Further Results on Probabilistic Model Validation in Wasserstein Metric

Abhishek Halder and Raktim Bhattacharya

***Abstract*— In a recent work [1], we have introduced a probabilistic formulation for the model validation problem to provide a unifying framework for (in)validating nonlinear deterministic and stochastic models, in both discrete and continuous time. As an extension to that work, this paper provides rigorous performance bounds for the model validation algorithms presented in [1]. Further, it is shown that the existing method of barrier certificate based nonlinear invalidation oracle, can be recovered as a special case of the proposed formulation. Some results are derived to quantify the effects of initial uncertainty on the Wasserstein gap. And finally, for discrete-time LTI and LTV systems, upper bounds on Wasserstein distance are derived in terms of the parameters of the systems under comparison, thus providing an offline estimate of the gap.**

I. Introduction

Recently, a probabilistic formulation of the model validation problem was proposed in [1] for nonlinear systems from the perspective of Monge-Kantorovich optimal transport [2]–[4]. Instead of binary invalidation oracle, this framework allows a relaxed notion of validation in probability. As the block diagram in Fig. 1 shows, in this formulation, the systems under comparison are excited with a *known* input signal $u(t)$, and an initial probability density function (PDF) $\xi_0(\widetilde{x})$, supported over the extended state space $\widetilde{x} := \{x,\ p\}^T$, where the states $x \in \mathbb{R}^{n_s}$, and the parameters $p \in \mathbb{R}^{n_p}$. Given the PDF $\eta(y(t))$ supported over the true output space $y \in \mathbb{R}^{n_o}$, and a candidate model, we compute[1] and then compare the model predicted output PDF $\widehat{\eta}(\widehat{y}(t))$, with $\eta(y(t))$ at each instances of measurement availability $\{t_j\}_{j=1}^{\tau}$. In [1], it was argued that the suitable metric for such comparison is L_2 *Wasserstein distance of order two*, denoted by ${}_2W_2$, and computational method for the same was provided therein, for comparing two general nonlinear non-Gaussian systems. The end result is a gap trajectory ${}_2W_2(t)$, which should remain within the user-specified tolerance levels $\{\gamma_j\}_{j=1}^{\tau}$ for validation. Due to finite sample computation, a probabilistic robust validation certificate is computed to guarantee the accuracy of the validation/invalidation oracle.

The main merit of this approach is its flexibility to handle both deterministic and stochastic dynamics, in both discrete and continuous time, without making any assumption about the structure of the nonlinearity (e.g. semi-algebraic) or about the uncertainty (e.g. interval-valued structured uncertainty as in robust control based methods [9]–[11], or set-valued uncertainty as in Barrier certificate method [12]). Also,

Abhishek Halder and Raktim Bhattacharya are with the Department of Aerospace Engineering, Texas A&M University, College Station, TX 77843, USA, {ahalder,raktim}@tamu.edu

[1]The computation of $\widehat{\eta}(\widehat{y}(t))$ calls for uncertainty propagation [5]–[8] in nonlinear systems.

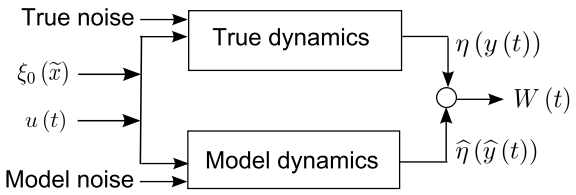

Fig. 1. Block diagram for the probabilistic model validation formulation.

the nature (structured or unstructured) and sources (initial condition, parameters, model error) of uncertainties can be heterogeneous. Since all these could be tackled in a unifying non-parametric framework, it motivates us to take a rigorous look at the sample and storage complexity of the Wasserstein computation for practical applicability. Further, from [1] it was not clear how the proposed formulation relates with existing nonlinear validation methods like [12]. Moreover, given a model dynamics, it remains to quantify how the initial PDF affects ${}_2W_2$. And lastly, given a model pair, it might be of interest to see if one could upper bound ${}_2W_2$ for simple dynamics like discrete-time linear systems, in terms of the parameters of the systems under comparison, as that would enable an offline estimate of the gap. These are the questions we set to answer in this paper.

The contributions of this paper, beyond the prior work [1] are as follows.

1) We provide sample and storage complexity results for computing ${}_2W_2$ at each instance of measurement availability $\{t_j\}_{j=1}^{\tau}$.

2) Using an example from [12], we show that our formulation indeed recovers the invalidation oracles predicted by barrier certificate approach. Notice that, unlike barrier certificates, the computation in the proposed formulation does not suffer from any conservatism when the nonlinearity is not semi-algebraic.

3) We provide a framework to characterize the effect of initial uncertainty on the Wasserstein gap for scalar dynamics, and show in particular, that for scalar linear systems, given a set of admissible initial PDFs, the initial PDF that maximizes Wasserstein gap (at all times), is the one with maximum second raw moment. All higher order moments have no effect on Wasserstein gap.

4) For discrete-time linear Gaussian systems, we derive

upper bounds for ${}_2W_2(k)$, $k \in \mathbb{N} \cup \{0\}$, in terms of system parameters.

The paper is structured as follows. Section 2 provides a short background on Wasserstein distance. The sample complexity and storage complexity results for Wasserstein based comparison of output PDFs, are derived in Section 3. Section 4 is intended to show that the results of existing nonlinear invalidation methods like barrier certificate-based deductive inference, can be recovered in this probabilistic setting. In Section 5, we derive some preliminary results on how the initial PDF can affect the Wasserstein gap depending on the structure of the dynamics. Section 6 provides upper bounds for Wasserstein distance in discrete-time linear model validation setting for both LTI and LTV models; and Section 7 concludes the paper.

Notation: Most notations are standard. $||.||_F$ denotes the matrix Frobenius norm while $||.||_p$ stands for the standard L_p norm. By $\mathrm{tr}\,(.)$, and $\lambda_{\max}(.)$, we denote trace and maximum eigenvalue, respectively. The symbol $\mathrm{vol}\,(.)$ denotes Lebesgue volume.

II. Background on Wasserstein Distance

Let M_1 and M_2 be complete, separable metric (Polish) spaces equipped with p^{th} order distance metric, say the L^p norm. Then the Wasserstein distance of order q, denoted as ${}_pW_q$, between two probability measures μ_1 and μ_2, supported on M_1 and M_2, is defined as

$${}_pW_q(\mu_1,\mu_2) := \left[\inf_{\mu\in\mathcal{M}(\mu_1,\mu_2)} \int_{M_1\times M_2} \|x-y\|_p^q \; d\mu(x,y)\right]^{1/q},$$

where $\mathcal{M}(\mu_1,\mu_2)$ is the set of all probability measures on $M_1 \times M_2$ with first marginal μ_1 and second marginal μ_2. It's well known [13] that on the set of Borel measures on $\mathbb{R}^d$ having finite second moments, ${}_pW_q$ defines a metric. 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 ${}_pW_q(\mu_1,\mu_2)$. This is assumed to hold for all subsequent analysis.

The Wasserstein metric is an integral notion of distance, as opposed to a pointwise notion (e.g. Hellinger distance, Kullback-Leibler (KL) divergence etc.), on the manifold of probability densities. This makes Wasserstein distance a good candidate for model validation, since the supports of the PDFs under consideration, evolving under two different dynamics, are often not identical. However, computation of Wasserstein metric is not straightforward. For real line, a closed form solution exists [14] in terms of the cumulative distribution functions (CDFs) of the test PDFs. Let F and G be the corresponding CDFs of the univariate PDFs ρ_1 and ρ_2 respectively. Then

$${}_pW_q^q(\rho_1,\rho_2) = \int_0^1 \|F^{-1}(\varsigma) - G^{-1}(\varsigma)\|_p^q \, d\varsigma. \tag{1}$$

For multivariate case, in general, one has to compute ${}_pW_q$ from its definition, which can be cast as a linear program (LP) in mn variables with $(m+n+mn)$ constraints, where the respective PDFs have m and n-sample representations. This was detailed in [1], with $p=q=2$. The choice $p=2$ is due to the fact that we measure inter-sample distance in Euclidean metric. The choice $q=2$ will be motivated in the following section.

III. Computational Complexity of Comparing Output PDFs in Wasserstein Distance

A. Sample Complexity

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

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

Similar consistency and sample complexity results are available in the literature (see Corollary 9(i) and Corollary 12(i) in [15]) for Wasserstein distance of order $q=1$. From Hölder's inequality, $W_{q_2} > W_{q_1}$ for $q_2 > q_1$, and hence that sample complexity bound, in general, does not hold for $q=2$.

To proceed, we need the following results.

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

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

Definition 1: (**Transportation cost inequality**) A probability measure μ is said to satisfy the *L^p-transportation cost inequality* (TCI) of order q, if there exists some constant $C>0$ such that for any probability measure ν, ${}_pW_q(\mu,\nu) \leqslant \sqrt{2CD_{KL}(\nu,\mu)}$, where D_{KL} denotes the KL divergence. In short, we write $\mu \in T_q(C)$.

We will need TCI results independent of dimensions. It was observed that [16], T_1 is not well adapted for dimension free bounds but T_2 is. Also, [17] have demonstrated that uncertainty evolution can be seen as a gradient flux of free energy with respect to the Wasserstein metric of order 2. For these reasons, we choose $q=2$.

Theorem 1: (**Rate-of-convergence of empirical measure in Wasserstein metric**)(Thm. 5.3, [18]) For a probability measure $\rho\in T_q(\mathscr{C})$, $1\leqslant q\leqslant 2$, and its n-sample estimate ρ_n, we have

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

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

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

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

$$\mathbb{P}\left(\left|{}_2W_2(\eta_m,\widehat{\eta}_n) - {}_2W_2(\eta,\widehat{\eta})\right| > \epsilon\right) \leqslant K_1 \exp\left(-\frac{m\epsilon^2}{32\mathscr{C}_1}\right) + K_2 \exp\left(-\frac{n\epsilon^2}{32\mathscr{C}_2}\right). \quad (3)$$

Remark 1: At a fixed time, K_1, K_2, $\mathscr{C}_1$ and $\mathscr{C}_2$ are constants in a given model validation problem, i.e. for a given pair of experimental data and proposed model. However, values of these constants depend on true and model dynamics. In particular, the TCI constants $\mathscr{C}_1$ and $\mathscr{C}_2$ depend on the dynamics via respective PDF evolution operators. The constants K_1 and K_2 depend on η and $\widehat{\eta}$, which in turn depend on the dynamics. For pedagogical purpose, we next illustrate the simplifying case $K_1 = K_2 = K$, $\mathscr{C}_1 = \mathscr{C}_2 = \mathscr{C}$, to compare the nature of the bound in (3) vis-a-vis with Lemma 1 in [1].

Corollary 3: (**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

$$N_{\text{wass}} = \left(\frac{32\mathscr{C}}{\epsilon^2}\right) \log\left(\frac{2K}{\delta}\right). \quad (4)$$

B. Storage complexity

For $m = n$, the constraint matrix for the LP described in [1], is a binary matrix of size $2n \times n^2$, whose each row has n ones. Consequently, there are total $2n^2$ ones in the constraint matrix and the remaining $2n^2(n-1)$ elements are zero. Hence at any fixed time, the sparse representation of the constraint matrix needs # non-zero elements $\times 3 = 6n^2$ storage. The probability mass function (PMF) vectors are, in general, fully populated. In addition, we need to store the model and true sample coordinates, each of them being a n_o-tuple. Hence at any fixed time, constructing cost matrix requires storing $2n_o n$ values. Thus total storage complexity at any given snapshot, is $2n(3n+n_o+1) = O(n^2)$, assuming $n > n_o$. However, if the sparsity of constraint matrix is not exploited by the solver, then storage complexity rises to $2n(n^2+n_o+1) = O(n^3)$. For example, if we take $n = 1000$ 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. We have used MOSEK[2] as the LP solver.

[2]available at `www.mosek.com`

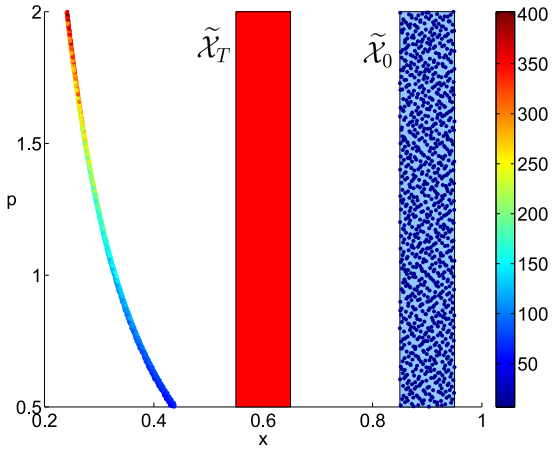

Fig. 2. This plot illustrates how Prajna's barrier certificate-based invalidation result (shown for Example 1), can be recovered in our probabilistic model validation framework. To show $\widetilde{\mathcal{X}}_T$ (*red* rectangle) is not reachable from the set $\widetilde{\mathcal{X}}_0$ (*blue* rectangle) in time $T = 4$, we sample $\widetilde{\mathcal{X}}_0$ uniformly and propagate that uniform ensemble subject to the given dynamics till $T = 4$. The samples are color coded (*red* = high probability, *blue* = low probability) according to the value of the joint PDF at that location. Here, the model is invalidated since the support of the joint PDF at final time and the set $\widetilde{\mathcal{X}}_T$ are disjoint.

IV. Comparison with Barrier Certificate Method

Barrier certificates were introduced by Prajna [12] as a tool for deductive inference based model invalidation. The method is attractive for two reasons. First, it is a non-simulation based invalidation method. If a barrier function, with some desired properties, can be constructed, then the model is invalidated and the existence of such function provides a proof/certificate for the invalidation oracle. Secondly, the applicability of this method has been shown to a broad class of dynamical systems. Naturally, we want to investigate how the present method relates with barrier certificate-based invalidation.

Given the initial probability density and final time, we can write the density at the final time using transfer operator that depends on the prescribed model dynamics. Given initial and final measurements as sets, we can transcribe them to uniform densities supported on those sets. These two uniform densities would then constitute a pair, which must satisfy the transfer operator equation. If not, then the model is invalidated. This probabilistic method is illustrated below through an example.

Example 1: Consider the nonlinear model validation problem Example 4 in [12], 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 [20] 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 shown to be inconsistent with measurements $\{\mathcal{X}_0, \mathcal{X}_T, T\}$.

To tackle this problem in our model validation framework,

consider the spatio-temporal evolution of the joint PDF $\xi(x,p,t)$ over the extended state space $\widetilde{x} = [x \;\; p]^T$, with initial support $\widetilde{\mathcal{X}}_0 := \mathcal{X}_0 \times \mathcal{P}$. Method-of-characteristics implementation of the Liouville equation [5] yields

$$\xi(x,p,t) = \xi_0(x_0,p) \exp\left(-\int_0^t \operatorname{div}\left(\widetilde{f}(x(\tau),\tau)\right) d\tau\right). \quad (5)$$

For the model dynamics $\dot{x} = -px^3$, we have $\operatorname{div}\left(\widetilde{f}(x(\tau))\right) = -3p(x(\tau))^2$ and $\frac{1}{x^2} = \frac{1}{x_0^2} + 2pt$. Consequently, (5) results

$$\begin{aligned}\xi(x,p,t) &= \xi_0(x_0,p)\left(1+2x_0^2pt\right)^{3/2} \\ &= \frac{1}{(1-2x^2pt)^{3/2}}\,\xi_0\left(\pm\frac{x}{\sqrt{1-2x^2pt}},p\right) \quad (6)\end{aligned}$$

In particular, for $\xi_0(x_0,p) \sim \mathcal{U}(x_0,p) = \frac{1}{\operatorname{vol}\left(\widetilde{\mathcal{X}}_0\right)}$, $\xi_T(x_T,p,T) \sim \mathcal{U}(x_T,p) = \frac{1}{\operatorname{vol}\left(\widetilde{\mathcal{X}}_T\right)}$ and $T = 4$, (6) requires us to satisfy

$$\left(1-8x_T^2p\right) = \left(\frac{\operatorname{vol}\left(\widetilde{X}_T\right)}{\operatorname{vol}\left(\widetilde{X}_0\right)}\right)^{2/3} > 0 \Rightarrow 1 > 8x_T^2p. \quad (7)$$

Since $8x_T^2p$ is an increasing function in both $x_T \in \mathcal{X}_T$ and $p \in \mathcal{P}$, we need at least $1 > 8(x_T)^2_{\min} p_{\min} = 8 \times (0.55)^2 \times 0.5 = 1.21$, which is incorrect. Thus the PDF $\xi_T(x_T,p,T) \sim \mathcal{U}(x_T,p)$ is not finite-time reachable (Fig. 2) 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 [12] as a special case. Instead of binary validation/invalidation oracle, we can now measure the degree of validation by computing the Wasserstein distance

$${}_2W_2\left(\frac{1}{(1-2x_T^2pT)^{3/2}}\frac{1}{\operatorname{vol}\left(\widetilde{\mathcal{X}}_0\right)},\; \frac{1}{\operatorname{vol}\left(\widetilde{\mathcal{X}}_T\right)}\right)$$

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 $\xi(x_T,p,T)$ supported over $\widetilde{\mathcal{X}}_T$ at $T = 4$, from (6) we can explicitly compute the initial PDF $\xi_0(x_0,p)$ supported over $\widetilde{\mathcal{X}}_0$ that, under the proposed model dynamics, yields the prescribed PDF, i.e.

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

In other words, if the measurements find the initial density given by (8) and final density $\xi(x_T,p,T)$ at $T = 4$, then the Wasserstein distance at $T = 4$ will be zero, thereby perfectly validating the model.

V. Effect of Initial Uncertainty on ${}_2W_2$

The inference for probabilistic model validation depends on the initial PDF $\xi_0(x_0)$. To account robust inference in presence of initial PDF uncertainty, a notion of *probabilistically robust validation certificate* (PRVC) was introduced in [1]. However, it does not resolve the sensitivity of the gap on the choice of initial PDF. This issue is important, for example, in model discrimination, where one looks for an initial PDF that *maximizes* the gap between two models, which seem to exhibit similar performance. The notion is similar to optimal input design for system identification. In particular, we show that for linear dynamics, the gap is oblivious beyond the first two moments. We restrict ourselves to scalar dynamics for this analysis.

A. Tools for Analysis

The main machinery for univariate analysis is the quantile function $Q_y(\varsigma)$ of the output process, defined as the inverse of the CDF for y. Here $\varsigma \in [0,1]$ denotes probability mass. (1) enables us to write squared Wasserstein distance as the integral of the squared difference of the respective output quantile functions. Further, instead of propagating densities and then transforming them back to quantiles, it would be convenient to directly work with the *quantile Fokker-Planck equation* (QFPE) [22], given by

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

associated with scalar dynamics $dx(t) = f(x)\,dt + g(x)\,d\beta$, where β is the standard Wiener process. Moreover, the quantile transformation rule [23] states that for an algebraic map $y = h(x)$, we have

$$Q_y(\varsigma) = \begin{cases} h \circ Q_x(\varsigma) & \text{if } h \text{ is non-decreasing,} \\ h \circ Q_x(1-\varsigma) & \text{if } h \text{ is non-increasing.} \end{cases} \quad (9)$$

For brevity, we only work out few simple cases to illustrate the idea.

B. Continuous Time Linear Systems

1) Deterministic dynamics: Let the dynamics of the two systems be

$$\dot{x_i} = a_i x, \quad y_i = c_i x, \qquad i = 1,2. \quad (10)$$

Theorem 4: For any initial density $\rho_0(x_0)$, the Wasserstein gap between the systems in (10), is given by

$${}_2W_2(t) = \sqrt{m_{20}}\left|c_1e^{a_1t} - c_2e^{a_2t}\right|, \quad (11)$$

where $m_{20} = \mu_0^2 + \sigma_0^2$, is the second raw moment of $\rho_0(x_0)$, while μ_0 and σ_0 are its mean and standard deviation, respectively.

Proof: For (10), $Q_{y_i} = c_iQ_{x_i}$, and the QFPE reduces to a linear PDE $\partial_t Q_{x_i} = a_iQ_{x_i}$, yielding $Q_{x_i}(\varsigma,t) =$

$Q_0(\varsigma)\, e^{a_i t}$, where Q_0 is the initial quantile function corresponding to ρ_0. Thus, we have

$$\begin{aligned}({}_2W_2(t))^2 &= \int_0^1 (Q_{y_1}(\varsigma,t) - Q_{y_2}(\varsigma,t))^2\, d\varsigma \\ &= \left(c_1 e^{a_1 t} - c_2 e^{a_2 t}\right)^2 \int_0^1 (Q_0(\varsigma))^2\, d\varsigma. \quad (12)\end{aligned}$$

Since the quantile function maps probability to the sample space, hence $x_0 = Q_0(\varsigma)$, and $d\varsigma = \rho_0(x_0)\, dx_0$. Consequently, we can rewrite (12) as

$$({}_2W_2(t))^2 = \left(c_1 e^{a_1 t} - c_2 e^{a_2 t}\right)^2 \underbrace{\int_{-\infty}^{\infty} x_0^2\, \rho_0(x_0)\, dx_0}_{m_{20}}.$$

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

Remark 2: (${}_2W_2$ **has limited dependence on** ρ_0) The above result shows that the Wasserstein gap between scalar linear systems, depends on the initial density up to mean and variance. Any other aspect (skewness, kurtosis etc.) of ρ_0, even when it's non-Gaussian, has no effect on ${}_2W_2(t)$.

Remark 3: (**Linear Gaussian systems**) For the linear Gaussian case, one can verify (20) without resorting to the QFPE. To see this, notice that if $\rho_0(x_0) = \mathcal{N}(\mu_0, \sigma_0^2)$, then the state PDFs evolve as $\rho_{x_i}(x_i,t) = \mathcal{N}(\mu_{x_i}(t), \sigma_{x_i}^2(t))$, where $\mu_{x_i}(t)$ and $\sigma_{x_i}^2(t)$ satisfy their respective state and Lyapunov equations, which, in this case, can be solved in closed form. Since $\rho_{y_i}(y_i,t) = \mathcal{N}(c_i\mu_{x_i}(t), c_i^2\sigma_{x_i}^2(t))$, and ${}_2W_2$ between two Gaussian PDFs is known [24] to be $\sqrt{(\mu_{y_1}-\mu_{y_2})^2 + (\sigma_{y_1}-\sigma_{y_2})^2}$, the result follows.

Remark 4: (**Affine dynamics**) Instead of (10), if the dynamics are given by $\dot{x}_i = a_i x + b_i,\ y_i = c_i x + d_i,\ i = 1,2$, then by variable substitution, one can derive that $Q_{x_i}(\varsigma,t) = Q_0(\varsigma)\, e^{a_i t} + \dfrac{b_i}{a_i}\left(e^{a_i t} - 1\right)$. Hence, we get

$${}_2W_2(t) = \sqrt{(p(t))^2\, m_{20} + 2p(t)\, q(t)\, m_{10} + (q(t))^2}, \quad (13)$$

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

2) Stochastic dynamics: Consider two stochastic dynamical systems with linear drift and constant diffusion coefficients, given by

$$dx_i = a_i x\, dt + b_i\, d\beta, \quad y_i = c_i x, \qquad i = 1,2, \quad (14)$$

where β is the standard Wiener process.

Theorem 5: For any initial density $\rho_0(x_0)$, the Wasserstein gap ${}_2W_2(t)$ between the systems in (14), is given by

$${}_2W_2(t) = \sqrt{(p(t))^2\, m_{20} + 2p(t)\, r(t)\, s(F_0) + (r(t))^2}, \quad (15)$$

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

Proof: For systems (14), quantile functions for the states evolve as (p. 102, [22])

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

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

$$\begin{aligned}({}_2W_2(t))^2 &= \int_0^1 (c_1 Q_{x_1}(\varsigma,t) - c_2 Q_{x_2}(\varsigma,t))^2\, d\varsigma \\ &= (p(t))^2 \int_0^1 (Q_0(\varsigma))^2\, d\varsigma \\ &+ 2p(t)\, r(t) \int_0^1 Q_0(\varsigma)\, Q_N(\varsigma)\, d\varsigma \\ &+ (r(t))^2 \int_0^1 (Q_N(\varsigma))^2\, d\varsigma. \quad (16)\end{aligned}$$

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

$$\begin{aligned}&\int_{-\infty}^{\infty} x_0\, F_N^{-1} \circ F_0(x_0)\ \rho_0(x_0)\, dx_0 \\ &= \sqrt{2}\, \mathbb{E}\left[x_0\, \mathrm{erf}^{-1}(2F_0(x_0) - 1)\right] = s(F_0). \quad (17)\end{aligned}$$

This completes the proof. ∎

Remark 5: (**Gaussian case**) Consider the special case when $\rho_0(x_0) = \mathcal{N}(\mu_0, \sigma_0^2)$. Then $Q_0(\varsigma) = \mu_0 + \sigma_0 Q_N(\varsigma)$, and hence the second integral equals σ_0. Thus, if the initial density is normal, then

$${}_2W_2(t) = \sqrt{(p(t))^2\, m_{20} + 2p(t)\, r(t)\, \sigma_0 + (r(t))^2}, \quad (18)$$

a function of μ_0 and σ_0, a result which can be verified otherwise by solving the mean and variance propagation equations. Hence in stochastic linear case, ${}_2W_2$ depends on the entire initial distribution, unlike its deterministic counterpart.

C. Discrete Time Linear Systems

1) Deterministic dynamics: Let the dynamics of the two systems be given by the maps

$$x_i^{(k+1)} = a_i x_i^{(k)}, \quad y_i^{(k)} = c_i x_i^{(k)}, \qquad i = 1,2, \quad (19)$$

where $k \in \mathbb{N} \cup \{0\}$, denotes the discrete time index.

Theorem 6: For any initial density $\rho_0(x_0)$, the Wasserstein gap between the systems in (19), is given by

$$W(k) = \sqrt{m_{20}}\left|c_1 a_1^k - c_2 a_2^k\right|. \quad (20)$$

Proof: The proof is immediate from linear recursion. ∎

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

A. LTI Bound

Theorem 7: Consider two discrete-time LTI systems $x_{k+1} = Ax_k$, and $\widehat{x}_{k+1} = \widehat{A}\widehat{x}_k$, $k \in \mathbb{N} \cup \{0\}$. Let

the initial PDF $\xi_0(x_0) = \mathcal{N}(0, P_0)$. Then, ${}_2W_2(k) \leqslant \sqrt{2}\left(\text{tr}\left(P_0\right)\right)^{1/2} ||\widehat{A}^{-k}||_F\, \Omega_{\text{LTI}}(k)$, where

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

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

Proof: We know that

$$\begin{aligned} \xi_k &= \mathcal{N}\left(0, A^k P_0 A^{k^T}\right) = \mathcal{N}\left(0, P_k\right), \\ \widehat{\xi}_k &= \mathcal{N}\left(0, \widehat{A}^k P_0 \widehat{A}^{k^T}\right) = \mathcal{N}\left(0, \widehat{P}_k\right). \end{aligned} \tag{21}$$

Therefore,

$$\begin{aligned} &D_{KL}\left(\xi_k || \widehat{\xi}_k\right) = D_{KL}\left(P_k || \widehat{P}_k\right) \\ &= \text{tr}\left(\widehat{P}_k^{-1} P_k - I\right) - \log\det\left(\widehat{P}_k^{-1} P_k\right). \end{aligned} \tag{22}$$

Now if we assume that the spectrum for P_0 is $\{\rho_i\}_{i=1}^{n_s}$, then from (22), $\det\left(P_k\right) = \prod_{i=1}^{n_s} \left(\rho_i \vartheta_i^{2k}\right) \Rightarrow \log\det\left(\widehat{P}_k^{-1} P_k\right) = \log \prod_{i=1}^{n_s} \frac{\vartheta_i^{2k}}{\widehat{\vartheta}_i^{2k}}$. Thus, $D_{KL}\left(P_k || \widehat{P}_k\right) = \text{tr}\left(\widehat{P}_k^{-1} P_k\right) - \log \prod_{i=1}^{n_s} \frac{\vartheta_i^{2k}}{\widehat{\vartheta}_i^{2k}} - n_s$.

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

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

Now to relate D_{KL} with ${}_2W_2$, we invoke the TCI for Gaussian case [21], which states ${}_2W_2(k) \leqslant \sqrt{2\lambda_{\max}\left(\widehat{P}_k^{-1}\right) D_{KL}(k)}$. But $\lambda_{\max}\left(\widehat{P}_k^{-1}\right) \leqslant \text{tr}\left(\widehat{P}_k^{-1}\right) \leqslant ||\widehat{A}^{-k}||_F^2 \text{tr}\left(P_0\right)$. These two, coupled with TCI, results

$${}_2W_2(k) \leqslant \sqrt{2}\left(\text{tr}\left(P_0\right)\right)^{1/2} ||\widehat{A}^{-k}||_F\, \Omega_{\text{LTI}}(k). \tag{23}$$

■

Corollary 8: (**A relaxed bound**) Since $\Omega_{\text{LTI}}(k) \leqslant ||A^k||_F ||\widehat{A}^{-k}||_F \text{tr}\left(P_0\right)$, the above Theorem can be relaxed to

$${}_2W_2(k) \leqslant \sqrt{2}\left(\text{tr}\left(P_0\right)\right)^{3/2} ||\widehat{A}^{-k}||_F^2 ||A^k||_F. \tag{24}$$

B. LTV Bound

Theorem 9: Consider two discrete-time LTV systems $x_{k+1} = A_k x_k$, and $\widehat{x}_{k+1} = \widehat{A}_k \widehat{x}_k$, $k \in \mathbb{N} \cup \{0\}$. Let the initial PDF $\xi_0(x_0) = \mathcal{N}(0, P_0)$. Then, ${}_2W_2(k) \leqslant \sqrt{2}\left(\text{tr}\left(P_0\right)\right)^{1/2} ||\widehat{A}_k^{-1} \widehat{A}_{k-1}^{-1} \dots \widehat{A}_1^{-1}||_F\, \Omega_{\text{LTV}}(k)$, where

$$\Omega_{\text{LTV}}(k) := \Big(||A_k A_{k-1} \dots A_1||_F^2 ||\widehat{A}_k^{-1} \widehat{A}_{k-1}^{-1} \dots \widehat{A}_1^{-1}||_F^2 \left(\text{tr}\left(P_0\right)\right)^2 - \log\left(\prod_{i=1}^{n_s} \frac{(\vartheta_{1i}\vartheta_{2i}\dots\vartheta_{ki})^2}{\left(\widehat{\vartheta}_{1i}\widehat{\vartheta}_{2i}\dots\widehat{\vartheta}_{ki}\right)^2} \right) - n_s \Big)^{1/2},$$

where ϑ_{ki} and $\widehat{\vartheta}_{ki}$ are the i^{th} eigenvalues of A_k and $\widehat{A}_k$, respectively.

Proof: The proof is similar to the LTI case and follows by showing $D_{KL}\left(P_k || \widehat{P}_k\right) \leqslant \left(\Omega_{\text{LTV}}(k)\right)^2$. We skip the details here. ■

VII. Conclusions

As an extension of our earlier work [1], this paper formalizes the probabilistic model validation framework proposed therein. First, to ensure the practical applicability, sample complexity and storage complexity bounds are derived. Secondly, in addition to providing a relaxed notion of validation in probability, it is shown to recover the invalidation oracle from barrier certificate formulation, as a special case. Thirdly, some results on gap sensitivity to the initial uncertainty, are presented. And finally, bounds are derived for the Wasserstein gap in discrete-time linear model validation scenario.

Appendix

A. Proof for Lemma 1

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

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

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

$$\mathbb{P}(A) = \mathbb{P}\left((A \cap B^c \cap C^c) \cup B^c \cup C^c\right). \tag{25}$$

Taking $\mathcal{E}_1 := A \cap B^c \cap C^c$, $\mathcal{E}_2 := B^c$, $\mathcal{E}_3 := C^c$, and noting that $\mathbb{P}(\mathcal{E}_1) = \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = \mathbb{P}(\mathcal{E}_3 \cap \mathcal{E}_1) = \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3)$, from Boole-Bonferroni inequality (Appendix C, [19]), (25) yields

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

B. Proof for Theorem 2

The main idea is to use triangle inequality and symmetry of Wasserstein distance to reduce the problem of rate-of-convergence of empirical ${}_2W_2$ to its true value, to the "easier" problem of bounding rate-of-convergence of empirical density to the respective true density, measured in Wasserstein distance. Since Wasserstein distance is a metric, from triangle inequality

$$\begin{aligned} {}_2W_2\left(\eta_m, \widehat{\eta}_n\right) &\leqslant {}_2W_2\left(\eta_m, \eta\right) + {}_2W_2\left(\widehat{\eta}_n, \eta\right) \\ &\leqslant {}_2W_2\left(\eta_m, \eta\right) + {}_2W_2\left(\widehat{\eta}_n, \widehat{\eta}\right) + {}_2W_2\left(\eta, \widehat{\eta}\right) \\ \Rightarrow {}_2W_2\left(\eta_m, \widehat{\eta}_n\right) - {}_2W_2\left(\eta, \widehat{\eta}\right) &\leqslant {}_2W_2\left(\eta_m, \eta\right) + {}_2W_2\left(\widehat{\eta}_n, \widehat{\eta}\right). \end{aligned} \tag{26}$$

Since ${}_2W_2\left(\eta_m, \widehat{\eta}_n\right)$ is a random variable, the LHS of (26) is a random variable, which we denote as X. Further, if we denote the random variables ${}_2W_2\left(\eta_m, \eta\right)$ as Y, and ${}_2W_2\left(\widehat{\eta}_n, \widehat{\eta}\right)$ as Z, then (26) can be seen as a probabilistic inequality, i.e. $X\left(\omega\right) \leqslant Y\left(\omega\right) + Z\left(\omega\right)\ \forall\, \omega \in \Omega$. 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 (26) holds for all $\omega \in \Omega$, we can relabel X as the absolute value of the LHS of (26). Otherwise, if X is negative, (26) is trivially satisfied. Now we are in a position to invoke Lemma 1.

Combining (26) with Lemma 1, we have

$$\begin{aligned} &\mathbb{P}\left(\left| {}_2W_2\left(\eta_m, \widehat{\eta}_n\right) - {}_2W_2\left(\eta, \widehat{\eta}\right) \right| > \epsilon\right) \leqslant \\ &\mathbb{P}\left({}_2W_2\left(\eta_m, \eta\right) > \frac{\epsilon}{2}\right) + \mathbb{P}\left({}_2W_2\left(\widehat{\eta}_n, \widehat{\eta}\right) > \frac{\epsilon}{2}\right), \end{aligned} \tag{27}$$

where each term in the RHS of (27) can be separately upper-bounded using Theorem 1 with $\theta \mapsto \dfrac{\epsilon}{2}$, i.e.

$$\begin{aligned} \mathbb{P}\left({}_2W_2\left(\eta_m, \eta\right) > \frac{\epsilon}{2}\right) &\leqslant K_1 \exp\left(-\frac{m\epsilon^2}{32\mathscr{C}_1}\right), \\ \mathbb{P}\left({}_2W_2\left(\widehat{\eta}_n, \widehat{\eta}\right) > \frac{\epsilon}{2}\right) &\leqslant K_2 \exp\left(-\frac{n\epsilon^2}{32\mathscr{C}_2}\right). \end{aligned} \tag{28}$$

Hence the result.

ACKNOWLEDGEMENT

This research was supported by NSF award # 1016299 with D. Helen Gill as the Program Manager.

REFERENCES

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

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

[3] C. Villani, *Topics in Optimal Transportation*, Graduate Studies in Mathematics, First ed., American Mathematical Society; 2003.

[4] C. Villani, *Optimal Transportation: Old and New*, First ed., Springer; 2008.

[5] 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, pp. 459–474.

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

[7] P. Dutta, and R. Bhattacharya, "Hypersonic State Estimation using the Frobenius-Perron Operator", *Journal of Guidance, Control and Dynamics*, Vol. 34, No. 2, 2011, pp. 325–344.

[8] 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, 2010, pp. 323–331.

[9] 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, 1992, pp. 942–952.

[10] 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, 1994, pp. 951–959.

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

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

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

[14] 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.

[15] 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.

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

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

[18] 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.

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

[20] 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.

[21] 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, 2004, pp. 2702–2732.

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

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

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