# Approximate Inference with Wasserstein Gradient Flows

**Charlie Frogner**
CSAIL
Massachusetts Institute of Technology

**Tomaso Poggio**
Center for Brains, Minds, and Machines
Massachusetts Institute of Technology

## Abstract

We present a novel approximate inference method for diffusion processes, based on the Wasserstein gradient flow formulation of the diffusion. In this formulation, the time-dependent density of the diffusion is derived as the limit of implicit Euler steps that follow the gradients of a particular free energy functional. Existing methods for computing Wasserstein gradient flows rely on discretization of the domain of the diffusion, prohibiting their application to domains in more than several dimensions. We propose instead a discretization-free inference method that computes the Wasserstein gradient flow directly in a space of continuous functions. We characterize approximation properties of the proposed method and evaluate it on a nonlinear filtering task, finding performance comparable to the state-of-the-art for filtering diffusions.

## 1 INTRODUCTION

Diffusion processes are ubiquitous in science and engineering. They arise when modeling dynamical systems driven by random fluctuations, such as action potentials in neuroscience, interest rates and asset prices in finance, reaction dynamics in chemistry, and population dynamics in ecology. In signal processing and machine learning, diffusion processes provide the dynamics underlying classic filtering methods such as the Kalman filter (Kalman and Bucy, 1961).

Inference for general diffusions is an outstanding challenge. Each diffusion process defines a probability distribution that evolves in continuous time; inference involves solving for the distribution at a future time given an initial distribution at the current time. Exact, closed-form solutions are typically unavailable, and numerous approximations have been proposed, including parametric approximations (Kalman and Bucy, 1961; Julier et al., 1995), particle or sequential Monte Carlo methods (Crisan and Lyons, 1999; Fearnhead et al., 2008), MCMC methods (Roberts and Stramer, 2001; Golightly and Wilkinson, 2008) and variational approximations (Archambeau et al., 2007; Vrettas et al., 2015; Sutter et al., 2016). Each poses a different tradeoff between fidelity of the approximation and computational burden.

In this paper, we investigate a novel approximate inference method for nonlinear diffusions. It is based on a characterization, due to Jordan, Kinderlehrer and Otto (Jordan et al., 1998), of the diffusion process as following a **gradient flow** with respect to a Wasserstein metric on probability measures. Concretely, they define a time discretization of the diffusion process in which the approximate probability density $\rho_k$ at the $k$th timestep solves a variational problem,

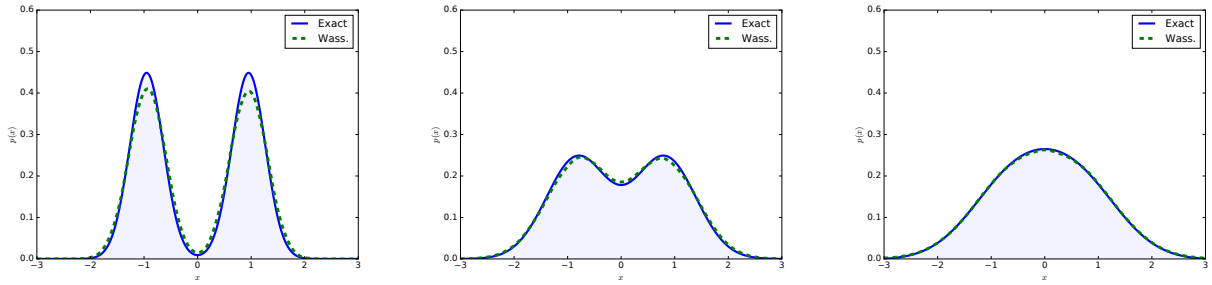
$$\rho_k = \operatorname*{argmin}_{\rho \in \mathcal{P}(\mathcal{X})} \mathcal{W}_2^2(\rho, \rho_{k-1}) + 2\tau f(\rho) \qquad (1)$$

with $\mathcal{W}_2 : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ being the 2−Wasserstein distance, $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ a free energy functional defining the diffusion process, and $\tau > 0$ the size of the timestep [1]. This discrete process is shown to converge, as $\tau \to 0$, to the exact diffusion process.

Exact computation of the time-discretized gradient step in (1) is intractable in general. Existing numerical methods rely on discretization of the domain of the diffusion, which restricts their application to spaces with very few dimensions – typically three or fewer. In this work, we propose a novel method for computing the gradient flow that avoids discretization, opting instead to operate directly on continuous functions lying in a reproducing kernel Hilbert space. Specifically, we derive a dual problem to (1) that uses a regularized Wasserstein distance in place of the unregularized one in (1). We

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. 

[1] $\mathcal{P}(\mathcal{X})$ is the space of probability measures defined on domain $\mathcal{X}$.




(a) $t = 0.05$. (b) $t = 0.2$ (c) $t = 0.5$.

Figure 1: Regularized Wasserstein gradient flow (Section 3) approximates closely an Ornstein-Uhlenbeck diffusion, initialized with a bimodal density. Both the regularization and the discrete timestep are sources of error. Shaded region is the true density.

show that, for a general strictly convex, smooth regularizer, this dual problem is an unconstrained stochastic program, which admits a tractable finite-dimensional RKHS approximation. This approach is motivated by a similar observation for the case of entropic regularization of optimal transport in (Genevay et al., 2016). Our proposed approximation yields an approximate inference method for diffusions that is computationally tractable in settings where domain discretization is impractical.

For reasonable values of the timestep $\tau$, the approximate inference method described in this paper can give a close approximation to the density of the diffusion. In Figure 1, for example, we compute the Wasserstein gradient flow for an Ornstein-Uhlenbeck diffusion, initialized with a bimodal density. We see that it follows the exact density closely.

The rest of this paper is organized as follows. In Section 2 we review diffusion processes and discuss related work. In Section 3 we derive a smoothed dual formulation of the Wasserstein gradient flow, and in Section 4 we use this dual formulation to derive a novel inference algorithm. In Section 5 we investigate theoretical properties. In Section 6 we characterize empirical performance of the proposed algorithm, before concluding.

## 2 BACKGROUND AND RELATED WORK

### 2.1 Diffusions, Free Energy, and the Fokker-Planck Equation

We consider a continuous-time stochastic process $X_t$ taking values in a smooth manifold $\mathcal{X}$, for $t \in [t_i, t_f]$, and having single-time marginal densities $\rho_t : \mathcal{X} \to \mathbb{R}$ with respect to a reference measure on $\mathcal{X}$. We are specifically interested in diffusion processes whose single-time marginal densities obey a diffusive partial differential equation,

$$\frac{\partial \rho_t}{\partial t} = \operatorname{div}\left[\rho_t \nabla f'(\rho_t)\right], \tag{2}$$

with $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ a functional on densities and $f'$ its gradient for the $L^2(\mathcal{X})$ metric.

$f$ is the *free energy* and defines the diffusion entirely. An important example, which will be our primary focus, is the *advection-diffusion process*, which is typically characterized as obeying an Itô stochastic differential equation,

$$dX_t = -\nabla w(X_t)dt + \beta^{-1/2} d\mathbf{W}_t \tag{3}$$

with $\nabla w$ being the gradient of a potential function $w : \mathcal{X} \to \mathbb{R}$, determining the advection or drift of the system, and $\beta^{-1/2} > 0$ the magnitude of the diffusion, which is driven by a Wiener process having stochastic increments $d\mathbf{W}_t$ (see (Kloeden and Platen, 2013) for a formal introduction) [2]. The advection-diffusion has marginal densities obeying a *Fokker-Planck* equation,

$$\frac{\partial \rho_t}{\partial t} = \beta^{-1} \Delta \rho_t + \operatorname{div}[\rho_t \nabla w], \tag{4}$$

which is a diffusive PDE with free energy functional $f(\rho) = \langle w, \rho \rangle_{L^2(\mathcal{X})} + \beta^{-1} \langle \rho, \log \rho \rangle_{L^2(\mathcal{X})}$, for scalar potential $w \in L^2(\mathcal{X})$. The advection-diffusion is *linear* whenever $\nabla w$ is linear in its argument.

We note that the current work applies to those diffusions that can be rendered into the form (2) via a change of variables. In particular, in the case of advection-diffusion, these are the *reducible* diffusions and include nearly all diffusions in one dimension (Aït-Sahalia, 2008).

[2] We assume sufficient conditions for existence of a strong solution to (3) are fulfilled (Oksendal, 2013) Thm. 5.2.1.

### 2.2 Approximate Inference for Diffusions

Inference for a nonlinear diffusion is generally intractable. Given an initial density at time $t_i$, the goal is to determine the single-time marginal density $\rho_t$ at some time $t > t_i$. Exact inference entails solving the foward PDE (2), for which closed-form solutions are seldom available.

**Domain discretization.** In certain cases, an Eulerian discretization of the domain, i.e. a fixed mesh, is available. Here one can apply standard numerical integration methods such as Chang and Cooper's (Chang and Cooper, 1970) or entropic averaging (Pareschi and Zanella, 2017) for integrating the Fokker-Planck PDE. A number of Eulerian methods have been proposed for Wasserstein gradient flows, as well, including finite element (Burger et al., 2009) and finite volume methods (Carrillo et al., 2015). Entropic regularization of the problem yields an efficient iterative method (Peyré, 2015). Lagrangian discretizations, which follow moving particles or meshes, have also been explored (Carrillo and Moll, 2009; Westdickenberg and Wilkening, 2010; Budd et al., 2013; Benamou et al., 2016).

**Particle simulation.** One approach to inference approximates the target density by a weighted sum of delta functions, $\rho_t(\mathbf{x}) = \sum_{i=1}^{N} \mathbf{w}_i \delta_{\mathbf{x}_t^{(i)}=\mathbf{x}}$, at locations $\mathbf{x}_t^{(i)} \in \mathcal{X}$. Each delta function represents a "particle," and can be obtained by sampling an initial location $\mathbf{x}_{t_i}$ according to $\rho_{t_i}$, then forward simulating a trajectory from that location, according to the diffusion. Standard simulation methods such as Euler-Maruyama discretize the time interval $[t_i, t]$ and update the particle's location recursively (Kloeden and Platen, 2013). For a fixed time discretization, such methods are biased in the sense that, with increasing number of particles, they converge only to an approximation of the true predictive density. To address this, one can use a rejection sampling method (Beskos et al., 2005, 2008) to sample exactly (with no bias) from the distribution over trajectories. Density estimation can be used to extrapolate the inferred density beyond the particle locations (Durham and Gallant, 2002; Hurn et al., 2003).

In the context of generative modeling, Liutkus et al. (2019) investigates an Euler-Maruyama-type particle method to simulate gradient flows for the sliced Wasserstein metric, which is not the same as the Wasserstein metric investigated here.

**Parametric approximations.** One can also approximate the predictive density by a member of a parametric class of distributions. This parametric density might be chosen by matching moments or another criterion. The extended Kalman filter (Kalman and Bucy, 1961; Kushner, 1967), for example, chooses a Gaussian density whose mean and covariance evolve according to a first order Taylor approximation of the dynamics. Sigma point methods such as the unscented Kalman filter (Julier et al., 1995, 2000; Sarkka, 2007) select a deterministic set of points $\mathbf{x}_t^{(i)} \in \mathcal{X}$ that evolve according to the exact dynamics of the process, such that the mean and covariance of the true predictive density is well-approximated by finite sums involving only these points. The mean and covariance so computed then define a Gaussian approximation. Gauss-Hermite (Singer, 2008), Gaussian quadrature and cubature methods (Särkkä and Solin, 2012; Särkkä and Sarmavuori, 2013) correspond to different mechanisms for choosing the sigma points $\mathbf{x}_t^{(i)}$.

Beyond Gaussian approximations, mixtures of Gaussians have been used as well to approximate the predictive density (Alspach and Sorenson, 1972; Terejanu et al., 2008, 2011). Variational methods attempt to minimize a divergence between the chosen approximate density and the true predictive density. These can include Gaussian approximations (Archambeau et al., 2007; Ala-Luhtala et al., 2015) as well as more general exponential families and mixtures (Vrettas et al., 2015; Sutter et al., 2016). And for a broad class of diffusions, closed-form series expansions are available (Aït-Sahalia, 2008).

## 3 SMOOTHED DUAL FORMULATION FOR WASSERSTEIN GRADIENT FLOW

Our target is the predictive distribution of a diffusion: given an initial density $\rho_t$, we want to evolve it forward by a time increment $\Delta t$, to obtain the solution for the diffusion (2) at time $t+\Delta t$. We propose to approximate this by $m$ steps of the Wasserstein gradient flow (1), with stepsize $\tau = \Delta t/m$. The problem is to compute approximately this gradient step.

### 3.1 Regularized Wasserstein Gradient Flow

We start by introducing a proximal operator for the gradient step, which uses a regularized Wasserstein distance. The regularizer enforces strict convexity of the distance with respect to each of the input measures, which will be critical for tractability of the inference problem in coming sections. For measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$, we define the squared, regularized $2-$Wasserstein distance as

$$\mathcal{W}_\gamma^2(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d^2(\mathbf{x},\mathbf{y}) d\pi(\mathbf{x},\mathbf{y}) + \gamma R(\pi). \tag{5}$$

| |
|---|
| $f(\mu) = \langle w, \mu\rangle + \beta^{-1}\langle \mu, \log\mu - 1\rangle$ |
| $f^*(z) = \beta^{-1}\int_{\mathcal{X}} \exp\left(\beta(z(\mathbf{x}) - w(\mathbf{x}))\right)$ |
| $\left(\nabla f^*(z)\right)(\mathbf{x}) = \exp\left(\beta(z(\mathbf{x}) - w(\mathbf{x}))\right)$ |
| $\left(\nabla^2 f^*(z)\right)(\mathbf{x}) = \beta\exp\left(\beta(z(\mathbf{x}) - w(\mathbf{x}))\right)$ |

Table 1: Free energy expressions for advection-diffusion

with $d : \mathcal{X}\times\mathcal{X} \to [0, +\infty)$ the distance in $\mathcal{X}$, $\Pi(\mu,\nu)$ the set of joint measures on $\mathcal{X}\times\mathcal{X}$ having marginals $\mu$ and $\nu$, and $R : \mathcal{P}(\mathcal{X}\times\mathcal{X}) \to \mathbb{R}$ a regularizer. We assume $R$ is Legendre-type (extending (Bauschke et al., 1997, Def. 2.8)), implying it is closed, strictly convex, smooth, and proper. We also assume $R$ is separable, in the sense that

$$R(\pi) = \int_{\mathcal{X}\times\mathcal{X}} \bar{R}(d\pi(\mathbf{x},\mathbf{y})), \tag{6}$$

for $\bar{R} : \mathbb{R} \to \mathbb{R}$ the component function and $d\pi$ the Radon-Nikodym derivative with respect to the Lebesgue measure on $\mathcal{X}\times\mathcal{X}$. In the case of an entropy regularizer, for example, this is $\bar{R} : u \mapsto u(\log u - 1)$. For an $L^2$ regularizer, this is $\bar{R} : u \mapsto u^2$.

Given a free energy functional $f$ (Section 2.1), we define the primal objective $P_\nu^{\gamma,\tau} : \mathcal{P}(\mathcal{X}) \to [0, +\infty)$,

$$P_\nu^{\gamma,\tau}(\mu) \triangleq \mathcal{W}_\gamma^2(\mu,\nu) + 2\tau f(\mu), \tag{7}$$

for $\gamma \geq 0$, and $\tau > 0$. The primal formulation for the regularized Wasserstein gradient flow is

$$\operatorname{prox}_{\tau f}^{\mathcal{W}_\gamma}\nu = \operatorname*{argmin}_{\mu\in\mathcal{P}(\mathcal{X})} P_\nu^{\gamma,\tau}(\mu). \tag{8}$$

For $\gamma > 0$, the map $\mu \mapsto \mathcal{W}_\gamma(\mu,\nu)$ is strictly convex and coercive such that, assuming a convex functional $f$ in (7), the proximal operator is uniquely defined.

Note that we give all formulas in terms of a general free energy $f$. Table 1 gives concrete expressions for the free energy and its conjugate, in the case of an advection-diffusion system.

### 3.2 Smoothed Dual Formulation

Computing the proximal operator (8) directly entails solving an infinite program over the set of possible joint measures $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$ having $\nu$ as the second marginal. As a step towards a tractable approximation, we will derive a dual formulation that is unconstrained.

Let $\mathcal{C}_b(\mathcal{X})$ be the set of continuous, bounded functions on $\mathcal{X}$. The dual objective $D_\nu^{\gamma,\tau} : \mathcal{C}_b(\mathcal{X})\times\mathcal{C}_b(\mathcal{X}) \to \mathbb{R}$ is

$$\begin{aligned} D_\nu^{\gamma,\tau}(g,h) \triangleq &-\tau f^*\left(-\frac{1}{\tau}g\right) + \int_{\mathcal{X}} h(\mathbf{x})d\nu(\mathbf{x}) \\ &- \gamma R^*\left(\max\left\{\frac{1}{\gamma}\left(g + h - d^2\right), \nabla R(\mathbf{0})\right\}\right), \end{aligned} \tag{9}$$

with $f^*$ and $R^*$ the convex conjugates [3]. We have the following.

**Proposition 1** (Strong duality)**.** *Let $\nu \in \mathcal{P}(\mathcal{X})$ and $f : \mathcal{P}(\mathcal{X}) \to [0, +\infty)$ a convex, lower semicontinuous and proper functional. Define $P_\nu^{\gamma,\tau}$ as in* (7) *and $D_\nu^{\gamma,\tau}$ as in* (9)*. Assume $\gamma > 0$. Then*

$$\inf_{\mu\in\mathcal{P}(\mathcal{X})} P_\nu^{\gamma,\tau}(\mu) = \sup_{g,h\in\mathcal{C}_b(\mathcal{X})} D_\nu^{\gamma,\tau}(g,h). \tag{10}$$

*Suppose $f$ is strictly convex and let $g_*, h_*$ maximize $D_\nu^{\gamma,\tau}$. Then*

$$\mu_* = \nabla f^*(-\frac{1}{\tau}g_*) \tag{11}$$

*minimizes $P_\nu^{\gamma,\tau}$.*

Importantly, we have replaced the linearly-constrained optimization in the primal (8) with an unconstrained problem (10).

## 4 INFERENCE VIA STOCHASTIC PROGRAMMING

### 4.1 Stochastic Programming Formulation

The unconstrained dual problem (9) is not directly computable in general. To construct an approximation, we start by noting that the dual has an interpretation as a stochastic program. Specifically, let $\mu_0, \nu_0 \in \mathcal{P}(\mathcal{X})$ be arbitrarily chosen probability measures, supported everywhere in $\mathcal{X}$. We can express the dual objective (9) as

$$D_\nu^{\gamma,\tau}(g,h) = \mathbf{E}_{X,Y}\, d_\nu^{\gamma,\tau}(X,Y,g,h) \tag{12}$$

for random variables $X, Y$ distributed as $\mu_0$ and $\nu_0$, respectively, where the integrand $d_\nu^{\gamma,\tau}$ is

$$\begin{aligned} d_\nu^{\gamma,\tau}(\mathbf{x},\mathbf{y},g,h) = &-\tau\frac{\bar{f}^*(-\frac{1}{\tau}g(\mathbf{x}))}{\mu_0(\mathbf{x})} + h(\mathbf{y})\frac{\nu(\mathbf{y})}{\nu_0(\mathbf{y})} \\ &- \frac{\gamma}{\mu_0(\mathbf{x})\nu_0(\mathbf{y})}\bar{R}^*\Big(\max\Big\{\frac{1}{\gamma}(g(\mathbf{x}) + h(\mathbf{y}) - d^2(\mathbf{x},\mathbf{y})), \\ &\qquad \nabla R(\mathbf{0})(\mathbf{x},\mathbf{y})\Big\}\Big). \end{aligned} \tag{13}$$

Here, the terms $\bar{f}^*$ and $\bar{R}^*$ arise when we express the conjugate functionals $f^*$ and $R^*$ in integral form,

$$f^*(z) = \int_{\mathcal{X}} \bar{f}^*(z(\mathbf{x})), \qquad R^*(\xi) = \int_{\mathcal{X}\times\mathcal{X}} \bar{R}^*(\xi(\mathbf{x},\mathbf{y})).$$

In the case of an advection-diffusion, for example, the former is

$$\bar{f}^*(z(\mathbf{x})) = \beta^{-1}\exp\left(\beta(z(\mathbf{x}) - w(\mathbf{x}))\right)$$

for $w : \mathcal{X} \to [0, +\infty)$ the advection potential.

[3] $f^*(z) = \sup_{\mu\in\mathcal{P}(\mathcal{X})}\int_{\mathcal{X}} z(\mathbf{x})d\mu(\mathbf{x}) - f(\mu)$, $R^*(\xi) = \sup_{\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{X})}\int_{\mathcal{X}\times\mathcal{X}}\xi(\mathbf{x},\mathbf{y})d\pi(\mathbf{x},\mathbf{y}) - R(\pi)$.

### 4.2 Monte Carlo Approximation

The stochastic programming formulation (12) suggests a Monte Carlo approximation. If we sample $N$ pairs $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \in \mathcal{X} \times \mathcal{X}$ independently according to $\mu_0 \otimes \nu_0$, we can approximate $D_\nu^{\gamma,\tau}$ by the empirical mean,

$$D_{\nu,N}^{\gamma,\tau}(g,h) = \frac{1}{N}\sum_{i=1}^{N} d_\nu^{\gamma,\tau}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}, g, h). \qquad (14)$$

This converges to $D_\nu^{\gamma,\tau}(g,h)$ in the limit of large $N$.

The measure $\mu_0 \otimes \nu_0$ functions similarly to the importance distribution in importance sampling. Here, we expect a low variance approximation requires $\mu_0 \otimes \nu_0$ to be similar to $\mu_* \otimes \nu$, with $\mu_*$ the exact primal solution for the gradient step. In practice, it suffices to choose a hypercube containing the effective support of $\mu_* \otimes \nu$, and sample uniformly. This effective support can be determined by a Gaussian approximation to the process, such as underlies the extended or unscented Kalman filter.

### 4.3 RKHS Approximation

There is one more step to obtain a tractable problem: we need to restrict the domain of the dual, to ensure a finite-dimensional solution. We choose a domain $\mathcal{G} \times \mathcal{G}$, with $\mathcal{G}$ a compact, convex subset of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ defined on $\mathcal{X}$. From a practical standpoint, this encompasses two settings: the first is the case in which we choose a finite set of basis functions $\{\phi_k\}_{k=1}^p \subset L^2(\mathcal{X})$ and let $\mathcal{G}$ be contained in their linear span; the second is the case in which we choose a reproducing kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ associated to an RKHS $\mathcal{H}$ and assume $\mathcal{G} \subset \mathcal{H}$. In the second case, the fact of a finite-dimensional representation arises from a representer theorem (Proposition 2). In either case we assume the coefficients are restricted to a compact, convex set.

**Proposition 2** (Representation for general RKHS)**.** *Let $\nu \in \mathcal{P}(\mathcal{X})$ and $\gamma, \tau, N > 0$. Let $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^N \subset \mathcal{X} \times \mathcal{X}$. Then there exist $g_*, h_* \in \mathcal{H}$ maximizing* (14) *such that*

$$(g_*, h_*) = \sum_{i=1}^{N} \left( \alpha_g^{(i)} \kappa(\mathbf{x}^{(i)}, \cdot), \alpha_h^{(i)} \kappa(\mathbf{y}^{(i)}, \cdot) \right),$$

*for some sequences of scalar coefficients $\{\alpha_g^{(i)}\}_{i=1}^N$ and $\{\alpha_h^{(i)}\}_{i=1}^N$, with $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ the reproducing kernel for $\mathcal{H}$.*

### 4.4 Optimization

The Monte Carlo stochastic program can be solved by a standard iterative methods for convex optimization.

**Algorithm 1** Stochastic program approximating Wasserstein gradient flow

**Given**: initial density $\rho_t$, constant $\gamma > 0$, timestep $\tau > 0$.
**Choose** sampling densities $\mu_0, \nu_0$ on $\mathcal{X}$.
**Sample** independently $N$ pairs $(\mathbf{x}_i, \mathbf{y}_i) \sim \mu_0 \otimes \nu_0$.
**Solve** $g_*, h_* = \operatorname{argmax}_{g,h \in \mathcal{G}} D_{\rho_t,N}^{\gamma,\tau}(g,h)$.
The evolved density is $\rho_{t+\tau} = \nabla f^* \left(-\frac{1}{\tau} g_*\right)$.

Algorithm 1 outlines the resulting inference method. Note that conditioning of the problem depends on the regularization parameter $\gamma$, which presents a tradeoff between accuracy of the Wasserstein approximation (smaller $\gamma$) and fast optimization (larger $\gamma$).

## 5 PROPERTIES

### 5.1 Consistency

The Monte Carlo stochastic program (14) yields a consistent approximation to the regularized Wasserstein gradient step (8), in the sense that, as we increase the number of samples, the solution converges to that of the original dual program (12). This holds under a set of assumptions including compactness of $\mathcal{X} \times \mathcal{X}$ and conditions on $\mu_0, \nu_0$ and $\mathcal{G}$ (Appendix C). The assumptions guarantee that the stochastic dual objective (14) is $L$-Lipschitz. Under the assumptions, we get uniform convergence of the Monte Carlo dual objective (14) to its expectation (12), and this suffices to guarantee consistency.

**Proposition 3** (Consistency of Monte Carlo approx.)**.** *Let $D_\nu^{\gamma,\tau}$ and $D_{\nu,N}^{\gamma,\tau}$ be defined as in* (12) *and* (14)*, respectively, with $\gamma, \tau, N > 0$, and suppose Assumptions* **A1**-**A6** *(Appendix C) hold. Let $(g_N, h_N)$ optimize $D_{\nu,N}$ and $(g_\infty, h_\infty)$ optimize $D_\nu^{\gamma,\tau}$. Then for any $\delta > 0$, with probability at least $1 - \delta$ over the sample of size $N$,*

$$D_\nu^{\gamma,\tau}(g_\infty, h_\infty) - D_\nu^{\gamma,\tau}(g_N, h_N) \leq \mathcal{O}\left( \sqrt{\frac{(HKL)^2 \log(1/\delta)}{N}} \right), \qquad (15)$$

*with $\|g\|_{\mathcal{H}} \leq H$ for all $g \in \mathcal{G}$ and $K = \max_{\mathbf{x} \in \mathcal{X}} \sqrt{\kappa(\mathbf{x}, \mathbf{x})}$.*

As we discuss in Section 6.1, consistency of the Monte Carlo approximation is just one piece of the total approximation of the true diffusion. While convergence of the exact, time-discretized Wasserstein gradient step (1) is classical (Jordan et al., 1998), the authors are not aware of similar characterizations for the regularization of the Wasserstein metric or the RKHS approximation.

### 5.2 Computational Complexity

Complexity of first order descent methods for the stochastic dual problem is dominated by evaluation of the functions $g$ and $h$ at each iteration, for each sample $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^N$. Each pointwise evaluation of $g$ at a point $\mathbf{x}$ (and analogously for $h$ at $\mathbf{y}$) requires evaluating the sum $\sum_{k=1}^p \phi_k(\mathbf{x})\alpha_k$, with $\alpha_k$ being the coefficients parameterizing the function [4]. Hence straightforward serial evaluation of $g$ and $h$ at each iteration is $\mathcal{O}(Np)$, with $p$ the dimension of $\mathcal{G}$. These sums, however, are trivially parallelizable. Moreover, for certain kernels (notably Gaussian kernels), the serial complexity can be reduced to $\mathcal{O}(N)$, by applying a fast multipole method such as the fast Gauss transform (Greengard and Strain, 1991).

## 6 EMPIRICAL PERFORMANCE

### 6.1 Discussion

We note that accuracy of the proposed method can vary significantly, depending on several factors, including the particular density being approximated. Even given an unlimited number of Monte Carlo samples, our method gives a biased approximation of the exact diffusion process. There are three sources of bias. First is the discrepancy between the exact Wasserstein gradient step and the exact diffusion process, which only vanishes when the timestep is taken to zero. The second is the regularization applied to the Wasserstein distance, which can move the solution away from the exact Wasserstein gradient step. And the third source is the space $\mathcal{G}$ within which we optimize the dual variables $g$ and $h$, which may not contain the true solution. All three present tradeoffs in accuracy vs. computational complexity of optimization, with smaller $\tau$ and $\gamma$ and more expressive $\mathcal{G}$ generally degrading the conditioning of the optimization problem. These represent design choices when applying the method.

### 6.2 Performance in High Dimensions: Ornstein-Uhlenbeck Process

We study the accuracy of our proposed inference method as the dimension of the domain increases. As we have sidestepped the need for discretization of the domain, our approximation is at least computable in arbitrary dimensions. The question is how the accuracy degrades with the dimension.

As a target, we use the only diffusion process of the form (4) known to have a computable closed form solution in high dimensions. This is the Ornstein-Uhlenbeck process, which is a diffusion with a quadratic potential $w(\mathbf{x}) = (\mathbf{x} - \mathbf{b})^\intercal \mathbf{A}(\mathbf{x} - \mathbf{b})$, parameterized by matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and offset $\mathbf{b} \in \mathbb{R}^d$. Given a deterministic initial condition, the exact solution at time $t$ is Gaussian with mean and covariance evolving in time towards their long-time stationary values. We fix $\beta = 1$ and generate random forcing matrices $\mathbf{A}$ and offsets $\mathbf{b}$.

As a baseline for comparison, we use the only other approach for high-dimensional inference that doesn't rely on a parametric assumption. This is a standard particle simulation method [5], coupled with Gaussian kernel density estimation to obtain the full inferred distribution.

Figure 2a shows the accuracy of the two methods as we increase the dimension of the underlying domain $\mathcal{X}$ [6], for a timestep of $\Delta t = 1$. The figure shows median and 95% interval over 20 replicates. We see that our method scales with the dimension roughly equivalently to the simulation method, achieving accuracy (in symmetric KL divergence) comparable to simulation with 1000 particles, through dimension 7.

### 6.3 Application: Nonlinear Filtering

We demonstrate filtering of a nonlinear diffusion, which is observed at discrete times via a noisy measurement process. This is a discrete-time stochastic process $Y_k$, taking values at times $t_k$, which is related to the underlying diffusion $X_t$ by

$$Y_k = X_{t_k} + \mathbf{v}_k$$

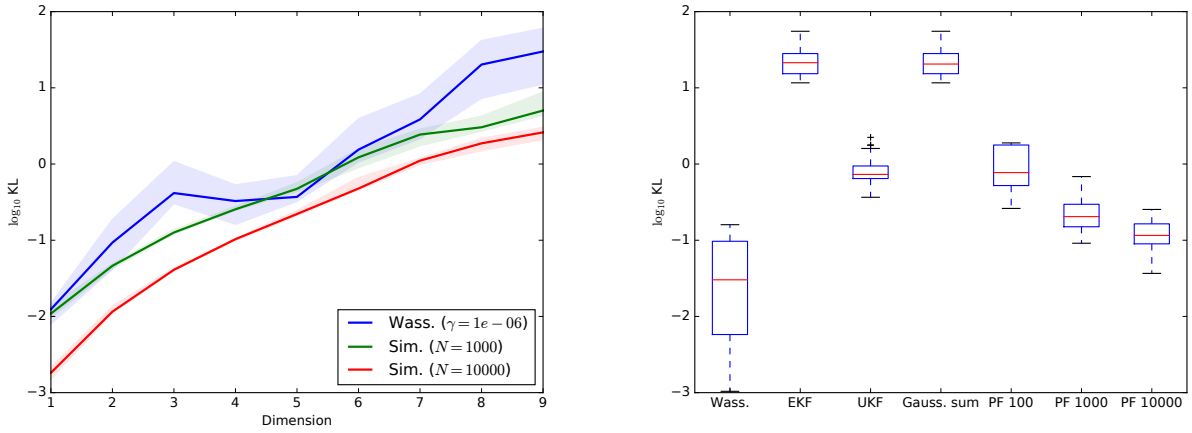
with $\mathbf{v}_k \sim \mathcal{N}(0, \sigma_Y^2)$ noise. Given a sequence of such measurements $\mathbf{y}_{0:K}$ up to time $t_K$, the **continuous-discrete filtering** problem is that of determining the corresponding distribution over the underlying state, $\Pr(X_t = \mathbf{x}_t | \mathbf{y}_{0:K})$, at some future time $t \geq t_K$. For future times $t > t_K$, this is the **marginal prior** or **predictive distribution** over states, defined by the dynamics of the diffusion process, satisfying the forward PDE (2) with initial density $\Pr(X_{t_K} = \mathbf{x}_{t_K} | \mathbf{y}_{0:K})$. At the measurement time $t = t_K$, this is the **marginal posterior**, conditional upon the measurements, and is defined by a recursive update equation

$$\Pr(X_{t_K} = \mathbf{x}_{t_K} | \mathbf{y}_{0:K}) = \frac{\Pr(Y_K = \mathbf{y}_K | X_{t_K} = \mathbf{x}_{t_K}) \Pr(X_{t_K} = \mathbf{x}_{t_K} | \mathbf{y}_{0:K-1})}{\Pr(Y_K = \mathbf{y}_K)}.$$

[4] In the case of a kernel parameterization, we have $p = N$ and $\phi_k(\mathbf{x}) = \kappa(\mathbf{x}, \mathbf{x}^{(k)})$.

[5] We use the Euler-Maruyama method for simulation, with timestep $10^{-3}$.

[6] We use an L2 regularizer and set $\gamma = 10^{-6}$. We use a third degree polynomial kernel for approximating $g$ and $h$ and approximate the objective using $2 \cdot 10^4$ sample points. We use a timestep of $\tau = 1/5$.




(a) Accuracy with increasing dimension of the domain.

(b) Posterior accuracy in nonlinear filtering.

Figure 2: Empirical performance. (Symmetric KL divergence to the true density.)

The term $\Pr(X_{t_K} = \mathbf{x}_{t_K} | \mathbf{y}_{0:K-1})$ is the predictive distribution given the measurements up to time $t_{K-1}$. We assume an initial distribution $\Pr(X_{t_0} = \mathbf{x}_{t_0})$ is given.

We assume the underlying state evolves according to a diffusion in the potential $w(x) = \frac{1}{\pi}\sin(2\pi x) + \frac{1}{4}x^2$, having unit diffusion coefficient $\beta = 1$. This is a highly nonlinear process and yields multimodal posteriors, which will present a challenge for most existing filtering methods. Measurements are made with noise $\sigma = 1$. We apply the Wasserstein gradient flow to approximate the predictive density of the diffusion, which at measurement times is multiplied pointwise with the likelihood $\Pr(\mathbf{y}_k | \mathbf{x}_{t_k})$ to obtain an unnormalized posterior density [7].

We use five methods as baselines for comparison. The first computes the exact predictive density by numerically integrating the Fokker-Planck equation (4) on a fine grid – this allows us to compare computed posteriors to the exact, true posterior. The second and third are the Extended and Unscented Kalman filters, which maintain Gaussian approximations to the posterior. The fourth method is a Gaussian sum filter (Alspach and Sorenson, 1972), which approximates the posterior by a mixture of Gaussians. And the fifth baseline is a bootstrap particle filter, which samples particles according to the transition density $\Pr(\mathbf{x}_{t_k} | \mathbf{x}_{t_{k-1}})$, by numerical forward simulation of the SDE (3) [8].

We simulate 20 observations at a time interval of $\Delta t = 1$, and compute the posterior density by each of the methods. Figure 2b shows quantitatively the fidelity of the estimated posterior to that computed by exact numerical integration, repeating the filtering experiment 100 times. Appendix F shows examples of the estimated posterior density of the diffusion. The Wasserstein gradient flow consistently outperforms the baselines, both qualitatively and quantitatively, achieving smaller symmetric KL divergence to the true posterior. Whereas the multimodality of the posterior presents a challenge for the baseline methods, the Wasserstein gradient flow captures it almost exactly.

## 7 DISCUSSION

We have described an approximate inference method for diffusion processes that circumvents the need for discretization of the domain of the diffusion by operating directly in a space of continuous functions. The method consists of a novel discrete-time approximation of a Wasserstein gradient flow in a space of probability measures. In addition to enabling inference in higher dimensions than are typically accessible with discretization-based methods, the proposed method might motivate new approaches for lower-dimensional settings where discretization is nevertheless difficult, such as when modeling complex physical systems (Náprstek and Král, 2016; Kolobov et al., 2019; Bar-Sinai et al., 2019).

As discussed in Section 6.1, performance of the method depends on the choice of timestep $\tau$, regularization parameter $\gamma$, and Hilbert space $\mathcal{H}$. The choice of $\mathcal{H}$, in particular, might offer a fruitful target for incorporating prior constraints on the inferred density, as an alternative to the standard parametric assumptions in the literature. Such constraints might derive from the system dynamics, for example, analogously to mesh-refinement strategies used in discretization-based inference.

[7] We use an L2 regularizer and set $\gamma = 10^{-6}$. We use a Gaussian kernel with bandwidth 0.1 and approximate the objective with $10^4$ samples. We use a timestep of $\tau = 1/4$.

[8] For foward simulation, we use an Euler's method with timestep $10^{-3}$.

## References


Yacine Aït-Sahalia. Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, 36(2):906–937, April 2008.

Juha Ala-Luhtala, Simo Särkkä, and Robert Piché. Gaussian filtering and variational approximations for Bayesian smoothing in continuous-discrete stochastic dynamic systems. *Signal Processing*, 2015.

Daniel Alspach and Harold Sorenson. Nonlinear bayesian estimation using gaussian sum approximations. *IEEE transactions on automatic control*, 17 (4):439–448, 1972.

Cédric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, and John Shawe-Taylor. Variational Inference for Diffusion Processes. *NIPS*, 2007.

Yohai Bar-Sinai, Stephan Hoyer, Jason Hickey, and Michael P Brenner. Learning data-driven discretizations for partial differential equations. *Proceedings of the National Academy of Sciences*, 116(31):15344–15349, 2019.

Heinz H Bauschke, Jonathan M Borwein, et al. Legendre functions and the method of random bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.

Jean-David Benamou, Guillaume Carlier, Quentin Mérigot, and Edouard Oudet. Discretization of functionals involving the monge–ampère operator. *Numerische Mathematik*, 134(3):611–636, 2016.

Pierre Bernhard and Alain Rapaport. On a theorem of danskin with an application to a theorem of von neumann-sion. *Nonlinear analysis*, 24(8):1163–1182, 1995.

Alexandros Beskos, Gareth O Roberts, et al. Exact simulation of diffusions. *The Annals of Applied Probability*, 15(4):2422–2444, 2005.

Alexandros Beskos, Omiros Papaspiliopoulos, and Gareth O Roberts. A factorisation of diffusion measure and finite sample path constructions. *Methodology and Computing in Applied Probability*, 10(1): 85–104, 2008.

Robert Grover Brown and Patrick Y. C. Hwang. *Introduction to Random Signals and Applied Kalman Filtering*. John Wiley and Sons, 1997.

Chris J Budd, MJP Cullen, and EJ Walsh. Monge–ampére based moving mesh methods for numerical weather prediction, with applications to the eady problem. *Journal of Computational Physics*, 236: 247–270, 2013.

Martin Burger, José Antonio Carrillo de la Plata, and Marie-Therese Wolfram. A mixed finite element method for nonlinear diffusion equations. 2009.

José A Carrillo and J Salvador Moll. Numerical simulation of diffusive and aggregation phenomena in nonlinear continuity equations by evolving diffeomorphisms. *SIAM Journal on Scientific Computing*, 31 (6):4305–4329, 2009.

José A Carrillo, Alina Chertock, and Yanghong Huang. A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Communications in Computational Physics*, 17(1):233–258, 2015.

JS Chang and G Cooper. A practical difference scheme for fokker-planck equations. *Journal of Computational Physics*, 6(1):1–16, 1970.

Dan Crisan and Terry Lyons. A particle approximation of the solution of the kushner–stratonovitch equation. *Probability Theory and Related Fields*, 115(4):549–578, 1999.

Garland B Durham and A Ronald Gallant. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics*, 20(3):297–338, 2002.

Paul Fearnhead, Omiros Papaspiliopoulos, and Gareth O Roberts. Particle filters for partially observed diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4): 755–777, 2008.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

Andrew Golightly and Darren J Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52(3):1674–1693, 2008.

Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET, 1993.

Leslie Greengard and John Strain. The fast gauss transform. *SIAM Journal on Scientific and Statistical Computing*, 12(1):79–94, 1991.

A Stan Hurn, Kenneth A Lindsay, and Vance L Martin. On the efficacy of simulated maximum likelihood for estimating the parameters of stochastic differential equations. *Journal of Time Series Analysis*, 24(1): 45–63, 2003.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker-Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998.

Simon Julier, Jeffrey Uhlmann, and Hugh F Durrant-Whyte. A new method for the nonlinear transforma-

tion of means and covariances in filters and estimators. *IEEE Transactions on automatic control*, 45 (3):477–482, 2000.

Simon J Julier, Jeffrey K Uhlmann, and Hugh F Durrant-Whyte. A new approach for filtering nonlinear systems. In *American Control Conference, Proceedings of the 1995*, volume 3, pages 1628–1632. IEEE, 1995.

Rudolph E Kalman and Richard S Bucy. New results in linear filtering and prediction theory. *Journal of basic engineering*, 83(1):95–108, 1961.

Peter E Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Science & Business Media, April 2013.

Vladimir Kolobov, Robert Arslanbekov, and Dmitry Levko. Boltzmann-fokker-planck kinetic solver with adaptive mesh in phase space. In *AIP Conference Proceedings*, volume 2132, page 060011. AIP Publishing, 2019.

Harold Kushner. Approximations to optimal nonlinear filters. *IEEE Transactions on Automatic Control*, 12 (5):546–556, 1967.

Antoine Liutkus, Umut Şimşekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. Sliced-wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. *International Conference on Machine Learning*, 2019.

David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

J Náprstek and R Král. Multi-dimensional fokker-planck equation analysis using the modified finite element method. In *Journal of Physics: Conference Series*, volume 744, page 012177. IOP Publishing, 2016.

Bernt Oksendal. *Stochastic Differential Equations*. An Introduction with Applications. Springer Science & Business Media, April 2013.

Lorenzo Pareschi and Mattia Zanella. Structure preserving schemes for nonlinear fokker-planck equations and applications. *arXiv preprint arXiv:1702.00088*, 2017.

Gabriel Peyré. Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

Gareth O Roberts and Osnat Stramer. On inference for partially observed nonlinear diffusion models using the metropolis–hastings algorithm. *Biometrika*, 88 (3):603–621, 2001.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1970.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55:58–63, 2015.

Simo Sarkka. On unscented kalman filtering for state estimation of continuous-time nonlinear systems. *IEEE Transactions on automatic control*, 52(9):1631–1641, 2007.

Simo Särkkä and Juha Sarmavuori. Gaussian filtering and smoothing for continuous-discrete dynamic systems. *Signal Processing*, 93(2):500–510, 2013.

Simo Särkkä and Arno Solin. On continuous-discrete cubature kalman filtering. *IFAC Proceedings Volumes*, 45(16):1221–1226, 2012.

S Shalev-Shwartz, O Shamir, N Srebro, and K Sridharan. Stochastic Convex Optimization. *COLT*, 2009.

Hermann Singer. Generalized gauss–hermite filtering. *AStA Advances in Statistical Analysis*, 92(2):179–195, 2008.

T Sutter, A Ganguly, and H Koeppl. A variational approach to path estimation and parameter inference of hidden diffusion processes. *J Mach Learn Res*, 2016.

Gabriel Terejanu, Puneet Singla, Tarunraj Singh, and Peter D Scott. A novel gaussian sum filter method for accurate solution to the nonlinear filtering problem. In *Information Fusion, 2008 11th International Conference on*, pages 1–8. IEEE, 2008.

Gabriel Terejanu, Puneet Singla, Tarunraj Singh, and Peter D Scott. Adaptive Gaussian Sum Filter for Nonlinear Bayesian Estimation. *IEEE Trans. Automat. Contr.*, 2011.

Michail D Vrettas, Manfred Opper, and Dan Cornford. Variational mean-field algorithm for efficient inference in large systems of stochastic differential equations. *Physical Review E*, 91(1):012148, 2015.

Michael Westdickenberg and Jon Wilkening. Variational particle schemes for the porous medium equation and for the system of isentropic euler equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(1):133–166, 2010.