2008 American Control Conference
Westin Seattle Hotel, Seattle, Washington, USA
June 11-13, 2008

Model Reduction of Stochastic Processes Using Wasserstein Pseudometrics

David Thorsley and Eric Klavins

Abstract— **We consider the problem of finding reduced models of stochastic processes. We use Wasserstein pseudometrics to quantify the difference between processes. The method proposed in this paper is applicable to any continuous-time stochastic process with output, and pseudometrics between processes are defined only in terms of the available outputs. We demonstrate how to approximate a wide class of behavioral pseudometrics and how to optimize parameter values to minimize Wasserstein pseudometrics between processes. In particular, we introduce an algorithm that allows for the approximation of Wasserstein pseudometrics from sampled data, even in the absence of models for the processes. We illustrate the approach with an example from systems biology.**

I. Introduction

In this paper, we consider the problem of model reduction of stochastic processes. In most areas of scientific or engineering interest, there are processes that are too complex to model precisely. Without reasonable models, it is not possible to develop methodologies to predict future behavior of such processes, verify that these processes have desirable properties of interest, or accurately control such processes. At present, there is no general framework for determining how complexity can safely be eliminated from process models to produce simple models that are amenable to analysis and reasonable approximations to the real-world process.

In particular, we consider a general approach to the problem of approximate equivalence of stochastic processes, motivated by problems in systems biology. Stochastic interpretations of the chemical master equation are often used when modeling chemical processes inside a single cell, because a deterministic approach can be rendered incorrect when the copy numbers of individual molecules are low [1]. The state spaces generated by these stochastic process models are often intractably large. As a result, these processes are usually studied via the stochastic simulation algorithm [2], which does not require a full enumeration of the state space, but does not lend itself to analysis of the underlying process. Recently, the finite state projection method [3] has been proposed to eliminate a large number of unlikely states from the state space, but it requires that model reduction start from a full and complete Markov process model.

Our approach to the problem of dealing with complex processes is based on defining a pseudometric on the space of stochastic processes. This pseudometric quantifies the similarity between two processes and can be approximated using sampled data for large processes. Given the basic structure of a simple model, our method can be used to determine if this model is close in behavior to a more complex process and to optimize parameters so as to minimize differences between the simple and complex models. Furthermore, our method does not require a model of the underlying complex process when sampled data is available.

The method we propose for calculating pseudometrics between two processes is based on the well-known Wasserstein metric ([4], [5]) and is applicable to any continuous-time stochastic process. Other approaches to the problem of process approximation have appeared in the computer science ([6]–[8]) and control [9] literature. In contrast to the cited literature, our method for approximating pseudometrics between processes is based on taking sample data from the processes to be compared, instead of performing an analytic calculation. Thus, stochastic processes described by models can be compared to each other and also compared to processes described by sample data generated from simulations or experiments. Furthermore, our method for defining pseudometrics is output-based and does not depend on unobservable or difficult to enumerate internal states; we do not require perfect knowledge of the process in order to approximate our pseudometrics, as opposed to the algorithms proposed in [6] and [7]. We can compare processes according to various notions of similarity; for example, we can define pseudometrics that can capture interesting aspects of either equilibrium or non-equilibrium behavior.

We organize this paper by first defining Wasserstein pseudometrics and illustrating their use in processes modeled by stochastic reaction networks and continuous-time Markov processes. We then present a method for approximating Wasserstein pseudometrics from sampled data and show how to estimate confidence intervals in which the true values of Wasserstein pseudometrics are likely to reside. We then state model reduction problems as stochastic optimization problems where Wasserstein pseudometrics are used as performance criteria. The results of this paper are illustrated with an example modeling gene expression.

II. Wasserstein Pseudometrics

This paper considers experiments whose outcomes are trajectories $\omega : \mathbb{R}^{\geq 0} \to Y$, where Y is a set of outputs. The sample space of such an experiment is denoted by $\Omega = (\mathbb{R}^{\geq 0} \to Y)$. Let $d : \Omega \times \Omega \to \mathbb{R}^{\geq 0}$ be a pseudometric on Ω; that is, we require that d satisfy the properties $d(\omega, \eta) \geq 0$ and $d(\omega, \varphi) + d(\varphi, \eta) \geq d(\omega, \eta)$ for all $\omega, \varphi, \eta \in \Omega$. However, we do not require that $d(\omega, \eta) = 0$ implies $\omega = \eta$.

This research is partially supported by the 2006 AFOSR MURI award "High Confidence Design for Distributed Embedded Systems."

D. Thorsley and E. Klavins are with Department of Electrical Engineering, University of Washington, Seattle, WA, 98195, USA. E-mail: {thorsley,klavins}@u.washington.edu

Let $\mathcal{P}_1$ and $\mathcal{P}_2$ denote two probability measures on Ω. We use the following psuedometric to quantify the difference between $\mathcal{P}_1$ and $\mathcal{P}_2$.

Definition 2.1: (From [4], Ch. 11) The *Wasserstein pseudometric* W_d between two probability measures $\mathcal{P}_1$ and $\mathcal{P}_2$ on a sample space Ω equipped with a pseudometric d is

$$W_d(\mathcal{P}_1, \mathcal{P}_2) = \inf_{\mathcal{Q} \in J(Z_1, Z_2)} E_{\mathcal{Q}} \left[d(Z_1, Z_2) \right], \qquad (1)$$

where Z_1 is a random variable with distribution $\mathcal{P}_1$, Z_2 is a random variable with distribution $\mathcal{P}_2$, and $J(Z_1, Z_2)$ is the set of all possible joint distributions of Z_1 and Z_2. □

A common interpretation of Wasserstein pseudometrics comes from economics ([4], Ch. 11). If goods are produced at locations distributed according to $\mathcal{P}_1$ and consumed at locations distributed according to $\mathcal{P}_2$, a Wasserstein pseudometric $W_d(\mathcal{P}_1, \mathcal{P}_2)$ represents the infimal cost necessary to transport the goods from the locations where they are produced to the locations where they are consumed.

A distinct Wasserstein pseudometric can be defined with respect to each d on Ω. For example, the Wasserstein pseudometric with respect to the discrete metric ($d(\omega, \eta) = 1$ if $\omega \neq \eta$ and 0 otherwise) is equal to the *total variation distance* on Ω [5]. However, the Wasserstein pseudometric is a more general definition that admits multiple methods for quantifying the differences between processes.

In this paper, we consider pseudodistances of the form

$$d(\omega, \eta) = \left| Z(\omega) - Z(\eta) \right|, \qquad (2)$$

where $Z : \Omega \to \mathbb{R}$ is a *reporter* random variable defined to capture some interesting feature of trajectories. In order to be a well-defined random variable, Z must be measurable with respect to a σ-field on Ω. To ensure the measurability of Z, we equip Ω with a σ-field $\mathcal{F}$ satisfying

$$\mathcal{F} \supseteq \bigcup_{B \in \mathcal{B}(\mathbb{R})} Z^{-1}(B),$$

where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-field on the real line.

Each probability measure $\mathcal{P}$ defines a cumulative distribution function (CDF) of Z

$$F_{\mathcal{P},Z}(z) \triangleq \mathcal{P}(Z < z).$$

The inverse CDF of Z is

$$F^{-1}_{\mathcal{P},Z}(y) \triangleq \inf\{z : F_{\mathcal{P},Z}(z) \geq y).$$

For pseudometrics of the form specified in Equation 2, a Wasserstein pseudometric between $\mathcal{P}_1$ and $\mathcal{P}_2$ can be expressed in terms of the CDFs of Z.

Theorem 1: (From [10]) The Wasserstein pseudometric between two probability measures $\mathcal{P}_1$ and $\mathcal{P}_2$ on a sample space Ω equipped with a pseudometric d defined according to Equation 2 is

$$W_d(\mathcal{P}_1, \mathcal{P}_2) = \int_{-\infty}^{\infty} \left| F_{\mathcal{P}_1,Z}(z) - F_{\mathcal{P}_2,Z}(z) \right| dz. \qquad (3)$$

■

Equivalently, a Wasserstein pseudometric can be calculated using the inverse CDFs [5] by

$$W_d(\mathcal{P}_1, \mathcal{P}_2) = \int_0^1 \left| F^{-1}_{\mathcal{P}_1,Z}(y) - F^{-1}_{\mathcal{P}_2,Z}(y) \right| dy, \qquad (4)$$

which is useful when numerically approximating Wasserstein pseudometrics.

III. Classes of Probability Measures

A. Reaction Networks and Continuous-Time Markov Processes

Any stochastic process with outputs generates a probability measure $\mathcal{P}$ on the space $(\Omega, \mathcal{F})$. We are particularly interested in those corresponding to stochastic chemical reactions, which we now describe.

Definition 3.1: A *reaction network* $RN = (\mathcal{S}, \mathcal{R})$ consists of a set of species $\mathcal{S}$ and a set of reactions $\mathcal{R}$. Each reaction in $\mathcal{R}$ takes the form

$$n_{1,L} S_1 + \cdots + n_{p,L} S_p \xrightarrow{k} n_{1,R} S_1 + \cdots + n_{q,R} S_p,$$

where $S_i \in \mathcal{S}$, $n_{i,L}, n_{i,R} \in \mathbb{Z}^{\geq 0}$, $i = 1 \ldots p$. □

Following [11], a reaction network can be interpreted stochastically. The state of a reaction network RN is a p-dimensional vector $x(t) = [N_1(t) \ldots N_p(t)]^T$, where $N_i(t)$ denotes the number of the species S_i at time t. The firing of a reaction network produces a state transition

$$x \mapsto x - [n_{1,L} \ldots n_{p,L}]^T + [n_{1,R} \ldots n_{p,R}]^T$$

corresponding to the consumption of the reactants and the creation of the products. The propensity function for a reaction $R \in \mathcal{R}$ is $a_R(x) = k \prod_{i=1}^{p} N_i^{n_{i,L}}$. The probability that a reaction R will fire in the interval $[t, t + dt)$ is defined as $a_R(x)dt$, given that x is the state of RN at time t.

Interpreting a reaction network in this manner produces a continuous time Markov process.

Definition 3.2: A *continuous time Markov process* (CTMP) is a tuple $S = (X, \mathbf{Q}, \pi_0, Y, h)$, where X is a countable set of states, $\mathbf{Q}$ is a transition rate matrix, π_0 is the initial probability distribution on X, and $h : X \to Y$ is a state output function. □

If $\mathbf{Q}_{ij}$ is a non-diagonal element of $\mathbf{Q}$, then the probability of a transition from x_i to x_j in the interval $[t, t + dt)$ is defined as $\mathbf{Q}_{ij}dt$; if $\mathbf{Q}_{ii}$ is a diagonal element of $\mathbf{Q}$, we require that $\mathbf{Q}_{ii} = -\sum_{j \neq i} \mathbf{Q}_{ij}$. Following ([12], Ch. 15), a CTMP S defines a probability measure $\mathcal{P} : \mathcal{F} \to [0, 1]$ on $(\Omega, \mathcal{F})$.

Example 3.1: The following set of reactions defines a generic model for gene expression:

$$\begin{aligned} \emptyset &\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{mR} \\ \mathrm{mR} &\xrightarrow{k_2} \mathrm{mR} + \mathrm{P} \\ \mathrm{P} &\xrightarrow{k_3} \emptyset, \end{aligned} \qquad (5)$$

where mR denotes messenger RNA, P denotes protein, and $\emptyset$ denotes the null species. A graphical version of this reaction network is shown in Figure 1(a). The state space of the

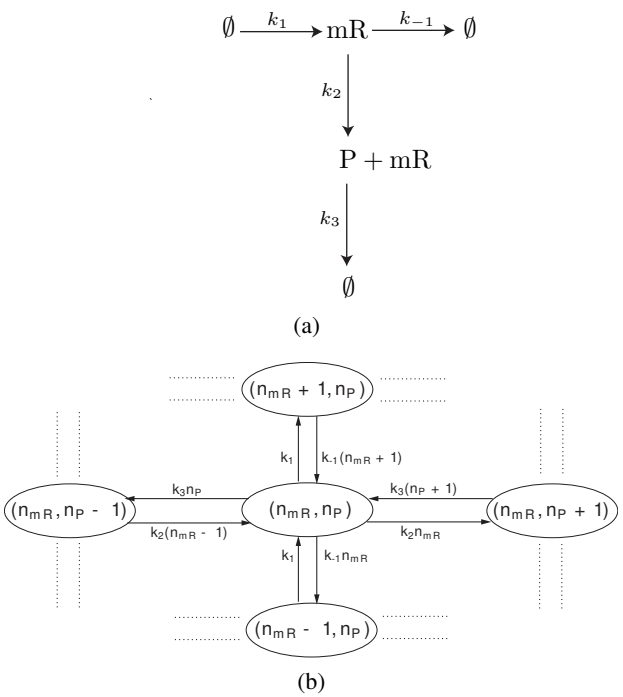


Fig. 1. (a) A graphical version of a reaction network modeling gene expression. (b) A typical state in the CTMP generated by the stochastic interpretation of the reaction network in (a). Each state is indicated by a pair (n_{mR}, n_P), and the labels on each arrow indicate the rate propensity for each transition in or out of (n_{mR}, n_P). Transitions to or from states bordering (n_{mR}, n_P) from states other than the typical state are not shown.

CTMP derived from this reaction network is $\mathbb{Z}^{\geq 0} \times \mathbb{Z}^{\geq 0}$, and a state is of the form $\begin{bmatrix} n_{mR} & n_P \end{bmatrix}$. A typical state of this CTMP and the rate propensities for each reaction entering or leaving that state are shown in Figure 1(b). The output function of this CTMP is $h(\begin{bmatrix} n_{mR} & n_P \end{bmatrix}) = n_P$, which corresponds to the ability to observe the level of protein in the system; experimentally, this can often be accomplished by incorporating a fluorescent marker into the protein.

We define reporter random variables to capture the following aspects of a trajectory ω.

- Z can represent the amount of a protein at a given time t: $Z(\omega) \triangleq \omega(t)$.
- Z can represent the amount of time necessary for n proteins to be present: $Z(\omega) \triangleq \min\{\omega^{-1}(n)\}$.
- Z can represent the average amount of protein over an interval (t_s, t_f): $Z(\omega) \triangleq \frac{1}{t_f - t_s} \int_{t_s}^{t_f} \omega(t)dt$.
- Z can indicate if more than n proteins are ever present: $Z(\omega) \triangleq 1$ if $n_P(t) > n$ for some $t \in \mathbb{R}^{\geq 0}$ and $Z(\omega) = 0$ otherwise.

There are many valid choices for the reporter random variable Z beyond those listed above. Each different choice of Z gives a different notion of the difference between processes and the Wasserstein pseudometric with respect to $d(\omega, \eta) = |Z(\omega) - Z(\eta)|$ is different as well. The choice of the reporter random variable is motivated by what questions are being asked about the system.

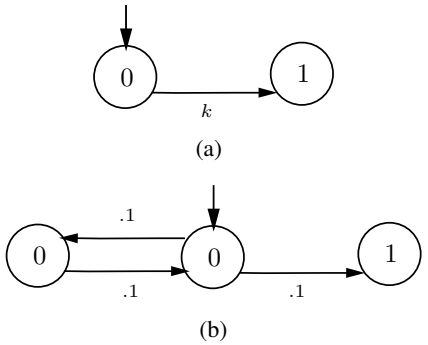


Fig. 2. (a) A two-state continuous-time Markov Process with an unknown rate propensity k. (b) A three-state continuous-time Markov Process.

B. Example: Analytic Calculation of a Wasserstein Pseudometric

Two simple CTMPs, S_1 and S_2, are shown in Figures 2(a) and 2(b), respectively. We will denote the probability distributions on $(\Omega, \mathcal{F})$ generated by these CTMPs as $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively. The initial distributions for S_1 and S_2 are defined so that the initial probability of the state marked with a vertical arrow is 1, and the initial probability of all other states is zero. The labels within each state denote the value of the output function in that state.

We define a reporter random variable Z by

$$Z(\omega) = \inf \omega^{-1}(1).$$

In this example, Z denotes the hitting time of the state with output label 1. The pseudodistance between two trajectories is thus $d(\omega, \eta) = |\inf \omega^{-1}(1) - \inf \eta^{-1}(1)|$.

For the system S_1 in Figure 2(a), the transition rate matrix is:

$$\mathbf{Q}_1 = \begin{bmatrix} -k & k \\ 0 & 0 \end{bmatrix}.$$

Because the state with output label 1 is an absorbing state, the event that $Z < t$ is equal to the event that S_1 is in the state with output 1 at time t. We compute the transition probabilities from time 0 to time t by calculating the matrix exponential of $\mathbf{Q}_1 t$:

$$e^{\mathbf{Q}_1 t} = \begin{bmatrix} e^{kt} & 1 - e^{kt} \\ 0 & 1 \end{bmatrix}.$$

The distribution function of Z in $\mathcal{P}_1(k)$ is the element of $e^{\mathbf{Q}_1 t}$ corresponding to starting in the initial state and ending in the absorbing state. Therefore,

$$\begin{aligned} F_{\mathcal{P}_1, Z}(t) &= \begin{bmatrix} 1 & 0 \end{bmatrix} e^{\mathbf{Q}_1 t} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= 1 - e^{-kt}. \end{aligned}$$

Following a similar procedure for S_2, the transition rate matrix is

$$\mathbf{Q}_2 = \begin{bmatrix} -0.1 & 0.1 & 0 \\ 0.1 & -0.2 & 0.1 \\ 0 & 0 & 0 \end{bmatrix}.$$

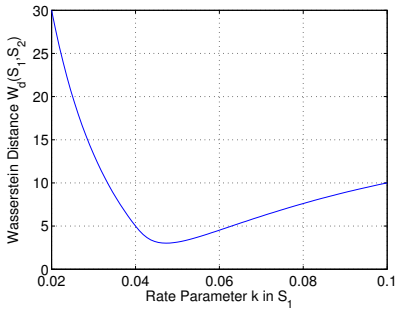


Fig. 3. The Wasserstein pseudometric given by Equation 6 varies as the parameter k in S_1 varies.

The distribution function of Z in $\mathcal{P}_2$ is

$$F_{\mathcal{P}_2,Z}(t) = \begin{bmatrix}0 & 1 & 0\end{bmatrix} e^{\mathbf{Q}_2 t} \begin{bmatrix}0\\0\\1\end{bmatrix}$$
$$= 1 - \frac{5-\sqrt{5}}{10} e^{-\frac{1}{20}(3+\sqrt{5})t} - \frac{5+\sqrt{5}}{10} e^{-\frac{1}{20}(3-\sqrt{5})t}$$

From Equation 3, the Wasserstein pseudometric with respect to d between the CTMPs S_1 and S_2 is

$$W_d(\mathcal{P}_1(k), \mathcal{P}_2) = \int_0^\infty \left| e^{-kt} - \frac{5-\sqrt{5}}{10} e^{-\frac{1}{20}(3+\sqrt{5})t} - \frac{5+\sqrt{5}}{10} e^{-\frac{1}{20}(3-\sqrt{5})t} \right| dt \quad (6)$$

This expression can be evaluated numerically. Figure 3 shows the values of $W_d(\mathcal{P}(S_1), \mathcal{P}(S_2))$ as k varies from 0.02 to 0.1.

This example motivates the use of Wasserstein pseudometrics as performance criteria in optimization problems. Figure 3 indicates that the optimal value of k that minimizes the Wasserstein pseudometric between the two and three-state CTMPs is $k^* = .047$, and that the value of this pseudometric is $W_d(\mathcal{P}_1(k^*), \mathcal{P}_2) = 3.0\ s$. For this example, we define the error introduced by approximating the process $\mathcal{P}_2$ with the process $\mathcal{P}_1(k)$ by

$$\varepsilon \triangleq \frac{W_d(\mathcal{P}_1(k), \mathcal{P}_2)}{E_{\mathcal{P}_2}(Z)}.$$

The error introduced by using $\mathcal{P}_1(k^*)$ as a reduced model for $\mathcal{P}_2$ is $\varepsilon = 3.0/20.0 = 15\%$. This indicates that for any choice of k, using a model of the form in Figure 2(a) as a reduction of the process in Figure 2(b) results in significant error.

This example also illustrates that, even for very simple processes, analytical calculation of a Wasserstein pseudometric can be very difficult. For processes of the complexity generated by reaction networks modeling biochemical processes, analytically calculating a Wasserstein pseudometric is most likely intractable. In the next section, we present a technique for approximating a Wasserstein pseudometric from sample data taken from a complex or unknown probability distribution.

IV. Wasserstein Pseudometrics from Sampled Data

The probability measure $\mathcal{P}$ on $(\Omega, \mathcal{F})$ generated by a CTMP may be too complex to calculate exactly and thus must remain unknown. Furthermore, the CTMP is merely an approximation of an actual physical process that is also an unknown probability measure on $(\Omega, \mathcal{F})$.

In practice, we can usually approximate an unknown probability measure $\mathcal{P}$ by taking n independent samples of Ω according to $\mathcal{P}$ and generating an empirical probability measure $\hat{\mathcal{P}}_n$ from the sample data. We calculate a Wasserstein pseudometric between empirical probability distributions to approximate the pseudometric between the unknown underlying distributions. The sample data used to make this approximation can be generated from physical processes by performing experiments or obtained from models by using the stochastic simulation algorithm (SSA) [2].

Given a set of n samples $\{\omega_1, \ldots, \omega_n\}$, the empirical probability measure of Z is defined as

$$\hat{\mathcal{P}}_n(Z^{-1}(B)) \triangleq \frac{|\{\omega : Z(\omega) \in B\}|}{n}, \quad (7)$$

for all $B \in \mathcal{B}(\mathbb{R})$. The empirical CDF of Z is thus

$$F_{\hat{\mathcal{P}}_n,Z}(z) = \frac{|\{\omega : Z(\omega) < z\}|}{n}. \quad (8)$$

Consider two unknown probability distributions $\mathcal{P}_1$ and $\mathcal{P}_2$, and suppose that we take n independent samples $\{\omega_1, \ldots, \omega_n\}$ from $\mathcal{P}_1$ and ℓn independent samples $\{\eta_1, \ldots, \eta_{\ell n}\}$ from $\mathcal{P}_2$, where $\ell \in \mathbb{N}$. Without loss of generality, we will assume that these sets of samples are sorted so that $Z(\omega_1) \leq Z(\omega_2) \leq \cdots \leq Z(\omega_n)$ and $Z(\eta_1) \leq Z(\eta_2) \leq \cdots \leq Z(\eta_{\ell n})$. From these two sets of samples we generate the empirical probability measures $\hat{\mathcal{P}}_{1,n}$ and $\hat{\mathcal{P}}_{2,\ell n}$, respectively. Because the samples are sorted, the inverse empirical CDFs of $\mathcal{P}_1$ and $\mathcal{P}_2$ are of the form

$$F^{-1}_{\hat{\mathcal{P}}_n,Z}(y) = Z(\omega_i), \quad \text{where } \frac{i-1}{n} < y \leq \frac{i}{n}. \quad (9)$$

The empirical probability distributions can be used to quickly approximate a Wasserstein pseudometric between the underlying probability distributions according to the following theorem.

Theorem 2: Suppose $E_{\mathcal{P}_i}(|Z|) < \infty$ for $i = 1, 2$. The Wasserstein pseudometric $W_d(\mathcal{P}_1, \mathcal{P}_2)$ between two probability distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ with respect to a pseudodistance function $d(\omega, \eta) = |Z(\omega) - Z(\eta)|$ on Ω is equal to

$$W_d(\mathcal{P}_1, \mathcal{P}_2) = \lim_{n\to\infty} \frac{1}{\ell n} \sum_{i=1}^{\ell n} \left| Z(\omega_{\lceil \frac{i}{\ell} \rceil}) - Z(\eta_i) \right| \quad (10)$$

almost surely.

Proof: The proof is a straightforward application of the Glivenko-Cantelli Theorem and the strong law of large numbers sketched in part in [13]. For the benefit of readers with only a basic knowledge of probability theory, we present all the details. The proof consists of two steps. First we show that

i) $W_d(\mathcal{P}_1, \mathcal{P}_2) = \lim_{n\to\infty} W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n})$

almost surely. We then show that

$$\text{ii)} \quad W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n}) = \frac{1}{\ell n} \sum_{i=1}^{\ell n} \left| Z(\omega_{\lceil \frac{i}{\ell} \rceil}) - Z(\eta_i) \right|$$

for all n, and by combining these two statements the theorem is proven.

(Proof of Statement i) We apply the Glivenko-Cantelli Theorem [4] to determine that

$$\lim_{n\to\infty} \left| F_{\hat{\mathcal{P}}_{1,n},Z}(z) - F_{\mathcal{P}_1,Z}(z) \right| = 0$$

almost surely for all $z \in \mathbb{R}$. It immediately follows that

$$\int_{-\infty}^{\infty} \lim_{n\to\infty} \left| F_{\hat{\mathcal{P}}_{1,n},Z}(z) - F_{\mathcal{P}_1,Z}(z) \right| dz = 0. \tag{11}$$

Since $E_{\mathcal{P}_i}(|Z|) < \infty$, the dominated convergence theorem is applicable (see appendix) and thus

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \left| F_{\hat{\mathcal{P}}_{1,n},Z}(z) - F_{\mathcal{P}_1,Z}(z) \right| dz = 0$$
$$\lim_{n\to\infty} W_d(\hat{\mathcal{P}}_{1,n}, \mathcal{P}_1) = 0.$$

Similarly we show that $\lim_{n\to\infty} W_d(\hat{\mathcal{P}}_{2,\ell n}, \mathcal{P}_2) = 0$.

Since W_d is a pseudometric, we apply the triangle inequality to get

$$W_d(\mathcal{P}_1, \mathcal{P}_2) \le W_d(\mathcal{P}_1, \hat{\mathcal{P}}_{1,n}) + W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n}) + W_d(\hat{\mathcal{P}}_{2,\ell n}, \mathcal{P}_2).$$

Taking the limit as $n \to \infty$, the first and third terms on the right-hand side of the inequality vanish, yielding

$$W_d(\mathcal{P}_1, \mathcal{P}_2) \le \lim_{n\to\infty} W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n}).$$

Similarly, re-applying the triangle inequality yields

$$W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n}) \le W_d(\hat{\mathcal{P}}_{1,n}, \mathcal{P}_1) + W_d(\mathcal{P}_1, \mathcal{P}_2) + W_d(\mathcal{P}_2, \hat{\mathcal{P}}_{2,\ell n}),$$

from which it follows that

$$W_d(\mathcal{P}_1, \mathcal{P}_2) \ge \lim_{n\to\infty} W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n}).$$

Therefore

$$W_d(\mathcal{P}_1, \mathcal{P}_2) = \lim_{n\to\infty} W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n}). \tag{12}$$

(Proof of Statement ii) From Equation 4, for any $n \in \mathbb{N}$, a Wasserstein pseudometric between $\hat{\mathcal{P}}_{1,n}$ and $\hat{\mathcal{P}}_{2,\ell n}$ is

$$W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n}) = \int_0^1 \left| F^{-1}_{\hat{\mathcal{P}}_{1,n},Z}(y) - F^{-1}_{\hat{\mathcal{P}}_{2,\ell n},Z}(y) \right| dy.$$

We partition the interval $[0, 1]$ into ℓn intervals of size $\frac{1}{n}$, yielding

$$W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n}) = \sum_{i=1}^{\ell n} \left(\int_{\frac{i-1}{\ell n}}^{\frac{i}{\ell n}} \left| F^{-1}_{\hat{\mathcal{P}}_{1,n},Z}(y) - F^{-1}_{\hat{\mathcal{P}}_{2,\ell n},Z}(y) \right| dy \right).$$

From Equation 9, on each interval $(\frac{i-1}{\ell n}, \frac{i}{\ell n})$, the values of both inverse empirical distributions are constant and are $F^{-1}_{\hat{\mathcal{P}}_{1,n},Z}(y) = Z(\omega_{\lceil \frac{i}{\ell} \rceil})$ and $F^{-1}_{\hat{\mathcal{P}}_{2,\ell n},Z}(y) = Z(\eta_i)$, respectively. The integral on each such interval is therefore $\frac{1}{n} \left| Z(\omega_{\lceil \frac{i}{\ell} \rceil}) - Z(\eta_i) \right|$, and thus

$$W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n}) = \frac{1}{\ell n} \sum_{i=1}^{\ell n} |Z(\omega_{\lceil \frac{i}{\ell} \rceil}) - Z(\eta_i)|. \tag{13}$$

Thus statement ii) is proven. Substituting Equation 13 into Equation 12 proves the theorem. ■

Theorem 2 suggests an algorithm for approximating a Wasserstein pseudometric between two probability measures $\mathcal{P}_1$ and $\mathcal{P}_2$.

Algorithm WP: Wasserstein Pseudometric Computation

Input n samples $\{\omega_1, \dots \omega_n\}$ generated according to $\mathcal{P}_1$.
Input ℓn samples $\{\eta_1, \dots \eta_{\ell n}\}$ generated according to $\mathcal{P}_2$.

1) **Calculate** $Z(\omega_i)$ for each ω_i, $i = 1 \dots n$.
2) **Sort** and re-index $\{\omega_1, \dots \omega_n\}$ so that $Z(\omega_1) \le Z(\omega_2) \le \dots \le Z(\omega_n)$.
3) **Calculate** $Z(\eta_i)$ for each η_i, $i = 1 \dots \ell n$.
4) **Sort** and re-index $\{\eta_1, \dots \eta_{\ell n}\}$ so that $Z(\eta_1) \le Z(\eta_2) \le \dots \le Z(\eta_{\ell n})$.
5) **Calculate** $|Z(\omega_{\lceil \frac{i}{\ell} \rceil}) - Z(\eta_i)|$ for $i = 1 \dots \ell n$.
6) **Calculate** $W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n}) = \frac{1}{n} \sum_{i=1}^{n} |Z(\omega_{\lceil \frac{i}{\ell} \rceil}) - Z(\eta_i)|$.

The complexity of this algorithm is $O(\ell n \log \ell n)$, as the rate determining step is the sorting of the outcomes $\{\eta_1, \dots \eta_{\ell n}\}$. Since generating the samples $\{\omega_1, \dots \omega_n\}$ and $\{\eta_1, \dots \eta_{\ell n}\}$ involves either performing a significant number of experiments or running the SSA a large number of times, the complexity of approximating a Wasserstein pseudometric from empirical probability measures is much less than the complexity of generating the data used to construct those measures.

V. Bootstrap Confidence Intervals

The values of Wasserstein pseudometrics $W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n})$ computed according to Algorithm WP are approximations of the true Wasserstein pseudometrics $W_d(\mathcal{P}_1, \mathcal{P}_2)$. Because this approximation is not exact, we need to determine a confidence interval CI that satisfies

$$\Pr\left(W_d(\mathcal{P}_1, \mathcal{P}_2) \in CI\right) \ge \alpha,$$

where α the desired coverage probability. Without requiring additional structure on the random variable Z, it is difficult to determine any parameters of the Wasserstein estimator $W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,\ell n})$, such as its variance, that can be used to calculate CI. Furthermore, the underlying probability distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ are either unknown or extremely difficult to characterize, and taking additional samples from these distributions is computationally expensive. A standard technique for estimating confidence intervals when there is little knowledge of the underlying probability distributions is the resampling technique *bootstrapping* [14].

Bootstrap estimates $\hat{\mathcal{P}}_{1,BOOT}$ of $\mathcal{P}_1$ are generated from $\hat{\mathcal{P}}_{1,n}$. With replacement, n independent samples are taken from $\hat{\mathcal{P}}_{1,n}$ and the bootstrap probability distribution on $(\Omega, \mathcal{F})$ is defined analogously to Equation 7 by

$$\hat{\mathcal{P}}_{1,BOOT}(Z^{-1}(B)) \triangleq \frac{|\{\omega : Z(\omega) \in B\}|}{n},$$

for all $B \in \mathcal{B}(\mathbb{R})$. Many bootstrap estimates of the probability distribution can be taken quickly, as resampling from $\hat{\mathcal{P}}_{1,n}$ is much faster than taking new samples from either the SSA or from experiment.

A simple bootstrapping technique for estimating a confidence interval is the *bootstrap percentile* method [15] presented in the following algorithm.

Algorithm BP: Bootstrap Percentile

Input n samples $\{\omega_1, \ldots \omega_n\}$ generated according to $\mathcal{P}_1$.
Input ℓn samples $\{\eta_1, \ldots \eta_{\ell n}\}$ generated according to $\mathcal{P}_2$.

1) **For** $i = 1 \ldots m$
   a) **Let** $\{\omega_{1,BOOT}, \ldots, \omega_{n,BOOT}\}$ be n independent samples, chosen with replacement, from the set $\{\omega_1, \ldots \omega_n\}$.
   b) **Let** $\{\eta_{1,BOOT}, \ldots, \eta_{\ell n,BOOT}\}$ be ℓn independent samples, chosen with replacement, from the set $\{\eta_1, \ldots \eta_{\ell n}\}$.
   c) **Let** $W(i) = WP(\{\omega_{1,BOOT}, \ldots, \omega_{n,BOOT}\}, \{\eta_{1,BOOT}, \ldots, \eta_{\ell n,BOOT}\})$.
2) **Sort** the list $W(i)$ and **let** $X(i)$ denote the sorted list.
3) **Let** $CI = (X(\lfloor (1-\alpha/2)m \rfloor), X(\lceil (1+\alpha/2)m \rceil))$ be the α-confidence interval.

The confidence interval generated by this algorithm is the centered confidence interval that lies between the $1-\alpha/2$ and $1+\alpha/2$ percentiles. One-sided confidence intervals for Wasserstein pseudometrics $(0, X(\lceil \alpha m \rceil))$ and$(X(\lfloor (1-\alpha)m \rfloor), \infty)$ can also be constructed by modifying the last step of the algorithm.

If more knowledge of the random variable Z or the underlying probability distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ is available, more advanced bootstrapping techniques may improve on the performance of this algorithm [14].

VI. Empirical Model Reduction

Suppose we have a stochastic process that defines a probability measure $\mathcal{P}_1$ on $(\Omega, \mathcal{F})$. If this process is difficult to analyze, we seek a reduced model that captures the properties of interest of the process but is easier to interpret.

Suppose the structure of the reduced model is known except for an unknown parameter vector $\mathbf{k}$ of length p. If our reduced model is a reaction network, these parameters may be unknown rate propensities; if it is a CTMP, these parameters may be transitions rates that are elements of the $\mathbf{Q}$ matrix. The reduced model defines a probability measure that depends on $\mathbf{k}$; we denote this measure by $\mathcal{P}_2(\mathbf{k})$.

In order to find the reduced model that most closely matches the original process, we use a Wasserstein pseudometric as a criterion for optimization. We choose a reporter random variable Z that captures the behavior of the system that we want to preserve under model reduction. Different choices of Z result in different criteria for optimization and thus different reduced models.

The model reduction problem is then

$$\arg\min_{\mathbf{k}} W_d\left(\mathcal{P}_1, \mathcal{P}_2(\mathbf{k})\right).$$

It is not practical to determine exact values of the performance criterion $W_d(\mathcal{P}_1, \mathcal{P}_2(\mathbf{k}))$; however, we can approximate the performance criterion for a given $\mathbf{k}$ by using Algorithm WP. Because Algorithm WP uses random samples from $\mathcal{P}_1$ and $\mathcal{P}_2(\mathbf{k})$ as inputs, our measurements of the performance criterion are stochastic and thus we frame this problem as a stochastic optimization problem.

The optimal values of $\mathbf{k}$ can be approximated through the use of the stochastic gradient descent algorithm [16]. The stochastic gradient descent algorithm is initialized with an initial estimate $\mathbf{k}_0$ of the optimal parameters and *gain sequences* $\langle a_0, a_1, \ldots \rangle$ and $\langle c_0, c_1, \ldots \rangle$. Each iteration of the algorithm updates the estimate of the optimal parameters according to the equation

$$\mathbf{k}_{n+1} = \mathbf{k}_n - a_n \hat{g}_n(\mathbf{k}_n),$$

where $\hat{g}_n(\mathbf{k}_n)$ is an estimate of the gradient of the Wasserstein pseudometric at $\mathbf{k}_n$. This estimate can be made using the finite difference method, yielding

$$\hat{g}_n(\mathbf{k}_n) = \frac{1}{2c_n} \times \begin{bmatrix} W_d(\mathcal{P}_1, \mathcal{P}_2(\mathbf{k}_n + c_n\xi_1)) - W_d(\mathcal{P}_1, \mathcal{P}_2(\mathbf{k}_n - c_n\xi_1)) \\ \vdots \\ W_d(\mathcal{P}_1, \mathcal{P}_2(\mathbf{k}_n + c_n\xi_p)) - W_d(\mathcal{P}_1, \mathcal{P}_2(\mathbf{k}_n - c_n\xi_p)) \end{bmatrix},$$

where ξ_i is a p-dimensional vector with 1 in the ith position and 0 in all other positions.

An alternative approach is to use the α-confidence interval as a performance criterion and find $\mathbf{k}$ that minimizes the upper bound of the one-sided confidence interval $(0, X(\lceil \alpha m \rceil))$. This optimization problem is

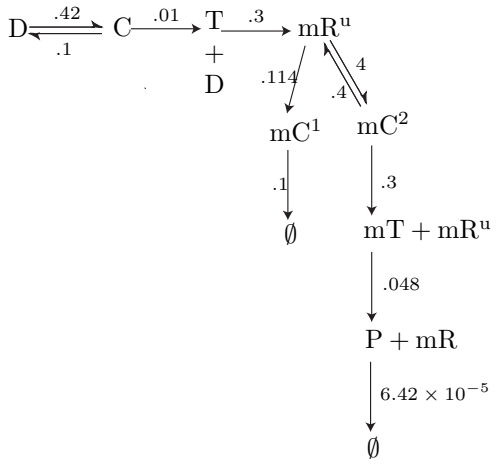
$$\arg\min_{\mathbf{k}} \left\{ \inf_U \Pr\left(W_d\left(\mathcal{P}_1, \mathcal{P}_2(\mathbf{k})\right) < U\right) \geq \alpha \right\}.$$

Stochastic estimates of this criterion can be made using Algorithm BP, and this problem can also be solved by the stochastic gradient descent method.

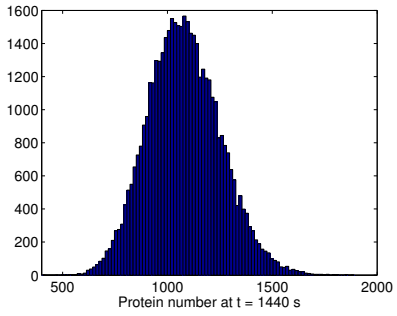
VII. Example: Gene Expression

A. Comparing Two Existing Models

Figure 4(a) is the reaction network model of gene expression found in [17]. In this model, RNA polymerase binds to a promoter (D) creating a complex (C) that in turn produces a transcribing polymerase (T). From the transcribing polymerase, unbound messenger RNA strands ($\mathrm{mR^u}$) are produced. The mRNA can bind to either a degradasome ($\mathrm{mC^1}$) or a ribosome ($\mathrm{mC^2}$). If the mRNA strand binds to a ribosome, it produces a transcribing complex (mT) and then a protein (P). The mRNA strand is preserved until it binds to a degradasome which consumes the mRNA. In contrast to



(a)



(b)

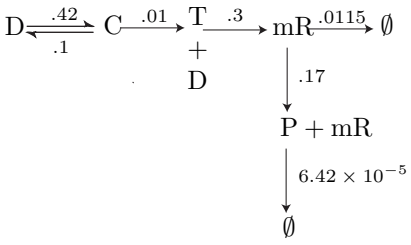
Fig. 4. (a) Gene expression model from [17]. (b) Histogram showing the values of $Z(\omega)$ for 45000 stochastic simulations ω of the reaction network in (a).

[17], we assume the rates are fixed throughout the process and do not vary with changes in cell size.

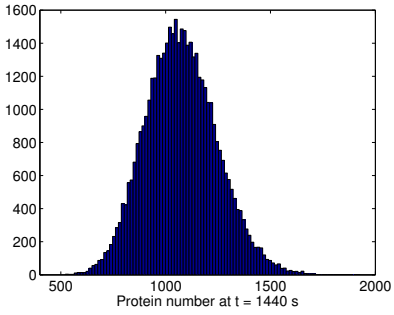
Suppose the output of this reaction network is the protein number, which could be estimated if the protein were, for example, fluorescent. We define a reporter random variable $Z(\omega) = \omega(1440)$, the number of proteins present in the system at $t = 1440$ seconds. This time is the time at which cell division first occurs in the model proposed in [17] and is a reasonable value for the length of the cell cycle in *E. coli*. Using the SSA, we generated $n = 45000$ samples from the reaction network and computed the value $Z(\omega)$ for each sample. The distribution of the data is shown in Figure 4(b).

A reduced model of gene expression, shown in Figure 5(a) is also considered in [17]. The translation process has been simplified. The transcribing polymerase T produces an mRNA strand (mR) that either decays or produces a protein (P). The complexes containing ribosomes and degradasomes are abstracted away in this model. Again using the SSA, we generated $n = 45000$ samples from this reaction network and computed $Z(\eta)$ for each sample; the distribution of the data for this reduced system is shown in Figure 5(b).

Using Algorithm WP, we approximated a Wasserstein pseudometric between the $\mathcal{P}_1$, the probability measure generated by the full model, and $\mathcal{P}_2$, the probability measure generated by the reduced model. The value of this approximation is



(a)



(b)

Fig. 5. (a) Reduced gene expression model from [17]. (b) Histogram showing the values of $Z(\eta)$ of 45000 stochastic simulations η of the reaction network in (a).

$$W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{2,n}) = 5.44 \text{ proteins.}$$

As the average protein number is the full system is $E_{\hat{\mathcal{P}}_{1,n}}(\omega) = 1080.2$, the error introduced by the reduced model is $\varepsilon = 5.44/1080.2$, or $\varepsilon = 0.50\%$.

A 95% confidence interval for $W_d(\mathcal{P}_1, \mathcal{P}_2)$ was estimated using the bootstrap percentile method. 5000 estimates of $W_d(\mathcal{P}_1, \mathcal{P}_2)$ were taken by resampling from the probability measures $\hat{\mathcal{P}}_{1,n}$ and $\hat{\mathcal{P}}_{2,n}$. The 95% confidence interval was estimated to be $(3.63, 7.50)$. Thus there is approximately a 97.5% probability that the difference between the full model and the reduced model is $\varepsilon < 7.50/1080.2 = 0.69\%$. These numbers demonstrate a close agreement between the full and reduced models.

B. Finding Optimal Parameters for a Model

The simple model in Figure 1(a) can, with an appropriate choice of parameters, also serve as a reduced model for the reaction network in Figure 4(a). Following [17], we assumed that the protein decay rate $k_3 = 6.42 \times 10^{-5}\ s^{-1}$ and that $k_2 = 15k_{-1}$.

Under these constraints, we performed a stochastic gradient descent to approximate the values of k_1 and k_2 to minimize the Wasserstein pseudometric $W_d(\mathcal{P}_1, \mathcal{P}_3(k_1, k_2))$, where $\mathcal{P}_3(k_1, k_2)$ is the probability measure generated by the reaction network in Figure 1(a). At each step of the algorithm, the gradient was estimated using the finite difference method. The optimal values for the parameters are approximately

$$k_1^* = .0554, k_2^* = .17,$$

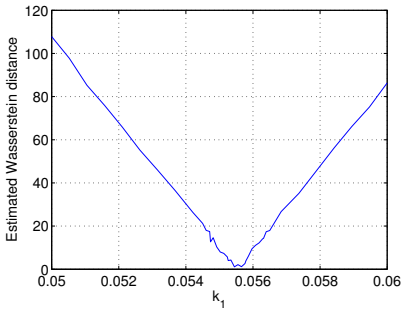


Fig. 6. Sensitivity of the Wasserstein pseudometric $W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{3,n}(k_1, k_2))$ to change in transcription parameter k_1. The translation parameter is held constant at $k_2 = .17$.

and the Wasserstein pseudometric between the full model and the optimized reduced model is approximately

$$W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{3,n}(k_1^*, k_2^*)) = 1.08 \text{ proteins}.$$

The difference between the optimized reduced model and the full model is $1.08/1080.2 = 0.10\%$. Thus, not only does the reaction network in Figure 1 have fewer species and reactions than the model in Figure 5(a), it is also a closer approximation to the model in Figure 4(a) with respect to the Wasserstein pseudometric W_d.

The optimal value of this Wasserstein pseudometric is very sensitive to changes in k_1, as shown in Figure 6. A 10% change in the value of k_1 increases the value of $W_d(\hat{\mathcal{P}}_{1,n}, \hat{\mathcal{P}}_{3,n}(k_1, k_2))$ to approximately 100, or causes approximately a 10% error. This example illustrates the difficulty in finding reduced models by hand, as small changes in the parameter k_1 quickly reduce the accuracy of the reduced model.

VIII. Discussion

In this paper, we propose the use of Wasserstein pseudometrics as criteria for comparing the behaviors of stochastic processes. With only a limited number of assumptions, efficient algorithms were developed to determine Wasserstein pseudometrics between processes and apply these results to find reduced models of gene expression. There are no restrictions on the types of stochastic processes under consideration and on the reporter random variable used to derive the pseudometric between trajectories.

In the future, we will extend these results to general Wasserstein pseudometrics that are defined with respect to trajectory pseudodistances not of the form $d(\omega, \eta) = |Z(\omega) - Z(\eta)|$. We also plan to study the rates of convergence of Algorithms WP and BP and developing further theoretical justifications for the use of bootstrapping and stochastic gradient descent. We are in the process of applying these results to other examples in systems biology and to several other areas of engineering interest, including nuclear power generation and robotics. By investigating these examples, we intend to extend the framework proposed in this paper beyond model reduction and study such problems as control synthesis and approximate verification in stochastic processes.

References

[1] J. Mettetal and A. van Oudenaarden, "Necessary noise," *Science*, vol. 317, no. 5837, pp. 463–464, July 2007.

[2] D. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *J. of Phys. Chem.*, vol. 81, pp. 2340–2360, 1977.

[3] B. Munsky and M. Khammash, "The finite state projection algorithm for the solution of the chemical master equation," *J. Chemical Physics*, vol. 124, no. 044104, 2006.

[4] R. Dudley, *Real Analysis and Probability*. Cambridge University Press, 2002.

[5] A. Gibbs and F. Su, "On choosing and bounding probability metrics," *Intl. Stat. Review*, vol. 70, no. 3, pp. 419–435, 2002.

[6] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden, "Metrics for labelled Markov processes," *J. Theoretical Computer Science*, vol. 318, pp. 323–354, June 2004.

[7] F. van Bruegel and J. Worrell, "Approximating and computing behavioural distances in probabilistic transition systems," *J. Theoretical Computer Science*, vol. 360, pp. 373–385, 2006.

[8] S. Mitra and N. Lynch, "Proving approximate implementations for probabilistic I/O automata," *Electron. Notes Theor. Comput. Sci.*, vol. 174, no. 8, pp. 71–93, 2007.

[9] A. Girard and G. J. Pappas, "Approximate bisimulation relations for constrained linear systems," *Automatica*, vol. 43, no. 8, pp. 1307–1317, 2007.

[10] S. Vallender, "Calculation of the Wasserstein distance between probability distributions on the line," *Theory of Prob. and its Applications*, vol. 18, no. 4, pp. 784–786, 1974.

[11] D. McQuarrie, "Stochastic approach to chemical kinetics," *J. Applied Probability*, vol. 4, pp. 413–478, 1967.

[12] L. Breiman, *Probability*. SIAM, 1992.

[13] E. del Barrio, E. Giné, and C. Matrán, "Central limit theorems for the Wasserstein distance between the empirical and true distributions," *Annals of Probability*, vol. 27, pp. 1009–1071, 1999.

[14] J. Shao and D. Tu, *The Jackknife and Bootstrap*. Springer, 1995.

[15] B. Efron, "Nonparametic standard errors and confidence intervals (with discussions)," *Canadian J. Statistics*, vol. 9, pp. 139–172, 1982.

[16] J. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley Interscience, 2003.

[17] P. Swain, M. Elowitz, and E. Siggia, "Intrinsic and extrinsic contributions to stochasticity in gene expression," *Science*, vol. 99, no. 20, pp. 12 795–12 800, 2002.

Appendix

A. Justification of the Use of Dominated Convergence Theorem in Theorem 2

The integrand on the right hand side of Equation 11 is the limit as $n \to \infty$ of

$$\begin{aligned}\left|F_{\hat{\mathcal{P}}_{1,n},Z}(z) - F_{\mathcal{P}_1,Z}(z)\right| &= \left|\hat{\mathcal{P}}_{1,n}(Z > z) - \mathcal{P}_1(Z > z)\right| \\ &\leq \hat{\mathcal{P}}_{1,n}(Z > z) + \mathcal{P}_1(Z > z) \\ &\leq \hat{\mathcal{P}}_{1,n}(|Z| > z) + \mathcal{P}_1(|Z| > z),\end{aligned}$$

where the last inequality follows because $Z \leq |Z|$. Because $|Z|$ is a non-negative random variable,

$$\int_{-\infty}^{\infty} \mathcal{P}_1(|Z| > z) = E_{\mathcal{P}_1}|Z|,$$

which is finite by assumption. Similarly, $\int_{-\infty}^{\infty} \hat{\mathcal{P}}_{1,n}(|Z| > z) = E_{\hat{\mathcal{P}}_{1,n}}|Z|$. By the strong law of large numbers $E_{\hat{\mathcal{P}}_{1,n}}|Z|$ converges to $E_{\mathcal{P}_1}|Z|$ almost surely, which indicates that $E_{\hat{\mathcal{P}}_{1,n}}|Z|$ is finite for all n. The supremum over all n of $\hat{\mathcal{P}}_{1,n}(|Z| > z) + \mathcal{P}_1(|Z| > z)$ is therefore an integrable dominating function for $\left|F_{\hat{\mathcal{P}}_{1,n},Z}(z) - F_{\mathcal{P}_1,Z}(z)\right|$, and thus the use of dominated convergence in Theorem 2 is justified.