annalen der physik

www.ann-phys.org

Variational Inference for Stochastic Differential Equations

Manfred Opper

The statistical inference of the state variable and the drift function of stochastic differential equations (SDE) from sparsely sampled observations are discussed herein. A variational approach is used to approximate the distribution over the unknown path of the SDE conditioned on the observations. This approach also provides approximations for the intractable likelihood of the drift. The method is combined with a nonparametric Bayesian approach which is based on a Gaussian process prior over drift functions.

1. Introduction

Stochastic differential equations play an important role in the modeling of dynamical systems which are driven by internal or external noise.[1] They find applications in various fields such as physics, chemistry, biology, or economy. Predictions with such models require the knowledge of the drift function and the diffusion term which are the deterministic and stochastic parts of the driving force of the dynamics. If precise information about these quantities is not available for a given problem, one may try to infer, that is, to *learn* them from observations of the stochastic path. If observations are only sparsely sampled over time or contain noise, the statistics of the unobserved path of the state variable *conditioned* on the data is a quantity which naturally appears in the calculations of drift estimators. We will discuss ideas that are based on variational approaches for the tractable approximations of this conditional statistics and for the estimation of the drift, when the diffusion is known. Variational methods have played a traditional role in statistical physics for the approximate computation of thermal averages. In recent years, they have also become powerful tools for approximate inference in the field of statistical machine learning.[2] Motivated by the successful approximation of path integrals in quantum statistics[3,4] we discuss the application of variational methods to infinite dimensional inference problems defined by the paths of stochastic differential equations. A combination of such methods with a Bayesian approach, assuming a prior probability distribution over drift functions, leads to a nonparametric estimation of the drift.

The article is structured as follows. In the second section, we introduce *Ito* stochastic differential equations and the inference problem for drift functions. The third section reviews a formulation of conditional path probabilities in terms of a variational free energy and discusses exact and approximate solutions. The fourth section addresses approximate inference of the drift function using a nonparametric Bayesian approach based on a Gaussian process prior distribution. We conclude with a discussion and extensions of the approach.

Prof. M. Opper
Articial Intelligence Group
Technische Universität Berlin
Marchstraße 23, Berlin 10587, Germany
E-mail: manfred.opper@tu-berlin.de

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/andp.201800233

DOI: 10.1002/andp.201800233

2. Stochastic Differential Equations

We consider stochastic differential equations (SDE) of the form

$$dX_t = f(X_t)dt + \sigma(X_t)dW_t \tag{1}$$

which describes the dynamics of a d-dimensional *state variable* $X_t \in R^d$ in time t. The drift function $f(\cdot) \in R^d$ is the deterministic part of the driving force and dW_t is a k-dimensional ($k \leq d$) vector of independent white noise sources given by the infinitesimal increments of Wiener processes. The strength of the noise is determined by the $d \times k$ dimensional diffusion matrix $\sigma(\cdot)$ which could, in general, be dependent on X_t. This formulation assumes the *Ito*– version of the SDE which can be understood as the limit $\Delta t \to 0$ of the time discretized equation

$$X_{t+\Delta t} - X_t = f(X_t)\Delta t + \sigma(X_t)\sqrt{\Delta t}\, z_t \tag{2}$$

The z_t are Gaussian random vectors with zero means and unit covariance matrix which are independent for different discrete time steps t. In contrast to the Stratonovich approach,[1] $\sigma(X_t)$ depends only on X_t and not $X_{t+\Delta t}$.

Let us assume that we have access to a set of observations $\mathbf{y} = (y_1, \ldots, y_n)$ which are noise corrupted versions of the process X_t sampled at discrete times $t_k \in [0, T]$ for $k = 1, \ldots, n$. As an example, we may think of $y_k = X_{t_k} + \nu_k$, with independent Gaussian noise variables ν_k.

In this contribution, we will discuss the problem of inferring the drift function $f(\cdot)$ using the observations $\mathbf{y}$. We will work with a probabilistic, Bayesian approach for its solution. This approach assumes both a likelihood of the drift (which is obtained by the observations and the dynamics) and a probability distribution over the drift which encodes the prior knowledge. The computation of the likelihood will use the statistics of the unobserved state variable X_t given the observations. For a broader overview of SDE inference methods, see ref. [5].

2.1. Bayes Drift Inference: Dense Observations

For the case, when a large amount of data is available and there is no parametric model for the drift of the SDE at hand, one might attempt a nonparametric estimation of the function $f(\cdot)$ in (1). This is relatively easy, when the data is free of noise and is observed at high frequency.[6] This means, we can assume that the (for simplicity constant) time τ between observations is small enough such that the discretized dynamics (2) applies to consecutive data points. Hence, the data is assumed to be generated as

$$y_{k+1} - y_k = f(y_k)\tau + \sigma(y_k)\sqrt{\tau}\, z_k\ ,\ k = 1, \ldots, n-1 \tag{3}$$

This defines a nonparametric *regression problem* with additive Gaussian noise for the function f. The solution of such problems has been extensively discussed in the field of machine learning, where a powerful Bayesian method based on Gaussian processes (GP) has been advocated.[7] The assumption is that prior knowledge about the smoothness and variability of the unknown function f can be encoded in a Gaussian prior measure $P_{GP}(f)$ which (assuming zero mean, for simplicity) is completely determined by its covariance *kernel*

$$K(x_1, x_2) = \mathrm{E}_f[f(x_1) f(x_2)] \tag{4}$$

The case of a scalar function f is easily extended to vectors $f(\cdot) \in R^d$. Hence, we can represent this prior measure by the quadratic functional

$$-\ln\left(P_{GP}(f)\right) = \frac{1}{2}\int\int f(x) K^{-1}(x, x') f(x') dx\, dx' + \text{const} \tag{5}$$

where K^{-1} is the inverse kernel operator. It is also straightforward to write down the likelihood $P(\mathbf{y}|f)$ for the drift f which corresponds to the Gaussian noise in (3). Its logarithm is also quadratic in f. This shows that the posterior distribution given by *Bayes rule*

$$P(f|\mathbf{y}) = \frac{P(\mathbf{y}|f)\, P_{GP}(f)}{P(\mathbf{y})} \tag{6}$$

is also a Gaussian process. An explicit computation of the posterior mean prediction for the drift function is then provided by linear algebra.[7] While a large amount of data makes the necessary inversions of large matrices nontrivial, a tractable solution is possible for the case, where K^{-1} is a differential operator.[6] The practical inversion of such large linear systems can be achieved with implicit schemes, see ref. [8]. As an alternative, the application of tractable sparse GP approximations for more general kernels K have been discussed in ref. [9]. However, it was also shown on simulated data[10] that with increasing time τ between observations, the prediction error caused by the simple discretization (3) rapidly increases.

This makes the method eventually useless for sparsely observed data. Nevertheless, a related quadratic functional reappears in an iterative approximate inference method, where the densely observed path of the diffusion process is replaced by a probability distribution over such paths given the sparse observations.

Manfred Opper is a professor for artificial intelligence at Technical University, Berlin since 2006. He has a Ph.D. in physics and a habilitation degree in theoretical physics from the University of Giessen, Germany. Before the appointment to Berlin, he has held faculty positions at Aston University in Birmingham, and at Southampton University in UK. His main research interest is in the development of methods for statistical inference. He has worked especially in the area of stochastic dynamical systems where he has developed approximations for inference in stochastic differential equations, Markov jump processes, and dynamical processes on large networks.

2.2. Bayes Drift Inference: Sparse Observations

In the general case of noisy (and not necessarily Gaussian) observations which are taken at larger time intervals, the likelihood of the drift can no longer be obtained analytically in closed form. A formal expression is given by the *path integral*

$$P(\mathbf{y}|f) = \int D(X_{0:T})\, P(\mathbf{y}|X_{0:T})\, P(X_{0:T}|f) \tag{7}$$

where $P(X_{0:T}|f)$ is the prior probability over paths $X_{0:T}$ (in the time interval $[0:T]$) induced by the dynamics (1) and

$$P(\mathbf{y}|X_{0:T}) = \prod_{k=1}^{n} p(y_k|X_{t_k}) \tag{8}$$

is a likelihood term for independent observations. To deal with the problem that paths $X_{0:T}$ cannot be integrated out exactly for arbitrary f in (7), one can treat a path as a latent random variable. A full Bayesian inference approach has to then deal with a posterior measure $P(X_{0:T}, f|\mathbf{y})$ over the augmented process defined by the unknown drift function f and the unobserved path $X_{0:T}$. This is possible, for example, within a Monte Carlo sampling approach. Papaspiliopoulos et al.[6] have developed a *Gibbs sampler* which iteratively updates samples from SDE paths and from drift functions. Using Bayes rule again, one can represent the conditional path measure as

$$P(X_{0:T}|\mathbf{y}, f) = \frac{P(X_{0:T}|f)\, P(\mathbf{y}|X_{0:T})}{P(\mathbf{y}|f)} \tag{9}$$

However, the sampling of (time discretized) paths from the probability (9) is nontrivial and requires Markov chain Monte Carlo

methods. For applications of such methods to inference in SDE, see refs. [11–14]. For an application to nonparametric drift estimation, see also ref. [15].

In Section 5, we will discuss a simpler and usually faster approximate approach, which is not based on sampling but uses variational techniques to deal with the path measure (9). This will also give us an approximation to the likelihood $P(\mathbf{y}|f)$. We will only discuss the computation of the simpler MAP (*maximum a-posteriori*) estimator, which represents the most likely drift function f given the observations. This is defined as

$$f_{MAP} = \arg\min_f \left(-\ln P(\mathbf{y}|f) - \ln P_{GP}(f)\right) \tag{10}$$

Before we discuss the application of the variational method to path measures, we will give a short summary of the main ideas in the following section.

3. Variational Formulation of Inference

Variational techniques play an important role in the field of probabilistic machine learning, where they serve as a basis for approximate inference, see ref. [2]. These methods are based on the identification of a conditional distribution, such as $P(X|y) = \frac{P(X,y)}{P(y)}$, as the solution of an optimization problem

$$P(\cdot|\mathbf{y}) = \arg\min_Q \mathcal{F}[Q] \tag{11}$$

The *variational free energy* $\mathcal{F}[Q]$ is defined as a functional of trial distributions Q in terms of the *Kullback–Leibler divergence* $\mathcal{D}(Q\|P)$

$$\mathcal{D}(Q\|P) = E_Q\left(\ln \frac{Q(X)}{P(X|y)}\right) \tag{12}$$

Its explicit form is

$$\mathcal{F}[Q] \doteq \mathcal{D}(Q\|P) - \ln P(y) = E_Q\left(\ln \frac{Q(X)}{P(X,y)}\right) \tag{13}$$

The result (11) follows from the fact that $\mathcal{D}(Q\|P) \geq 0$, and $\mathcal{D}(Q\|P) = 0$ only if $Q = P$. Variational approximations to P are obtained by restricting the minimization to tractable families of distributions Q. Note, that the subtraction of the intractable likelihood term $\ln P(y)$ in (13) does not change the minimizer. One might argue that the reversed Kullback–Leibler divergence $\mathcal{D}(P\|Q)$ might serve as a more sensible measure for the quality of approximations, because it involves the expectation over the correct rather than the approximate distribution. However, the assumed intractability of the probability P makes this choice technically difficult or even impossible, see ref. [16].

The variational method was originally developed in the field of statistical physics, see refs. [3,4] as the thermodynamic variational principle. This is obtained by setting $P(X|y) = \frac{e^{-H(X)}}{Z}$ (omitting the dependency on y) and $Q(x) = \frac{e^{-H_0(X)}}{Z_0}$ where $H(X)$ and $H_0(X)$ are exact and trial Hamiltonian functions of a classical system in statistical mechanics. We can also identify the partition function as $Z = P(y)$. In this case, the exact equilibrium thermodynamic free energy $-\ln Z$ is the minimum of the variational free energy

$$\min_Q \mathcal{F}[Q] = \min_Q \left\{-\ln Z_0 + E_Q(H - H_0)\right\} = -\ln Z \tag{14}$$

This fact is used in the inference framework to approximate the intractable likelihood. Since

$$\min_Q \mathcal{F}[Q] = -\ln P(\mathbf{y}|f) \tag{15}$$

we obtain lower bounds on the likelihood by restricting the optimization to tractable probability measures Q.

4. Variational Path Inference

We will next apply the variational approach to path measures of the type (9). Our approach will be somewhat different from Feynman's original treatment of path integrals by the variational approach.[3,4] We will first use a convenient representation of the posterior process defined by (9). Our description of the prior process is Markovian and given by the stochastic differential equation (2). Remarkably, under this assumption, it can be shown that also the posterior process, that is, the dynamics conditioned on the observations is also a Markov process. We will explain the main idea behind this result for the example of a single observation only. The general case can be proved in a similar way. The Markov property means that conditioned on the present state, the future of the process becomes independent of its past. Hence, for the case of the posterior process with a single observation y obtained at time T, and for times $0 \leq s \leq t \leq T$ we would like to show that $P(X_t|X_{0:s}, y, f) = P(X_t|X_s, y, f)$, where $X_{0:s}$ denotes the path from time 0 to time s. But this is easily obtained from the equalities

$$\begin{aligned} P(X_t|X_{0:s}, y, f) &= \frac{P(y|X_t, X_{0:s}, f)\,P(X_t|X_{0:s}, f)}{P(y|X_{0:s}, f)} \\ &= \frac{P(y|X_t)\,P(X_t|X_s, f)}{P(y|X_s, f)} = P(X_t|X_s, y, f) \end{aligned} \tag{16}$$

The first equality follows from elementary properties of conditional expectations and the second one uses the Markov property of the prior process to show that $P(X_t|X_{0:s}, f) = P(X_t|X_s, f)$ and

$$\begin{aligned} P(y|X_t, X_{0:s}, f) &= \int P(y|X_T)\,P(X_T|X_t, X_{0:s}, f)\,dX_T \\ &= \int P(y|X_T)\,P(X_T|X_t, f)\,dX_T = P(y|X_t, f) \end{aligned} \tag{17}$$

The denominator in (16) is simplified in a similar way. It can be further shown (see refs. [17,18]) that the conditioned process is a diffusion process described by an effective SDE of the form

$$dX_t = g(X_t, t)dt + \sigma(X_t)dW_t \tag{18}$$

Note that the diffusion terms for both processes are the same. In fact, it can be shown that a change of the diffusion term would lead to an infinite KL-divergence $\mathcal{D}(Q\|P)$ between the two processes in the continuous time limit. Hence, the only change caused by the conditioning on the observations is that the prior drift f is replaced by the posterior drift g. The latter is explicitly time dependent which takes the inhomogeneity caused by the observed data $\mathbf{y}$ into account. We will next discuss how g can be computed from the variational approach.

To compute the variational free energy (13) for path measures (9), we decompose it as

$$\mathcal{F}[Q] = \mathcal{D}_T(Q\|P_f) - \sum_{k=1}^{n} E_Q[\ln p(y_k|X_{t_k})] \tag{19}$$

$\mathcal{D}_T$ is the Kullback–Leibler divergence between path measures over a time window of length T and $P_f \equiv P(X_{0:T}|f)$. Specializing to trial distributions Q which correspond to diffusions of the type (18), we need to compute the KL-divergence between two path probabilities with different drift functions. This can be obtained in a heuristic way from the time discretized model (2). We can see that, conditioned on X_t, the variable $X_{t+\Delta}$ is Gaussian distributed with mean $X_t + f(X_t)\Delta t$ and covariance matrix given by $\Delta t\, D(X_t)$ where $D(x) \doteq \sigma(x)\sigma(x)^\top$ defines the diffusion matrix. By computing the product of the corresponding transition probabilities, we obtain the probability density of the discretized path of the prior process in the form

$$P(X_{0:T}|f) \propto P(X_{0:T}|f=0) \times \tag{20}$$
$$\times \exp\left[-\frac{1}{2}\sum_t \left\|f(X_t)\right\|_D^2 \Delta t + \left(f(X_t),\, X_{t+\Delta t} - X_t\right)_D\right]$$

where $P(X_{0:T}|f=0) \propto \exp[-\frac{1}{2\Delta t}\sum_t ||X_{t+\Delta t} - X_t||_D^2]$ represents free Brownian motion. To declutter notation, we have used the abbreviations $(u, v)_D \doteq u \cdot D^{-1} v$ and $\|u\|_D^2 \doteq u \cdot D^{-1} u$. Using (20) and the corresponding expression for the process with drift g, one obtains (taking the limit $\Delta t \to 0$ to convert a Riemann sum into an integral over time) the simple expression

$$\mathcal{D}_T(Q\|P_f) = \frac{1}{2}\int_0^T \int q(x,t)\left\|g(x,t) - f(x)\right\|_D^2 dx dt \tag{21}$$

where $q(x,t)$ is the marginal density of X_t (the process at time t) with respect to the drift g. A rigorous derivation of this result can be obtained by *Girsanov's* change of measure theorem. Hence, the free energy functional becomes

$$\mathcal{F}[Q] = \frac{1}{2}\int_0^T \int q(x,t)\left\{\left\|g(x,t) - f(x)\right\|_D^2 + U(x,t)\right\} dx dt \tag{22}$$

with

$$U(x,t) = -\sum_i \ln p(y_i|x)\, \delta(t - t_i) \tag{23}$$

The variational formulation for the conditional path probability can be interpreted as a specific stochastic control problem, where the goal is to find an extra drift, that is, a control force $u(x,t) \doteq g(x,t) - f(x)$ in an SDE such that the functional $\mathcal{F}[Q]$ is minimal. The first term in the free energy can be understood as an "energy" term which penalizes controls that are too large. The second term is a cost which strongly penalizes paths if they are too unlikely to explain the observations $\mathbf{y}$. For more details on such stochastic optimal control problems and their relations to inference, see refs. [18–20].

We will now discuss the exact minimization of (22). Unfortunately, we cannot perform a variation of (22) with respect to the drift $g(x,t)$ and the density $q(x,t)$ independently. Both are coupled through the time-dependent *Fokker–Planck equation*

$$\frac{\partial q(x,t)}{\partial t} = \mathcal{L}_g q(x,t) \tag{24}$$

where the operator $\mathcal{L}_g$ is given by

$$\mathcal{L}_g q(x,t) = \nabla \cdot \left[-g(x,t)q(x,t) + \frac{1}{2}\nabla \cdot (D(x)q(x,t))\right] \tag{25}$$

and ∇ denotes derivatives with respect to x. This constrained optimization problem can be solved in terms of a *Lagrange* function

$$L = \frac{1}{2}\int_0^T \int q(x,t)\left\{\left\|g(x,t) - f(x)\right\|_D^2 + U(x,t)\right\} dx dt$$
$$- \int_0^T \int \lambda(x,t)\left(\frac{\partial q(x,t)}{\partial t} - \mathcal{L}_g q(x,t)\right) dx dt$$

with a Lagrange-multiplier function $\lambda(x,t)$. After the logarithmic transformation $\lambda(x,t) \doteq -\ln \phi(x,t)$, the *Euler–Lagrange* equations can be written as

$$0 = \frac{\partial \phi(x,t)}{\partial t} + \mathcal{L}_f^\dagger \phi(x,t) - U(x,t)\phi(x,t) \tag{26}$$

$$g(x,t) = f(x) + D(x)\nabla \ln \phi(x,t) \tag{27}$$

where $\mathcal{L}_f^\dagger$ is the adjoint operator to (25). Note that in the context of optimal control, $-\ln \phi(x,t)$ is the *value function* (or optimal cost to go) obeying the Hamilton–Jacobi–Bellmann equation, see ref. [19], which is equivalent to Equation (26). To illustrate this result for a simple case, let us assume that we start the process at $y_0 \doteq X_0$ and have only a single, perfect (noise free) observation at time T, that is, $y_1 \doteq X_T$. The more general problem where the *distributions* of both X_0 and X_T are fixed, is known as the *Schrödinger bridge problem*. For more details and relations to *optimal mass transport* problems, see ref. [21]. This problem was originally introduced by E. Schrödinger[22] and deals with finding the *most likely* (in the sense of maximum entropy or minimal relative entropy) random evolution of the state variable X_t which connects prescribed initial and final densities.

To treat the problem with perfect observations, one can take the zero noise limit in the density $p(y|x)$ which defines the potential U in Equation (23). One can show that this results in the end condition $\phi(x,T) = \delta(x - y_1)$. For $t < T$, we then have $U = 0$ and the partial differential equation (PDE) (26) reduces to the *Kolmogorov backward* equation.[1] This has the solution

$$\phi(x,t) = \int p_{T-t}(x'|x)\phi(x',T)dx' \tag{28}$$

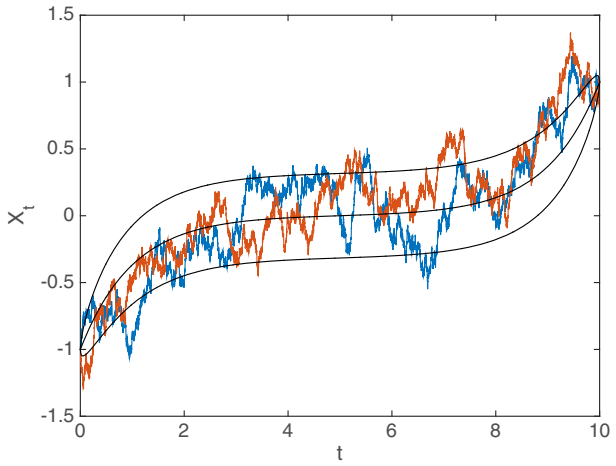

Figure 1. Two sample paths of an OU-process with $\alpha = 1$ and $\sigma^2 = 0.2$, conditioned on $X_0 = -1$ and $X_{10} = 1$. The simulations were based on the drift (30) and the discretized dynamics (2) with $\Delta t = 10^{-3}$. The solid lines gives the mean of the conditional path distribution together with $\pm$ the corresponding standard deviation.

where $p_{T-t}(x'|x)$ is the transition density of going from x to x' in the time $T - t$ for the process (1). Hence, we obtain $\phi(x, t) = p_{T-t}(y_1|x)$ and finally

$$g(x, t) = f(x) + D(x)\nabla \ln p_{T-t}(y_1|x) \tag{29}$$

We have shown that exact inference for the unobserved path reduces to the solution of a linear PDE. An analytically solvable example is given by the 1D *Ornstein–Uhlenbeck* process for which the drift $f(x) = -\alpha x$ is a linear function and $D = \sigma^2$ is constant. This linear problem has been studied in more general settings in the control literature by ref. [23,24]. Using a well known result for its transition density, we get

$$g(x, t) = -\alpha x + \alpha \frac{y_1 - xe^{-\alpha(T-t)}}{\sinh\left(\alpha(T-t)\right)} \tag{30}$$

Figure 1 shows two random sample paths from the SDE with drift (30). Initial and final conditions are $y_0 \doteq X_0 = -1$ and $y_1 = X_{10} = 1$. The process has parameters $\alpha = 1$ and $\sigma^2 = 0.2$ and $T = 10$. The divergence of the drift for $x = y_1$ at $t = T$ forces the paths to converge to the desired end condition. On the other hand, for times $0 \ll t \ll T$, the path fluctuations are close to the equilibrium of the OU-process.

To deal with nonlinear SDE models one can attempt an approximate minimization of the variational free energy (22) by assuming a simpler class of processes Q defined by their drift functions g. If the diffusion D is independent of X_t, a linear approximation $g(x, t) = A(t)x + b(t)$ with a time dependent matrix $A(t)$ and a vector $b(t)$ is a natural choice and leads to a time dependent Ornstein–Uhlenbeck process with marginal densities $q(x, t)$ which are Gaussian with mean $m(t)$ and covariance matrix $S(t)$. For these, it is no longer necessary to consider the full FP Equation (24), but only the time evolution of mean and covariance matrix which are given by [1]

$$\frac{dm}{dt} = Am + b\,, \qquad \frac{dS}{dt} = AS + SA^\top + D \tag{31}$$

Introducing appropriate Lagrange multipliers, the resulting equations are no longer PDEs but *ordinary differential equations* instead. These can be solved numerically by a forward–backward algorithm. The resulting variational free energy has been used to estimate parameters of drift and diffusion in ref. [25]. A great advantage of the variational method comes from the fact that one has the freedom to tune the complexity of the approximation to the complexity of the problem and the available computational resources. For example, by applying an additional mean field approximation (for diagonal diffusion matrices D) for the components of the vector X_t, it is possible to perform inference on the unobserved state X_t for large systems of coupled nonlinear SDE,[26] where a fraction of the dimensions of X_t are not observed. The mean field approximation leads to effective 1D problems, where $A(t)$ and $S(t)$ are diagonal matrices. For this case, we can also eliminate A and b explicitly using (31) and express the KL-divergence (21) in terms of m and S and its time derivatives. As an example, we apply this elimination to the 1D OU-process $f^{OU}(x) = -\alpha x$ and $D = \sigma^2$, and arrive at the the action functional

$$\mathcal{D}_T(Q\|P_{f^{OU}}) = \frac{1}{2\sigma^2}\int_0^T \left(\frac{1}{4S}\left(\frac{dS}{dt}\right)^2 + \frac{\sigma^4}{4S} + \alpha^2 S\right)dt + \frac{1}{2\sigma^2}\int_0^T \left(\left(\frac{dm}{dt}\right)^2 + \alpha^2 m^2\right)dt + \text{const} \tag{32}$$

for the dynamics of the mean $m(t)$ and variance $S(t)$. The constant contains only boundary terms. The difference to Equation (27) of ref. [26], is due to a typo in this equation: The term $s_i^2(t)$ in the denominator should be replaced by $s_i(t)$. From the minimization of the functional (32) with fixed boundary conditions $m(0) = y_0$, $m(T) = y_1$ together with $S(0) = S(T) = 0$, we can recover the mean and variance for the OU-bridge process with drift (30). The Euler–Lagrange equations can be solved analytically and provide the mean and standard deviations shown in Figure 1.

The variational approach is not the only way to construct an approximating Gaussian measure over paths of an SDE. The so-called *Expectation Propagation* (EP) method provides an iterative algorithm for approximating non-Gaussian terms in probability distributions by Gaussian ones using a local moment matching procedure. By taking the continuous time limit of the EP approximation to the free energy,[27] an efficient algorithm for state inference and parameter estimation is derived which is closely related to the variational approach.

An alternative computation of an approximation of the conditional drift (29) combines a variational formulation with a sampling approach. In the so-called *cross-entropy* method, one tries to minimize the reverse KL-divergence $\mathcal{D}_T(P_g\|Q_{g_\theta})$ between the optimal path measure P_g and a family of distributions with parametrized drifts g_θ. It turns out, that the gradient with respect to θ can be estimated iteratively by a Monte Carlo method using adaptive importance sampling.[28,29]

5. Nonparametric Estimation of the Drift

5.1. An Approximate EM Algorithm

In the following, we will show how a simplified variational technique can be used to perform approximate drift estimation using a GP prior. If we use the variational representation of the log-likelihood (15) for the MAP estimator (10), we obtain the double variational formulation

$$f_{MAP}(\cdot) = \arg\min_{f,Q} \left(\mathcal{F}[Q, f] - \ln P_{GP}(f) \right) \tag{33}$$

with the free energy

$$\mathcal{F}[Q, f] = E_Q \left(\ln \frac{Q(X_{0:T})}{P(X_{0:T}, \mathbf{y}|f)} \right) \tag{34}$$

An iterative solution of the optimization problem (33) is obtained by alternating between minimization with respect to path measures Q and to drift functions f. This is equivalent to the well known EM algorithm for inference on models with latent variables.[30] Assuming that observations of the path occur at regular intervals $k\tau$ and that they are free of noise for simplicity, we have

$$P(X_{0:T}, \mathbf{y}|f) \propto P(X_{0:T}|f) \prod_{k=1}^{n} \delta(y_k - X_{k\tau}) \tag{35}$$

We restrict approximate path measures Q to be of the form

$$Q(X_{0:T}) \propto P\left(X_{0:T}|f^{OU}\right) \prod_{k=1}^{n} \delta(y_k - X_{k\tau}) \tag{36}$$

where $P(X_{0:T}|f^{OU})$ corresponds to an Ornstein–Uhlenbeck (OU) process. Hence, the algorithm cycles between two steps. In the E-step, given a OU-path approximation Q_{old} which is based on a previous drift estimate f_{old}, we compute

$$\begin{aligned} \mathcal{L}[f] &\doteq E_{Q_{old}} \ln \left[\frac{P(X_{0:T}|f=0)}{P(X_{0:T}|f)} \right] \\ &= \mathcal{D}(Q_{old} \| P_f) - \mathcal{D}(Q_{old} \| P_0) \end{aligned}$$

which equals the free energy $\mathcal{F}[Q_{old}, f] = \mathcal{L}[f] + \text{const}$ up to a constant that is independent of f. To keep expressions simple, we have subtracted a reference measure with drift 0. In the M-Step of the algorithm, we recompute the most likely drift function as

$$f_{new} = \arg\min_f \left(\mathcal{L}[f] - \ln P_{GP}[f] \right) \tag{37}$$

A short computation, using (21) yields the quadratic functional

$$\mathcal{L}[f] = \frac{1}{2} \int \| f^2(x) \|_D A(x) dx - \int (f(x), b(x))_D dx + \text{const}$$

with the functions

$$A(x) = \int_0^T q(x,t) dt \qquad b(x) = \int_0^T g(x,t) q(x,t) dt \tag{38}$$

$A(x)$ and $b(x)$ are computed using an Ornstein Uhlenbeck bridge with an effective drift similar to (30). Rather than optimizing its parameters using the variational free energy, we have used a simpler approach and computed the bridge corresponding to an OU process that is based on a local linearization[9,10] of $f(x)$. This is defined by $f^{OU}(x) = f_{old}(\gamma_k) + \nabla f_{old}(\gamma_k)(x - \gamma_k)$ for times t in the interval between two consecutive observations. Since the resulting optimization problem is quadratic in f, the M-step can again be solved by linear methods as for the case of dense observations. For more details and applications, see ref. [9,10]. This method has a better behavior when the time τ between observations grows. Nevertheless, the Gaussian OU approximation still introduces a growing error in the estimate. In the following section, we will discuss a method which does not rely on a Gaussian approximation but can be applied to specific classes of SDE models only.

5.2. Pseudo- Bayesian Drift Estimation Using the Stationary Density

The method for drift estimation discussed in the last section used the information given by the temporal order of the observations. One would expect that under certain additional assumptions, simpler approaches could be possible, which only use an ergodic sample of the stationary density of the process. For example, if the drift is derived from a potential energy, that is, $f(x) = -\nabla V(x)$ and the diffusion $D = \sigma^2 I$ is constant and isotropic, we know that the stationary density equals $p(x) \propto e^{-\frac{2}{\sigma^2} V(x)}$. Hence, the drift could be obtained from the density, for example, by a kernel density estimator. As an alternative to such a method, we will generalize score function approaches[31,32] (see also ref. [33] for relations to maximum likelihood estimation), which directly estimate the *logarithm* of the density (up to a constant). Again, this can be derived by a variational approach.

Suppose we know that the total drift f of the process can be composed as a sum of an unknown drift g and a known reference drift r, that is, $f(x) = g(x) + r(x)$. We also assume that we know the stationary density $p(x)$ of the process. Of course, for dimensions $d > 1$, there is not enough information contained in the stationary Fokker–Planck equation $\mathcal{L}_f p(x) = 0$ for computing g. To make the least assumptions on g, we could aim for a kind of *maximum entropy* solution to the estimation problem by assuming that among all processes with stationary density p, we look for the one, which is closest in *relative entropy* to the process with the reference drift r. We will measure this entropic distance via the *relative entropy rate* defined as

$$d(P_f, P_r) \doteq \lim_{T\to\infty} \frac{1}{T} \mathcal{D}_T(P_f, P_r) = \frac{1}{2} \int p(x) \left\| g(x) \right\|_D^2 dx \tag{39}$$

where $P_f \equiv P(X_{0:T}|f)$ and $P_r \equiv P(X_{0:T}|r)$ are path measures for the two processes. Here, we have used Equation (21) assuming that for large times, the density of X_t becomes the stationary density $p(x)$. We will minimize (39) under the condition that $p(x)$ fulfils the stationary Fokker–Planck equation $\mathcal{L}_f p(x) = \mathcal{L}_r p(x) - \nabla(g(x)p(x)) = 0$. Similar to Section 4, we introduce a

Lagrange-multiplier function $\psi(x)$ for the constraint. A straightforward variation of the Langrange function

$$L = \frac{1}{2}\int p(x)\|g(x)\|_D^2 - \psi(x)\{\mathcal{L}_r\, p(x) - \nabla(g(x)\,p(x))\}dx \tag{40}$$

leads to

$$g(x) = D(x)\nabla\psi(x) \tag{41}$$

Inserting (41) back into the Lagrangean (40) shows that $\psi(x)$ is the minimizer of the functional

$$\varepsilon[\psi] \doteq \int \left\{\frac{1}{2}\big(\nabla\psi(x)\cdot D(x)\nabla\psi(x)\big) + \mathcal{L}_r^\dagger\psi(x)\right\} p(x)dx \tag{42}$$

Hence, if we know that the drift of the process is of the form $f(x) = r(x) + D(x)\nabla\psi(x)$, we can use (42) to estimate f. For $r = 0$ and $D = \sigma^2 I$, the optimal $\psi(x)$ equals $\ln p(x)$ up to a constant. Our generalization allows us to treat nontrivial models such as Langevin dynamics, when both positions and velocities are measured simultaneously.[34] The basic idea for applying (42) to data is to approximate the true stationary density using a large sample $x_1, \ldots, x_n$ drawn at random from the process and to minimize the empirical functional

$$\hat{\varepsilon}[\psi] \doteq \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{2}\big(\nabla\psi(x_i)\cdot D(x_i)\nabla\psi(x_i)\big) + \mathcal{L}_r^\dagger\psi(x_i)\right\} \tag{43}$$

as an estimator of (42). Since the functional is at most quadratic in ψ and its derivatives, its minimization leads again to linear equations. One can, for example, expand ψ in a finite set of basis functions and optimize the coefficients. A second possibility is a nonparametric (pseudo-) Bayesian approach where one views $\hat{\varepsilon}[\psi]$ as a proxy for a proper negative log-likelihood. A Gaussian process of the form (5) can serve as prior over functions ψ. In contrast to a vanilla GP regression problem, the resulting posterior functional contains both ψ and its derivatives. Nevertheless, one can derive an explicit result for the posterior mean of ψ and the resulting drift f in terms of matrix inversions. This is possible, because GPs and their derivatives are jointly Gaussian processes. This fact has been used before in dealing with the mathematically similar problem of solving noisy linear operator equations by ref. [35]. For explicit details and applications of our method, see ref. [34]. As an illustration, we show in **Figure 2** the results of inferring the drift of an SDE from observations generated with exact drift (blue curves) given by $f(x) = 4x(1 - x^2)$ and diffusion $\sigma^2 = 1$. This corresponds to a double well potential with local minima at $x = \pm 1$. We used a GP prior over the drift function f defined by the kernel $K(x, x') \propto e^{-(x-x')^2}$. For the example of $n = 200$ observations, we can see that the estimator (red curve) is biased toward smaller absolute values of $f(x)$ (favored by the prior) when compared to the exact drift for larger $|x|$ where data is scarce.

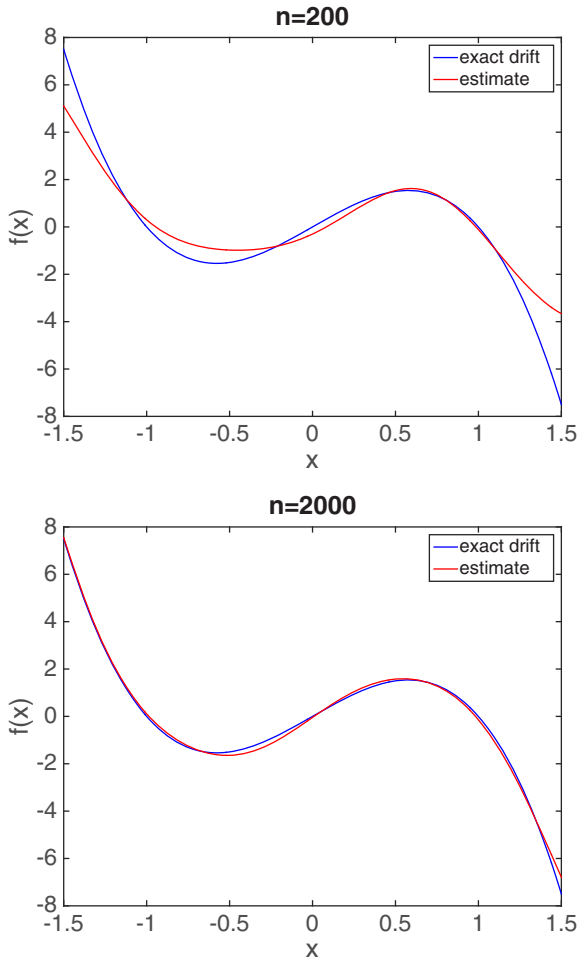

Figure 2. Estimates of the drift f for diffusion in a double well with $n = 200$ and $n = 2000$ observations. Inference is based on the Pseudo–Bayesian method with a GP prior over f. Simulations of the SDE were based on (2) with step–size $\Delta t = 10^{-3}$ and time $\tau = 100$ between observations.

6. Outlook and Discussion

In this contribution, we have discussed variational approaches for the inference of the the state variable and drift function of an SDE from observations. We have relied much on Gaussian approximations to path probability measures conditioned on observations. Such a class of approximations is of course restrictive. But there are some ways to improve on this method. One can, for example, use Gaussian path measures as a proposal for a Monte Carlo sampling approach.[36,37] This would lead, at least asymptotically, to exact samples for the conditioned paths. Sutter et al.[38] have recently generalized the variational method to allow for non-Gaussian families of marginal densities $q(x, t)$ and applied it to models with state dependent diffusion. Another possibility to go beyond a Gaussian variational ansatz would be to view this approximation as a first order variational perturbation theory. It should be possible to extend this to higher orders, see ref. [4], to increase (or assess) the quality of the variational approximation. It would also be interesting to explore alternative methods for treating the variational path problems which have been developed within the optimal control community for solving the *Schrödinger bridge* problem (see ref. [21]).

The approximation of the distribution of diffusion paths should not be confused with recent path approximations using

GPs which were introduced for Bayesian parameter estimation of *deterministic* dynamical systems. Here, one deals with ordinary differential equations (ODE) rather than SDE. In these approaches (see refs. [39–42] for a critical discussion of such methods), the computational complexity of solving nonlinear ODEs for the evaluation of likelihoods is avoided by gradient matching methods. The local relation (at observation times) between a state variable and its time derivative defined by the ODE is matched to GP approximations to the *deterministic* (but unobserved) path. In contrast, our method would have no uncertainty on the path left (for given drift and observations) when the deterministic limit of zero diffusion is taken in the SDE (1).

So far, we have only applied our methods to the computation of the most likely drift. For dense observations, the GP approach also provides the full posterior over drift functions. To deal with sparsely sampled observations, one could calculate at least the quadratic fluctuations around the MAP solution given by the EM-algorithm. A different possibility would be to replace the EM algorithm by a mean field variational approximation, which would factorize the posterior probability measure into a product of distributions over paths and drift functions. Those would then be optimized by cyclical updates similar to the EM algorithm. It is however unclear, if also the simpler "pseudo-likelihood" approach of the previous section can be turned into a proper Bayesian method which provides meaningful measures of uncertainty.

Another important problem is to obtain reliable estimators for the diffusion. In the case of perfect and densely sampled observations of the path, this seems possible. Our results in ref. [10] indicate that one can use GP regression to estimate even state dependent diffusions as the conditional mean square displacement, at least in data rich regions of state space. The case of sparse observations seems to be more complicated. Constant (state independent) diffusions can be estimated by an approximate maximum likelihood method which is based on the variational approximation. An extension to state dependent diffusions might be more involved. It is interesting to note that a solution of diffusion estimation by an EM algorithm is not possible. This is because two path measures with different diffusions would have infinite KL-divergence. Thus, the updated diffusion would equal the old one. For an interesting approach to solving a related problem for a corresponding Gibbs sampler, see ref. [12].

Acknowledgements

The work was supported by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1294 "Data Assimilation", Project (A06) "Approximative Bayesian inference and model selection for stochastic differential equations (SDEs)".

Conflict of Interest

The authors declare no conflict of interest.

Keywords

nonparametric Bayesian methods, statistical inference, stochastic differential equations

Received: July 6, 2018
Revised: November 3, 2018
Published online: January 25, 2019

[1] C. W. Gardiner, *Handbook of Stochastic Methods*, 2nd ed., Springer, Berlin **1996**.
[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Heidelberg **2006**.
[3] R. P. R. P. Feynman, A. R. Hibbs, *Quantum Mechanics and Path Integrals*, 1st ed., International Series in Pure and Applied Physics, MacGraw-Hill, New York **1965**, pp. xiv+365.
[4] H. Kleinert, *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets*, EBL-Schweitzer, World Scientific, **2009**. ISBN 9789814273572.
[5] S. M. Iacus, *Simulation and Inference for Stochastic Differential Equations: With R Examples*, 1st ed., Springer Series in Statistics, Springer, Heidelberg **2008**.
[6] O. Papaspiliopoulos, Y. Pokern, G. O. Roberts, A. M. Stuart, *Biometrika* **2012**, *99*, 511.
[7] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge **2006**.
[8] F. S. Kitaura, T. A. Enßlin, *Mon. Not. R. Astron. Soc.* **2008**, *389*, 497.
[9] A. Ruttor, P. Batz, M. Opper, *Advances in Neural Information Processing Systems*, Curran Associates, New York **2013**, pp. 2040–2048.
[10] P. Batz, A. Ruttor, M. Opper, *Phys. Rev. E* **2018**, *98*, 022109.
[11] M. Johannes, N. Polson, *Handbook of Financial Econometrics*, Vol. 2, Elsevier, Amsterdam **2009**.
[12] A. Golightly, D. J. Wilkinson, *Learning and Inference for Computational Systems Biology*, MIT Press, Cambridge **2010**, p. 253.
[13] H. Wu, F. Noe, *Phys. Rev. E* **2011**, *83*, 836705.
[14] G. O. Roberts, O. Stramer, *Biometrika* **2001**, *88*, 603.
[15] F. van der Meulen, M. Schauer, H. van Zanten, *Comput. Stat. Data Anal.* **2014**, *71*, 615.
[16] R. Leike, T. A. Enßlin, *Entropy* **2017**, *19*, 402.
[17] S. N. Majumdar, H. Orland, *J. Stat. Mech.: Theory Exp.* **2015**, *2015*, P06039.
[18] R. Chetrite, H. Touchette, *J. Stat. Mech.: Theory Exp.* **2015**, *2015*, P2005.
[19] H. J. Kappen, *Phys. Rev. Lett.* **2005**, *95*, 200201.
[20] H. J. Kappen, V. Gómez, M. Opper, *Mach. Learn.* **2012**, *87*, 159.
[21] Y. Chen, T. Georgiou, M. Pavon, *J. Optim. Theory and Applic.* **2016**, *169*, 671.
[22] E. Schrödinger, *Phys. Math. Klasse* **1931**, *169*, 144.
[23] Y. Chen, T. Georgiou, M. Pavon, *IEEE Trans. Aut. Control* **2016**, *61*, 1158.
[24] Y. Chen, T. Georgiou, M. Pavon, *IEEE Trans. Aut. Control* **2016**, *61*, 1170.
[25] C. Archambeau, D. Cornford, M. Opper, J. Shawe-Taylor, *J. Mach. Learn. Res.* **2007**, *1*, 1.
[26] M. Vrettas, M. Opper, D. Cornford, *Phys. Rev. E* **2015**, *91*, 012148.
[27] B. Cseke, D. Schnoerr, M. Opper, G. Sanguinetti, *J. Phys. A: Math. Theor.* **2016**, *49*, 494002.
[28] W. Zhang, H. Wang, C. Hartmann, M. Weber, C. Schütte, *Siam J. Sci. Comput.* **2014**, *36*, A2654.
[29] H. J. Kappen, H. C. Ruiz, *J. Stat. Phys.* **2016**, *162*, 1244.
[30] A. P. Dempster, N. M. Laird, D. B. Rubin, *J. Royal Stat. Soc. Series B (Methodological)* **1977**, *39*, 1.
[31] A. Hyvärinen, *J. Mach. Learn. Res.* **2005**, *6*, 695.
[32] B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, R. Kumar, *J. Mach. Learn. Res.* **2017**, *18*, 1.
[33] Y. Pokern, A. Stuart, E. V. Eijnden, *SIAM Multiscale Model. Simul.* **2009**, *8*, 69.

[34] P. Batz, A. Ruttor, M. Opper, *J. Stat. Mech.: Theory Exp.* **2016**, *2016*, 083404.
[35] T. Graepel, *Proc. of the Twentieth Int. Conf. on Machine Learning (ICML 2003)*, AAAI Press, Menlo Park **2003**.
[36] Y. Shen, C. Archambeau, D. Cornford, M. Opper, J. Shawe-Taylor, R. Barillec, *J. Signal Process. Syst.* **2010**, *61*, 51.
[37] F. J. Pinski, G. Simpson, A. M. Stuart, H. Weber, *SIAM J. Sci. Comput.* **2015**, *37*, A2733.
[38] T. Sutter, A. Ganguly, H. Koeppl, *J. Mach. Learn. Res.* **2016**, *17*, 190:1.
[39] B. Calderhead, M. Girolami, N. D. Lawrence, in *Advances in Neural Information Processing Systems 21* (Eds: D. Koller, D. Schuurmans, Y. Bengio, L. Bottou), Curran Associates, Inc., New York **2009**, pp. 217–224.
[40] D. Barber, Y. Wang, in *Proc. of the 31st Int. Conf. on Machine Learning* (Eds: Eric P. Xing, T. Jebara), Vol. 32, PMLR, Bejing, China **2014**, pp. 1485–1493.
[41] N. S. Gorbach, S. Bauer, J. M. Buhmann, in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA*, Curran Associates, New York **2018**, pp. 4809–4818.
[42] B. Macdonald, C. Higham, D. Husmeier, *J. Mach. Learn. Res.: Workshop Conf. Proc.* **2015**, *37*, 1539.