Probabilistic Model Validation for Uncertain Nonlinear Systems

Abhishek Halder, *Student Member, IEEE,* Raktim Bhattacharya, *Member, IEEE.*

Abstract

This paper presents a probabilistic model validation methodology for nonlinear systems. 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. Some examples are worked out to illustrate its use.

Index Terms

Model validation, stochastic nonlinear systems, uncertainty propagation, Wasserstein distance.

I. Introduction

A model serves as a mathematical abstraction of the physical system, that provides 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 even more imperative for applications where the model needs to be used for prediction (e.g. weather forecasting, stock market) or for safety-critical control design (e.g. aerospace, nuclear, systems biology) purposes.

Here it's 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. Halder and R. Bhattacharya are with the Department of Aerospace Engineering, Texas A&M University, College Station, TX 77843-3141 USA (email: ahalder@tamu.edu, raktim@tamu.edu).

a candidate model and the experimentally observed measurements of the physical system, how well does the model replicate the experimental measurements? As outlined in [1], a different but related notion is that of *model discrimination problem*, which reads: *given two different proposed models for the same physical system, design an input that can best discriminate between the two models, thus enabling a relative assessment between them.* Notice that, model validation problem concerns with whether *a model* is *good enough or not*; model discrimination problem concerns with which of the *two models* is *better*. The former is an analysis problem while the latter is a synthesis problem. Only the first problem is addressed in this paper. Furthermore, it has been argued in the literature [2]–[5] 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' is preferred. In this paper, instead of hard invalidation, we will consider the validation/invalidation problem in a probabilistically relaxed sense which can guide model refinement.

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 [6] which has been extended to account probabilistic models [6], [7]. Second is the H_∞ control framework where time-domain [4], [8], frequency domain [3] and mixed domain [9] model validation methods have been studied extensively assuming structured norm-bounded uncertainty. The third framework involves deductive inference based on barrier certificates [5] which was shown to encompass a large class of nonlinear models including differential-algebraic equations (DAEs) [10], dynamic uncertainties described by integral quadratic constraints (IQCs) [11], stochastic [12] and hybrid dynamics [13].

In statistical setting, model validation has been looked from system identification perspective [14], [15] 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 [16] for model validation. [17] have connected the robust control framework with prediction error based identification for frequency-domain validation of linear systems. In another vein, using Bayesian conditioning, [18] showed that for *parametric* uncertainty models, the statistical validation problem may be reduced to the computation of relative weighted volumes of convex sets. However, for *nonparametric* models: "the situation is significantly more complicated" [18] and to the best of our knowledge, has not been addressed in the literature.

Recently, in the spirit of weak stochastic realization problem [19], Ugrinovskii [20] investigated

the conditions for which the output of a stochastic nonlinear system can be realized through perturbation of a nominal stochastic *linear* system. However in practice, one often encounters the situation where a model is either proposed based on 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 a region of validation, one must also be able to demonstrate that the answer is *provably correct* on the face of uncertainty. This brings forth the notion of *probabilistically robust model validation* as opposed to *probabilistically worst-case model validation*. In this paper, we will show how to construct such a *robust validation certificate* guaranteeing the performance of probabilistic model validation algorithm, to be detailed in the sequel.

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

1) Instead of interval-valued structured uncertainty (as in H_∞ control framework) or moment based uncertainty (as in parametric statistics framework), we deal with model validation in the sense of nonparametric statistics by considering aleatoric uncertainty. In other words, the uncertainty in the model is quantified in terms of the probability density functions (PDFs) of the associated random variables. We argue that such a formulation offers some advantages. *Firstly*, we show that model uncertainty in the parameters and initial states can be propagated accurately by spatio-temporally evolving their joint PDF. Since experimental data usually come in the form of histograms, it's a more natural quantification of uncertainty than specifying sets [5] 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 [5] dealt with nonlinearities specified by semialgebraic sets and relied on sum of squares (SOS) decomposition [21] for computational tractability. From an implementation point of view, the approach presented in this paper doesn't require

such conservatism.

2) Due to the uncertainties in initial conditions, parameters, process and measurement 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 use *Wasserstein distance* to compare the output PDFs and argue why common information-theoretic quantities like *Kullback-Leibler (KL) divergence* [22] are not appropriate for this purpose and justify our choice of metric.
3) We show that the uncertainty propagation through continuous or discrete-time dynamics can be done in a numerically efficient way, 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 model validation. Thus the proposed framework is not only flexible, but also provides computational performance guarantees.

The paper is organized as follows. In Section II, we describe the problem setup by providing some intuitions followed by the methodology. Then we expound on the three steps of our validation framework: uncertainty propagation, distributional comparison and construction of validation certificates in Section III, IV and V, respectively. Examples are given in Section VI to illustarte the ideas presented in this paper, followed by conclusions in Section VII.

Notation

Most notations are standard. We use the superscript † to denote matrix transpose. The notation ${}_pF_q\left(a_1,\ldots,a_p;b_1,\ldots,b_q;x\right)$ stands for generalized hypergeometric function. The symbols $\mathcal{N}\left(.,.\right)$ and $\mathcal{U}\left(.\right)$ are used for normal and uniform distributions, respectively.

II. Problem Setup

A. Intuitive Idea

Given the experimental measurements of the physical system in the form of a distribution, our goal is to compare the shape of this measured output distribution with that predicted by the model. At every instant of time, if the model-predicted distribution matches with the experimental one "reasonably well" (to be made precise later in the paper), we conclude that the model is validated to be a good candidate with some quantification of such "goodness".

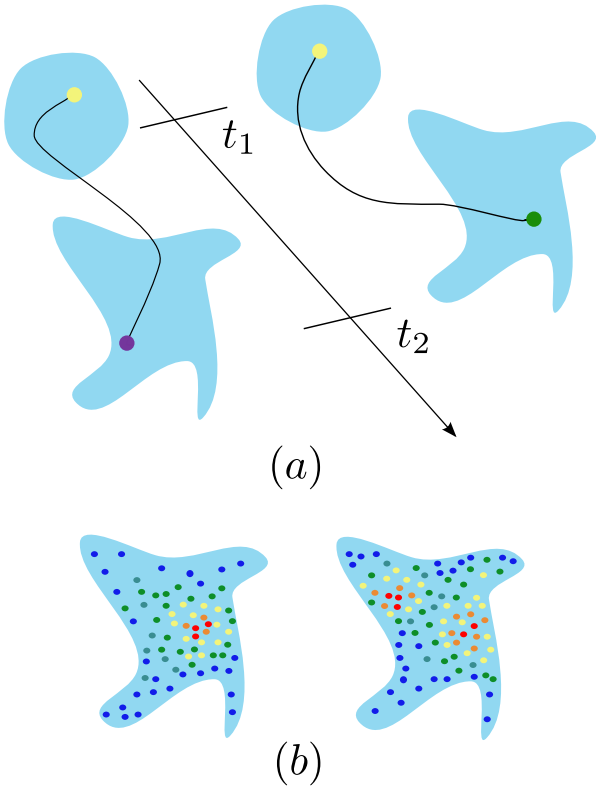

Fig. 1. Supports of the output distribution at time t_1 and t_2, as observed from experiments (right) and as predicted by the model (left). (a) The supports may match at finite instances, but the same sample (yellow dot) at t_1 may evolve in different trajectories and at t_2, may land at different locations in the supports of same shape. (b) At t_2, though supports are identical, the concentrations of trajectories are different. The experiment predicts bimodal distribution while the model forecasts it to be unimodal at this instant.

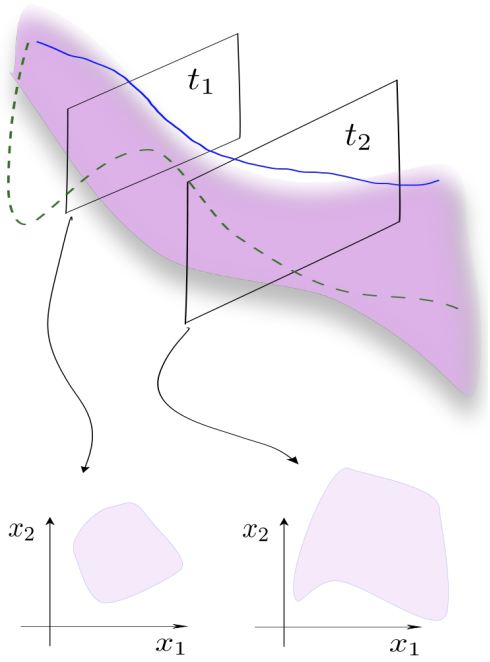

Fig. 2. Margin-of-safety "tube" wrapped around experimentally observed output ensemble. Cross-sections of this tube contain support snapshots as in Fig. 1(b). If the initial (left) and final (right) segments of the tube are safety-critical, then a model that predicts the dashed trajectory will be invalidated with high probability. On contrary, a model that predicts the solid trajectory may be considered as valid since it grazes the outer surface of the tube for a non-critical segment. Note that, the solid line stays outside the tube longer than the dashed one.

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. Further, it's more accurate than simply matching the set-containment [5] since two identical supports may have different trajectory concentrations (Fig. 1).

For real-life applications in general, and in the context of probabilistic validation in particular, the margin-of-safety (and hence tolerance level) is not fixed during the course of operation of the system. For example, take-off and landing are critical operational segments during the flight of a commercial aircraft. These two segments have very high margin-of-safety and it's unacceptable to have a controller which does not guarantee the nominal performance with very high probability. However, for cruise segment, the requirement is less stringent and a controller may be validated with some small but finite probability of departure from the safety-margin for this segment, provided it meets the take-off and landing tolerance level. Thus we need to account three factors for model validation. One is the *extent of departure* from the tolerance "tube" (Fig. 2), the *criticality of the segment* where the departure happens and the *duration of such departure*. For instance, large departure in a critical segment even for a short duration, should result invalidation with high probability but small departure for a prolonged duration may be acceptable provided that takes place in non-critical operational segment(s) (see Fig. 2). Such practical considerations can't be accommodated by hard invalidation methods.

B. Methodology

In this section, we formalize the ideas presented above. Fig. 3 shows the outline of the model validation framework proposed here. The experiment is carried out with the physical plant taking some initial PDF $\xi_0(\widetilde{x})$. Given the data for experimentally observed output PDF $\eta(y,t)$, one starts propagating $\xi_0(\widetilde{x})$ through the proposed model's state dynamics, thereby computing state PDF $\xi(\widetilde{x},t)$ and from it, obtains the output PDF $\widehat{\eta}(y,t)$ using the output dynamics prescribed by the model. If the output PDFs $\eta(y,t)$ and $\widehat{\eta}(y,t)$ are "close" in the sense that a suitable distance metric on the space of probability densities, $J(\eta,\widehat{\eta})$ remains small (within the specified tolerance level) at all times t when the experimental data are available, then it will be concluded that the model is "close" to the physical plant with some quantitative measure.

Since the basic idea relies on comparing the concentration of output trajectories at each instant of experimental observation, one can think of three distinct segments of such a model validation

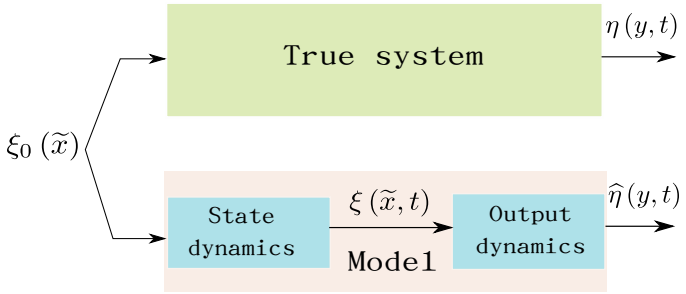

Fig. 3. The proposed model validation framework compares experimentally observed output PDF $\eta(y,t)$ with the model-predicted one $\widehat{\eta}(y,t)$, the comparison being made with respect to some suitable metric $J(\eta,\widehat{\eta})$.

framework. These are

1) Uncertainty propagation: evolving state and output PDFs using the proposed model.
2) Distributional comparison: measuring distance between the experimentally observed and model-predicted output PDFs and computing the margin by which the model-prediction obeys/violates the specified tolerance level.
3) Construction of validation certificates: probabilistic quantification of provably correct inference in this framework and providing sample complexity bounds for the same.

Now we will elicit each of these segments.

III. Uncertainty Propagation

A. Continuous-time Systems

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

$$\dot{\widetilde{x}} = \widetilde{f}(\widetilde{x}), \qquad \widetilde{x} := \begin{Bmatrix} x \\ p \end{Bmatrix} \in \mathcal{X} \times \mathcal{P} \subseteq \mathbb{R}^{n+p}, \quad \text{and} \quad \widetilde{f} = \begin{Bmatrix} f_{n\times 1} \\ \mathbf{0}_{p\times 1} \end{Bmatrix}. \tag{1}$$

The output dynamics can be written as

$$y = h(\widetilde{x}) \tag{2}$$

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

$$\frac{\partial \xi}{\partial t} \;=\; \mathscr{L}\xi = D_1\xi = -\nabla.(\xi f) = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}(\xi f_i), \tag{3}$$

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

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

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

2) Uncertainty Propagation through Stochastic Flow: Consider the continuous-time nonlinear system with state dynamics given by the Itô SDE

$$d\widetilde{x} = \widetilde{f}(\widetilde{x})\; dt + g(\widetilde{x})\; dW, \tag{5}$$

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

$$\mathbb{E}[dW_i] = 0 \quad \text{and} \quad \mathbb{E}[dW_i dW_j] = Q_{ij} = \alpha_i\, \delta_{ij} \qquad \forall\, i, j = 1, 2, \ldots, \omega \tag{6}$$

where $\mathbb{E}[.]$ 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 dynamics is still assumed to be given by (2). In such a setting, the evolution of the state PDF $\xi(\widetilde{x}, t)$ subject to (5) is governed by the *Fokker-Planck equation*, also known as *forward Kolmogorov equation*

$$\frac{\partial \xi}{\partial t} = \mathscr{L}\xi = (D_1 + D_2)\,\xi = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}(\xi f_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j}\left(\left(gQg^T\right)_{ij}\xi\right), \tag{7}$$

which is a homogeneous parabolic PDE, second order in space and first order in time. In this case, the spatial operator $\mathscr{L}$ 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).

Remark 1: The output model can be generalized to account additive measurement noise by modifying the change-of-variable formula given by (4).

Remark 2: For both eqn. (3) and (7), the operators $(\mathscr{L})$ are infinite dimensional and non-self-adjoint. This severely complicates the applicability of operator eigenfunction approach to solve (3) or (7) unless the dynamics (f) is very simple [23]. On the other hand, the method of eigenfunction expansion is known (p. 139, [24]) to have poor convergence.

Remark 3: Due to the above reason, non-spectral methods are preferred to solve (3) and (7) for general nonlinear systems. In particular, the authors have earlier shown [25] that (3) can be solved directly using the method-of-characteristics (MOC), even for high-dimensional nonlinear systems. For solving (7), one can choose a recently-developed meshless semi-analytic method [26] that efficiently constructs the PDF as a global-local approximation and outperforms traditional FEM based methods, which are hardly scalable beyond three dimensions. Alternatively, one can approximate the noise term in (5) by finite-term Karhunen-Loève expansion and apply MOC to solve (3) for the resulting approximate deterministic flow. It has been shown [27] that the distribution computed from this approximate dynamics converges to that of the true stochastic dynamics in mean square sense. To get the computational leverage, these methods are adopted in this paper for numerically solving (3) and (7), respectively.

B. Discrete-time Systems

1) Uncertainty Propagation through Deterministic Maps: We start with the following two definitions.

Definition 1: Let $\mathcal{X} \times \mathcal{P} \subseteq \mathbb{R}^{n+p}$ be a compact set and let $\mathcal{B}(\mathcal{X} \times \mathcal{P})$ be the Borel-σ algebra defined on it. With respect to the measure space $(\mathcal{X} \times \mathcal{P}, \mathcal{B}, \mu)$, a transformation $\mathcal{T} : \mathcal{X} \times \mathcal{P} \mapsto \mathcal{X} \times \mathcal{P}$ is called measurable if $\mathcal{T}^{-1}(B) \in \mathcal{B} \quad \forall B \in \mathcal{B}$.

Definition 2: A measurable transformation $\mathcal{T} : \mathcal{X} \times \mathcal{P} \mapsto \mathcal{X} \times \mathcal{P}$ is said to be nonsingular on the measure space $(\mathcal{X} \times \mathcal{P}, \mathcal{B}, \mu)$, if $\mu(\mathcal{T}^{-1}(B)) = 0 \quad \forall B \in \mathcal{B}$ such that $\mu(B) = 0$.
Consider the discrete-time nonlinear system with state dynamics given by the vector recurrence relation

$$\widetilde{x}_{k+1} = \mathcal{T}(\widetilde{x}_k) \tag{8}$$

where $\mathcal{T} : \mathcal{X} \times \mathcal{P} \mapsto \mathcal{X} \times \mathcal{P}$ 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 $\xi\left(\widetilde{x}_k\right)$ is dictated by the *Perron-Frobenius operator* $\mathscr{P}$, given by

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

for $B \in \mathcal{B}$. Following properties (Chap. 3, [28]) of Perron-Frobenius operator are important from computational standpoint.

Property 1: (Linearity) $\mathscr{P}\left(\beta_1\xi_1 + \beta_2\xi_2\right) = \beta_1\mathscr{P}\xi_1 + \beta_2\mathscr{P}\xi_2 \qquad \forall\, \xi_1, \xi_2 \in L^1,\ \beta_1, \beta_2 \in \mathbb{R}.$

Property 2: (Non-negativity) $\xi \geqslant 0 \Rightarrow \mathscr{P}\xi \geqslant 0$.

Property 3: (Composition) If we denote $\mathscr{P}_k$ as the Perron-Frobenius operator corresponding to the k^{th} iterate of the map $\mathcal{T}$ given by $\mathcal{T}_k = \mathcal{T} \circ \ldots \circ \mathcal{T}$ (composed k times), then $\mathscr{P}_k = \mathscr{P}^k$.

Property 4: (Change-of-variable) On the measure space $(\mathcal{X} \times \mathcal{P}, \mathcal{B}, \mu)$, let $\mathcal{T} : \mathcal{X} \times \mathcal{P} \mapsto \mathcal{X} \times \mathcal{P}$ be a measurable, invertible, nonsingular transformation (i.e. $\mathcal{T}^{-1}$ is nonsingular) and $\mathscr{P}$ be the Perron-Frobenius operator associated with $\mathcal{T}$. Then $\mathscr{P}\xi\left(\widetilde{x}_k\right) = \xi\left(\mathcal{T}^{-1}\left(\widetilde{x}_k\right)\right)\left|det\left(\mathcal{J}^{-1}\left(\widetilde{x}_k\right)\right)\right| \quad \forall \xi \in L^1\left(\mathbb{R}^{n+p}\right)$ where $\mathcal{J}^{-1}\left(\widetilde{x}_k\right)$ is the Jacobian of the inverse map $\mathcal{T}^{-1}\left(\widetilde{x}_k\right)$.

The proofs for the first three properties are straightforward from (9). For a proof of Property 4, see Corollary 3.2.1 in [28]. Further, assuming the output dynamics as $y_k = h\left(\widetilde{x}_k\right)$, one can derive $\widehat{\eta}\left(y_k\right)$ from $\xi\left(\widetilde{x}_k\right)$ using the discrete analogue of (4).

2) Uncertainty Propagation through Stochastic Maps: In this case, we consider the nonlinear state space representation given by the stochastic maps of general form

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

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 $\mathcal{T}$ is not required to be a non-singular transformation (Chap. 10, [28]). Since $\mathcal{T}$ defines a Markov Chain on $\mathcal{X} \times \mathcal{P}$, it can be shown that [28], [29] evolution of the joint PDFs follow

$$\begin{aligned}
\xi_{k+1} := \xi_{\widetilde{x}_{k+1}}\left(\widetilde{x}\right) &= \int_{\mathcal{X}\times\mathcal{P}} \mathcal{K}_{\mathcal{T}}\left(\widetilde{x}|z\right)\xi_{\widetilde{x}_k}\left(z\right)dz = \int_{\mathcal{X}\times\mathcal{P}} \mathcal{K}_{\mathcal{T}}\left(\widetilde{x}|z\right)\xi_k\left(z\right)dz, \\
\widehat{\eta}_k := \widehat{\eta}_{y_k}\left(y\right) &= \int_{\mathcal{X}\times\mathcal{P}} \mathcal{K}_h\left(y|z\right)\xi_{\widetilde{x}_k}\left(z\right)dz = \int_{\mathcal{X}\times\mathcal{P}} \mathcal{K}_h\left(y|z\right)\xi_k\left(z\right)dz,
\end{aligned} \tag{11}$$

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

IV. Distributional Comparison

Given $\xi_0(\widetilde{x})$ and $\eta(y,t)$, in the previous section we have put machineries in place to compute the lower branch of the block diagram (Fig. 3). Now, we need a metric J that can compare the shapes of the two distributions $\eta(y,t)$ and $\widehat{\eta}(y,t)$, at any fixed time t. We argue that the suitable metric for our purpose is the Wasserstein distance, formally introduced below.

A. Wasserstein Distance

Let M be a complete, separable metric (Polish) space with a p^{th} order distance metric d_p. For simplicity, we let M to be $\mathbb{R}^n$ and take d_p as the L^p norm. Then the Wasserstein distance of order q, denoted as ${}_pW_q$, between two Borel probability measures μ_1 and μ_2 on $\mathbb{R}^n$ is defined as

$$ {}_pW_q(\mu_1,\mu_2) := \left[\inf_{\mu\in\mathcal{M}(\mu_1,\mu_2)}\int_{\mathbb{R}^n\times\mathbb{R}^n}\|\underline{x}-\underline{y}\|_p^q\ d\mu\left(\underline{x},\underline{y}\right)\right]^{1/q} = \left(\inf_{\mu\in\mathcal{M}(\mu_1,\mu_2)}\mathbb{E}\left[\|\underline{x}-\underline{y}\|_p^q\right]\right)^{1/q}, \tag{12}$$

where $\mathcal{M}(\mu_1,\mu_2)$ is the set of all probability measures on $\mathbb{R}^n\times\mathbb{R}^n$ with first marginal μ_1 and second marginal μ_2. It's well known [31] that on the set of Borel measures on $\mathbb{R}^n$ having finite second moments, ${}_pW_q$ defines a metric that admits a dual representation via Kantorovich-Rubinstein theorem (§11.8, [32]). If the measures μ_1 and μ_2 are absolutely continuous w.r.t. the Lebesgue measure, with densities ρ_1 and ρ_2, then we can write $\mathcal{M}(\rho_1,\rho_2)$ for the set $\mathcal{M}(\mu_1,\mu_2)$ and accordingly ${}_pW_q(\rho_1,\rho_2)$ in lieu of ${}_pW_q(\mu_1,\mu_2)$ in (12). This is assumed to hold for all subsequent analysis.

The Wasserstein metric has many nice properties. For example, it's insensitive to oscillations and it metrizes weak convergence (Theorem 7.12, p. 212, [33]). Further, the supports of the PDFs under consideration, need not be same. However, as one may guess, computation of Wasserstein metric is not straightforward. For real line, a closed form solution exists [34] in terms of the cumulative distribution functions (CDFs). Let F and G be the corresponding CDFs of the univariate PDFs ρ_1 and ρ_2 respectively. Then

$$ {}_pW_q^q(\rho_1,\rho_2) = \int_{\mathbb{R}}\|F(x)-G(x)\|_p^q\,dx = \int_0^1\|F^{-1}(t)-G^{-1}(t)\|_p^q\,dt. \tag{13}$$

For multivariate case, in general, one has to seek a numerical solution of (12). Before addressing the computational method, we first justify the choice of metric.

B. Justification for the Choice of Metric

One must realize that (12) defines the Wasserstein metric as the infimum distance between two random vectors with fixed distributions. This infimum occurs when the correlation between the random vectors (in this case, experimental observation vector and model-predicted output vector) is as large as possible, given their marginals. Since the stochastic dependence between these random vectors are not known a priori, characterizing the difference between *shapes* rather than the difference between *realizations*, is meaningful for choosing a validation metric [35].

Gibbs and Su [36] have listed a host of distances on the space of probability densities and relations among them. Amid these, the information-theoretic KL divergence (D_{KL}) has been prominent in the dynamical systems and control literature [37]–[39]. Recently, [22] have used KL rate for comparing dynamical systems and for nonlinear model reduction, both in discrete-time setting. It's natural to ask why KL divergence can not be used for the present purpose and in what respects Wasserstein distance suits better than KL as a validation metric.

1) KL Divergence and Wasserstein Distance:

Proposition 1: If $\mathcal{D}$ is a bounded domain, ${}_pW_q \leqslant \operatorname{diam}(\mathcal{D}) \sqrt{\dfrac{D_{KL}}{2}}$.

Proof: This result can be obtained in two steps. First by relating ${}_pW_q$ with the total variation distance d_{TV}, which is half of the L^1 norm between two probability measures. This yields ${}_pW_q \leqslant \operatorname{diam}(\mathcal{D})\ d_{TV}$ (see Theorem 4, [36]). In the second step, we recall a result due to Kullback [40] that says $d_{TV} \leqslant \sqrt{\dfrac{D_{KL}}{2}}$. Together, the statement follows. ■

Remark 4: In an uncertain nonlinear system, since the domain and the PDF are both evolving with time, the above proposition shows convergence in one metric may not imply convergence in other. For example, even if ${}_pW_q$ becomes progressively small, it may not imply so for D_{KL}.

Remark 5: Naturally, it's of interest to investigate the conditions for which the distance of a probability measure μ from *any* other measure ν computed through ${}_pW_q$ can be related with that computed through D_{KL}, independent of the domain-size. Such a characterization is given by the following [41].

Definition 3: 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 all probability measure ν, ${}_pW_q(\mu, \nu) \leqslant \sqrt{2CD_{KL}(\nu, \mu)}$. In short, we write $\mu \in T_q(C)$.

Notice that, we are interested in TCI results independent of dimensions. It was observed [42]

that T_1 is not well adapted for dimension free bounds but T_2 is. Also, [43] demonstrates that uncertainty evolution can be seen as a gradient flux of free energy with respect to the Wasserstein metric of order two. For these reasons, we choose $q = 2$ in this study. We have the following $T_2(C)$ result when $\mu \sim \mathcal{N}(\mathbf{0}_{\kappa\times 1}, I_{\kappa\times\kappa})$.

Proposition 2: [44] If $\mu(dx) = \dfrac{e^{-\frac{1}{2}\|x\|_2^2}}{(2\pi)^{\kappa/2}}\,dx$, then ${}_pW_2 \leqslant \sqrt{2D_{KL}}$, i.e. $\mu \in T_2(1)$.

For $L^2(\mathbb{R}^n)$, it can be further generalized [41] for $\mu \sim \mathcal{N}(m_{\kappa\times 1}, \Sigma_{\kappa\times\kappa})$ as $\mu \in T_2(C)$ with $C = \lambda_{\max}(\Sigma)$. In the model validation context, an often encountered situation is one where a linear model is proposed for a nonlinear plant. In such a situation, an immediate consequence of the above results, is the characterization of the comparative behavior of KL divergence and Wasserstein distance in terms of model parameters.

Proposition 3: Consider a linear time-invariant (LTI) model given by $\dot{\widetilde{x}} = A\widetilde{x}$, $y = C\widetilde{x}$, proposed for a nonlinear plant, where (A, C) is an observable pair. If $\xi_0(\widetilde{x}) \sim \mathcal{N}(\mu_0, P_0)$, then

$$ {}_2W_2(\eta(y,t), \widehat{\eta}(y,t)) \leqslant \sqrt{2\lambda_{\max}(\Sigma(t))\, D_{KL}(\eta(y,t), \widehat{\eta}(y,t))} \qquad \forall\, t \in [0,\infty) $$

where the time-varying matrix $\Sigma(t)$ is given by $\Sigma(t) = CP(t)C^T$, where $P(t)$ satisfies the Lyapunov equation $\dot{P} = AP + PA^T$.

Proof: Since the initial PDF over the extended state space is Gaussian, a model with linear state evolution $\dot{\widetilde{x}} = A\widetilde{x}$, retains the state PDF Gaussian at all times. In other words, $\xi_0(\widetilde{x}) \sim \mathcal{N}(\mu_0, P_0) \Rightarrow \xi(\widetilde{x}, t) \sim \mathcal{N}(\mu, P)$ where

$$ \begin{aligned} \dot{\mu} &= A\mu, & \mu(0) &= \mu_0, \\ \dot{P} &= AP + PA^T, & P(0) &= P_0. \end{aligned} $$

Since $y = C\widetilde{x}$ and linear transformation of a Gaussian PDF remains Gaussian, it's straightforward to verify that $\widehat{\eta}(y,t) \sim \mathcal{N}\left(C\mu, CPC^T\right)$. However, the original plant being nonlinear, $\eta(y,t)$ does not remain Gaussian in general, even though $\xi_0(\widetilde{x})$ is Gaussian. This requires us to compare a Gaussian PDF $\widehat{\eta}(y,t)$ with a non-Gaussian PDF $\eta(y,t)$ at all times. Now we can apply the $T_2(C)$ result for L^2 as discussed above to arrive at the inequality. ■

Remark 6: If the proposed model includes process noise and is given by the Itô SDE $d\widetilde{x} = A\widetilde{x}dt + BdW$ such that W is a ω-dimensional Wiener process with $\mathbb{E}[dW_i] = 0$, and $\mathbb{E}[dW_i dW_j] = Q_{ij},\ \forall\, i,j = 1,2,\ldots,\omega$, then the above result holds with a modified covariance propagation equation $\dot{P} = AP + PA^T + BQB^T$ (Riccati equation).

TABLE I

WASSERSTEIN DISTANCE $({}_2W_2)$ AND KL DIVERGENCE (D_{KL}) BETWEEN TWO GAUSSIANS, THE FIRST ROW IS FOR THE UNIVARIATE CASE WHILE THE SECOND ROW IS FOR THE MULTIVARIATE CASE.

${}_2W_2$	D_{KL}	TCI
$\sqrt{(m_1-m_2)^2+(\sigma_1-\sigma_2)^2}$	$\frac{1}{2}\left[\frac{(m_1-m_2)^2}{\sigma_2^2}+\left(\frac{\sigma_1^2}{\sigma_2^2}-1-\ln\frac{\sigma_1^2}{\sigma_2^2}\right)\right]$	${}_2W_2 \leqslant \sqrt{2\sigma_2^2 D_{KL}}$ equality for $\sigma_1=\sigma_2$
$\sqrt{\lVert m_1-m_2\rVert_2^2+\operatorname{tr}(\Sigma_1)+\operatorname{tr}(\Sigma_2)-2\operatorname{tr}\left(\left(\sqrt{\Sigma_1}\Sigma_2\sqrt{\Sigma_1}\right)^{1/2}\right)}$	$\frac{1}{2}\left[\ln\frac{\det(\Sigma_2)}{\det(\Sigma_1)}+\operatorname{tr}\left(\Sigma_2^{-1}\Sigma_1\right)+(m_2-m_1)^T\Sigma_2^{-1}(m_2-m_1)-\nu\right]$	${}_2W_2 \leqslant \sqrt{2\lambda_{\max}(\Sigma_2)\, D_{KL}}$ equality for $\Sigma_1=\Sigma_2$

2) Comparing Gaussians: a Shape Comparison Case Study: To illustrate the inadequacy of KL divergence for the present context, we now construct an example in Lemma 1 below, where the Wasserstein distance captures the shape difference while KL does not. Consider a special case of Proposition 3 where both the actual plant and the proposed model are linear. In that case, if $\xi_0(\widetilde{x})$ is Gaussian, then both $\eta(y,t)$ and $\widehat{\eta}(y,t)$ remain Gaussian at all times. The Wasserstein distance and KL divergence between two Gaussian PDFs can be computed in closed form in terms of the respective mean vectors and covariance matrices, as summarized in Table 1. For ${}_2W_2$, the multivariate formula is due to [45] and the univariate formula can either be obtained as a special case of the same or can be derived independently from (13) using the normal CDFs. For D_{KL}, the multivariate formula can be found in [46] and the univariate formula follows from there. The T_2 results are shown in the third column of Table 1. The following lemma illustrates that for the case of two Gaussians of identical shape except the mean, D_{KL} does not capture the shape difference while ${}_2W_2$ does.

Lemma 1: Consider two Gaussian PDFs, both supported over $\mathbb{R}^\nu$ such that they have different mean vectors m_1 and m_2, but same covariance $\Sigma_1=\Sigma_2$. Then D_{KL} is not a shape measure unless $\Sigma_1=\Sigma_2=\dfrac{\lVert m_1-m_2\rVert_2}{2}I_{\nu\times\nu}$ for $\nu>1$, and $\sigma_1^2=\sigma_2^2=\dfrac{1}{2}(m_1-m_2)$ for $\nu=1$.

Proof: Let's prove the univariate version first. Since $\sigma_1=\sigma_2$, from the first row of Table 1, it's evident that ${}_2W_2=(m_1-m_2)$ and $D_{KL}=\dfrac{(m_1-m_2)^2}{2\sigma_2^2}$. Since $m_1\neq m_2$, D_{KL} is not a shape measure unless $\dfrac{m_1-m_2}{2\sigma_2^2}-1=0\Rightarrow\sigma_1^2=\sigma_2^2=\dfrac{1}{2}(m_1-m_2)$. This completes the proof for univariate case $(\nu=1)$.

For the multivariate case $(\nu>1)$, the second row of Table 1 yields ${}_2W_2=\lVert m_1-m_2\rVert_2$ and $D_{KL}=\dfrac{1}{2}(m_2-m_1)^T\Sigma_2^{-1}(m_2-m_1)$. If we introduce $m:=m_2-m_1$, then $\dfrac{D_{KL}}{{}_2W_2}=$

$\frac{\|m\|_2}{2}\, r$, where $r := \frac{m^T \Sigma_2^{-1} m}{m^T m}$ is the Rayleigh quotient corresponding to the positive semi-definite precision matrix Σ_2^{-1}. It's known (Chap. 7, [47]) that if we denote

$$\mathscr{C} := \textbf{ConvHull}\{\lambda_1, \lambda_2, \ldots, \lambda_\nu\} = \{\lambda : \lambda = \sum_{i=1}^{\nu} \alpha_i \lambda_i,\ \sum_{i=1}^{\nu} \alpha_i = 1,\ \alpha_i \geqslant 0,\ \forall i = 1, 2, \ldots, \nu\}$$

as the convex hull of the eigenvalues of the precision matrix Σ_2^{-1}, then $r(m) \in \mathscr{C}$. In particular,

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

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 $\frac{D_{KL}}{{}_2W_2}$, even when $\|m\|$ is kept fixed. In particular, the ratio assumes unity iff $r = \frac{2}{\|m\|} \Rightarrow \Sigma_2^{-1} = \frac{2}{\|m\|} I_{\nu\times\nu} \Rightarrow \Sigma_1 = \Sigma_2 = \frac{\|m\|}{2} I_{\nu\times\nu}$. ■

The inadequacy of KL divergence for capturing shape characteristics and the utility of Wasserstein distance for the same, have been observed numerically in [48]. Also, the reason why Wasserstein distance can overcome the shortcomings of pointwise (pseudo)metrics like D_{KL} have been discussed in the context of image processing [49]. Further discussions along these lines can be found in [50], [51]. Below we provide two univariate examples showing PDFs of same entropy may have different shapes, resulting non-zero value of ${}_2W_2$.

3) Example 1: PDFs from Same Family: Consider the two parametric family of beta densities $f_b(x; \alpha, \beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, $\alpha, \beta > 0$, $x \in [0, 1]$, where $B(\alpha, \beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, is the (complete) beta function and $\Gamma(z)$ denotes the gamma function. The differential entropy for the beta family can be computed as [52]

$$\begin{aligned} H_b(\alpha, \beta) &:= -\int_0^1 f_X(x; \alpha, \beta) \log f_X(x; \alpha, \beta)\ dx \\ &= \log B(\alpha, \beta) - (\alpha - 1)(\Psi(\alpha) - \Psi(\alpha + \beta)) - (\beta - 1)(\Psi(\beta) - \Psi(\alpha + \beta)) \end{aligned} \quad (14)$$

where $\Psi(z) := \frac{d}{dz} \log \Gamma(z)$, is the digamma function. Since (14) remains invariant under $(\alpha, \beta) \mapsto (\beta, \alpha)$, $\alpha \neq \beta$, $f_b(x; \alpha, \beta)$ and $f_b(x; \beta, \alpha)$ have same entropy but one is skewed to right and the other to left, as shown in Fig. 4. This may seem intuitive since both densities in Fig. 4, because of the above symmetry, have same information content compared to uniform density. But Fig. 5 shows that in general, any two asymmetric pairs (α_1, β_1) and (α_2, β_2), $\alpha_1 \neq \beta_1$, $\alpha_2 \neq \beta_2$, satisfying $H_b(\alpha, \beta) =$ constant, have same entropy.

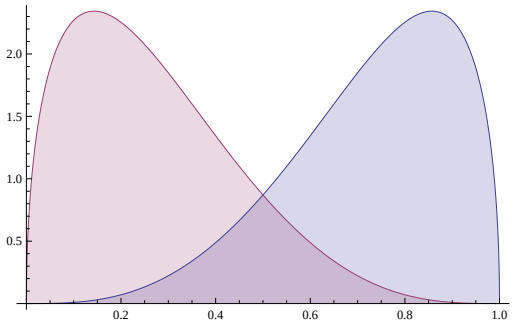

Fig. 4. The two beta densities $f_b(x;\alpha,\beta)$ (right-skewed) and $f_b(x;\beta,\alpha)$ (left-skewed) with $\alpha = 4$, $\beta = \frac{3}{2}$, have same entropy $H_b(\alpha,\beta)$ but have different shapes.

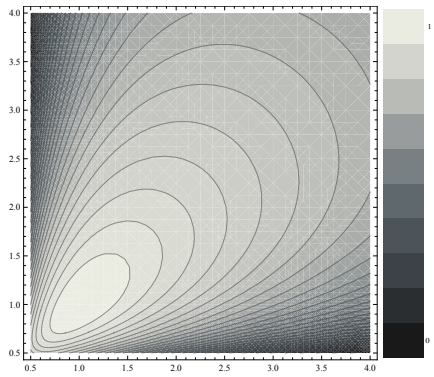

Fig. 5. Isentropic contours of beta family in (α,β) space. Notice the symmetry about $\alpha = \beta$ line. It also shows uniform distribution $(\alpha = \beta = 1)$ to be of maximum entropy.

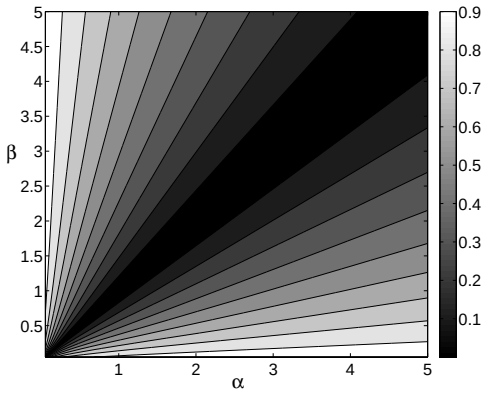

Fig. 6. Contours of ${}_2W_2\left(f_b(x;\alpha,\beta), f_b(x;\beta,\alpha)\right)$ in (α,β) space. Since ${}_2W_2$ is a metric, it has symmetry about $\alpha = \beta$ line. For same reason, it vanishes along this line.

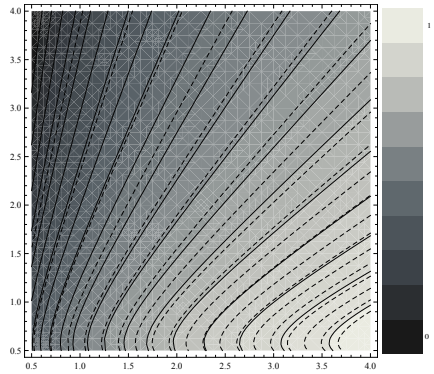

Fig. 7. Solid curves show the isentropic contours of the Weibull family in (α, c) space. The dotted curves show the contours given by (17).

Now we provide a closed-form expression for the Wasserstein distance between $f_b(x;\alpha,\beta)$ and $f_b(x;\beta,\alpha)$. Here, $I_t^{-1}(\alpha,\beta)$ is the inverse of the beta CDF $I_x(\alpha,\beta) := \dfrac{B(x;\alpha,\beta)}{B(\alpha,\beta)}$, the regularized (incomplete) beta function, and $B(x;\alpha,\beta) := \displaystyle\int_0^x z^{\alpha-1}(1-z)^{\beta-1}\,dz$ is the incomplete beta function.

Theorem 1: ${}_2W_2\left(f_b(x;\alpha,\beta), f_b(x;\beta,\alpha)\right) = \sqrt{\dfrac{\alpha(\alpha+1)+\beta(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)} - 2\left(\dfrac{\beta}{\alpha+\beta} - \mathcal{J}\right)}$,

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

4) Example 2: PDFs from Different Families: Consider the uniparametric Rayliegh family with density $f_r(x;\beta) = \dfrac{x}{\beta^2}\exp\left(-\dfrac{x^2}{2\beta^2}\right)$; $x,\beta>0$, and the two parametric Weibull family with density $f_w(x;c,\alpha) = \left(\dfrac{c}{\alpha}\right)\left(\dfrac{x}{\alpha}\right)^{c-1}\exp\left(-\left(\dfrac{x}{\alpha}\right)^{c}\right)$; $x,\alpha,c>0$. Their respective differential entropies are given by [52]

$$H_r(\beta) = 1+\log\left(\frac{\beta}{\sqrt{2}}\right)+\frac{\gamma}{2}, \tag{15}$$

$$H_w(\alpha,c) = 1+\log\left(\frac{\alpha}{c}\right)+\left(\frac{c-1}{c}\right)\gamma, \tag{16}$$

where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant. Equating the above two entropies and solving for β yields

$$\beta = \left(\frac{\sqrt{2}\,\alpha}{c}\right)\exp\left(\gamma\left(\frac{1}{2}-\frac{1}{c}\right)\right). \tag{17}$$

Whenever $\beta=$ constant curves intersect with the Weibull contours given by (17), the associated Rayleigh and Weibull densities admit same entropy. This is illustrated in Fig. 7, where the dotted contours are of the form $\left(\dfrac{\sqrt{2}\,\alpha}{c}\right)\exp\left(\gamma\left(\dfrac{1}{2}-\dfrac{1}{c}\right)\right)=$ constant $=k$ (say). Any Weibull PDF $f_w(x;c,\alpha)$ along this curve, have entropy equal to that of a Rayleigh PDF $f_r(x;k)$.

To show that Wasserstein distance between $f_r(x;\beta)$ and $f_w(x;c,\alpha)$ need not be zero (implying shape difference) even when the respective differential entropies match, we first compute ${}_2W_2\left(f_r(x;\beta), f_w(x;c,\alpha)\right)$ as a function of β, c and α.

Theorem 2: ${}_2W_2\left(f_r(x;\beta), f_w(x;c,\alpha)\right) = \sqrt{\left(\dfrac{2\alpha^2}{c}\right)\Gamma\left(\dfrac{2}{c}\right)+2\beta^2-2\sqrt{2}\alpha\beta\left(\dfrac{1}{c}+\dfrac{1}{2}\right)\Gamma\left(\dfrac{1}{c}+\dfrac{1}{2}\right)}$.

Corollary 1: If $f_r(x;\beta)$ and $f_w(x;c,\alpha)$ are isentropic, then the Wasserstein distance between them is non-zero except for $c=2$.

C. Computing Multivariate Wasserstein Distance of Order Two

Computing Wasserstein distance from (12) calls for solving *Monge-Kantorovich optimal transportation plan* [53]. This problem has a rich history [54], [55] and subsequent developments led to nobel prize (1975) in economics. 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* [56]–[58], 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 $c_{ij} := \| \ u_i - v_j \ \|_{\ell_2}^2$ denotes the cost of transporting unit mass from vertex u_i to v_j. Then, according to (12), computing ${}_2W_2^2$ translates to

$$\text{minimize} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} \, \varphi_{ij} \tag{18}$$

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 $\varphi_{ij} := \varphi\left(u_i \rightarrow v_j\right)$ associated with cost c_{ij}. Clearly, in addition to constraints (C1)–(C3), (18) 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, (18) strives to find that particular joint PMF which minimizes the total cost for transporting the probability mass while respecting the normality condition.

The LP set up described above does not require the PDFs under comparison ($\eta\left(y, t\right)$ and $\widehat{\eta}\left(y, t\right)$), to be represented by the same number of samples. For the purpose of model validation, this flexibility of Wasserstein distance offers practical advantages since experimental data are often expensive to gather. For example, such situations occur for gene expression data in biomedical research and atmospheric data in planetary research, which entail '*large dimension small sample problem*'. However, model based simulation can harness the computational resources and hence simulation sample size is often larger than that of experimental data. Notice that pointwise pseudo-metrics like D_{KL} must work with same number of samples for the two distributions.

V. Construction of Validation Certificates

Till now, we have described a validation framework that propagates initial condition, parametric and model uncertainty by evolving the distribution over extended state space and compares the model-predicted instantaneous joint distribution over the output space with that predicted by experiments. We argued that such a comparison aims to measure the shape difference between the transient output PDFs and hence Wasserstein distance, endowed with a suitable metric over the state space, was introduced as a tool for the same. Notice that uncertainty need not be an intrinsic attribute of the proposed model, rather it's a construct imposed by the validation framework. A natural question to ask: how robust are the conclusions of this method? If the initial density is altered, does the inference change drastically? If yes, then how many initial densities are enough to make a provably correct validation guarantee? And finally, can we provide quantitative validation certificates? The vehicle to answer these questions is the ϵ δ analysis of randomized algorithms [59], as we demonstrate next.

A. 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 it's exact shape (e.g. normal, semi-circular etc.) may not be known. Even when the distribution-type is known (e.g. normal), it's 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 $\#(E) = 1$. Δ can be thought of as a categorical random variable which 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 $\xi_0 \sim \mathcal{N}(\mu, \sigma^2)$ with a given joint distribution over the μ σ^2 space, then in our model validation framework, one sample from this space will return one distance measure between the instantaneous output PDFs. How many such (μ, σ^2) samples are necessary to guarantee the robustness of the model validation oracle? The Chernoff bound [60] provides such an estimate for finite sample complexity.

At time step k, let the *validation probability* be $p(\gamma_k) := \mathbb{P}({}_2W_2(\eta_k(y), \widehat{\eta}_k(y)) \leqslant \gamma_k)$. Here $\gamma_k \in \mathbb{R}^+$ is the instantaneous tolerance level (margin of safety). If the model validation is performed by drawing finite N samples from Ω, then the *empirical validation probability* is

$\widehat{p}_N(\gamma_k) := \frac{1}{N}\sum_{i=1}^{N} \chi_{V_k^{(i)}}$ where $V_k^{(i)} := \{\widehat{\eta}_k^{(i)}(y) \,:\, {}_2W_2\left(\eta_k^{(i)}(y), \widehat{\eta}_k^{(i)}(y)\right) \leqslant \gamma_k\}$. Consider $\epsilon, \delta \in (0,1)$ as the desired accuracy and confidence, respectively.

Lemma 2: (**Chernoff bound**) (p. 123, [59]) For any $\epsilon, \delta \in (0,1)$, if $N \geqslant N_{ch} := \frac{1}{2\epsilon^2}\log\frac{2}{\delta}$, 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

Algorithm 1 Construct PRVC

Require: $\epsilon, \delta \in (0,1)$, T, ν, law of Δ, experimental data $\{\eta_k(y)\}_{k=1}^T$, model, tolerance vector $\{\gamma_k\}_{k=1}^T$

1: $N \leftarrow N_{\text{ch}}(\epsilon, \delta)$ ▷ Determine sample size of initial densities using lemma 2
2: 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 Δ ▷ Use MCMC
3: **for** $k = 1$ to T **do** ▷ Index for time step
4: **for** $i = 1$ to N **do** ▷ Index for initial density
5: **for** $j = 1$ to ν **do** ▷ Index for samples in the extended state space, drawn from $\xi_0^{(i)}(\widetilde{x})$
6: Propagate states using dynamics
7: Propagate measurements
8: **end for**
9: Propagate state PDF ▷ Use (3), (7), (9) or (11)
10: Compute instantaneous output PDF ▷ Algebraic transformation
11: Compute ${}_2W_2\left(\eta_k^{(i)}(y), \widehat{\eta}_k^{(i)}(y)\right)$ ▷ Distributional comparison by solving LP (18) s.t. (C0)–(C3)
12: sum $\leftarrow 0$ ▷ Initialize
13: **if** ${}_2W_2\left(\eta_k^{(i)}(y), \widehat{\eta}_k^{(i)}(y)\right) \leqslant \gamma_k$ **then** ▷ Check if valid
14: sum $\leftarrow$ sum + 1
15: **else**
16: do nothing
17: **end if**
18: **end for**
19: $\widehat{p}_N(\gamma_k) \leftarrow \frac{\text{sum}}{N}$ ▷ Construct PRVC vector of length $T \times 1$
20: **end for**

that the model is valid at time 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. In practice, $\{\gamma_k\}_{k=1}^T$ is often specified as percentage tolerance. Since ${}_2W_2\left(\eta_k^{(i)}(y), \widehat{\eta}_k^{(i)}(y)\right) \in [0, \text{diam}(\mathcal{D}_k)]$, where $\mathcal{D}_k := \mathcal{D}_k^{\text{experiment}} \times$

$\mathcal{D}_k^{\text{model}}$ is the product of the experimental and model state space at time step k, a normalized Wasserstein distance $\dfrac{{}_2W_2}{\text{diam}\,(\mathcal{D}_k)}$ can be employed for comparison purposes. Thus the framework enables us to compute a provably correct validation certificate on the face of uncertainty with finite sample complexity.

B. Computational Complexity for Wasserstein Distance

1) Sample Complexity: In algorithm 1, we provided sample complexity bounds for number of initial PDFs (N) to guarantee performance on the face of initial uncertainty. However, we did not specify the bounds for number of samples ν for a given initial PDF. Since the finite ν estimate of Wasserstein distance is a random variable, we need to answer how large should ν be, in order to guarantee that the empirical estimate of Wasserstein distance obtained by solving the LP (18), (C1)–(C3) with $m = n = \nu$, is close to (12) in probability. In other words, given $\epsilon, \delta \in (0,1)$ and a fixed $i = 1, \ldots, N$, we want to estimate a lower bound of ν as a function of ϵ and δ, such that

$$\mathbb{P}\left(\left|{}_2W_2\left(\eta_k^{(i)}(y), \widehat{\eta}_k^{(i)}(y)\right) - \left[\inf_{\mu\in\mathcal{M}(\eta,\widehat{\eta})}\int_{\mathbb{R}^n\times\mathbb{R}^n}\|\underline{x}-\underline{y}\|_2^2\; d\mu\left(\underline{x},\underline{y}\right)\right]^{1/2}\right| < \epsilon\right) > 1-\delta. \quad (19)$$

One significant work in this direction is due to [51] who proved the consistency and sample complexity results (see Corollary 9(i) and Corollary 12(i) in that paper) for Wasserstein distance of order $q = 1$ using the Kantorovich-Rubinstein duality. From Hölder's inequality, $W_{q_2} > W_{q_1}$ for $q_2 > q_1$, and hence their sample complexity bound, in general, does not hold for $q = 2$.

Before deriving a bound of the form (19), we first make few notational simplifications. In this subsection, we will denote $\eta_k^{(i)}(y)$ and $\widehat{\eta}_k^{(i)}(y)$ by η and $\widehat{\eta}$, and their finite sample representations by η_m and $\widehat{\eta}_n$ respectively. Since Wasserstein distance is a metric, from triangle inequality

$$\begin{aligned}
{}_2W_2(\eta_m,\widehat{\eta}_n) &\leqslant {}_2W_2(\eta_m,\eta) + {}_2W_2(\widehat{\eta}_n,\eta)\;\; (\because {}_2W_2(\eta,\widehat{\eta}_n) = {}_2W_2(\widehat{\eta}_n,\eta)) \\
&\leqslant {}_2W_2(\eta_m,\eta) + {}_2W_2(\widehat{\eta}_n,\widehat{\eta}) + {}_2W_2(\eta,\widehat{\eta}) \\
\Rightarrow {}_2W_2(\eta_m,\widehat{\eta}_n) - {}_2W_2(\eta,\widehat{\eta}) &\leqslant {}_2W_2(\eta_m,\eta) + {}_2W_2(\widehat{\eta}_n,\widehat{\eta}). \qquad (20)
\end{aligned}$$

Since ${}_2W_2(\eta_m,\widehat{\eta}_n)$ is a random variable, the LHS of (20) is a random variable, which we denote as X. Further, if we denote the random variables ${}_2W_2(\eta_m,\eta)$ as Y, and ${}_2W_2(\widehat{\eta}_n,\widehat{\eta})$ as Z, then (20) can be seen as a probabilistic inequality, i.e. $X(\omega) \leqslant Y(\omega) + Z(\omega)\;\forall\,\omega\in\Omega$. It can be noted that X, Y and Z are independent random variables. Further, observe that Y and Z are

non-negative but X need not be. However, since (20) holds for all $\omega \in \Omega$, we can relabel X as the absolute value of the LHS of (20). Otherwise, if X is negative, (20) is trivially satisfied.

Lemma 3: If X, Y, Z are non-negative independent random variables such that $X \leqslant Y + Z$, then for $\epsilon > 0$, we have

$$\mathbb{P}\left(X > \epsilon\right) \leqslant \mathbb{P}\left(Y + Z > \epsilon\right) \leqslant \mathbb{P}\left(Y > \frac{\epsilon}{2}\right) + \mathbb{P}\left(Z > \frac{\epsilon}{2}\right). \tag{21}$$

To proceed next, we need the following result that goes beyond the traditional asymptotic convergence [61] of empirical probability measure in Wasserstein sense.

Theorem 3: (**Rate-of-convergence of empirical measure in Wasserstein metric**)(Thm. 5.3, [62]) For a probability measure $\rho \in T_q\left(C\right)$, $1 \leqslant q \leqslant 2$, and its n-sample estimate ρ_n, we have

$$\mathbb{P}\left({}_pW_q\left(\rho, \rho_n\right) > \theta\right) \leqslant K_\theta \, \exp\left(-\frac{n\theta^2}{8C}\right), \tag{22}$$

where $\theta > 0$, and the constant K_θ is obtained by solving the optimization problem $\log K_\theta := \frac{1}{C} \inf_{\nu} \text{card}\left(\text{supp}\, \nu\right) \left(\text{diam}\left(\text{supp}\, \nu\right)\right)^2$. The optimization takes place over all probability measures ν of finite support, such that ${}_pW_q\left(\rho, \nu\right) \leqslant \theta/4$.

Theorem 4: (**Rate-of-convergence of empirical Wasserstein estimate**)

$$\mathbb{P}\left(\left| {}_2W_2\left(\eta_m, \widehat{\eta}_n\right) - {}_2W_2\left(\eta, \widehat{\eta}\right) \right| > \epsilon\right) \leqslant K_1 \, \exp\left(-\frac{m\epsilon^2}{32C_1}\right) + K_2 \, \exp\left(-\frac{n\epsilon^2}{32C_2}\right). \tag{23}$$

Remark 7: At a fixed time, K_1, K_2, C_1 and 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 C_1 and C_2 depend on the dynamics via respective PDF evolution operators, described in Section 3. 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 $K_1 = K_2 = K$, $C_1 = C_2 = C$, to compare the nature of the bound in (23) vis-a-vis with Lemma 2.

Corollary 2: (**Sample complexity for empirical Wasserstein estimate**) For desired accuracy $\epsilon \in (0, 1)$, and confidence $1 - \delta$, $\delta \in (0, 1)$, the sample complexity $m = n = N_{\text{wass}}$, for finite sample Wasserstein computation is given by

$$N_{\text{wass}} = \left(\frac{32C}{\epsilon^2}\right) \log\left(\frac{2K}{\delta}\right). \tag{24}$$

Proof: Assuming $K_1 = K_2 = K$, $C_1 = C_2 = C$, from (19) and (23), we get

$$2K \exp\left(-\frac{N_{\text{wass}}\epsilon^2}{32C}\right) = \delta \Rightarrow N_{\text{wass}} = \left(\frac{32C}{\epsilon^2}\right) \log\left(\frac{2K}{\delta}\right). \tag{25}$$

■

Remark 8: In view of the sample complexity for computing Wasserstein distance, in Algorithm 1, we must set $\nu \leftarrow N_{\text{wass}}(\epsilon, \delta)$. Further, since (24) can be rewritten as $N_{\text{wass}} = 32C\left(2N_{\text{ch}} + \frac{1}{\epsilon^2}\log K\right)$, the numerical value of $C > 0$ governs whether $N_{\text{wass}} > N_{\text{ch}}$.

2) Runtime Complexity: For $m = n = \nu$, the LP formulation (18), (C1)–(C3) requires solving for ν^2 unknowns subject to $(\nu^2 + 2\nu)$ constraints. It can be shown that [63], [64] the runtime complexity for solving the LP is $O(d\,\nu^{2.5}\log\nu)$. Notice that the dimension d only enters through the cost c_{ij} in (18) and hence affects the computational time linearly.

3) Storage Complexity: For $m = n = \nu$, the constraint matrix for the linear program (18), (C1)–(C3), is a binary matrix of size $2\nu \times \nu^2$, whose each row has ν ones (i.e. each column has 2 ones). Consequently, there are total $2\nu^2$ ones in the constraint matrix and the remaining $2\nu^2(\nu - 1)$ elements are zero. Hence at any fixed time, the sparse representation of the constraint matrix needs # non-zero elements $\times 3 = 6\nu^2$ storage. The PMF vectors $\{\alpha_i\}_{i=1}^m$ and $\{\beta_j\}_{j=1}^n$ are, in general, fully populated. In addition, we need to store the model and true sample coordinates, each of them being a d-tuple. Hence at any fixed time, constructing c_{ij} requires storing $2d\nu$ values. Thus total storage complexity at any given snapshot, is $2\nu(3\nu + d + 1) = O(\nu^2)$, assuming $\nu > d$. Notice however that if the sparsity of constraint matrix is not exploited by the solver, then storage complexity rises to $2\nu(\nu^2 + d + 1) = O(\nu^3)$. For example, if we take $\nu = 1000$ samples and use IEEE 754 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 LP solver.

C. From Distributed Temporal Inference to Validation Oracle

Let $\mathbf{Valid}_k$ be the robust validation event at time k that $|p(\gamma_k) - \widehat{p}_N(\gamma_k)| < \epsilon$. Since $\mathbb{P}(|p(\gamma_k) - \widehat{p}_N(\gamma_k)| < \epsilon) > 1 - \delta$, we have $\mathbb{P}(\mathbf{Valid}_k) > 1 - \delta \Rightarrow \mathbb{P}(\mathbf{Invalid}_k) < \delta$. From union bound

$$\mathbb{P}\left(\bigcup_{k=1}^{T} \mathbf{Valid}_k\right) \leqslant \sum_{k=1}^{T} \mathbb{P}(\mathbf{Invalid}_k) < T\delta. \tag{26}$$

Since $\mathbb{P}\left(\bigcap_{k=1}^{T} \textbf{Valid}_k\right) + \mathbb{P}\left(\bigcup_{k=1}^{T} \textbf{Invalid}_k\right) = 1$, (26) results $\mathbb{P}\left(\bigcap_{k=1}^{T} \textbf{Valid}_k\right) > 1 - T\delta$, which provides a lower bound on the event that the model is validated at all instances of data availability. However, this crude bound is informative only when $T\delta \leqslant 1$. For large T, even $T\delta = 1$ may not be guaranteed without making δ prohibitively small, i.e. it's not obvious whether there is a non-zero probability that the model is validated at all times. To obtain a practically meaningful bound, we resort to Lovász local lemma [65], which is useful in our context since the validation/invalidation events at time k and $k+1$ have Markov dependence.

Lemma 4: (**Symmetric version of Lovász local lemma**) [65] If $B_1, B_2, \ldots, B_n$ are events on an arbitrary probability space such that

1) $\mathbb{P}(B_i) \leqslant p, \quad \forall\, i = 1, \ldots, n,$
2) each event is dependent on **at most** D other events (i.e. degree of the dependency graph (p. 139, [66]) is upper bounded by D),
3) $e\,p\,(D+1) \leqslant 1$, where $e = 2.718...$ is the base of the natural logarithm,

then $\mathbb{P}\left(\bigcap_{i=1}^{n} B_i^c\right) > 0$.

Remark 9: A stronger version [67] of Lemma 4, weakened the inequality in condition 3 to $\dfrac{p\,D^D}{(D-1)^{D-1}} \leqslant 1$ for $D \neq 1$, and $p \leqslant \dfrac{1}{2}$ otherwise. This result is optimal.

Next we show that for memoryless dynamics, under mild and intuitive condition on δ, the model is validated with non-zero probability.

Theorem 5: (**Instantaneous validation with** 50% **confidence is enough**) If $\mathbb{P}(\textbf{Invalid}_k) < \delta \leqslant \dfrac{1}{2},\ \forall\, k = 1, \ldots, T$, then $\mathbb{P}\left(\bigcap_{k=1}^{T} \textbf{Valid}_k\right) > 0$.

Proof: Since the dynamics and hence the density evolution is Markov, the dependency graph over events $\textbf{Invalid}_k$, $k = 1, \ldots, T$, is an irreducible tree with chain structure, and hence has degree $D = 1$. From Lemma 4 and Remark 9, the result follows. ■

VI. Illustrative Examples

A. Example 3: Continuous-time Deterministic Model

Consider the following nonlinear dynamical system

$$\ddot{x} = -ax - b\sin 2x - c\dot{x} \tag{27}$$

with parametric values $a = 0.1$, $b = 0.5$, and $c = 1$. The system has five fixed points of the form $(x_1^*, 0)$, where $b \sin 2x_1^* = -ax_1^*$. Solution to this transcendental equation can be found 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 8. It is easy to verify that the origin is a stable focus while the remaining fixed points are saddles (Fig. 9). To illustrate the model validation framework, let's assume that 'true

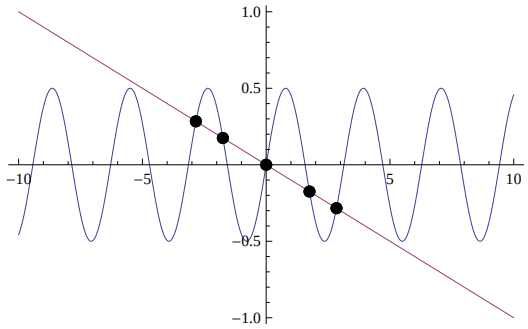

Fig. 8. Points of intersection of the curve $f(x) = b \sin 2x$ and the line $g(x) = -ax$.

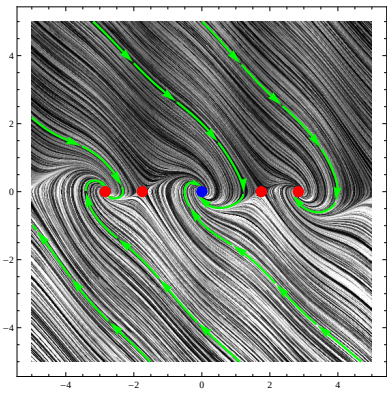

Fig. 9. Phase portrait of the vector field (27) and its one stable (origin) and four saddle fixed points.

data' is generated by the dynamics (27) and the proposed model is a linearization of (27) about the origin. For simplicity, we take the states as the outputs, in both true and model dynamics. Starting from the bivariate uniform distribution $\mathcal{U}([-4, 6] \times [-4, 6])$, we evolve the respective joint PDFs through true and model dynamics via MOC implementation of Liouville equation [25]. The distributional shape discrepancy is captured via the normalized Wasserstein distance $\frac{{}_2W_2(\eta_k, \widehat{\eta}_k)}{\text{diam}(\mathcal{D}_k)}$, between these instantaneous joint PDFs, as shown in Fig. 10. As the individual joint PDFs converge toward their respective stationary densities, the slope of the Wasserstein time-history decreases progressively.

B. Example 4: Continuous-time Stochastic Model

For the sake of demonstration, we assume the true data being generated by (27) with an additive white noise. Letting $x_1 = x$ and $x_2 = \dot{x}$, the associated Itô SDE can be written in state-space form similar to (5)

$$\begin{Bmatrix} dx_1 \\ dx_2 \end{Bmatrix} = \begin{Bmatrix} x_2 \\ (-ax_1 - b \sin 2x_1 - cx_2) \end{Bmatrix} dt + \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} dW, \tag{28}$$

where the Wiener process $W(t)$ satisfies $\mathbb{E}[dW] = 0$ and $\mathbb{E}[dW^2] = Q = \dfrac{\sqrt{2\pi}}{10}$ (say). The stationary Fokker-Planck equation for (28) can be solved in closed form

$$\eta_\infty(x_1, x_2) = \xi_\infty(x_1, x_2) \propto \exp\left(-\frac{c}{Q}\left(ax_1^2 + x_2^2 - b\cos 2x_1\right)\right), \tag{29}$$

and one can verify that peaks of (29) appear at the fixed points of the nonlinear dynamics.

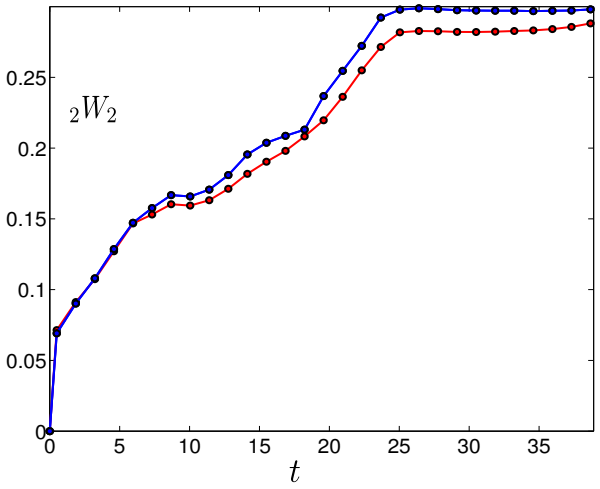

Fig. 10. Time history of normalized Wasserstein distance measured between the joint PDFs for (27) and its linearization about the origin. The curve on top results when α_i and β_j in LP constraints (C1) and (C2) are computed from gridded PMFs. The curve below results by constructing the same from scattered joint PDF data reslting from meshless propagation.

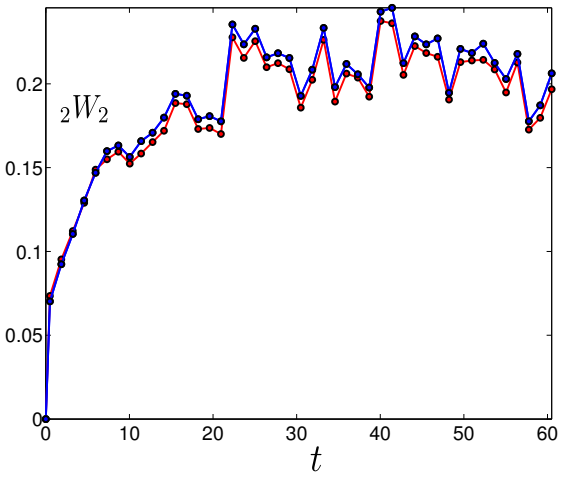

Fig. 11. Time history of normalized Wasserstein distance measured between the joint PDFs for (28) and (30). The convention for two curves are same as in Fig. 10. Notice that in presence of process noise, $_2W_2$ takes longer to settle down compared to Fig. 10.

Let the proposed model is a linearization of (28) about the origin. It is well-known [68] that the stationary density of a linear SDE of the form

$$dx = Ax\,dt + B\,dW, \qquad x \in \mathbb{R}^n, \quad W \in \mathbb{R}^m, \tag{30}$$

is given by

$$\widehat{\eta}_\infty(x) = \mathcal{N}(\mathbf{0}, \Sigma_\infty) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_\infty)}}\, e^{-1/2\; x^\dagger \Sigma_\infty^{-1} x}, \tag{31}$$

provided A is Hurwitz and (A, B) is a controllable pair. The steady-state covariance matrix Σ_∞ satisfies

$$A\Sigma_\infty + \Sigma_\infty A^\dagger + BB^\dagger = 0. \tag{32}$$

For the linearized version of (28), $A = \begin{bmatrix} 0 & 1 \\ (-a-2b) & c \end{bmatrix}$ and $B = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix}$ satisfy the aforementioned conditions and the stationary density is obtained from (31) and (32).

Taking the initial density same as in Example 1, we propagated the joint PDFs for (28) and (30) using the KLPF method described in [27]. Fig. 11 shows that the addition of noise delays the saturation of Wasserstein distance due to slowing down of the rate-of-convergence for the transient PDFs toward their respective stationary densities.

C. Example 5: Discrete-time Deterministic Model

Let the true data being generated by the Chebyshev map [69]

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

where $\mathcal{T} : [-1,1] \mapsto [-1,1]$. The Perron-Frobenius operator for (33) is given by [70]

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

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

$$x_{k+1} = \widehat{\mathcal{T}}(x_k) = 4x_k(1-x_k), \tag{35}$$

where $\widehat{\mathcal{T}} : [0,1] \mapsto [0,1]$. The Perron-Frobenius operator for (35) is given by [28]

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

with stationary density $\widehat{\xi}_\infty(x) = \dfrac{1}{\pi\sqrt{x(1-x)}}$, and CDF $G_\infty(x) = \dfrac{2}{\pi}\sin^{-1}\left(\sqrt{x}\right) = \dfrac{1}{2} + \dfrac{\sin^{-1}(2x-1)}{\pi}$. Taking the outputs identical to states, the asymptotic Wasserstein distance between (33) and (35) becomes

$$\begin{aligned} {}_2W_2\left(\xi_\infty(x), \widehat{\xi}_\infty(x)\right) &= \sqrt{\int_0^1 \left(F_\infty^{-1}(t) - G_\infty^{-1}(t)\right)^2 dt} = \sqrt{\int_0^1 \left(\frac{1}{2} - \frac{1}{2}\cos(\pi t) + \cos t\right)^2 dt} \\ &= \sqrt{\frac{7}{8} + \sin(1) - \frac{\sin(1)}{\pi^2-1} + \frac{\sin(2)}{4}} \approx 1.36. \end{aligned} \tag{37}$$

Given an initial density, ${}_2W_2\left(\xi(x,t), \widehat{\xi}(x,t)\right)$ can be computed between (34) and (36).

D. Example 6: Discrete-time Stochastic Model

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 [71], [72]. The Perron-Frobenius operator for (38) is given by (p. 330-331, [28])

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

with the (multiplicative) stochastic kernel

$$\mathcal{K}_{\text{mul}}(x_k, y) = \frac{1}{\mathcal{S}(y)}\, \phi\left(\frac{x_k}{\mathcal{S}(y)}\right). \tag{40}$$

In particular, $\zeta_k \sim \mathcal{N}(0,1)$ results $\mathscr{P}\xi(x_k) = \displaystyle\int_0^\infty \xi(y) \frac{1}{\sqrt{2\pi}\, x_k (1 - x_k)}\, e^{-\frac{1}{2(1-x_k)^2}}\, dy$. The asymptotic behavior of (38) is known [71] to depend on the noise density $\phi(.)$. Specifically, $\mathbb{E}[\log \zeta_0] < 0, = 0$, and > 0 implies $x_k \xrightarrow{\text{a.s.}} 0$, $x_k \xrightarrow{\text{p}} 0$, 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

$$x_{k+1} = \widehat{\mathcal{T}}(x_k, \zeta_k) = \widehat{\mathcal{S}}(x_k) + \zeta_k = x_k + \zeta_k, \tag{41}$$

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

$$\mathscr{P}\widehat{\xi}(x_k) = \int_{-\infty}^\infty \widehat{\xi}(y)\, \mathcal{K}_{\text{add}}(x_k, y)\ dy, \tag{42}$$

with the (additive) stochastic kernel

$$\mathcal{K}_{\text{add}}(x_k, y) = \phi\left(x_k - \widehat{\mathcal{S}}(y)\right). \tag{43}$$

Consequently, the Perron-Frobenius operator for (41) is $\mathscr{P}\widehat{\xi}(x_k) = \displaystyle\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_k - y)^2}{2}\right) \widehat{\xi}(y)\, dy$. It can be verified (p. 325, [28]) that $\mathscr{P}^k\widehat{\xi}$ 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 (42). This example demonstrates that (in)validating a stochastic map has sensitive dependence on noise density.

E. Example 7: Comparison with Barrier Certificate based Model Validation

Consider the nonlinear model validation problem Example 4 in [5], where the model is $\dot{x} = -px^3$, with parameter $p \in \mathcal{P} = [0.5, 2]$. The measurement data are $\mathcal{X}_0 = [0.85, 0.95]$ 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 through sum-of-squares optimization [73] 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 invalidated by the existence of such certificate, i.e. the model $\dot{x} = -px^3$, with parameter $p \in \mathcal{P}$ was deemed 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 $\xi(x,p,t)$ over the extended state space $\widetilde{x} = [x \;\; p]^\dagger$, with initial support $\widetilde{\mathcal{X}}_0 := \mathcal{X}_0 \times \mathcal{P}$. MOC implementation of the Liouville equation [25] yields

$$\xi(x,p,t) = \xi_0(x_0,p) \exp\left(- \int_0^t \text{div}\left(\widetilde{f}(x(\tau),\tau) \right) d\tau \right). \tag{44}$$

For the model dynamics $\dot{x} = -px^3$, we have $\text{div}\left(\widetilde{f}(x(\tau)) \right) = -3p(x(\tau))^2$ and $\frac{1}{x^2} = \frac{1}{x_0^2} + 2pt$. Consequently (44) results

$$\xi(x,p,t) = \xi_0(x_0,p)\left(1 + 2x_0^2 pt\right)^{3/2} = \frac{1}{(1-2x^2pt)^{3/2}} \xi_0\left(\pm \frac{x}{\sqrt{1-2x^2pt}}, p \right). \tag{45}$$

In particular, for $\xi_0(x_0,p) \sim \mathcal{U}(x_0,p) = \frac{1}{\text{vol}\left(\widetilde{\mathcal{X}}_0\right)}$, $\xi_T(x_T,p,T) \sim \mathcal{U}(x_T,p) = \frac{1}{\text{vol}\left(\widetilde{\mathcal{X}}_T\right)}$ and $T = 4$, (45) requires us to satisfy

$$\left(1 - 8x_T^2 p\right) = \left(\frac{\text{vol}\left(\widetilde{X}_T\right)}{\text{vol}\left(\widetilde{X}_0\right)} \right)^{2/3} > 0 \Rightarrow 1 > 8x_T^2 p. \tag{46}$$

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 $1 > 8(x_T)_{\min}^2 p_{\min} = 8 \times (0.55)^2 \times 0.5 = 1.21$, which is absurd. Thus the PDF $\xi_T(x_T,p,T) \sim \mathcal{U}(x_T,p)$ is not finite-time reachable from $\xi_0(x_0,p) \sim \mathcal{U}(x_0,p)$ for $T = 4$, via the proposed model dynamics. Hence our measure-theoretic formulation recovers Prajna's invalidation result as a special case. Instead of binary validation/invalidation oracle, we can now measure the degree of

validation by computing the Wasserstein distance ${}_2W_2\left(\frac{1}{(1-2x_T^2pT)^{3/2}}\frac{1}{\text{vol}\left(\widetilde{\mathcal{X}}_0\right)},\frac{1}{\text{vol}\left(\widetilde{\mathcal{X}}_T\right)}\right)$ 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 same supports $\widetilde{\mathcal{X}}_0$ and $\widetilde{\mathcal{X}}_T$. Indeed, given a joint PDF $\xi\left(x_T,p,T\right)$ supported over $\widetilde{\mathcal{X}}_T$ at $T=4$, from (45) 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.

$$\xi_0\left(x_0,p\right)=\frac{1}{\left(1+8x_0^2p\right)^{3/2}}\,\xi\left(\pm\frac{x_0}{\sqrt{1+8x_0^2p}},p,4\right). \tag{47}$$

In other words, if the measurements find the initial density given by (47) 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.

VII. Conclusion

We have presented a probabilistic model validation framework for nonlinear systems, in both discrete and continuous-time. The notion of soft validation allows us to quantify the degree of mismatch 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. In Section IV.B, we provided some theoretical details with examples to bring forth the efficacy of the same compared to information-theoretic (pseudo)metrics like Kullback-Leibler divergence. These results have non-trivial implications from information geometric perspective, not pursued in this paper. For example, our results imply that geodesic distance on the manifold of output probability distributions, does not serve well as a validation metric.

From practical point of view, the framework presented here applies to any deterministic or stochastic nonlinearity, not necessarily semialgebraic type. The computational cost and quality of probabilistic inference have been guaranteed. Additional simulation results and proofs for Theorem 1, Theorem 2, Corollary 1, Lemma 3 and Theorem 4 can be found at `http://people.tamu.edu/~ahalder/ModelValidation`.

ACKNOWLEDGMENT

The authors would like to thank B. K. Sriperumbudur at University College London, and S. Chakravorty at Texas A&M University, for insightful discussions. This research was supported through National Science Foundation award # 1016299, with D. Helen Gill as the program manager.

REFERENCES

[1] D. Georgiev, and E. Klavins, "Model Discrimination of Polynomial Systems via Stochastic Inputs", *IEEE Conference on Decision and Control*, Cancun, Mexico, Dec. 9–11, 2008.

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

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

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

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

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

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

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] D. Liberzon, and R.W. Brockett, "Spectral Analysis of Fokker-Planck and Related Operators Arising from Linear Stochastic Differential Equations". *SIAM Journal of Control and Optimization*, Vol. 38, No. 5, pp. 1453–1467, 2000.

[24] H. Risken, *The Fokker-Planck Equation: Methods of Solution and Applications.* Springer, Second ed., 1989.

[25] 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.

[26] 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.

[27] P. Dutta, A. Halder, and R. Bhattacharya, "Uncertainty Quantification for Stochastic Nonlinear Systems using Perron-Frobenius Operator and Karhunen-Loève Expansion". *preprint*, 2011.

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

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

[32] R.M. Dudley, *Real Analysis and Probability.* Cambridge University Press, Second ed., 2002.

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

[34] 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.

[35] S. Ferson, and W.L. Oberkampf, "Validation of Imprecise Probability Models". *International Journal of Reliability and Safety*, Vol. 3, No. 1, pp. 3–22, 2009.

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

[37] G. Zang, and P.A. Iglesias, "Nonlinear Extension of Bode's Integral Based on An Information Theoretic Interpretation". *Systems & Control Letters*, Vol. 50, No. 1, pp. 11–19, 2003.

[38] Y. Sun, and P.G. Mehta, "Fundamental Performance Limitations via Entropy Estimates with Hidden Markov Models". *IEEE Conference on Decision and Control*, New Orleans, 2007.

[39] P.G. Mehta, U. Vaidya, and A. Banaszuk, "Markov Chains, Entropy and Fundamental Limitations in Nonlinear Stabilization". *IEEE Transactions on Automatic Control*, Vol. 53, No. 3, pp. 784–791, 2008.

[40] S. Kullback, "A Lower Bound for Discrimination in Terms of Variation". *IEEE Transactions on Information Theory*, Vol. 13, No. 1, pp. 126–127, 1967.

[41] 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.

[42] P. Cattiaux, and A. Guillin, "Criterion for Talagrand's Quadratic Transportation Cost Inequality". *preprint*, 2003, arXiv:math/0312081v3.

[43] 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.

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

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

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

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

[48] 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.

[49] 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.

[50] D. Zhou, and T. Shi, "Statistical Inference based on Distances between Empirical Distributions". *Preprint*, Available at `http://www.stat.osu.edu/~taoshi/research/papers/Zhou_and_Shi_TR_848_2010.pdf`, 2010.

[51] 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.

[52] 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.

[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] V.M. Zolotarev, "Probability Metrics". *Theory of Probability and its Applications*, Vol. 28, pp. 278–302, 1983.

[55] R.E. Burkard, B. Klinz, and R. Rudolf, "Perspectives of Monge Properties in Optimization". *Discrete Applied Mathematics*, Vol. 70, No. 2, pp. 95–161, 1996.

[56] 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.

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

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

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

[60] H. Chernoff, "A Measure of Asymptotic Efficiency for Tests of A Hypothesis Based on the Sum of Observations". *Annals of Mathematical Statistics*, Vol. 23, pp. 493–507, 1952.

[61] G. Biau, L. Devroye, and G. Lugosi, "On the Performance of Clustering in Hilbert Spaces". *IEEE Transactions on Information Theory*, Vol. 54, No. 2, 2008.

[62] 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.

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

[64] 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.

[65] N. Alon, and J.H. Spencer, *The Probabilistic Method.* Wiley-Interscience, Second ed., 2000.

[66] M. Mitzenmacher, and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis.* Cambridge University Press, First ed., 2005.

[67] J.B. Shearer, "On A Problem of Spencer". *Combinatorica*, Vol. 5, No. 3, pp. 241–245, 1985.

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

[70] 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.

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

[72] 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.

[73] 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.