributions and Indirect Observations", *Studia Scientiarium Mathematicarum Hungarica*, Vol. 2, 1967, pp. 299–318.

[40] A. Müller, "Integral Probability Metrics and Their Generating Classes of Functions", *Advances in Applied Probability*, Vol. 29, 1997, pp. 429–443.

[41] C. Villani, *Topics in Optimal Transportation.* American Mathematical Society, First ed., 2003.

[42] R. Jordan, D. Kinderlehrer, and F. Otto, "The Variational Formulation of the Fokker-Planck Equation". *SIAM Journal of Mathematical Analysis*, Vol. 29, No. 1, pp. 1–17, 1998.

[43] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models.* John Wiley, First ed., 1991.

[44] Q. Wang, S.R. Kulkarni, S. Verdú, "Divergence Estimation of Continuous Distributions Based on Data-dependent Partitions", *IEEE Transactions on Information Theory*, Vol. 51, No. 9, 2005, pp. 3064–3074.

[45] X. Nguyen, M.J. Wainwright, M.I. Jordan, "Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization", *IEEE Transactions on Information Theory*, Vol. 56, No. 11, 2010, pp. 5847–5861.

[46] A.V. Lazo, and P. Rathie, "On the Entropy of Continuous Probability Distributions". *IEEE Transactions on Information Theory*, Vol. 24, No. 1, pp. 120–122, 1978.

[47] C.R. Givens, and R.M. Shortt, "A Class of Wasserstein Metrics for Probability Distributions". *Michigan Mathematical Journal*, Vol. 31, No. 2, pp. 231–240, 1984.

[48] R. Kullhavý, *Recursive Nonlinear Estimation: A Geometric Approach.* Lecture Notes in Control and Information Sciences, Vol. 216, Springer-Verlag, 1996.

[49] A. Poznyak, *Advanced Mathematical Tools for Automatic Control Engineers.* Vol. 1: Deterministic Techniques, Elsevier Science, 2008.

[50] B.W. Hong, S. Soatto, K. Ni, and T. Chan, "The Scale of A Texture and Its application to Segmentation". *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, 2008.

[51] K. Ni, X. Bresson, T. Chan, and S. Esedoglu, "Local Histogram Based Segmentation Using the Wasserstein Distance". *International Journal of Computer Vision*, Vol. 84, No. 1, pp. 97–111, 2009.

[52] S.S. Vallander, "Calculation of the Wasserstein Distance between Distributions on the Line". *Theory of Probability and Its Applications*, Vol. 18, pp. 784–786, 1973.

[53] S.T. Rachev, "The Monge–Kantorovich Mass Transference Problem and Its Stochastic Applications". *Theory of Probability and its Applications*, Vol. 29, pp. 647–676, 1985.

[54] F. Hitchcock, "The Distribution of a Product from Several Sources to Numerous Localities". *Journal of Mathematics and Physics*, Vol. 20, No. 2, pp. 224–230, 1941.

[55] T.C. Koopmans, "Optimum Utilization of the Transportation System". *Econometrica: Journal of the Econometric Society*, Vol. 17, pp. 136–146, 1949.

[56] T.C. Koopmans, "Efficient Allocation of Resources". *Econometrica: Journal of the Econometric Society*, Vol. 19, No. 4, pp. 455–465, 1951.

[57] B.K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G.R.G. Lanckriet, "On Integral Probability Metrics, ϕ-Divergences and Binary Classification". *Preprint*, arXiv:0901.2698v4, Available at `http://arxiv.org/abs/0901.2698v4`, 2009.

[58] M. Talagrand, "Transportation Cost for Gaussian and Other Product Measures". *Geometric and Functional Analysis*, Vol. 6, No. 3, pp. 587–600, 1996.

[59] H. Djellout, A. Guillin, and L. Wu, "Transportation Cost-Information Inequalities and Applications to Random Dynamical Systems and Diffusions". *The Annals of Probability*, Vol. 32, No. 3B, pp. 2702–2732, 2004.

[60] E. Boissard, and T. le Gouic, ""Exact" Deviations in Wasserstein Distance for Empirical and Occupation Measures". *Preprint*, arXiv:1103.3188v1, Available at `http://arxiv.org/abs/1103.3188v1`, 2011.

[61] R.E. Burkard, M. Dell'Amico, and S. Martello, *Assignment Problems*, SIAM, PA; 2009.

[62] R. Julien, G. Peyré, J. Delon, and B. Marc, "Wasserstein Barycenter and its Application to Texture Mixing", *Preprint*, available at `http://hal.archives-ouvertes.fr/hal-00476064/fr/`, 2010.

[63] R. Tempo, G. Calafiore, and F. Dabbene, *Randomized Algorithms for Analysis and Control of Uncertain Systems*, Springer-Verlag, First Ed., 2004.

[64] P. Khargonekar, and A. Tikku, "Randomized Algorithms for Robust Control Analysis and Synthesis Have Polynomial Complexity". *IEEE Conference on Decision and Control*, Kobe, Japan, Dec. 11–13, 1996.

[65] R. Tempo, E.-W. Bai, and F. Dabbene, "Probabilistic Robustness Analysis: Explicit Bounds for the Minimum Number of Samples". *Systems & Control Letters*, Vol. 30, pp. 237–242, 1997.

[66] X. Chen, and K. Zhou, "Order Statistics and Probabilistic Robust Control". *Systems & Control Letters*, Vol. 35, pp. 175–182, 1998.

[67] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods.* CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, 1992.

[68] D. Liberzon, and R.W. Brockett, "Nonlinear Feedback Systems Perturbed by Noise: Steady-state Probability Distributions and Optimal Control". *IEEE Transactions on Automatic Control*, Vol. 45, No. 6, pp. 1116–1130, 2000.

[69] M. Vidyasagar, "Randomized Algorithms for Robust Controller Synthesis using Statistical Learning Theory". *Automatica*, Vol. 37, No. 10, pp. 1515–1528, 2001.

[70] T. Geisel, and V. Fairen, "Statistical Properties of Chaos in Chebyshev Maps". *Physics Letters A*, Vol. 105, No. 6, pp. 263–266, 1984.

[71] M. Mackey, and M. Tyran-Kamińska, "Deterministic Brownian Motion: The Effects of Perturbing a Dynamical System by A Chaotic Semi-dynamical System". *Physics reports*, Vol. 422, No. 5, pp. 167–222, 2006.

[72] K.B. Athreya, and J. Dai, "Random Logistic Maps. I". *Journal of Theoretical Probability*, Vol. 13, No. 2, pp. 595–608, 2000.

[73] F.C. Klebaner, "Population and Density Dependent Branching Processes". In K.B. Athreya, and P. Jagers (eds.), *Classical and Modern Branching Processes*, Vol. 84, IMA, Springer-Verlag, 1997.

[74] S. Prajna, A. Papachristodoulou, and P.A. Parrilo, "Introducing SOSTOOLS: A General Purpose Sum of Squares Programming Solver". *IEEE Conference on Decision and Control*, 2002.

[75] A. Rantzer, "A Dual to Lyapunov's Stability Theorem." *Systems & Control Letters*, Vol. 42, No. 3 ,pp. 161–168, 2001.

[76] A. Rantzer, and S. Prajna, "On Analysis and Synthesis of Safe Control Laws". *Proceedings of the Allerton Conference on Communication, Control, and Computing*, 2004.

[77] S. Prajna, and A. Rantzer, "Convex Programs for Temporal Verification of Nonlinear Dynamical Systems". *SIAM Journal on Control and Optimization*, Vol. 46, No. 3, pp. 999–1021, 2007.

[78] S. Prajna, and A. Rantzer, "On the Necessity of Barrier Certificates". *Proceedings of the IFAC World Congress*, 2005.

[79] D. Georgiev, and E. Klavins, "Model Discrimination of Polynomial Systems via Stochastic Inputs". *47th IEEE Conference on Decision and Control*, 2008.

[80] A. Kremling, S. Fischer, K. Gadkar, F.J. Doyle, T. Sauter, E. Bullinger, F. Allgöwer, and E.D. Gilles, "A Benchmark for Methods in Reverse Engineering and Model Discrimination: Problem Formulation and Solutions". *Genome Research*, Vol. 14, No. 9, pp. 1773–1785, 2004.

[81] G. Steinbrecher, and W.T. Shaw, "Quantile Mechanics", *European Journal of Applied Mathematics*, Vol. 19, No. 2, pp. 87–112, 2008.

[82] W.G. Gilchrist, *Statistical Modeling with Quantile Functions*, CRC Press; 2000.

[83] R. Motwani, and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, NY; 1995.

[84] V.I. Klyatskin, *Dynamics of Stochastic Systems.* Translated from Russian by A. Vinogradov, First Ed., Elsevier, 2005.

[85] D.S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas*, Second ed., Princeton University Press; 2009.

A Computing ${}_2W_2\left(\eta_b\left(x;\alpha,\beta\right),\eta_b\left(x;\beta,\alpha\right)\right)$

We denote $I_t^{-1}\left(\alpha,\beta\right)$ as the inverse of the beta CDF, $I_x\left(\alpha,\beta\right) := \dfrac{B\left(x;\alpha,\beta\right)}{B\left(\alpha,\beta\right)}$ as the regularized incomplete beta function, and $B\left(x;\alpha,\beta\right) := \displaystyle\int_0^x z^{\alpha-1}\left(1-z\right)^{\beta-1}dz$ as the incomplete beta function.

Theorem 7 ${}_2W_2\left(\eta_b\left(x;\alpha,\beta\right),\eta_b\left(x;\beta,\alpha\right)\right) = \sqrt{\dfrac{\alpha\left(\alpha+1\right)+\beta\left(\beta+1\right)}{\left(\alpha+\beta\right)\left(\alpha+\beta+1\right)} - 2\left(\dfrac{\beta}{\alpha+\beta}-\mathcal{J}\right)}$,
$\mathcal{J} := \dfrac{1}{\beta+1}\displaystyle\int_0^1 \left(I_t^{-1}\left(\alpha,\beta\right)\right)^{1-\alpha}\left(1-I_t^{-1}\left(\alpha,\beta\right)\right)^{1-\beta}$
$\left(I_t^{-1}\left(\beta,\alpha\right)\right)^{\beta+1}\ {}_2F_1\left(\beta+1,1-\alpha;\beta+2;I_t^{-1}\left(\beta,\alpha\right)\right)\,dt.$

Proof. From (15), we have

$$\begin{aligned}&{}_2W_2^2\left(f_b\left(x;\alpha,\beta\right),f_b\left(x;\beta,\alpha\right)\right)\\&=\int_0^1\left(I_t^{-1}\left(\alpha,\beta\right)-I_t^{-1}\left(\beta,\alpha\right)\right)^2\,dt.\end{aligned} \tag{A.1}$$

The following identities, stated without proof, will be useful for the evaluation of (A.1).

Property 1

$$\int I_t^{-1}\left(a,b\right)\,dt = \frac{1}{\left(a+1\right)B\left(a,b\right)}\left(I_t^{-1}\left(a,b\right)\right)^{a+1}\ {}_2F_1\left(a+1,1-b;a+2;I_t^{-1}\left(a,b\right)\right) + \text{ constant.}$$

Property 2

$$\int \left(I_t^{-1}\left(a,b\right)\right)^2\,dt = \frac{1}{\left(a+1\right)B\left(a,b\right)}\left(I_t^{-1}\left(a,b\right)\right)^{a+1}\left(\,{}_2F_1\left(a+1,1-b;a+2;I_t^{-1}\left(a,b\right)\right) - {}_2F_1\left(a+1,-b;a+2;I_t^{-1}\left(a,b\right)\right)\right) + \text{ constant.}$$

Property 3
$I_0^{-1}\left(a,b\right)=0$, and $I_1^{-1}\left(a,b\right)=1$.

Property 4 *(Gauss Theorem)*
${}_2F_1\left(A,B;C;1\right) = \dfrac{\Gamma\left(C\right)\Gamma\left(C-A-B\right)}{\Gamma\left(C-A\right)\Gamma\left(C-B\right)}$.

Property 5
$\dfrac{d}{dt}I_t^{-1}\left(a,b\right) = B\left(a,b\right)\left(I_t^{-1}\left(a,b\right)\right)^{1-a}\left(1-I_t^{-1}\left(a,b\right)\right)^{1-b}$.

Using Properties 2 and 3, we get

$$\int_0^1\left(I_t^{-1}\left(\alpha,\beta\right)\right)^2\,dt = \frac{1}{\left(\alpha+1\right)B\left(\alpha,\beta\right)}\left[\,{}_2F_1\left(\alpha+1,1-\beta;\alpha+2;1\right) - {}_2F_1\left(\alpha+1,-\beta;\alpha+2;1\right)\right]. \tag{A.2}$$

Recalling that $\Gamma(k+1) = k\Gamma(k)$, Property 4 results

$$ {}_2F_1(\alpha+1, 1-\beta; \alpha+2; 1) = \frac{\Gamma(\alpha+2)\,\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}, \tag{A.3} $$

$$ {}_2F_1(\alpha+1, -\beta; \alpha+2; 1) = \frac{\Gamma(\alpha+2)\ \beta\Gamma(\beta)}{(\alpha+\beta+1)\,\Gamma(\alpha+\beta+1)}. \tag{A.4} $$

Substituting the above expressions in (A.2), we obtain

$$ \int_0^1 \left(I_t^{-1}(\alpha,\beta)\right)^2\, dt = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}, \text{ similarly} $$
$$ \int_0^1 \left(I_t^{-1}(\beta,\alpha)\right)^2\, dt = \frac{\beta(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)}. \tag{A.5} $$

Thus, (A.1) simplifies to

$$ {}_2W_2^2\left(\eta_b(x;\alpha,\beta), \eta_b(x;\beta,\alpha)\right) = \frac{\alpha(\alpha+1)+\beta(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)} $$
$$ -2\int_0^1 I_t^{-1}(\alpha,\beta)\, I_t^{-1}(\beta,\alpha)\ dt. \tag{A.6} $$

To evaluate the remaining integral in (A.6), we employ integration-by-parts with $f(t) := I_t^{-1}(\alpha,\beta)$ as the first function and $g(t) := I_t^{-1}(\beta,\alpha)$ as the second. Now, we know that $\int_0^1 f(t)\, g(t)\, dt$ equals

$$ \underbrace{\left[f(t)\int g(t)\, dt\right]\Bigg|_{t=0}^{t=1}}_{\mathcal{I}} - \underbrace{\int_0^1 \left(f'(t)\int g(t)\ dt\right) dt}_{\mathcal{J}}. \tag{A.7} $$

From Properties 1 and 3, we get

$$ \mathcal{I} = \left[\frac{1}{(\beta+1)\,B(\alpha,\beta)} I_t^{-1}(\alpha,\beta)\left(I_t^{-1}(\beta,\alpha)\right)^{b+1} \right. $$
$$ \left. {}_2F_1(b+1, 1-a; b+2; 1)\right]\Bigg|_{t=0}^{t=1} $$
$$ = \frac{{}_2F_1(\beta+1, 1-\alpha; \beta+2; 1)}{(\beta+1)\,B(\alpha,\beta)} = \frac{\beta}{\alpha+\beta}. \tag{A.8} $$

Further, Properties (1) and (5) yield

$$ \mathcal{J} = \frac{1}{\beta+1}\int_0^1 \left(I_t^{-1}(\alpha,\beta)\right)^{1-\alpha}\left(1 - I_t^{-1}(\alpha,\beta)\right)^{1-\beta} $$
$$ \left(I_t^{-1}(\beta,\alpha)\right)^{\beta+1}\ {}_2F_1\left(\beta+1, 1-\alpha; \beta+2; I_t^{-1}(\beta,\alpha)\right)\, dt. \tag{A.9} $$

Combining (A.6), (A.7), (A.8) and (A.9), the result follows. ■

B On the stationary density of nonlinear systems with multiple stable equilibria

Proposition 5 *Consider a nonlinear dynamical system $\dot{x}(t) = f(x(t))$, having multiple stable equilibria $\{x_i^\star\}_{i=1}^{n^\star}$. Let us assume that the system does not admit any invariant set other than these stable equilibria. Also, let $\mathcal{R}_i$ be the region-of-attraction for the i^{th} equilibrium point. If the dynamics evolves from an initial PDF ξ_0, then its stationary PDF is given by*

$$ \xi_\infty(x) = \sum_{i=1}^{n^\star} m_i^\star \delta\left(x - x_i^\star\right), \tag{B.1} $$

where $m_i^\star = \displaystyle\int_{\mathrm{supp}(\xi_0)\cap\mathcal{R}_i} \xi_0(x_0)\, dx_0$.

Proof. Since $\{x_i^\star\}_{i=1}^{n^\star}$ is the unique set of attractors, it is easy to verify that the stationary PDF is of the form (B.1); however, it remains to determine the weights $m_i^\star$. We observe that **either** $\mathrm{supp}(\xi_0) \subseteq \mathcal{R}_i$, for some $i = 1, \ldots, n^\star$, **or** $\mathrm{supp}(\xi_0)$ intersects multiple $\mathcal{R}_i$.

Now, recall that $R_i \triangleq \{x_0 : \dot{x}(t) = f(x(t)), x(0) = x_0, \lim_{t\to\infty} x(t) = x_i^\star\}$. Thus, if $\mathrm{supp}(\xi_0) \subseteq \mathcal{R}_i$, then $m_i^\star = \int_{\mathrm{supp}(\xi_0)} dm_0 = \int_{\mathrm{supp}(\xi_0)} \xi(x_0)\, dx_0 = 1$, and consequently, $m_j^\star = 0$, $\forall j = 1, \ldots, n^\star$, $j \neq i$, since $\int \xi_\infty(x)\, dx = 1$. In this case, notice that $\mathrm{supp}(\xi_0) = \mathrm{supp}(\xi_0) \cap \mathcal{R}_i$.

On the other hand, if $\mathrm{supp}(\xi_0)$ intersects multiple $\mathcal{R}_i$, then only for $x_0 \in \mathrm{supp}(\xi_0) \cap \mathcal{R}_i$, the integral curves of $\dot{x}(t) = f(x(t)), x(0) = x_0$, will satisfy $\lim_{t\to\infty} x(t) = x_i^\star$. In other words, only the set $\mathrm{supp}(\xi_0) \cap \mathcal{R}_i$ contributes to $m_i^\star$, i.e. $m_i^\star = \int_{\mathrm{supp}(\xi_0)\cap\mathcal{R}_i} dm_0 = \int_{\mathrm{supp}(\xi_0)\cap\mathcal{R}_i} \xi(x_0)\, dx_0 < 1$.

Combining the above two cases, we conclude $m_i^\star = \displaystyle\int_{\mathrm{supp}(\xi_0)\cap\mathcal{R}_i} \xi_0(x_0)\, dx_0$. ■

C Proof for Lemma 1

(i) Proof of $\mathbb{P}(X > \epsilon) \leqslant \mathbb{P}(Y + Z > \epsilon)$: Let $A_1 := \{\omega : X(\omega) > \epsilon\}$ and $A_2 := \{\omega : Y(\omega) + Z(\omega) > \epsilon\}$. If we denote $B_1^\epsilon := \{\omega : X(\omega) \leqslant \epsilon\}$ and $B_2^\epsilon := \{\omega : Y(\omega) + Z(\omega) \leqslant \epsilon\}$, then

$$ X(\omega) \leqslant Y(\omega) + Z(\omega) < \epsilon \quad \forall\, \omega \in \Omega $$
$$ \Rightarrow B_2^\epsilon \subseteq B_1^\epsilon \Rightarrow \mathbb{P}(B_2^\epsilon) \leqslant \mathbb{P}(B_1^\epsilon) $$
$$ \Rightarrow 1 - \mathbb{P}(B_2^\epsilon) \geqslant 1 - \mathbb{P}(B_1^\epsilon) \Rightarrow \mathbb{P}(A_2) \geqslant \mathbb{P}(A_1). $$

(ii) Proof of $\mathbb{P}\left(Y+Z>\epsilon\right) \leqslant \mathbb{P}\left(Y>\dfrac{\epsilon}{2}\right)+\mathbb{P}\left(Z>\dfrac{\epsilon}{2}\right)$: Let $A := \{\omega : Y\left(\omega\right)+Z\left(\omega\right)>\epsilon\}$, $B := \{\omega : Y\left(\omega\right) \leqslant \epsilon/2\}$, and $C := \{\omega : Z\left(\omega\right) \leqslant \epsilon/2\}$. Next, we write

$$\mathbb{P}\left(A\right) = \mathbb{P}\left(\left(A \cap B^c \cap C^c\right) \cup B^c \cup C^c\right). \qquad \text{(C.1)}$$

Taking $\mathcal{E}_1 := A \cap B^c \cap C^c$, $\mathcal{E}_2 := B^c$, $\mathcal{E}_3 := C^c$, and noting that $\mathbb{P}\left(\mathcal{E}_1\right) = \mathbb{P}\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) = \mathbb{P}\left(\mathcal{E}_3 \cap \mathcal{E}_1\right) = \mathbb{P}\left(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3\right)$, from Boole-Bonferroni inequality (Appendix C, [83]), (C.1) yields

$$\begin{aligned}\mathbb{P}\left(A\right) = \mathbb{P}\left(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3\right) &= \mathbb{P}\left(\mathcal{E}_2\right) + \mathbb{P}\left(\mathcal{E}_3\right) - \mathbb{P}\left(\mathcal{E}_2 \cap \mathcal{E}_3\right)\\ &\leqslant \mathbb{P}\left(\mathcal{E}_2\right) + \mathbb{P}\left(\mathcal{E}_3\right).\end{aligned}$$

■

D Proof for Theorem 2

Since Wasserstein distance is a metric, from triangle inequality

$$\begin{aligned}&{}_2W_2\left(\eta_m, \widehat{\eta}_n\right) \leqslant {}_2W_2\left(\eta_m, \eta\right) + {}_2W_2\left(\widehat{\eta}_n, \eta\right)\\ &\leqslant {}_2W_2\left(\eta_m, \eta\right) + {}_2W_2\left(\widehat{\eta}_n, \widehat{\eta}\right) + {}_2W_2\left(\eta, \widehat{\eta}\right)\\ &\Rightarrow {}_2W_2\left(\eta_m, \widehat{\eta}_n\right) - {}_2W_2\left(\eta, \widehat{\eta}\right) \leqslant {}_2W_2\left(\eta_m, \eta\right) + {}_2W_2\left(\widehat{\eta}_n, \widehat{\eta}\right).\end{aligned}$$

Combining the above with Lemma 1, we have

$$\begin{aligned}&\mathbb{P}\left(\left\vert {}_2W_2\left(\eta_m, \widehat{\eta}_n\right) - {}_2W_2\left(\eta, \widehat{\eta}\right)\right\vert > \epsilon\right) \leqslant\\ &\mathbb{P}\left({}_2W_2\left(\eta_m, \eta\right) > \frac{\epsilon}{2}\right) + \mathbb{P}\left({}_2W_2\left(\widehat{\eta}_n, \widehat{\eta}\right) > \frac{\epsilon}{2}\right), \qquad \text{(D.1)}\end{aligned}$$

where each term in the RHS of (D.1) can be separately upper-bounded using Theorem 1 with $\theta \mapsto \dfrac{\epsilon}{2}$. Hence the result. ■

E Derivation of stationary PDF (31)

We re-write the Itô SDE (30) as

$$\begin{Bmatrix} dx_1 \\ dx_2 \end{Bmatrix} = \begin{Bmatrix} x_2 \\ -\dfrac{\partial}{\partial x_1} U\left(x_1\right) - c x_2 \end{Bmatrix} dt + \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} dW, \qquad \text{(E.1)}$$

with $U\left(x_1\right) := \frac{1}{2}\left(a x_1^2 - b \cos 2x_1\right)$. An Itô SDE with drift nonlinearity of the form (E.1), admits [84] stationary PDF $\eta_{\infty}\left(x_1, x_2\right) \propto \exp\left(-\frac{c}{Q} H\left(x_1, x_2\right)\right)$, where the Hamiltonian function $H\left(x_1, x_2\right) := U\left(x_1\right) + \frac{1}{2} x_2^2$.

F Proof for ${}_2W_2(k) \leq \parallel P_k^{1/2} - \widehat{P}_k^{1/2} \parallel_F$

It is known (Fact 8.19.21, [85]) that for $0 \leqslant p \leqslant 1$, $\operatorname{tr}\left(P_k^p \widehat{P}_k^p\right) \leqslant \operatorname{tr}\left(\widehat{P}_k^{1/2} P \widehat{P}_k^{1/2}\right)^p$. Taking $p = \frac{1}{2}$, we get

$$\operatorname{tr}\left(P_k^{\frac{1}{2}} \widehat{P}_k^{\frac{1}{2}}\right) \leqslant \operatorname{tr}\left(\widehat{P}_k^{\frac{1}{2}} P \widehat{P}_k^{\frac{1}{2}}\right)^{\frac{1}{2}} = \operatorname{tr}\left(P_k^{\frac{1}{2}} \widehat{P} P_k^{\frac{1}{2}}\right)^{\frac{1}{2}}, \qquad \text{(F.1)}$$

where the last equality follows from the symmetry of Wasserstein distance, and can be separately proved by noting that $\operatorname{tr}\left(\sqrt{MM^{\top}}\right) = \operatorname{tr}\left(\sqrt{M^{\top}M}\right)$ for $M = P_k^{1/2} \widehat{P}_k^{1/2}$.

Next, recall that square root of a positive definite matrix is unique, and matrix square root commutes with matrix transpose. Thus, we have

$$\begin{aligned}&\parallel P_k^{\frac{1}{2}} - \widehat{P}_k^{\frac{1}{2}} \parallel_F^2 \triangleq \operatorname{tr}\left[\left(P_k^{\frac{1}{2}} - \widehat{P}_k^{\frac{1}{2}}\right)^{\top}\left(P_k^{\frac{1}{2}} - \widehat{P}_k^{\frac{1}{2}}\right)\right]\\ &= \operatorname{tr}\left[\left(P_k^{\frac{1}{2}}\right)^{\top} P_k^{\frac{1}{2}} - \left(P_k^{\frac{1}{2}}\right)^{\top} \widehat{P}_k^{\frac{1}{2}} - \left(\widehat{P}_k^{\frac{1}{2}}\right)^{\top} \widehat{P}_k^{\frac{1}{2}} + \left(\widehat{P}_k^{\frac{1}{2}}\right)^{\top} \widehat{P}_k^{\frac{1}{2}}\right]\\ &= \operatorname{tr}\left[P_k\right] + \operatorname{tr}\left[\widehat{P}_k\right] - 2\operatorname{tr}\left[P_k^{\frac{1}{2}} \widehat{P}_k^{\frac{1}{2}}\right]\\ &\underbrace{\geq \operatorname{tr}\left[P_k\right] + \operatorname{tr}\left[\widehat{P}_k\right] - 2\operatorname{tr}\left(P_k^{\frac{1}{2}} \widehat{P} P_k^{\frac{1}{2}}\right)^{\frac{1}{2}}}_{\left({}_2W_2(k)\right)^2} \qquad \text{(using (F.1))}\end{aligned}$$

and hence, ${}_2W_2(k) \leq \parallel P_k^{1/2} - \widehat{P}_k^{1/2} \parallel_F$. From (F.1), the equality condition is $P_k \widehat{P}_k = \widehat{P}_k P_k$. ■