A Gaussian Surrogate of Partially Observed Stochastic Processes using Wasserstein Metric

Saroj Prasad Chhatoi Ibrahim Ramadan Aneel Tanwani

***Abstract*— Approximating the evolution of probability measures for nonlinear stochastic differential equations (SDEs) and the associated nonlinear filtering problems is a challenging problem as it involves solving high-dimensional differential equations. In contrast to classical variational inference methods which address this challenge by minimizing the Kullback-Leibler (KL) divergence between the true and approximate distributions, we propose a Wasserstein-based variational framework for approximating the laws of stochastic systems. In particular, instead of minimizing the KL divergence, our approach minimizes the Wasserstein-2 (W_2) distance between the joint probability distributions of the state and observation processes. This formulation respects the underlying transport geometry and results in evolution equations for Gaussian parameters that provide an approximation of the dynamics of the true measure. An illustration is provided for some of our results with the help of an academic example.**

I. Introduction

For nonlinear stochastic dynamical systems, describing the evolution of the underlying probability measure is a challenging task. The probability law of the system evolves according to the Kolmogorov forward equation (also known as the Fokker-Planck equation) [1], [2], a partial differential equation (PDE) whose analytical solution is available only in very special cases. In practice, the evolution of the probability measure is intractable for most nonlinear and non-Gaussian systems. A closely related problem arises in nonlinear filtering, where one seeks to infer the posterior distribution of the system state given noisy observations. The evolution of this conditional law is governed by stochastic partial differential equations (SPDEs) such as the Kushner-Stratonovich and Zakai equations [3]–[5]. Solving these SPDEs exactly is rarely feasible, which motivates the search for approximation techniques that can capture the essential behavior of the evolving probability measures.

Several methods have been proposed for computing approximate solutions to the evolution of probability measures associated with stochastic processes. Among these, Gaussian approximations–where the true measure is projected onto the family of Gaussian distributions–have proved particularly popular [6], [7]. Such approaches lead to moment closure schemes or variational formulations where the mean and covariance evolve according to ordinary differential equations derived from minimizing an appropriate divergence measure. In this context, variational inference (VI) provides a principled framework for approximating intractable probability distributions by tractable surrogates. In the classical setting, the approximation is obtained by minimizing the Kullback–Leibler (KL) divergence between the true distribution and the approximating family. The KL-based variational formulation has been applied to both stochastic differential equations (SDEs) and the nonlinear filtering problem [8]–[10], yielding efficient Gaussian and mixture-Gaussian approximations to the corresponding posterior or marginal distributions. These approaches have also found wide application in Bayesian inference, control, and signal processing [11].

However, the KL divergence has inherent limitations when used for approximating dynamical systems. It is asymmetric and primarily sensitive to differences in low-probability regions, which can lead to biased approximations–particularly in the tails or in multimodal settings [12]. Moreover, minimizing the KL divergence often corresponds to moment matching or energy functional minimization that does not directly reflect the geometry of the underlying probability space. As an alternative to finding approximations using KL divergence, we take a different perspective in this paper and consider a variational approach based on Wasserstein-2 (W_2) distance. That is, instead of the KL divergence, we measure the discrepancy between probability measures using the W_2 distance. The W_2 distance arises naturally while studying optimal transport problems where the transportation cost between two prespecified marginals is described by a quadratic function. In contrast to other distance measures between probability distributions, such as the total variation distance, the W_2 distance defines a geometrically meaningful metric on the space of probability measures. This property has made it particularly valuable for a wide range of data science applications (see, for example, [13]). We refer to [14]–[16] for textbook treatments on optimal transport.

Our contribution differs from classical VI-based approaches in two aspects. First, we propose to approximate the joint probability distribution of the state and observation processes, instead of concentrating ex-

The authors are with Laboratoire d'analyse et d'architecture des systèmes (LAAS)–CNRS, University of Toulouse, France. Email addresses: spchhatoi@laas.fr (Saroj Prasad Chhatoi), ibrahim.ramadan@laas.fr (Ibrahim Ramadan), aneel.tanwani@laas.fr (Aneel Tanwani).

This work was partly supported by the project CyPhAI, financed by ANR-JST CREST program with grant number ANR-20-JSTM-0001.

clusively on the conditional posterior, as in [17]. This formulation provides a more global view of the system dynamics and avoids separately handling normalization or marginalization inherent in conditional distributions. Second, we obtain closed form relations for linear surrogate by minimizing the W_2 distance between the joint distributions, leading to a Gaussian approximation that is consistent with the natural geometry of the space of probability measures. This is enabled by our continuous–discrete setup: an Euler–Maruyama step makes the conditional laws Gaussian even under nonlinear dynamics. Since W_2 distance between Gaussians has an explicit formula, we can solve the variational problem analytically, without solving a separate optimal-transport problem at each iteration. While the use of W_2 metric between joint distributions is not seen in the literature, we see that the recent work [18] provides Wasserstein-1 bounds on the gap between the true joint distribution and its moment-matched Gaussian approximation in nonlinear filters.

The remainder of the paper is organized as follows. Section II formulates the problem of designing a Gaussian surrogate using Wasserstein-based variational formulation and introduces some relevant notions from optimal transport. In Section III, we solve the problem for the fully observed SDEs case and present an illustrative example. Section IV addresses the partially observed SDEs and derives an approximation of the posterior distribution. The proofs are omitted in this article and can be found in the extended version [19].

II. Problem Setting

Let us consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the standard Wiener process $(\beta_t)_{t\geq 0} \in \mathbb{R}^m$ with infinitesimal covariance equal to identity, that is, $\mathsf{E}[\mathrm{d}\beta_t \mathrm{d}\beta_t^\top] = I_m \mathrm{d}t$. The state process of our interest is a continuous-time Markov process $(X_t)_{t\geq 0}$ in $\mathbb{R}^n$ whose evolution is described by the following stochastic differential equation (SDE):

$$\mathrm{d}X_t = f(t, X_t)\,\mathrm{d}t + G(t)\,\mathrm{d}\beta_t, \quad X_0 \sim \rho_0 \tag{1}$$

where $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ describes the drift term, and the diffusion term $G : \mathbb{R} \to \mathbb{R}^{n\times m}$ is assumed to be time-dependent only, and X_0 is independent of $(\beta_t)_{t\geq 0}$. Under standard regularity assumptions (e.g., global Lipschitz and linear growth in x and $G(\cdot)$ being locally bounded), there exists unique strong solution $(X_t)_{t\geq 0}$ [2, Theorem 5.2.1]. Note, due to the nonlinearity in f, the process need not be Gaussian even when ρ_0 is Gaussian.

As noted earlier, we consider the approximation with respect to the W_2 distance between distributions. In what follows, we first overview some fundamental notions from the theory of optimal transport to define W_2 distance. Then, we formulate the first optimization problem that provides an approximation of the distribution of the state process. We then associate an observation process with (1) and formulate an optimization problem to approximate the posterior of the state conditioned upon this observation process.

A. Optimal Transport Preliminaries

We briefly recall notations from optimal transport (see, e.g., [13], [14]).

Spaces and measures. Let $(\mathcal{X}, \|\cdot\|)$ be an Euclidean space (or a Polish metric space) and let $\mathcal{P}_2(\mathcal{X})$ be the set of Borel probability measures with finite second moment:

$$\mathcal{P}_2(\mathcal{X}) := \big\{\mu : \int_{\mathcal{X}} \|x\|^2 \,\mathrm{d}\mu(x) < \infty\big\}.$$

Pushforward. For a measurable map $\mathsf{T} : \mathcal{X} \to \mathcal{Y}$ and $\mu \in \mathcal{P}_2(\mathcal{X})$, the *pushforward* measure $\mathsf{T}_\sharp\mu \in \mathcal{P}_2(\mathcal{Y})$ is

$$\mathsf{T}_\sharp\mu(O) := \mu\big(\mathsf{T}^{-1}(O)\big), \qquad O \subset \mathcal{Y} \text{ Borel}.$$

Couplings and marginals. For $\mu \in \mathcal{P}_2(\mathcal{X})$ and $\nu \in \mathcal{P}_2(\mathcal{Y})$, a *coupling* (or transport plan) is a probability $\gamma \in \mathcal{P}_2(\mathcal{X}\times\mathcal{Y})$ whose marginals are μ and ν:

$$\Gamma(\mu,\nu) := \big\{\gamma : \ \pi_{1\sharp}\gamma = \mu,\ \pi_{2\sharp}\gamma = \nu\big\},$$

where $\pi_1(x,y) = x$ and $\pi_2(x,y) = y$ are the canonical projections, and $\pi_{i\sharp}\gamma$ denotes the pushforward (i.e., the i-th marginal).

Quadratic Wasserstein distance. For $\mu, \nu \in \mathcal{P}_2(\mathcal{X})$, the Wasserstein-2 ($W_2$) distance is defined as

$$W_2^2(\mu,\nu) \ := \ \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^2 \, d\gamma(x,y). \tag{2}$$

This defines a metric on $\mathcal{P}_2(\mathcal{X})$.

Product spaces and notation. For random vectors (U, V) on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ with law P_{UV}, we write

$$P_U := \pi_{1\sharp}P_{UV}, \qquad P_V := \pi_{2\sharp}P_{UV}.$$

Gaussian case (closed form and map). Let $\mu = \mathcal{N}(m_1, \Sigma_1)$ and $\nu = \mathcal{N}(m_2, \Sigma_2)$ be two non-degenerate Gaussian measures on $\mathbb{R}^n$. Then the squared W_2 distance between μ and ν admits the following closed form:

$$W_2^2(\mu,\nu) = \|m_1 - m_2\|^2 + \mathcal{B}^2(\Sigma_1, \Sigma_2), \tag{3}$$

where $\mathcal{B}^2(\Sigma_1,\Sigma_2) = \mathrm{Tr}\Big(\Sigma_1 + \Sigma_2 - 2\,(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}\Big)$, is the *squared Bures metric* between covariance matrices (see [20] for the closed form and [21] for modern treatment of Wasserstein-Bures metric). Moreover, when μ is absolutely continuous w.r.t. the Lebesgue measure (e.g., $\Sigma_1 \succ 0$), the unique W_2–optimal transport map $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ pushing μ to ν is affine:

$$\mathcal{T}(x) = m_2 + M\,(x - m_1), \qquad \text{where } M := \Sigma_2^{1/2}\left(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}\right)^{-1/2}\Sigma_2^{1/2} \tag{4}$$

(see [22] for the explicit map).

B. Gaussian Surrogate for Nonlinear SDE

To address the problem of computing a Gaussian approximation of process defined in (1) over an interval $[0, T]$, let us consider the Euler–Maruyama discretization scheme with step size τ. We choose a time instant $s \geq 0$, such that, for some $k \in \mathbb{N}$, we have $s = k\tau$. We use the notation x_s, or x_k when the time step τ is obvious from the context, to denote $x_{k\tau}$. The discretization, therefore, leads to the transition density

$$\rho(x_{s+\tau}|x_s) = \mathcal{N}\left(x_s + \tau f(x_s), \Sigma_x(s)\right), \quad \text{where } \Sigma_x(s) := G(s)G(s)^\top \tau. \tag{5}$$

To describe the family of approximating distributions σ, we consider a surrogate linear SDE that admits a Gaussian evolution, namely

$$\mathrm{d}Z_t = (A(t)Z_t + b(t))\,\mathrm{d}t + B(t)\mathrm{d}\beta_t \tag{6}$$

with deterministic time-varying design parameters $A : \mathbb{R}_{\geq 0} \to \mathbb{R}^{n\times n}$, $B : \mathbb{R}_{\geq 0} \to \mathbb{R}^{n\times m}$, and the vector $b : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$. Discretizing the linear system (6) with time step τ yields the Gaussian transition kernel

$$\sigma(z_{s+\tau}|z_s) = \mathcal{N}\left(z_s + \tau A(s)z_s + b(s)\tau, \Sigma_z(s)\right), \quad \text{such that } \Sigma_z(s) := B(s)B(s)^\top \tau. \tag{7}$$

The problem of finding an optimal linear surrogate for a SDE over a certain time interval can be formulated as finding optimal system parameters for each time step by minimizing the Wasserstein-2 distance defined in (2) between the nonlinear and linear joint laws at each time instant. In other words, let $\rho_x(x_{s+\tau}, x_s)$ denote the joint probability distribution between $x_{s+\tau}$ and x_s induced by the transition kernel in (5). The objective is to design the parameters in (6) so that the joint distribution $\sigma_z(z_{s+\tau}, z_s)$, induced by the transition kernel in (7), approximates $\rho_x(x_{s+\tau}, x_s)$ with respect to the W_2 metric. In particular, for every time $s = k\tau$, $k \in \mathbb{N}$, we determine $A(s), b(s)$ and $B(s)$ as minimizers of

$$\min_{A(s),b(s),B(s)} W_2^2\left(\rho_x(x_{s+\tau}, x_s), \sigma_z(z_{s+\tau}, z_s)\right)$$

as the discretization step $\tau \to 0$. The solution to this problem is provided in Section III, where we construct a Gaussian process (6) in which the parameters $A(s), b(s)$ and $B(s)$ are chosen to be optimal at each time step.

C. Gaussian Surrogate for Posterior

We next consider the problem of extending our approach for computing the Gaussian approximation of posterior distribution of the state conditioned upon an observation process. More precisely, with system (1), we associate an observation process

$$Y_{t_k} = h(X_{t_k}) + \epsilon_k, \tag{8}$$

where, for the sake of simplicity in this paper, we assume that $t_k = k\tau$, with $k \in \mathbb{N}$. The mapping $h : \mathbb{R}^n \to \mathbb{R}^p$ is assumed to be continuous, and $\epsilon_k \sim \mathcal{N}(0, R_k)$ is the mean-zero Gaussian noise with covariance $R_k \in \mathbb{R}^{p\times p}$ symmetric and positive definite, for each $k \in \mathbb{N}$.

The posterior distribution, which we want to approximate, is ideally determined using the Bayes rule:

$$\rho(x_{t_k}|y_{t_k}) \propto \rho(y_{t_k}|x_{t_k})\,\rho(x_{t_k}), \tag{9}$$

where $\rho(y_{t_k}|x_{t_k})$ is the likelihood function, and $\rho(x_{t_k})$ is the prior distribution at time t_k. We consider the joint distribution over states at consecutive sampling instants, and the next observation, that is,

$$\rho_{xy}(x_{s+\tau}, x_s, y_{s+\tau}) \in \mathcal{P}(\mathbb{R}^{2n+p}).$$

By the disintegration theorem [23, Corollary 10.4.13], this joint law can be written as

$$\rho_{xy}(x_{s+\tau}, x_s, y_{s+\tau}) = \rho(y_{s+\tau}|x_{s+\tau}, x_s)\,\rho_x(x_{s+\tau}, x_s).$$

Under the Markovian assumption on the state process, the dependence simplifies to

$$\rho_{xy}(x_{s+\tau}, x_s, y_{s+\tau}) = \rho(y_{s+\tau}|x_{s+\tau})\,\rho_x(x_{s+\tau}, x_s),$$

which can be further decomposed as

$$\rho_{xy}(x_{s+\tau}, x_s, y_{s+\tau}) = \rho(y_{s+\tau}|x_{s+\tau})\,\rho(x_{s+\tau}|x_s)\,\rho(x_s). \tag{10}$$

To obtain Gaussian approximation for this posterior, we append the linear surrogate (6) with a linear observation process:

$$w_k = H_k Z_k + d_k + \delta_k, \tag{11}$$

so that the measurement w_k is available at time $s = k\tau$. Here, for each $k \in \mathbb{N}$, $\delta_k \sim \mathcal{N}(0, \tilde{R}_k)$ and we want to choose the positive definite covariance matrix $\tilde{R}_k$, the matrix $H_k \in \mathbb{R}^{p\times n}$, and the vector $d_k \in \mathbb{R}^p$ to compute the best possible Gaussian approximation of the posterior using this model. We next consider the joint distribution coming from the surrogate model (6) and (11)

$$\sigma_{zw}(z_{s+\tau}, z_s, w_{k+1}) = \sigma(w_{k+1}|z_{s+\tau})\,\sigma(z_{s+\tau}|z_s)\,\sigma(z_s). \tag{12}$$

Define $\tilde{x} := (x_{s+\tau}, x_s)$ and $\tilde{z} := (z_{s+\tau}, z_s)$. The W_2 distance between the joint distributions σ_{zw} and ρ_{xy} is then given by

$$W_2^2(\rho_{xy}, \sigma_{zw}) = \inf_{\gamma\in\mathcal{P}(\mathbb{R}^{2n+p}\times\mathbb{R}^{2n+p})} \int \left(\|\tilde{x} - \tilde{z}\|^2 + \|y_{k+1} - w_{k+1}\|^2\right) \mathrm{d}\gamma(\tilde{x}, y_{k+1}; \tilde{z}, w_{k+1}) \quad \text{such that} \quad \pi_{1\sharp}\gamma = \rho_{xy},\ \pi_{2\sharp}\gamma = \sigma_{zw}. \tag{13}$$

We propose projecting ρ_{xy} onto the Gaussian manifold using the Wasserstein metric, that is,

$$\min_{\sigma_{zw}} W_2^2(\sigma_{zw}, \rho_{xy}), \tag{14}$$

such that σ_{zw} is a Gaussian distribution. The minimization above is carried out w.r.t. the system parameters $A(s), b(s), B(s), H_k, d_k, \tilde{R}_k$.

With the disintegration shown in (12), the problem in (14) naturally decomposes into a two-step prediction–update procedure. In particular, the projection decomposes into two parts: the first corresponds to the propagation step described earlier in Section III, while the second corresponds to the posterior projection step.

III. Gaussian Surrogate for State Process

In this section, we address the problem of calculating the Gaussian surrogate for the nonlinear SDE (1), as defined in Section II-B, without taking the observation process into consideration.

We begin by recalling the notation. Let $(X_t)_{t\in[0,T]}$ solve the nonlinear SDE (1). For a fixed $\tau \geq 0$ such that $s = k\tau$ for $k \in \mathbb{N}$, the joint distribution between $X_{s+\tau}$ and X_s is denoted by ρ_x. Similarly, let $(Z_t)_{t\in[0,T]}$ solve the linear surrogate SDE (6) with system parameters $(A(s), B(s), b(s))$, and let the joint distribution between $Z_{s+\tau}$ and Z_s be denoted by σ_z. The W_2 distance between ρ_x and σ_z is given by

$$W_2^2(\rho_x, \sigma_z) = \inf_{\kappa\in\mathcal{P}(\mathbb{R}^{2n}\times\mathbb{R}^{2n})} \int \big\|(x_{s+\tau}, x_s) - (z_{s+\tau}, z_s)\big\|^2 \,\mathrm{d}\kappa \quad \text{such that} \quad \pi_{1\sharp}\kappa = \rho_x,\ \pi_{2\sharp}\kappa = \sigma_z, \tag{15}$$

where $\pi_i : \mathbb{R}^{2n} \to \mathbb{R}^n$, $i = 1, 2$, are the canonical projection mappings such that $\pi_1(z, x) = z$, and $\pi_2(z, x) = x$. We are now ready to solve the surrogate design problem which is expressed as

$$\min_{A(s),\, b(s),\, B(s)} W_2^2\big(\rho_x, \sigma_z\big). \tag{16}$$

In order to obtain computable approximations for ρ, at each time step we approximate the true prior distribution $\rho \in \mathcal{P}(\mathbb{R}^n)$ with a Gaussian distribution $\sigma \in \mathcal{P}(\mathbb{R}^n)$. This approximation is characterized by a mean $\mu_s = \mathsf{E}_\sigma(Z_s)$ and a covariance $\Sigma_s = \mathsf{E}_\sigma((Z_s - \mu_s)(Z_s - \mu_s)^\top)$, so that $\sigma = \mathcal{N}(\mu_s, \Sigma_s)$.

A. Main result

We now state the main result. Its proof needs an intermediate technical lemma stated below.

Theorem 1. *The minimizer of* (16) *is unique and given by*

$$A^\star(s) = C(s)\,\Sigma_s^{-1}, \tag{17}$$

$$b^\star(s) = \phi(s) - A^\star(s)\,\mu_s, \tag{18}$$

$$B^\star(s)\,B^\star(s)^\top = G(s)\,G(s)^\top, \tag{19}$$

where $\phi(s) := \mathsf{E}_\sigma[f(s, X_s)]$, *and* $C(s) := \mathsf{E}_\sigma\big[(f(s, X_s) - \phi_s)(X_s - \mu_s)^\top\big]$.

Lemma 2. *The* W_2 *distance between* $\sigma_z \in \mathcal{P}(\mathbb{R}^{2n})$ *and* $\rho_x \in \mathcal{P}(\mathbb{R}^{2n})$ *in* (16) *is given by*

$$W_2^2(\rho_x(x_{s+\tau}, x_s),\, \sigma_z(z_{s+\tau}, z_s)) = \int_{\mathbb{R}^n} W_2^2\big(\rho(x_{s+\tau}|x_s), \sigma(z_{s+\tau}|x_s)\big)\mathrm{d}\sigma(x_s). \tag{20}$$

Proof of Theorem 1 follows from substitution of (3) in (20) and finding minimizers of the resulting expression.

B. Corollary and Illustration

The optimal system data $(A^\star(s), b^\star(s), B^\star(s))$ can then be used to propagate the mean and covariance of the linear model as follows:

$$\begin{aligned}
\mu_{s+\tau} &= \mu_s + \tau\,\mathsf{E}_\sigma[f(s, X_s)],\\
\Sigma_{s+\tau} &= \Sigma_s + \tau\,\mathsf{E}_\sigma\big[f(s, X_s)(X_s - \mu_s)^\top\big]\\
&\quad + \tau\,\mathsf{E}_\sigma\big[(X_s - \mu_s)f(s, X_s)^\top\big]\\
&\quad + \tau^2\,\mathsf{E}_\sigma\big[f(s, X_s)(X_s - \mu_s)^\top\big]\Sigma_s^{-1}\mathsf{E}_\sigma\big[(X_s - \mu_s)f(s, X_s)^\top\big]\\
&\quad + \tau\,G(s)G(s)^\top.
\end{aligned} \tag{21}$$

Taking the limit $\tau \to 0$, we find a set of two coupled ODEs:

$$\begin{aligned}
\dot\mu(t) &= \mathsf{E}_\sigma[f(t, X)],\\
\dot\Sigma(t) &= \mathsf{E}_\sigma[f(t, X)(X - \mu)^\top] + \mathbb{E}_\sigma[(X - \mu)f(t, X)^\top]\\
&\quad + G(t)G(t)^\top.
\end{aligned} \tag{22}$$

Example 1. We consider a one-dimensional nonlinear SDE of the form

$$\mathrm{d}X_t = \sin(X_t)\,\mathrm{d}t + 0.3\,\mathrm{d}\beta_t.$$

Using the results obtained in Section III, specifically (17), (18), (19), we obtain a linear surrogate. To write the expressions explicitly, we make use of the fact that, for a Gaussian $\sigma \sim \mathcal{N}(\mu_t, \Sigma_t)$, we have $\mathsf{E}_\sigma(f(X_t)) = \mathsf{E}_\sigma(\sin(X_t)) = e^{-\frac{1}{2}\Sigma_t}\sin(\mu_t)$. Similarly, $\mathsf{E}_\sigma(\cos(X_t)) = e^{-\frac{1}{2}\Sigma_t}\cos(\mu_t)$. Moreover, as $f(X_t) = \sin(X_t)$ is a differentiable function, then by Stein's Lemma [24], we have $\mathsf{E}_\sigma(\sin(X_t)(X_t - \mu_t)) = \Sigma_t\mathsf{E}_\sigma(\cos(X_t))$. Using these expressions, we arrive at $A(t) = \mathsf{E}_\sigma(\cos(X_t))$, $b(t) = \mathsf{E}_\sigma(\sin(X_t)) - A(t)\mu_t$, and $B(t) = G(t) = 0.3$. Thus, the linear approximating SDE becomes

$$\mathrm{d}Z_t = \big[\mathsf{E}_\sigma(\cos(X_t))Z_t + \mathsf{E}_\sigma(\sin(X_t)) - \mathsf{E}_\sigma(\cos(X_t))\mu_t\big]\,\mathrm{d}t + 0.09\,d\beta_t.$$

Using (21) we find the following expressions for the propagation of mean and covariance:

$$\begin{aligned}
\mu_{s+\tau} &= \mu_s + \big(\mathsf{E}_\sigma(\sin(X_s))\big)\,\tau,\\
\Sigma_{s+\tau} &= \Sigma_s + \tau\big[2\Sigma_s\mathsf{E}_\sigma(\cos(X_s)) + 0.09\big].
\end{aligned}$$

The simulation of different sample paths of the nonlinear SDE and its Gaussian approximation, with time step $\tau = 0.01$, are shown in Figure 1.

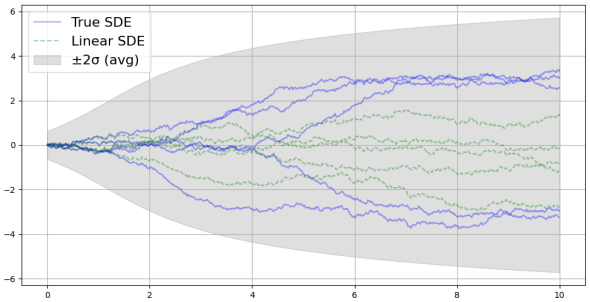

Fig. 1: Different sample paths of the nonlinear SDE and its Gaussian approximation.

IV. From Propagation to Posterior

Next, we address the problem described in Section II-C of obtaining a Gaussian approximation of the posterior distribution in (9). We focus our attention on time-discretized version of the problem, where we fix a step size $\tau > 0$ and use the subscript $k \in N$ to denote the time instant $s = k\tau$; similarly, $k+1$ refers to the time instant $s+\tau$. To this end, we formulate a minimization problem over the system parameters $(A(s),\, b(s),\, B(s),\, H_k,\, d_k,\, \widetilde{R}_k)$, with a squared W_2 objective measuring the distance between the joint distributions $\rho_{xy} \in \mathcal{P}(\mathbb{R}^{2n+p})$ and $\sigma_{zw} \in \mathcal{P}(\mathbb{R}^{2n+p})$. For notational convenience, let $\tilde{x} := (x_{k+1}, x_k)$ and $\tilde{z} := (z_{k+1}, z_k)$, and define $\rho_x(\tilde{x}) := \rho_x(x_{k+1}, x_k)$ and $\sigma_z(\tilde{z}) := \sigma_z(z_{k+1}, z_k)$. We disintegrate any admissible coupling γ as

$$\mathrm{d}\gamma(\tilde{x}, y;\, \tilde{z}, w) = \mathrm{d}\eta(y, w \mid \tilde{x}, \tilde{z})\; d\kappa(\tilde{x}, \tilde{z}), \tag{23}$$

where $\kappa \in \Gamma(\rho_x, \sigma_z)$, and for each $(\tilde{x}, \tilde{z})$, $\eta(\,\cdot\,, \cdot \mid \tilde{x}, \tilde{z})$ is a coupling between $\rho(y_{k+1}|\tilde{x})$ and $\sigma(w_{k+1}|\tilde{z})$.
Substituting the disintegrated measure (23) into (13), we obtain

$$\begin{aligned} W_2^2(\rho_{xy}, \sigma_{zw}) = \inf_{\kappa} \int_{\mathbb{R}^{2n}\times\mathbb{R}^{2n}} \Bigg\{ &\|\tilde{x} - \tilde{z}\|^2 \\ &+ \inf_{\eta} \int_{\mathbb{R}^p\times\mathbb{R}^p} \|y - w\|^2 \,\mathrm{d}\eta(y, w \mid \tilde{x}, \tilde{z}) \Bigg\} \mathrm{d}\kappa(\tilde{x}, \tilde{z}) \\ \text{such that } \;\; \kappa \in \Gamma(\rho_x, \sigma_z),\; &\eta \in \Gamma\big(\,\rho(y_{k+1}|\tilde{x}),\, \sigma(w_{k+1}|\tilde{z})\,\big). \end{aligned} \tag{24}$$

The second term inside the curly braces in (24) represents the conditional transport cost between the distributions of the observation models, i.e., $W_2^2\big(\rho(y_{k+1}|\tilde{x}),\, \sigma(w_{k+1}|\tilde{z})\big)$. Thus, we can equivalently write the problem as

$$\begin{aligned} W_2^2(\rho_{xy}, \sigma_{zw}) = \inf_{\kappa} \int_{\mathbb{R}^{2n}\times\mathbb{R}^{2n}} \Big\{ &\|\tilde{x} - \tilde{z}\|^2 + \\ &W_2^2\big(\rho(y_{k+1}|\tilde{x}),\, \sigma(w_{k+1}|\tilde{z})\big) \Big\} d\kappa(\tilde{x}, \tilde{z}). \end{aligned} \tag{25}$$

In order to find a tractable iterative solution for the problem obtained after the substitution of (25) into (14), we adopt a two–stage approximation. First, we find $\kappa^\star$ which minimizes the first term on the RHS of (25), i.e., $\inf_\kappa \int_{\mathbb{R}^{2n}\times\mathbb{R}^{2n}} \|\tilde{x} - \tilde{z}\|^2 \mathrm{d}\kappa(\tilde{x}, \tilde{z})$ which corresponds to $W_2^2(\rho_x, \sigma_z)$. We then use this $\kappa^\star$ as an approximation of the optimal coupling κ in (25), yielding

$$\begin{aligned} \min_{\Delta} W_2^2(\rho_{xy}, \sigma_{zw}) \le \min_{\Delta} \Big\{ &W_2^2(\rho_x, \sigma_z) + \\ &\int_{\mathbb{R}^{2n}\times\mathbb{R}^{2n}} W_2^2(\rho(y_{k+1}|\tilde{x}),\, \sigma(w_{k+1}|\tilde{z})) d\kappa^\star(\tilde{x}, \tilde{z}) \Big\}, \end{aligned} \tag{26}$$

where $\kappa^\star$ is the optimizer of the first term in (26), and $\Delta := (A(s), b(s), B(s), H_{k+1}, d_{k+1}, \tilde{R}_{k+1})$ denotes the collection of all design parameters.
The first term on the RHS of (26) depends only on $(A(s), b(s), B(s))$ and has already been calculated in Section III. This term can be seen as a propagation step, as it provides the optimal parameters for the linear system (6), which are then used to compute the updated mean and covariance as in (21).
In the next lemma, we derive the optimal coupling $\kappa^\star$, which will be crucial in handling the second term on the RHS of (25).

Lemma 3. *The optimal coupling $\kappa^\star$ is given by*

$$\begin{aligned} \mathrm{d}\kappa^\star(x_{k+1}, x_k; z_{k+1}, z_k) = \mathrm{d}\sigma_z(x_k)\, \delta_{x_k}(z_k) \\ \times \mathrm{d}\rho(x_{k+1}|x_k)\, \delta_{\mathcal{T}_{x_k}(x_{k+1})}(\mathrm{d}z_{k+1}). \end{aligned} \tag{27}$$

where $\mathcal{T}_{x_k}(x_{k+1}) = x_{k+1} - \big(A_k^\star x_k + b_k^\star - f(x_k)\big)\tau$ and $x_{k+1} \sim \rho(\cdot|x_k)$.

A. Update step

Next, we focus on the second term on the RHS of (26). The Markovian nature of the dynamics in (8) and (11) implies that the coupling η, conditioned on $(\tilde{x}, \tilde{z}) = (x_{k+1}, x_k, z_{k+1}, z_k)$, depends only on x_{k+1} and z_{k+1}. Hence, we obtain

$$\begin{aligned} &W_2^2(\rho(y_{k+1}|\tilde{x}),\, \sigma(w_{k+1}|\tilde{z})) \\ &= \inf_{\eta} \int \|y_{k+1} - w_{k+1}\|^2 \, d\eta(y_{k+1}, w_{k+1} \mid x_{k+1}, z_{k+1}) \\ &= W_2^2(\rho(y_{k+1}|x_{k+1}),\, \sigma(w_{k+1}|z_{k+1})) \\ &\text{such that } \eta \in \Gamma\big(\rho(y_{s+\tau} \mid x_{s+\tau}),\, \sigma(w_{s+\tau} \mid z_{s+\tau})\big). \end{aligned}$$

Thus, the problem of finding a linear surrogate for the partially observed system, as presented in (14), can be written as

$$\begin{aligned} \min_{\Delta} W_2^2(\rho_{xy}, \sigma_{zw}) \le \min_{\Delta} \Big\{ &W_2^2(\rho_x, \sigma_z) + \\ &\mathsf{E}_{\kappa^\star}\Big[W_2^2\big(\rho(y_{k+1}|x_{k+1}),\, \sigma(w_{k+1}|z_{k+1})\big)\Big] \Big\} \end{aligned} \tag{28}$$

where $\Delta := (A(s), b(s), B(s), H_{k+1}, d_{k+1}, \tilde{R}_{k+1})$.

For the discrete-time observation model (8), the conditional distribution $\rho(y_{k+1}|x_{k+1})$ is Gaussian and can be written as

$$\rho(y_{k+1}|x_{k+1}) = \mathcal{N}\big(h(x_{k+1}),\, R_{k+1}\big),\; R_{k+1} \succ 0, \tag{29}$$

whereas for the linear surrogate model (11), the conditional distribution is given by

$$\sigma(w_{k+1}|z_{k+1}) = \mathcal{N}\big(H_{k+1}z_{k+1} + d_{k+1},\, \widetilde{R}_{k+1}\big),\; \widetilde{R}_{k+1} \succ 0. \tag{30}$$

Now, to derive the surrogate linear observation model in (24), we consider the following optimization problem:

$$\begin{aligned} \min_{H_{k+1}, d_{k+1}, \tilde{R}_{k+1}} \mathsf{E}_{\kappa^*}\Big[\, W_2^2\big(\rho(y_{k+1}|x_{k+1})\,, \\ \sigma(w_{k+1}|z_{k+1})\big)\Big]. \end{aligned} \tag{31}$$

Proposition 4. *The optimization problem* (31) *admits a unique solution given by*

$$H^\star_{k+1} = \mathsf{E}_\psi\big[(h(X_{k+1}) - \mu_h)(Z - \mu_Z)^\top\big]\,\Sigma_{ZZ}^{-1},$$
$$d^\star_{k+1} = \mathsf{E}_\psi\left[h(X_{k+1}) - H_{k+1}\,\mathcal{T}_{X_k}(X_{k+1})\right],$$
$$\widetilde{R}^\star_{k+1} = R_{k+1},$$

where $\Sigma_{ZZ} := \mathsf{E}_\psi\big[(Z-\mu_Z)(Z-\mu_Z)^\top\big]$, $\mu_h = \mathsf{E}_\psi[h(X_{k+1})]$ *and* $\mu_Z = \mathsf{E}_\psi[Z]$ *with* $Z = \mathcal{T}_{x_k}(X_{k+1})$. *The expectation is taken w.r.t.* $\mathrm{d}\psi(x_{k+1}, x_k) = \mathrm{d}\sigma(x_k)\,\mathrm{d}\rho(x_{k+1}|x_k)$.

B. Surrogate Kalman Filter

Using the results from the previous sections, we have obtained a linear system which is a projected version of the nonlinear system using the W_2 metric. This yields the following Kalman-type filter.

a) Propagation step: Given the prior $(\mu_{s|s}, \Sigma_{s|s})$, and using the notation $A^\star_\tau = (\mathrm{Id} + \tau A^\star)$, the *prediction* step is

$$\mu_{s+\tau|s} = A^\star_\tau(s)\,\mu_{s|s} + \tau\, b^\star(s),$$
$$\Sigma_{s+\tau|s} = \Sigma_{s|s} + A^\star_\tau(s)\,\Sigma_{s|s} + \Sigma_{s|s}\,A^\star_\tau(s)^\top + A^\star_\tau(s)\,\Sigma_{s|s}\,A^\star_\tau(s)^\top + \tau\, B^\star(s)\,B^\star(s)^\top.$$

b) Update step: Upon observing $y_{s+\tau}$, the *update* step takes the Kalman form with the projected measurement:
Innovation covariance:

$$S_{s+\tau} = H^\star_{s+\tau}\,\Sigma_{s+\tau|s}\,{H^\star_{s+\tau}}^\top + \widetilde{R}^\star_{s+\tau}.$$

W_2–projected gain:

$$K_{s+\tau} = \Sigma_{s+\tau|s}\,{H^\star_{s+\tau}}^\top\,S^{-1}_{s+\tau}.$$

Update:

$$\mu_{s+\tau|s+\tau} = \mu_{s+\tau|s} + K_{s+\tau}\big(w_{s+\tau} - H^\star_{s+\tau}\,\mu_{s+\tau|s} - d^\star_{s+\tau}\big),$$
$$\Sigma_{s+\tau|s+\tau} = (I_n - K_{s+\tau}H^\star_{s+\tau})\,\Sigma_{s+\tau|s}\,(I_n - K_{s+\tau}H^\star_{s+\tau})^\top + K_{s+\tau}\,\widetilde{R}^\star_{s+\tau}\,K^\top_{s+\tau}.$$

V. Conclusions

We have proved two results about computing Gaussian approximations for stochastic dynamical systems. The first one relates to the probability measure obtained from Fokker-Planck equation. The second result corresponds to computing the approximation for the posterior distribution in continuous-discrete filtering problems.

To build on the results presented in this manuscript, an immediate direction is to overcome the suboptimal coupling used in Section IV for computing the posterior. In our current approach, the joint coupling is disintegrated into two parts, and the coupling from the propagation step is reused for the update step. While this is feasible, it is not necessarily optimal. The optimal update-step coupling remains to be computed, either analytically or via computationally tractable methods. Additionally, it would be valuable to compare the resulting filter, obtained by minimizing the distance between the joint state–output distributions, with KL-based variational inference approaches such as those in [9] and [25].

References

[1] A. Kolmogoroff. Über die analytischen methoden in der wahrscheinlichkeitsrechnung. *Mathematische Annalen*, 104(1):415–458, 1931.

[2] B. Øksendal. *Stochastic differential equations*. Springer, sixth edition, 2003.

[3] H.J. Kushner. Dynamical equations for optimal nonlinear filtering. *J. Differential Equations*, 3(2):179–190, 1967.

[4] M. Zakai. On the optimal filtering of diffusion processes. *Z. Wahrsch. Verw. Gebiete*, 11(3):230–243, 1969.

[5] M. Fujisaki, G. Kallianpur, and H. Kunita. Stochastic differential equations for the nonlinear filtering problem. *Osaka J. Math*, 9(1):19–40, 1972.

[6] D. Brigo. On SDEs with marginal laws evolving in finite-dimensional exponential families. *Statistics & probability letters*, 49(2):127–134, 2000.

[7] C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor. Gaussian process approximations of stochastic differential equations. In N.D. Lawrence, A. Schwaighofer, and J. Qui nonero Candela, editors, *Gaussian Processes in Practice*, volume 1 of *Proceedings of Machine Learning Research*, pages 1–16, 2007.

[8] C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Taylor. Variational inference for diffusion processes. *Advances in neural information processing systems*, 20, 2007.

[9] T. Sutter, A. Ganguly, and H. Koeppl. A variational approach to path estimation and parameter inference of hidden diffusion processes. *Journal of Machine Learning Research*, 17(190):1–37, 2016.

[10] M. Lambert, S. Bonnabel, and F. Bach. The continuous-discrete variational Kalman filter (CD-VKF). In *61st IEEE Conference on Decision and Control*, pages 6632–6639, 2022.

[11] S. Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *ArXiv preprint, arXiv:1805.00909*, 2018.

[12] A. Chan, H. Silva, S. Lim, T. Kozuno, A. R. Mahmood, and M. White. Greedification operators for policy optimization: Investigating forward and reverse KL divergences. *Journal of Machine Learning Research*, 23:1–79, 2022.

[13] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[14] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009.

[15] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Springer, Basel, 2nd edition, 2008.

[16] F. Santambrogio. *Optimal transport for applied mathematicians*. Birkhäuser Cham, 2015.

[17] M. Opper and C. Archambeau. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

[18] T. Karvonen and S. Särkkä. Wasserstein bounds for non-linear Gaussian filters. ArXiv preprint, arXiv:2503.21643.

[19] S.P. Chhatoi, I. Ramadan, and A. Tanwani. A Gaussian surrogate of partially observed stochastic processes using Wasserstein metric. Extended version. Available online: `https://hal.science/hal-05487303v1`.

[20] M. Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

[21] R. Bhatia, T. Jain, and Y. Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

[22] C. R. Givens and R.M. Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.

[23] V.I. Bogachev. *Measure Theory*. Springer-Verlag, Berlin, 2007.

[24] C.M. Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

[25] M. Lambert, S. Bonnabel, and F. Bach. The recursive variational Gaussian approximation (R-VGA). *Statistics and Computing*, 32(1):10, 2022.