Approximate Bayes learning of stochastic differential equations

Philipp Batz, Andreas Ruttor,* and Manfred Opper†
TU Berlin, Fakultät IV–MAR 4-2, Marchstrasse 23, 10587 Berlin, Germany

(Received 17 February 2017; revised manuscript received 21 June 2018; published 8 August 2018)

We introduce a nonparametric approach for estimating drift and diffusion functions in systems of stochastic differential equations from observations of the state vector. Gaussian processes are used as flexible models for these functions, and estimates are calculated directly from dense data sets using Gaussian process regression. We develop an approximate expectation maximization algorithm to deal with the unobserved, latent dynamics between sparse observations. The posterior over states is approximated by a piecewise linearized process of the Ornstein-Uhlenbeck type and the maximum *a posteriori* estimation of the drift is facilitated by a sparse Gaussian process approximation.

DOI: 10.1103/PhysRevE.98.022109

I. INTRODUCTION

Dynamical systems in the physical world evolve in continuous time and often the (noisy) dynamics is described naturally in terms of (stochastic) differential equations (SDE) [1]. In cases where the parameters (which determine the drift and the diffusion) of such a model cannot be computed from first principles, it is necessary to fit such parameters to a time series of observed data [2]. Since small changes in parameters could lead to large changes in global dynamical behavior when the model is nonlinear, a proper fit of parameters may be crucial to make good predictions or for allowing a system to be controlled [3] by external forces. The Bayesian approach provides an important method for parameter estimation when prior knowledge on typical values and uncertainties of parameters is available which can be encoded in a prior probability distribution. In many applications of Bayesian methods closed form analytical computations of parameter estimates are not possible. Hence, one often has to resort to Monte Carlo (MC) sampling from the posterior distribution of parameters (for a review see, e.g., Refs. [4–9]). While such sampling approaches are feasible for SDE models with a small number of parameters, their efficiency decreases when the number grows. This is of special relevance, when we deal with the nonparametric scenario discussed in this paper.

In the nonparametric case, the drift (and possibly the diffusion) as a function of the state of the system can no longer be expressed by a finite number of parameters. One possibility is to keep the number of parameters (e.g., the number of basis functions used to model the drift) as a random variable and use *reversible-jump Markov chain Monte Carlo* methods to sample from the posterior distribution [10]. In this paper we work with a different Bayesian approach, by specifying prior distributions over functions. This allows for a considerable freedom of modeling assumptions and is, at least in a statistical sense, feasible when one has enough data observations to allow for a good estimation of an entire function. But from a computational point of view, estimation becomes more complicated [11].

The simplest class of nonparametric priors are Gaussian processes (GP), which are completely determined by a mean function and a covariance kernel. The use of Gaussian random fields in modeling unobserved functions is well known in physical sciences. See, e.g., Ref. [12], where a permeability field is modelled by a Gaussian random function. In this and similar applications, the measurements are noisy versions of nonlinear functions of the unobserved fields, yielding analytical expressions for likelihoods which in turn allow for straightforward MC sampling.

The SDE case, however, turns out to be more complicated. As shown in Refs. [13,14], dense observations (in time) of the state variables lead to an exact analytically computable quadratic log-likelihood functional for the drift. This, together with a Gaussian process prior, yields an exact Gaussian posterior distribution. Unfortunately, this simplicity is lost, when observations are not dense, but separated by larger time intervals. In this sparse case, the likelihood is a functional integral without an explicit exact analytical solution. Hence, one has to resort to Monte Carlo Gibbs samplers [13], which are generalizations of the ones used for the parametric case [6]. These alternate between sampling complete trajectories of the SDE conditioned on observations and drift, and sampling a GP for the drift given a complete trajectory. One computational problem is the sampling of SDE trajectories conditioned on the observations. This requires good proposals for trajectories which are used in Metropolis-Hastings steps within the sampler. A second problem stems from the matrix inversions required by the GP predictions. For a densely sampled trajectory these matrices become very large which leads to a strong increase in computational complexity. Papaspiliopoulos *et al.* [13] have shown for the case of univariate SDEs that the latter numerical problem can be circumvented if one chooses a GP prior where the covariance operator is the inverse of a differential operator. In this case efficient predictions are possible in terms of solutions of ordinary differential equations.

*andreas.ruttor@tu-berlin.de
†manfred.opper@tu-berlin.de

In this paper, we develop an alternative, approximate method for Bayesian inference for SDEs using GP priors. The method is faster than the MC sampling approaches and can be applied to GPs with arbitrary covariance kernels and also multivariate SDEs. In case of dense observations the framework of GP regression is used to estimate both drift and diffusion in a nonparametric way. GP inference becomes feasible by using an additional variational GP approximation frequently used in the field of machine learning. With this method only small matrices have to be inverted. For sparse observations, we use an approximate expectation maximization (EM) algorithm [15] for estimating the most likely drift function. This extends our approach introduced in the conference publication [16]. The EM algorithm cycles between the computation of expectations over SDE paths which are approximated by those of a locally fitted linear model and the computation of the maximum posterior GP prediction of the drift.

The paper is organized as follows. Stochastic differential equations are introduced in Sec. II and Gaussian processes in Sec. III. Then Sec. IV explains GP based inference for completely observed paths and shows results on dense data sets. As large data sets slow down standard GP inference considerably, Sec. V reviews an efficient sparse GP method. In Sec. VI our approximate EM algorithm is derived and its performance is demonstrated on a variety of SDEs. Section VII presents a discussion and concludes with an outline of possible extensions to the method.

II. STOCHASTIC DIFFERENTIAL EQUATIONS AND LIKELIHOODS FOR DENSE OBSERVATIONS

We consider diffusion processes given by a SDE written in Ito form as

$$dX_t = f(X_t)dt + D^{1/2}(X_t)dW_t, \tag{1}$$

where the vector function $f(x) = [f^1(x), \ldots, f^d(x)]$ defines the deterministic drift depending on the current state $X_t \in \mathcal{R}^d$. W_t denotes a Wiener process, which models white noise, and $D(x)$ is the $d \times d$ diffusion matrix.

Suppose we observe a path $X_{0:T}$ of the process over a time interval $[0,T]$. Our goal is to estimate the drift function $f(x)$ based on the information contained in $X_{0:T}$. A well-known statistical approach to the estimation of unknown model parameters is the method of maximum likelihood [2]. This would maximize the probability of the observed path with respect to f. To derive an expression for such a path probability, we use the Euler time discretization of the SDE [17] given by

$$X_{t+\Delta t} - X_t = f(X_t)\Delta t + D(X_t)^{1/2}\sqrt{\Delta t}\,\epsilon_t, \tag{2}$$

where $\epsilon_t \sim \mathcal{N}(0,I)$ is a sequence of i.i.d. Gaussian noise vectors and Δt is a time discretization. We will later set $\Delta t \to 0$, when we compute explicit results for estimators. Since the short-time transition probabilities of the process are Gaussian, the probability density for the discretized path can be written as the product

$$p(X_{0:T}|f) = p_0(X_{0:T})L(X_{0:T}|f), \tag{3}$$

where

$$p_0(X_{0:T}) \propto \exp\left[-\frac{1}{2\Delta t}\sum_t ||X_{t+\Delta t} - X_t||^2\right] \tag{4}$$

is the measure over paths without drift, and a term

$$L(X_{0:T}|f) = \exp\left[-\frac{1}{2}\sum_t ||f(X_t)||^2\Delta t + (f(X_t), X_{t+\Delta t} - X_t)\right], \tag{5}$$

which is the relevant term for estimating the function f from the observations of the path. To avoid cluttered notation, we have introduced the inner product

$$(u,v) \doteq u^\top D^{-1} v \tag{6}$$

and the corresponding squared norm

$$||u||^2 \doteq u^\top D^{-1} u. \tag{7}$$

The estimation of f using the method of maximum likelihood can be motivated by the following heuristics: Consider the case of a very large observation time T. In this limit we may write

$$\begin{aligned} -\frac{1}{T}\ln L(X_{0:T}|f) &= \frac{1}{2T}\sum_t ||f(X_t)||^2\Delta t - 2(f(X_t), X_{t+\Delta t} - X_t) \\ &\simeq \frac{1}{2T}\int_0^T \mathrm{E}\left[||f(X_t)||^2\right] - 2\mathrm{E}[(f(X_t), f_*(X_t))]dt \\ &= \frac{1}{2}\int ||f(x)||^2 p(x)dx - \int (f(x), f_*(x))p(x)dx, \end{aligned} \tag{8}$$

where we have taken the limit $\Delta t \to 0$. The expectations are defined with respect to the true (but unknown) process from which the data points are generated and $p(x)$ denotes its stationary density. The true drift is given by the conditional expectation

$$f_*(x) = \lim_{\Delta t \to 0} \frac{1}{\Delta t}\mathrm{E}[X_{t+\Delta t} - X_t | X_t = x]. \tag{9}$$

Obviously, a minimization of the last term in Eq. (8) would lead to the estimator $\hat{f}(x) = f_*(x)$, which is the true drift indicating that asymptotically, for a long sequence of data we get a consistent estimate. Unfortunately, for finite sample time T, an unconstrained maximization of the likelihood Eq. (5) does not lead to sensible results [13]. One has to use a regularization approach which restricts the complexity of the drift function. The simplest possibility is to work with a parametric model, e.g., representing f by a polynomial and estimating its coefficients. However, in many cases it may not be clear in advance how many parameters such a model should have.

Another possibility for regularization is a nonparametric Bayesian approach which uses prior probability distributions $P_0(f)$ over drift functions. With different choices of the prior different statistical ensembles of typical drift functions can be selected. We denote probabilities over the drift f by upper case symbols to avoid confusion with path probabilities. We will

also denote expectations over functions f by the symbol E_f. Our Bayes estimator will be based on the posterior distribution

$$p(f|X_{0:T}) \propto P_0(f)L(X_{0:T}|f), \tag{10}$$

where the neglected constant of proportionality only contains terms which do not depend on f. To construct such a prior distribution, we note that the exponent in Eq. (5) contains the drift f at most quadratically. Hence, a natural (conjugate) prior to the drift for this model is given by a Gaussian measure over functions, i.e., a Gaussian process (GP) [13]. Although a more general model is possible, we will restrict ourselves to the case where the GP priors over the components $f^j(x)$, $j = 1,\ldots,d$ of the drift factorize and we also assume that we have a diagonal diffusion matrix $D(x) = \mathrm{diag}[D^1(x),\ldots,D^d(x)]$. In this case, the GP posteriors of $f^j(x)$ also factorize in the components j, and we can estimate drift components independently.

We will show, that for dense observations, Bayesian inference with GPs becomes equivalent to GP regression, a topic which has been studied extensively in the machine learning community [18]. We will also show later, that for the dense setting, also the nonparametric estimation of diffusion functions $D^j(x)$ can be mapped onto a GP regression problem. Since the topic of GP regression may not be well known in the physics community, we will give a short introduction into this problem in the following section.

III. BAYESIAN REGRESSION WITH GAUSSIAN PROCESSES

In the following, we will give a heuristic derivation of the analytical results for solving regression problems with Gaussian processes which will be later applied to both drift and diffusion estimation. A more detailed formulation can be found in Ref. [18]. In the basic regression setting, we assume that we have a set of n input-output data points (x_i,y_i) for $i = 1,\ldots,n$, where the y_i are modeled as noisy values of an unknown function $f(x)$, i.e.,

$$y_i = f(x_i) + \nu_i, \tag{11}$$

where the noise values ν_i are taken to be independent Gaussian random variables with zero mean and possibly different but known variances σ_i^2. Within a Bayesian setting, we assume that the unknown function f is treated as a random object, being the realization of a Gaussian process. Using a GP prior over functions f, we try to filter out the noise from the observations and learn to predict the unknown function $f(x)$ at arbitrary input values x. GPs are completely defined through a mean function $m(x) = \mathrm{E}_f[f(x)]$ (which we will set to zero throughout the paper) and a kernel function defined as

$$K(x_1,x_2) = \mathrm{E}_f[f(x_1)f(x_2)], \tag{12}$$

which specifies the correlation of function values at two arbitrary arguments x_1 and x_2. By the choice of the kernel K we can encode prior assumptions about typical realizations of such random functions.

A popular covariance kernel is the radial basis function (RBF) kernel

$$K_{\mathrm{RBF}}(x_1,x_2) = \tau_{\mathrm{RBF}}^2 \exp\left(-\frac{||x_1 - x_2||^2}{2l_{\mathrm{RBF}}^2}\right), \tag{13}$$

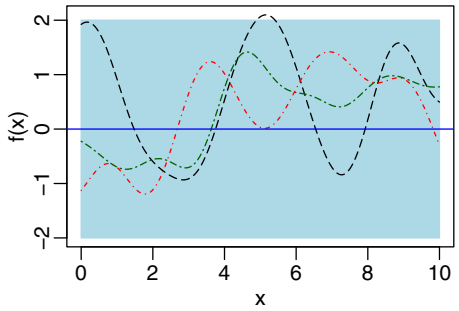

FIG. 1. Randomly drawn samples from a GP prior with RBF kernel and hyperparameters set to $\tau = 1$ and $l = 1$. The prior mean function is shown as blue solid line and blue shades denote the 95%-Bayes confidence bounds of the prior GP.

where the hyperparameters τ_{RBF}^2 and l_{RBF} denote the variance and the correlation length scale of the process. The RBF kernel assumes smooth, infinitely differentiable functions $f(\cdot)$. Samples from a GP using this kernel are shown in Fig. 1.

In some cases, the class of functional relationship in the data set is known beforehand, so that specialized kernel functions encoding this prior information can be applied. In our experiments, we use such kernels for the estimation of periodic and polynomial functions $f(\cdot)$. A (one-dimensional) periodic kernel is given by

$$K(x_1,x_2)_{\mathrm{Per}} = \tau_{\mathrm{Per}}^2 \exp\left[-\frac{2\sin\left(\frac{x_1-x_2}{2}\right)^2}{l_{\mathrm{Per}}^2}\right], \tag{14}$$

where the hyperparameters τ_{Per}^2 and l_{Per} denote the variance and the correlation length scale of the process. The polynomial kernel of degree p is given by

$$K_{\mathrm{Pol}}(x_1,x_2) = \left(1 + x_1^\top x_2\right)^p. \tag{15}$$

Since this may not immediately appear as a valid covariance function, we give a short proof in Appendix A that K_{Pol} is in fact a positive semidefinite kernel.

The probabilistic model for regression Eq. (11) corresponds to a likelihood

$$p(\mathbf{y}|f) \propto \exp\left\{-\sum_{i=1}^{n}\frac{1}{2\sigma_i^2}[f(x_i) - y_i]^2\right\}. \tag{16}$$

We will next give a derivation of the Bayes prediction $\hat{f}(x)$ for the function $f(x)$ given the observations $y_1,\ldots,y_n$. This prediction is given by the posterior mean of $f(x)$. Our derivation is a heuristic alternative to the standard approach given in Ref. [18], which is based on properties of conditional Gaussian distributions. We will instead use the fact that the mean of a Gaussian distribution equals the most likely value. Hence, we compute the most likely function f given the observations by minimizing the negative log-posterior functional

$$-\ln[P_0(f)p(\mathbf{y}|f)] \simeq \frac{1}{2}\iint f(x)K^{-1}(x,x')f(x')dx\,dx' + \sum_{j=1}^{n}\frac{1}{2\sigma_i^2}[f(x_j) - y_j]^2. \tag{17}$$

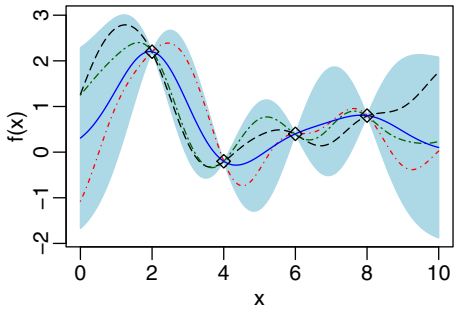

FIG. 2. Gaussian process regression using a GP with RBF kernel conditioned on four noise-free observations (GP posterior). Hyperparameters have been set to $\tau = 1$ and $l = 1$. The mean function of the posterior GP is shown as blue solid line and blue shades denote the 95%-Bayes confidence bounds. Dashed and dashed-dotted lines show randomly drawn samples from the GP posterior.

Here K^{-1} is the formal inverse of the kernel operator. Setting the functional derivative

$$\frac{\delta \ln[P_0(f)p(\mathbf{y}|f)]}{\delta f(x)} = 0 \tag{18}$$

and applying the kernel operator K to the resulting equation, we get

$$f(x) = \sum_{j=1}^{n} \frac{[y_j - f(x_j)]}{\sigma_j^2} K(x,x_j). \tag{19}$$

Evaluating this equation at each observation $x = x_i$ we obtain a system of linear equations for the $f(x_i)$, which is solved by

$$\frac{[y_i - f(x_i)]}{\sigma_i^2} = [(\mathbf{K} + \mathbf{\Sigma})^{-1}\mathbf{y}]_i. \tag{20}$$

Here $\mathbf{K} = [K(x_i,x_j)]_{i,j=1}^n$ denotes the kernel matrix and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$ is a diagonal matrix composed of the noise variances at the data points. Inserting this result back into Eq. (19) we get the following explicit expression (see also Ref. [18]) for the GP estimator of the function f:

$$\hat{f}(x) = [\mathbf{k}(x)]^\top (\mathbf{K} + \mathbf{\Sigma})^{-1}\mathbf{y}, \tag{21}$$

where $\mathbf{k}(x) = [K(x,x_i)]^\top$. A similar approach leads to the Bayesian uncertainty at x: the posterior variance

$$\hat{V}_f(x) = K(x,x) - [\mathbf{k}(x)]^\top (\mathbf{K} + \mathbf{\Sigma})^{-1}\mathbf{k}(x) \tag{22}$$

is used to calculate the 95% Bayesian confidence interval (credible interval)

$$\hat{C} = [\hat{f}(x) - 2\hat{V}_f(x)^{1/2}; \hat{f}(x) + 2\hat{V}_f(x)^{1/2}], \tag{23}$$

which contains the true function value $f(x)$ with probability

$$P[f(x) \in \hat{C}|\mathbf{y}] = 0.95 \tag{24}$$

if the assumptions made for Gaussian process regression are correct. An example of GP regression applied to four observations is shown in Fig. 2.

The GP predictions depend on a set of hyperparameters, which determine the shape of the underlying kernel function. The RBF and periodic kernel have variance and length scale parameters, where the latter denotes the smoothness of the process, the polynomial kernel has a variance and a degree parameter.

In the GP framework, the hyperparameters are usually found by optimizing the so-called *Bayes evidence*, which equals the probability of the path $p(\mathbf{y})$ (in its Euler discretization), with respect to the hyperparameter of interest. The evidence is defined as the nd-dimensional Gaussian integral

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p_0(\mathbf{f})d\mathbf{f}, \tag{25}$$

where $\mathbf{f}$ denotes the vector with components $f(X_{t_i})$ for $i = 1,\ldots,n$, and $p_0(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})$ is the prior Gaussian density induced by the GP prior over functions. Since both terms are Gaussian distributed, we can easily obtain the closed form expression

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0},\mathbf{K} + \mathbf{\Sigma}). \tag{26}$$

We note however, that finding hyperparameters by optimizing the evidence presupposes that we are dealing with the correct model and could lead to suboptimal results under a model mismatch. In this case, the hyperparameters can be found by using so called *cross validation* methods, which determine the optimal values by minimizing the expectation error based on a given target function. In contrast to the evidence ansatz, cross validation does not presuppose the correct model, but functions as a general black-box optimization tool. For our purposes, we will consider a twofold cross-validation scheme. This method randomly divides the observation into two subsets of equal size, and learns a GP estimator on each of the subsets. Then the goodness of fit is determined by computing the mean squared error of each estimator on the data of the remaining subset.

IV. DIRECT ESTIMATION FOR DENSE OBSERVATIONS

We first consider the case of dense observation, where we can apply Gaussian process regression directly to estimate drift and diffusion functions of stochastic differential equations.

A. Drift estimation

To apply GP regression to the drift estimation problem, we specialize to the jth drift component and identify $f(x) \equiv f^j(x)$. Setting $D(x) \equiv D^j(x)$, a comparison between the SDE Eq. (1) and the regression problem Eq. (11) shows that we can identify

$$y_i = \left(X_{t_i+\Delta t} - X_{t_i}\right)/\Delta t, \tag{27}$$

$$\sigma_i^2 = \frac{D\left(x_{t_i}\right)}{\Delta t}. \tag{28}$$

Note the scaling of the noise variance with $1/\Delta t$ reflecting the roughness of a diffusion path.

Hence, from Eq. (21) we can read off the GP prediction for the drift

$$\hat{f}^j(x) = [\mathbf{k}(x)^j]^\top \left(\mathbf{K}^j + \frac{1}{\Delta t}\mathbf{D}^j\right)^{-1}\mathbf{y}^j, \tag{29}$$

where $\mathbf{k}(x)^j = [K(x,x_i)^j]^\top$ and where $\mathbf{D}^j$ is a diagonal matrix composed of the diffusions $D^j(x_i)$ for $i = 1,\ldots,n$. The

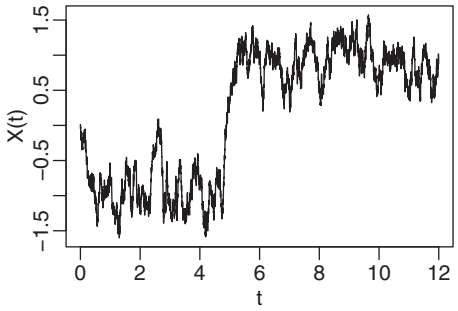

FIG. 3. Sample path with $n = 6000$ data points generated from a double-well model with time distance $\Delta_t = 0.002$.

Bayesian uncertainty at x is found from Eq. (22) as

$$\hat{V}_{f^j}(x) = \hat{K}^j - [\mathbf{k}(x)^j]^\top \left(\mathbf{K}^j + \frac{1}{\Delta t}\mathbf{D}^j\right)^{-1} \mathbf{k}(x)^j, \tag{30}$$

with $\hat{K}^j = K(x,x)^j$.

Bayesian methods such as GP regression are known to be typically fairly robust against choosing "wrong" hyperparameters when there is enough data. This corresponds to the limit $T \to \infty$ in the SDE case. In this limit we can estimate the drift well, even if the diffusion function is not known, but simply replaced by a constant.

As an example, we take observations from a double-well process shown in Fig. 3, which has a state-dependent diffusion displayed in Fig. 5, and compare the drift estimation which uses the true diffusion with the corresponding estimation assuming a model with constant noise. As Fig. 6 shows, both drift estimations practically agree around the data rich regions, despite the constant diffusion assumption only being a crude approximation to the true process noise.

Another point of view is to argue that one can use GP regression in the limit of a large amount of data as a tool for estimating the drift as the conditional expectation

$$f(x) = \mathrm{E}[X_{t+\Delta t} - X_t | X_t = x]/\Delta t \tag{31}$$

for $\Delta t \to 0$ without making precise assumptions on the noise. We will use this property of GP regression in the next section to derive a simple heuristic method for estimating also the diffusion function.

B. Diffusion estimation

For a smaller data length, one would expect that a good knowledge of the diffusion will also improve the estimation of the drift. To estimate the diffusion from data we distinguish between two cases in the following, namely models with constant and with state-dependent diffusion. If the diffusion matrix D is known to be constant, i.e., it does not depend on the state, we will use a Bayesian maximum likelihood approach and optimize the so-called *Bayes evidence*, which equals the probability of the path $p(X_{0:T})$ (in its Euler discretization), with respect to the diffusion constants $D = (D^1, \ldots, D^d)$. Again the probability factorizes in the components $j = 1, \ldots, d$. For component j of the process, the evidence is defined as the n-dimensional Gaussian integral,

$$p\left(X^j_{0:T}\right) = \int p\left(X^j_{0:T} | \mathbf{f}^j\right) p_0(\mathbf{f}^j) d\mathbf{f}^j, \tag{32}$$

where $\mathbf{f}^j$ denotes the vector with components $f^j(X_{t_i})$ for $i = 1, \ldots, n$ and $p_0(\mathbf{f}^j) = \mathcal{N}(\mathbf{f}^j | \mathbf{0}, \mathbf{K}^j)$ is the prior Gaussian density induced by the GP prior over functions. Introducing, as before, the notation

$$y^j_i = \frac{X^j_{t_i+\Delta t} - X^j_{t_i}}{\Delta t}, \tag{33}$$

we easily obtain the closed form expression

$$p\left(X^j_{0:T}\right) = \mathcal{N}(\mathbf{y}^j | \mathbf{0}, \mathbf{K}^j + \mathbf{\Sigma}^j) \tag{34}$$

from Eq. (32) with $\mathbf{\Sigma}^j = (D^j/\Delta t)\mathbf{I}$, and where $\mathbf{I}$ denotes the identity matrix. For the optimization, we use a quasi-Newton method.

For the case of state-dependent diffusions $D^j(x)$, we will again assume prior knowledge about functions encoded in a prior distribution. Since the diffusion must be nonnegative, one would have to use a nonlinear transformation of GPs, e.g., by exponentiation. A Bayesian approach, where drift and diffusion are jointly estimated using two GP priors is no longer analytically tractable. One might use a variational Bayesian technique [19], where regression with heteroscedastic noise is solved approximately by iteration. We will use a much simpler but more efficient heuristic technique assuming that with enough data, GP regression will be able to estimate conditional expectations of the type Eq. (11) fairly well even when the assumptions about the noise in the estimation problem are not entirely correct. Hence, we use the well known representation [1] for an arbitrary component of the exact diffusion

$$\begin{aligned} D^*(x) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{Var}(X_{t+\Delta t} - X_t | X_t = x) \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} (\mathrm{E}[(X_{t+\Delta t} - X_t)^2 | X_t = x] \\ &\quad - \mathrm{E}[X_{t+\Delta t} - X_t | X_t = x]^2) \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} (\mathrm{E}[(X_{t+\Delta t} - X_t)^2 | X_t = x] - \mathrm{E}[\Delta t f^*(x)]^2) \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{E}[(X_{t+\Delta t} - X_t)^2 | X_t = x]. \end{aligned} \tag{35}$$

In the third line, we use the fact that the second term on the right hand side equals the squared conditional drift Eq. (9). Then—by taking Δt out of the expectation $\mathrm{E}[\Delta t f^*(x)]$—we can easily see that the term vanishes in the limit $\Delta t \to 0$. Hence, the conditional variance does not depend on the drift. For its computation, we use again GP regression, but now on the data set $[(x_1, \tilde{y}_1), \ldots, (x_n, \tilde{y}_n)]$, where $\tilde{y}_i = (X_{t_i+\Delta t} - X_{t_i})^2/\Delta t$ are proportional to the squared observations of the drift estimation problem.

By taking the square of the time discretized SDE Eq. (2) we see that the dominant fluctuations of the data $\tilde{y}_i$ for $\Delta t \to 0$ are given by the non–Gaussian noise $D(X_t)(1 - \epsilon_t^2)$. In contrast to the fluctuations of the data for drift estimation, diffusion data are much smoother with a variance which remains finite as $\Delta t \to 0$. Hence, for dense data, it should be possible to filter out the noise even if we do not use the correct noise model

for regression. We expect that the following simple heuristics gives good results for densely sampled paths: We regard the GP framework as a regression tool for function estimation, which in our case happens to be the diffusion function. The regression curve is given by the GP mean Eq. (21) with $\mathbf{y}^j$ substituted by $\tilde{\mathbf{y}}^j$. Under the *GP as a regression toolbox* lense, we work with a constant Gaussian noise rate σ^2 in the likelihood which becomes a nuisance parameter without a direct interest to us. Still, we have to determine suitable variance values as well as possibly length scale parameters in the case of a RBF kernel, which might not be readily available. Finding hyperparameters by optimizing the evidence presupposes that the we are dealing with the correct model and could lead to suboptimal results in the case of model mismatch. Therefore, we resort to a twofold cross-validation scheme. This method randomly divides the observations into two subsets of equal size and learns a GP estimator on each of the subsets. Then the goodness of fit is determined by computing the mean squared error of each estimator on the data of the remaining subset.

C. Experiments

We consider drift and diffusion estimation in cases where the time grid Δt on which the data points are observed is small. This approach will be referred to as the *direct Gaussian Process (GP)* estimation with mean and variance given by Eqs. (29) and (30), respectively. We will treat drift and diffusion estimation in turn and start with the latter. The order is motivated by the fact that the diffusion estimation is independent of the drift. Hence, if both drift and diffusion are unknown, one should first learn the diffusion and then incorporate the estimation results into the drift learning procedure. But, as shown in Fig. 6, this typically leads to only a small correction in regions with sufficient data points.

Once we have diffusion values at the observations at our disposal, the estimation of the drift function becomes straightforward. All we have to do is to evaluate for each component j the diffusion at the observations $\mathbf{D}^j(x)$, which we then use as GP variance in the drift estimation. For the constant but unknown diffusion model, we insert the estimated value $\hat{D}^j$ into the diagonal of the matrix $\mathbf{D}^j$, in the state-dependent unknown diffusion model, we use the estimated value $\hat{D}^j(x_i)$ from the diffusion regression function described above. Then, running the GPs on the observations $\mathbf{y}^j$ leads to a drift estimation, which can once again be interpreted as Bayesian posterior.

In our experiments we found that the choice of the variance kernel parameter τ for both drift and diffusion estimation did not have a noticeable impact on the estimation results. Consequently, we fixed its value to $\tau = 1$. In the case of the length scale hyperparameter l the user usually has relevant prior expert knowledge about the specific problem at hand and is able to determine its value *a priori*. Similarly, if one knows that the underlying problem is of polynomial form, one should be able to specify its order p or at least an upper bound for p. We found that this approach usually works well in practice. We note, however, that in the case of dense data the kernel hyperparameters can also be automatically determined in a principled way (see Sec. III). Here we show the results for two experiments with unknown state-dependent diffusion.

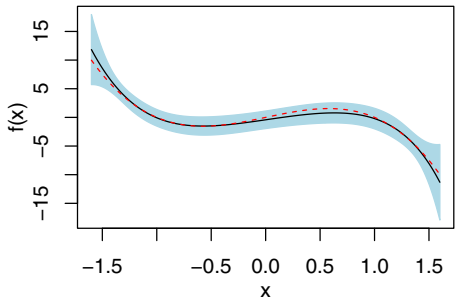

FIG. 4. Estimation for the double-well model based on the direct GP with the solid black line denoting the mean and the dashed red line the true drift function. The blue shades denote the 95%-Bayes confidence bounds.

First we look at synthetic data and then at a real-world data set used in climate research. The synthetic data sets analyzed are generated using the Euler method from the corresponding SDE with grid size $\Delta t = 0.002$.

1. *Double-well model with unknown state-dependent diffusion*

To evaluate the direct GP method, we generated a sample of size $n = 5000$ with step size $\Delta t = 0.002$ from the double-well process [20] with state-dependent diffusion,

$$dX = 4(X - X^3)dt + \sqrt{\max(4 - 1.25X^2, 0)}\,dW_t, \qquad (36)$$

which is shown in Fig. 3. The direct GP was run with a polynomial kernel function of order $p = 4$. The estimated functions for drift and diffusion are shown in Figs. 4 and 5, respectively. In both cases, we see a good fit between estimator and true function.

As an alternative using less prior knowledge the drift estimation with a RBF kernel is shown in Fig. 6. Comparing the results for polynomial and RBF kernel, one can see that the difference is most pronounced in the tail regions where few observations are located.

In a second experiment, we empirically checked the convergence rate of the diffusion estimator as a function of the time grid for this particular model. We generated the different data sets by first generating a sample path on a dense grid $\Delta t = 0.002$ and then selecting for $i = 1, \ldots, 20$, every ith sample point as observation point, yielding 20 data sets with

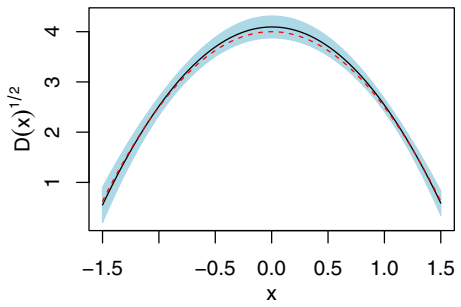

FIG. 5. Diffusion estimation of the double well based on the direct GP. The dashed red line denotes the square root of the diffusion $D(x)^{1/2}$ and the solid black line the estimator. The blue shades denote the 95%-Bayes confidence bounds.

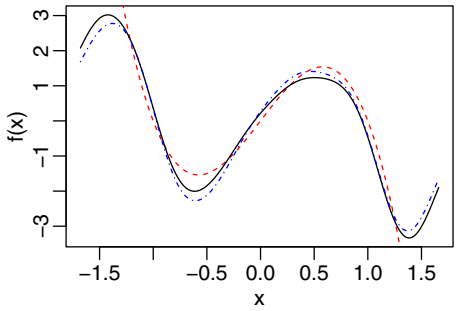

FIG. 6. The figure shows drift estimations based on $n = 5000$ dense observations from the double well with state-dependent diffusion. The dashed red line is the true drift, the black solid line the drift estimator using the true diffusion, and the dashed-dotted blue line the estimator which assumes a constant drift $\hat{\sigma} = 1.19$ determined by maximizing the evidence. In both cases, we used a RBF kernel with $l = 0.42$.

corresponding time steps from $\Delta t = 0.002$ to $\Delta t = 0.04$. We repeated this procedure for $M = 15$ dense sample paths and computed for each data set the GP estimator, using a polynomial kernel of order $p = 2$. To compare the estimation accuracy over the different time grids, we used the approximate mean-squared error (MSE)

$$\int p(z)(\hat{D}(z) - D(z))^2 dz \approx \frac{1}{S}\sum_{i=1}^{S}[\hat{D}(z_i) - D(z_i)]^2 \quad (37)$$

of the corresponding estimator. Here $\hat{D}(z)$ denotes the estimated diffusion function and $D(z)$ the true diffusion value, each evaluated on a set of $S = 100$ fixed points evenly spaced over the range of the samples.

As in the previous experiment, the M dense sample paths, each with size $n = 10\,000$, were generated from the process given in Eq. (36). Figure 7 shows the empirical result of the MSE as a function of the time grid. One can see that for very small time intervals, the model fit denoted by the dotted line in the graph roughly equals MSE $\approx n^{-1}$ as a function for the number of data points (remember that in our construction, the data set with twice the time grid contains half the number

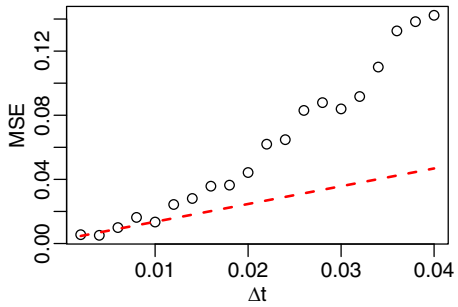

FIG. 7. The figure shows the mean MSE of the diffusion estimator as a function of the time grid. Each dot in the figure represents the mean of the $M = 15$ MSE values for the particular time grid. The dotted line is fitted by linear regression for the first five data points with time grid $\Delta t \leqslant 0.01$. The dotted red line denotes the model fit MSE $\approx n^{-1.1}$ based on the values of the smallest five time grids.

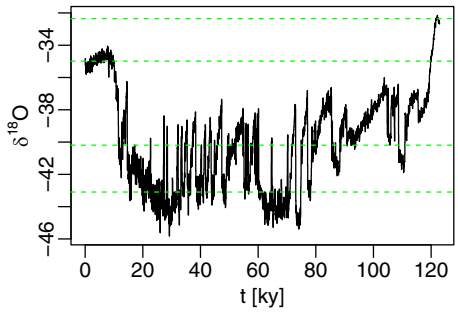

FIG. 8. Plot of the ice-core data (as solid black line) with metastable states marked by dashed green lines. These four minima of the potential function were identified by the direct GP algorithm with state-dependent diffusion.

observations). The error rate increases significantly for time grids bigger than $\Delta t = 0.01$, indicating that the estimation of the diffusion function becomes inexact for even moderately densely distributed observations. We return to this issue in our discussion of diffusion estimation for sparse observations.

2. Ice-core model

As an example of a real-world data set, we used the NGRIP ice-core data (provided by Niels-Bohr institute in Copenhagen [21]), which provides an undisturbed ice-core record containing climatic information stretching back into the last glacial. Specifically, this data set as shown in Fig. 8 contains 4918 observations of oxygen isotope concentration $\delta^{18}O$ over a time period from the present to roughly 1.23×10^5 years into the past. Since there are generally less isotopes in ice formed under cold conditions, the isotope concentration can be regarded as an indicator of the temperature at the time of the ice formation.

While this time series itself only shows the evolution of the average temperature in the past, finding a model explaining the dynamics behind the observed change is important to understanding climate transitions and to predict future changes of the temperature. Possible approaches for constructing such models include:

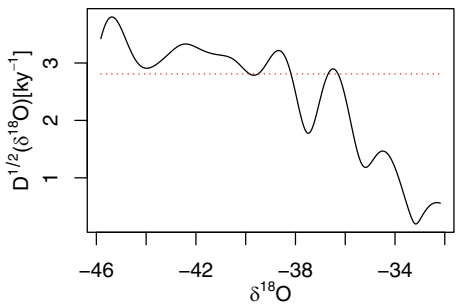

FIG. 9. Diffusion function estimators of the ice-core model for the state-dependent (solid black line) and the constant diffusion model (dotted red line). The constant value $D^{1/2} = 2.81$ was found by optimization of the marginal likelihood. For the GP in the state-dependent model we used a RBF kernel, whose length scale $l = 2.71$ and diffusion $D = 0.1$ was determined by twofold cross-validation.

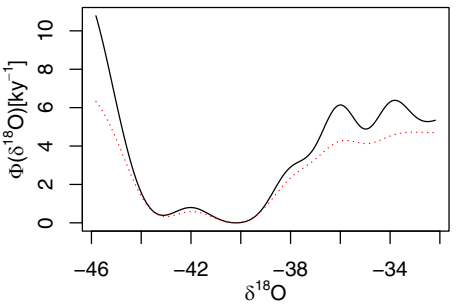

FIG. 10. The figure shows the estimated potentials of the ice-core data both from a model with state-dependent diffusion $D(x)$ (solid black line) and with constant diffusion D (dotted red). For both models we use a RBF kernel with length scale $l = 0.7$. The corresponding diffusion estimators are shown in Fig. 9.

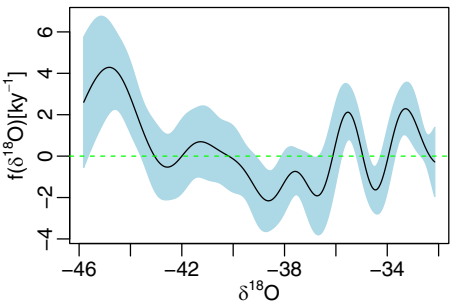

FIG. 11. Plot of the ice-core drift function corresponding to the potential function shown in Fig. 10 as solid black line together with the 95%-Bayes confidence bounds shaded in blue.

(1) A complex model is built starting from the basic laws of physics (*ab initio*). These models can become very accurate descriptions by including more and more details, but it is difficult to see the high-level properties of the dynamics directly. Instead one has to try out numerical simulations of different scenarios.

(2) Some assumed or observed properties of the dynamics lead to a parametric model which implements them. For example, a system switching between two meta-stable states could be described qualitatively by a polynomial drift of order 3. To reproduce the observations and to make quantitative predictions the model has to be fit to the data by parameter estimation. But a good fit alone does not guarantee that the model is correct.

(3) A nonparametric model—as used in this paper—is more flexible and can adapt to different data sets. It typically has only few built-in assumptions, e.g., that functions are smooth and differentiable. This reduces the risk of a model mismatch considerably.

Recent research [22,23] suggest to model the rapid paleoclimatic changes exhibited in the data set (Dansgaard-Oeschger events [24]) by a simple dynamical system with a drift function of order $p = 3$ as canonical model, which allows for bistability. This corresponds to a metastable state at higher temperatures close to marginal stability and a stable state at low values, which is consistent with other research on this data set linking a stable state of oxygen isotopes to a baseline temperature and a region at higher values corresponding to the occurrence of rapid temperature spikes. For this particular dataset the consecutive observations are spaced $\Delta t = 0.05\,\mathrm{ky}^{-1}$ apart. The underlying dynamics of the NGRIP data set is often modelled as a constant noise process in the literature [23].

Figure 9 shows that the estimated diffusion function changes significantly over the range of the observed isotope concentration, which seems to make the constant diffusion assumption in the model of Ref. [23] inadequate. Our data-driven and nonparametric approach not only reveals this multiplicative nature of the noise but also a richer structure of the learnt potential in comparison to the potential function of the constant diffusion model. Both functions Φ are shown in Fig. 10 and defined by their usual relation

$$\mathbf{f}(\mathbf{x}) = -\nabla_{\mathbf{x}}\Phi(\mathbf{x}) \tag{38}$$

to the corresponding drift function $\mathbf{f}(\mathbf{x})$.

Hence, choosing a state-dependent diffusion model is advisable even in cases where one is only interested in the qualitative form of the potential, since a wrong diffusion estimate can obscure it. This has happened here for $\delta^{18}O > -39$, where the two minima are barely visible in the result of the constant diffusion model.

In total, we find four local minima, but only two would be expected for a polynomial drift of order $p = 3$. Switches between the two lowest states at $\delta^{18}O \approx -43.1$ and $\delta^{18}O \approx -40.2$ occur quite frequently due to a low barrier and high diffusion. As indicated by the 95%-Bayes confidence bounds in Fig. 11, another possible but less likely explanation of the data points would be only one zero-crossing of the drift function in this region. A more obvious metastable state is found at $\delta^{18}O \approx -35.0$ because of an asymmetric barrier and lower noise levels. Around the last metastable state at $\delta^{18}O = -32.4$, the diffusion remains small, but the uncertainty about the drift function increases again, as there are only a few data points available there.

V. LARGE NUMBER OF OBSERVATIONS: THE NEED FOR A SPARSE GP

In practice, the number of observations can be large for a fine time discretization, and a fast computation of the matrix inverses in Eq. (29) could become infeasible. A possible way out of this problem—as suggested by [13]—could be a restriction to kernels for which the inverse kernel is a differential operator. We will now resort to a different approach which applies to arbitrary kernels and generalizes easily to multivariate SDEs.

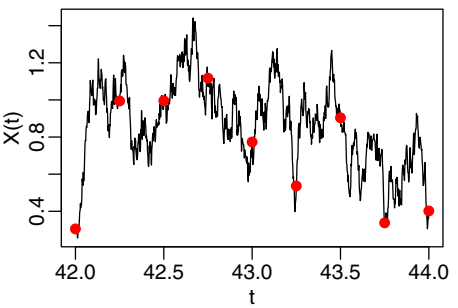

FIG. 12. Snippet of the double-well sample path in black with observations denoted as red dots.

Our method is based on a variational approximation to the GP posterior [25,26], where the likelihood term of the GP model Eq. (5) is replaced by another effective likelihood, which depends only on a smaller set of variables $\mathbf{f}_s$.

A. The general case

We assume a collection of random variables $f = \{f(x)\}_{x\in T}$, where the index variable $x \in T$ takes values in some possibly infinite index set T. We will assume a *prior measure* denoted by $P_0(f)$ and a *posterior measure* of the form

$$P(f) = \frac{1}{Z} P_0(f)\, e^{-U(f)}, \tag{39}$$

where $U(f)$ is a functional of f. The goal is to approximate P by another measure Q of the form

$$Q(f) = \frac{1}{Z_s} P_0(f)\, e^{-U_s(\mathbf{f}_s)}, \tag{40}$$

where the potential U_s depends only on a smaller *sparse* set $\mathbf{f}_s = \{f(x)\}_{x\in S}$ of dimension m. S is not necessarily a subset of T. While we keep the set S fixed, U_s will be optimized to minimize the variational free energy of the approximation

$$-\ln Z \leqslant -\ln Z_s + \mathrm{E}_s[U(f) - U_s(\mathbf{f}_s)]. \tag{41}$$

We write the joint probability of $\mathbf{f}$ and $\mathbf{f}_s$ as

$$Q(f,\mathbf{f}_s) = Q(f|\mathbf{f}_s)Q(\mathbf{f}_s) = P_0(f|\mathbf{f}_s)Q(\mathbf{f}_s), \tag{42}$$

where the last equality follows from the fact that fixing the sparse set $\mathbf{f}_s$, $U(\mathbf{f}_s)$ becomes non-random and the dependency on the random variables f is only via P_0 and we have

$$Q(\mathbf{f}_s) = \frac{P_0(\mathbf{f}_s)}{Z_s}\, e^{-U_s(\mathbf{f}_s)}. \tag{43}$$

Hence, we can integrate out all variables f except $\mathbf{f}_s$ using $P_0(f|\mathbf{f}_s)$ and rewrite the variational bound as the finite-dimensional integral

$$\begin{aligned} -\ln Z &\leqslant -\ln Z_s + \int Q(\mathbf{f}_s)\{\mathrm{E}_0[U(f|\mathbf{f}_s)] - U_s(\mathbf{f}_s)\}d\mathbf{f}_s \\ &= \int Q(\mathbf{f}_s) \ln\left(\frac{Q(\mathbf{f}_s)}{P_0(\mathbf{f}_s)\, e^{-\mathrm{E}_0[U(f|\mathbf{f}_s)]}}\right) d\mathbf{f}_s. \end{aligned} \tag{44}$$

$\mathrm{E}_0[U(f|\mathbf{f}_s)]$ is the conditional expectation with respect to P_0. Since this is of the form of a relative entropy, we conclude that the bound is minimized by the choice

$$Q(\mathbf{f}_s) \propto P_0(\mathbf{f}_s) e^{-\mathrm{E}_0[U(f|\mathbf{f}_s)]} \tag{45}$$

and thus $U_s(\mathbf{f}_s) = \mathrm{E}_0[U(f|\mathbf{f}_s)]$.

B. Gaussian random variables

We next specialize to a Gaussian measure P_0 with zero mean and covariance kernel K. If we assume (for notational simplicity) that the set $\{f\}$ is represented as a finite but high-dimensional vector $\mathbf{f}$ and

$$U(\mathbf{f}) = \tfrac{1}{2}\mathbf{f}^\top \mathbf{A}\mathbf{f} - \mathbf{b}^\top \mathbf{f} \tag{46}$$

is a quadratic form, we can then further simplify the conditional expectation Eq. (45) to

$$\mathrm{E}_0[U(\mathbf{f})|\mathbf{f}_s] = \tfrac{1}{2}(\mathrm{E}_0[\mathbf{f}|\mathbf{f}_s])^\top \mathbf{A}\mathrm{E}_0[\mathbf{f}|\mathbf{f}_s] - \mathbf{b}^\top \mathrm{E}_0[\mathbf{f}|\mathbf{f}^s] + C, \tag{47}$$

where

$$C = \tfrac{1}{2}\mathrm{tr}(\mathrm{Cov}_0[\mathbf{f}|\mathbf{f}_s]\mathbf{A}) \tag{48}$$

is a constant independent of $\mathbf{f}_s$. This follows from the fact that $\mathrm{E}_0[\mathbf{f}|\mathbf{f}_s]$ is the optimal mean square predictor of the vector $\mathbf{f}$ given $\mathbf{f}_s$ [27], the difference $\mathbf{f} - \mathrm{E}_0[\mathbf{f}|\mathbf{f}_s]$ is a random vector which is uncorrelated to the vector $\mathbf{f}_s$ and thus for jointly Gaussian random variables *independent* of $\mathbf{f}_s$. Hence the conditional covariance Cov_0 of $\mathbf{f}$ does not depend on $\mathbf{f}_s$. The explicit result for this predictor is given by

$$\mathrm{E}_0[\mathbf{f}|\mathbf{f}^s] = \mathbf{K}_{Ns}\mathbf{K}_s^{-1}\mathbf{f}_s, \tag{49}$$

where $\mathbf{K}_s$ is the kernel matrix for the sparse set and $\mathbf{K}_{Ns}$ is the $n \times m$ kernel matrix between the non-sparse and the sparse set. It is now easy to generalize to the infinite-dimensional case of the form

$$U(f) = \frac{1}{2}\int f^2(x)A(x)dx - \int f(x)b(x)dx, \tag{50}$$

for which we get

$$\mathrm{E}_0[f(x)|\mathbf{f}_s] = \mathbf{k}_s^\top(x)(\mathbf{K}_s)^{-1}\mathbf{f}_s \tag{51}$$

and thus

$$\begin{aligned} \mathrm{E}_0[U(\mathbf{f})|\mathbf{f}_s] = \frac{1}{2}\mathbf{f}_s^\top \mathbf{K}_s^{-1}\left\{\int \mathbf{k}_s(x)\, A(x)\, \mathbf{k}_s^\top(x)dx\right\}\mathbf{K}_s^{-1}\mathbf{f}_s \\ - \mathbf{f}_s^\top \mathbf{K}_s^{-1}\int \mathbf{k}_s(x)\, b(x)\, dx. \end{aligned} \tag{52}$$

C. Sparse GP drift and diffusion estimation

Now, setting

$$U(f) = -\ln[L(X_{0:T} \mid f)], \tag{53}$$

we can derive the drift estimator for the sparse representation analogously to Eq. (17). With definitions $\boldsymbol{\pi}^j = \mathbf{K}_{Ns}^j(\mathbf{K}_s^j)^{-1}$ and $\Omega^j = \Delta t(\boldsymbol{\pi}^j)^\top \mathbf{D}^{-1}\boldsymbol{\pi}^j$ we get for the jth component of the drift vector:

$$\hat{f}^j(x) = [\mathbf{k}(x)^j]^\top\left(\mathbf{I} + \Omega^j\mathbf{K}_s^j\right)^{-1}\Delta t(\boldsymbol{\pi}^j)^\top(\mathbf{D}^j)^{-1}\mathbf{y}^j, \tag{54}$$

where $\mathbf{k}(x)^j = [K(x,x_i)^j]^\top$.

The corresponding expression for the variance estimator is given by

$$\hat{V}_{f^j}(x) = K(x,x) - \mathbf{k}(x)^\top\left(\mathbf{I} + \Omega^j\mathbf{K}_s^j\right)^{-1}\Omega^j\mathbf{k}(x). \tag{55}$$

Notice that the inverted matrix inside the drift and variance estimators is no longer of the size of observations $n \times n$, but of the size of the sparse set $m \times m$.

While it is possible to also optimize the approximation with respect to the set of sparse points numerically [25], we use a simple heuristic, where we construct a histogram over the observations and select as our sparse set S the midpoints of all histogram hypercubes containing at least one observation. Here, the intuition is that a sparse point in a region of high empirical density is a good approximation to the data points in the respective hypercube. The number of histogram bins is determined by Sturges' formula [28], which is implicitly based on the range of the data. Note, that the cardinality m of the

sparse set is not set in advance, but automatically determined by the spatial structure of the data. This heuristic typically leads to $m \ll n$ and therefore to substantial computational gains compared to the full GP.

In practice, using the sparse GP for the drift and diffusion function estimation can be easily accomplished by first determining a sparse set S for the relevant data set and then substituting mean Eq. (29) and variance Eq. (30) with their sparse GP counterpart Eqs. (54) and (55), respectively.

One exception is the estimation of the constant diffusion $\mathbf{D}$, where we have to replace the marginal distribution Eq. (32) with a corresponding sparse approximation. Here, we follow [25] and optimize for each component j a lower bound to the evidence with respect to the diffusion constants:

$$F_V(X_{0:T}) = \log\left[\mathcal{N}\left(\mathbf{y}^j \middle| \mathbf{0}, \mathbf{Q}_N^j + \frac{1}{\Delta t}\mathbf{D}^j\right)\right] - \frac{\Delta t}{2}(\mathbf{D}^j)^{-1}\mathrm{tr}\left(\mathbf{K}^j - \mathbf{Q}_N^j\right), \tag{56}$$

where $\mathbf{Q}_N^j = \mathbf{K}_{Ns}^j(\mathbf{K}_s^j)^{-1}(\mathbf{K}_{Ns}^j)^\top$ and $\mathrm{tr}(\cdot)$ denotes the trace of the matrix.

D. Performance comparison

To get a feel for the performance differences between the standard GP and its sparse counterpart, we compare both versions in terms of accuracy and performance on the double-well model,

$$dX = 4(X - X^3)dt + D^{1/2}dW_t, \tag{57}$$

with constant and known variance $D = 1$. For the comparison, we analyzed the performance for data sets of different sizes, where we generated 10 data sets with $\Delta t = 0.002$ for each fixed number of observations and using the approximate MSE,

$$\frac{1}{S}\sum_{i=1}^{S}[\hat{f}(z_i) - f(z_i)]^2, \tag{58}$$

as performance measure. Here $\hat{f}(z)$ denotes the estimated drift and $f(z)$ the true drift value, evaluated on $S = 100$ evenly spaced points of the sample range. We then measured the run time and MSE of each data set based on the sparse GP and the standard GP estimation, each with a polynomial kernel of order $p = 4$. All MSE are computed for one fixed test set of size $n = 4000$, which we generated from the same model with $\Delta t = 0.5$.

Table I shows the mean values of the run time and MSE for each fixed observation number, respectively. One can see that the sparse GP algorithm leads to a significant reduction in computing time while exhibiting practically no loss in estimation accuracy. As expected, the efficiency gain grows with larger data sets and even allows us to to analyze big data sets which are computationally infeasible for the standard GP method.

TABLE I. Results of mean run times and MSEs of the standard GP and sparse GP algorithms for different sample sizes, run on a machine with Intel Core i3 processor. The size of the sparse sets varied between $m = 6$ and $m = 19$.

Sample size	Full GP runtime	Full GP MSE	Sparse GP runtime	Sparse GP MSE
300	0.077	1.507	0.005	1.507
500	0.104	1.384	0.008	1.384
1000	0.828	1.292	0.014	1.293
2500	4.19	1.157	0.028	1.157
5000	30.18	0.973	0.056	0.973
10 000	324.5	0.592	0.162	0.593
50 000	–	–	0.783	0.142

VI. ESTIMATION FOR SPARSE OBSERVATIONS

The direct GP approach outlined above leads to wrong estimates of the drift when observations are sparse in time. In the sparse setting, we assume that n observations $z_k \doteq X_{\tau_k}$, $k = 1, \ldots, n$ are obtained at (for simplicity) regular intervals $\tau_k = k\tau$, where $\tau \gg \Delta t$ is much larger than the microscopic time scale (for an example see Fig. 12). In this case, a straightforward discretization in Eq. (5), where the sum over microscopic times t_i would be replaced by a sum over *macroscopic* times τ_k and Δt by τ, would correspond to a discrete time dynamical model of the form Eq. (1) again replacing Δt by τ. But this discretization is a bad approximation to the true SDE dynamics. This is because the transition kernel over macroscopic times τ is simply not a Gaussian for a general f as was assumed in Eq. (17). The failure of the direct estimator for larger time distances can be seen in Fig. 13, where the red dashed line corresponds to the true drift of the double well (with constant, known diffusion) and the black solid line to its prediction based on observations with $\tau = 0.2$. The exact likelihood for the sparse problem would be obtained by integrating the probability Eq. (3) over all paths which are compatible with the observations. Unfortunately, an analytical computation of such functional integrals is not possible for a general drift function f.

To deal with this problem, one treats the process X_t for times t between consecutive observations $k\tau < t < (k+1)\tau$ as a hidden stochastic process with a conditional path probability

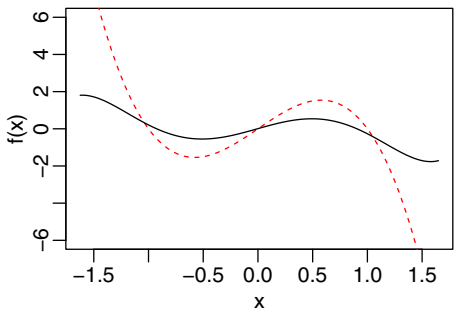

FIG. 13. Estimated drift function for the double well based on the direct approach, where the red dashed line denotes the true drift function and the solid black line the mean function. One can clearly see that the larger distance between the consecutive points leads to a wrong prediction.

given by

$$p(X_{0:T}|\mathbf{z},f) \propto p(X_{0:T}|f)\prod_{k=1}^{n}\delta(z_k - X_{k\tau}), \tag{59}$$

where $\mathbf{z}$ is the collection of observations z_k. A Monte Carlo approach to a full Bayesian estimation of the drift uses a Gibbs sampling method [13] which iteratively updates SDE paths and drift function samples. A short description of such a sampler is given in Appendix B.

As a much more efficient estimation procedure we will describe in the following a different iterative method based on an approximate EM algorithm [15], in which the unobserved complete paths are replaced by an appropriate expectation using the probability Eq. (59).

A. EM algorithm

The EM algorithm cycles between two steps

(1) In the E-step, we compute the expected negative logarithm of the complete data likelihood

$$\mathcal{L}(f,p) = -\mathrm{E}_p[\ln L(X_{0:T}|f)], \tag{60}$$

where p denotes the posterior $p(X_{0:T}|\mathbf{z},f_{\mathrm{old}})$ for the previous estimate f_{old} of the drift.

(2) In the M-Step, we recompute the most likely drift function by the minimization

$$f_{\mathrm{new}} = \arg\min_f [\mathcal{L}(f,p) - \ln P_0(f)]. \tag{61}$$

On can show [15] that the EM algorithm converges to a local maximum of the log-posterior. To compute the expectation in the E-step, we use Eq. (5) and take the limit $\Delta t \to 0$ at the end, when expectations have been computed. As $f(x)$ is a time-independent function, this yields

$$\begin{aligned}
&-\mathrm{E}_p[\ln L(X_{0:T}|f)]\\
&= \lim_{\Delta t\to 0}\frac{1}{2}\sum_t \mathrm{E}_p[||f(X_t)||^2]\Delta t\\
&\quad -2\mathrm{E}_p[(f(X_t), X_{t+\Delta t} - X_t)]\\
&= \frac{1}{2}\int_0^T \mathrm{E}_p[||f(X_t)||^2] - 2\mathrm{E}_p[(f(X_t), g_t(X_t))]dt\\
&= \frac{1}{2}\int ||f(x)||^2 A(x)dx - \int (f(x), b(x))dx.
\end{aligned} \tag{62}$$

We have defined the corresponding drift conditioned on data

$$g_t(x) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\mathrm{E}_p[X_{t+\Delta t} - X_t | X_t = x], \tag{63}$$

as well as the functions

$$A(x) = \int_0^T q_t(x)dt \tag{64}$$

and

$$b(x) = \int_0^T g_t(x)q_t(x)dt. \tag{65}$$

In contrast to Eq. (8), expectations are now over marginal densities $q_t(x)$ of X_t computed from the conditional path measure, not over the asymptotic stationary density. Hence, we end up again with a simple quadratic form in f to be minimized. Note that due to the smoothness of the kernel the prediction of Eq. (61) can be easily differentiated analytically, a fact that will be needed later.

However, there are two main problems for a practical realization of this EM algorithm:

(1) We cannot compute the expectation with respect to the conditional path measures exactly and need to find approximations applicable to *arbitrary* prior drift functions $f(x)$.

(2) Although real observations are sparse, the hidden path involves a continuum of values X_t. This will require (e.g., after some fine discretization of time) the inversion of large matrices in Eq. (29).

We can readily deal with the latter problem by resorting to the sparse GP representation introduced in Sec. V.

1. *Linear drift approximation: The Ornstein-Uhlenbeck bridge*

In this section we will look at the first problem of computing expectations in the E-step. For given drift $f(\cdot)$ and times $t \in I_k$ in the interval $I_k = [k\tau;(k+1)\tau]$ between two consecutive observations, the exact marginal $p_t(x)$ of the conditional path distribution equals the density of $X_t = x$ *conditioned on* the fact that $X_{k\tau} = z_k$ and $X_{(k+1)\tau} = z_{k+1}$. This is a so-called diffusion bridge. Using the Markov property, this density can be expressed by the transition densities $p_s(x_{t+s}|x_t)$ of the homogeneous Markov diffusion process with drift $f(x)$ as

$$p_t(x) \propto p_{(k+1)\tau - t}(z_{k+1}|x)p_{t-k\tau}(x|z_k) \quad \text{for } t \in I_k. \tag{66}$$

As functions of t and x, the second factor fulfills a forward Fokker-Planck equation and the first one a Kolmogorov backward equation [1]. Since exact computations are not feasible for general drift functions, we *approximate* the transition density $p_s(x|x_k)$ in each interval I_k by that of a homogeneous *Ornstein-Uhlenbeck process* [1], where the drift $f(x)$ is replaced by a local linearization. Hence, we consider the approximate process

$$dX_t = [f(z_k) - \Gamma_k(X_t - z_k)]dt + D_k^{1/2}dW \tag{67}$$

with $\Gamma_k = -\nabla f(z_k)$ and $D_k = D(z_k)$ for $t \in I_k$. For this process, the transition density is a multivariate Gaussian

$$q_s^{(k)}(x|z) = \mathcal{N}[x|\alpha_k + e^{-\Gamma_k s}(z - \alpha_k); S_s], \tag{68}$$

where $\alpha_k = z_k + \Gamma_k^{-1}f(z_k)$ is the stationary mean. The covariance $S_s = A_s B_s^{-1}$ is calculated in terms of the matrix exponential

$$\begin{bmatrix} A_s \\ B_s \end{bmatrix} = \exp\left(\begin{bmatrix} \Gamma_k & D_k \\ 0 & -\Gamma_k^\top \end{bmatrix} s\right)\begin{bmatrix} 0 \\ \mathbf{I} \end{bmatrix}. \tag{69}$$

Then we obtain the Gaussian approximation

$$q_t^{(k)}(x) = \mathcal{N}[x|m(t);C(t)] \tag{70}$$

of the marginal posterior for $t \in I_k$ by multiplying the two transition densities, where

$$\begin{aligned}
C(t) &= \left(e^{-\Gamma_k^\top(t_{k+1}-t)}S_{t_{k+1}-t}^{-1}e^{-\Gamma_k(t_{k+1}-t)} + S_{t-t_k}^{-1}\right)^{-1},\\
m(t) &= C(t)\, e^{-\Gamma_k^\top(t_{k+1}-t)}S_{t_{k+1}-t}^{-1}(z_{k+1} - \alpha_k\\
&\quad + e^{-\Gamma_k(t_{k+1}-t)}\alpha_k) + C(t)\, S_{t-t_k}^{-1}\\
&\quad \times [\alpha_k + e^{-\Gamma_k(t-t_k)}(z_k - \alpha_k)].
\end{aligned}$$

By inspecting mean and variance we see that the distribution is in fact equivalent to a bridge between the points $X = z_k$ and $X = z_{k+1}$ and collapses to point masses at these points.

Finally, in this approximation we obtain for the conditional drift

$$\begin{aligned} g_t(x) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{E}[X_{t+\Delta t} - X_t | X_t = x, X_\tau = z_{k+1}] \\ &= f(z_k) - \Gamma_k (x - z_k) + D_k e^{-\Gamma_k^\top (t_{k+1} - t)} S^{-1}_{t_{k+1} - t} \\ &\quad \times [z_{k+1} - \alpha_k - e^{-\Gamma_k (t_{k+1} - t)} (x - \alpha_k)] \end{aligned}$$

as shown in Appendix C.

2. Sparse M-Step approximation

For the M-Step approximation we use the sparse GP formalism of section V. The resulting sparse approximation to the likelihood Eq. (62) is given by

$$\begin{aligned} \mathcal{L}_s(\mathbf{f}, q) = \frac{1}{2} \int ||\mathrm{E}_0[f(x)|\mathbf{f}_s]||^2 \; A(x) \, dx \\ - \int (\mathrm{E}_0[f(x)|\mathbf{f}_s], b(x)) \, dx, \end{aligned} \tag{71}$$

where the conditional expectation is over the GP prior. While the exact likelihood does not contain interactions of the form $f(x)f(x')$ for $x \neq x'$, we allow for couplings of the type $\frac{1}{2}\mathbf{f}^\top \mathbf{\Lambda} \mathbf{f} - \mathbf{a}^\top \mathbf{f}$ in the effective log-likelihood.

To avoid cluttered notation, it should be noted that in the following results for a component f^j, the quantities $\mathbf{\Lambda}_s, \mathbf{f}_s, \mathbf{k}_s, \mathbf{K}_s^{-1}, z(x), \mathbf{D}(x)$ similar to Eq. (29) depend on the component j, but not $A(x)$.

We easily get

$$\mathrm{E}_0[f(x)|\mathbf{f}_s] = \mathbf{k}_s^\top(x) \mathbf{K}_s^{-1} \mathbf{f}_s. \tag{72}$$

Hence,

$$\mathcal{L}_s(\mathbf{f}, q) = \tfrac{1}{2} \mathbf{f}_s^\top \mathbf{\Lambda}_s \mathbf{f}^s - \mathbf{f}_s^\top \mathbf{y}_s, \tag{73}$$

with

$$\mathbf{\Lambda}_s = \mathbf{K}_s^{-1} \left\{ \int \mathbf{k}_s(x) \mathbf{D}(x)^{-1} \; A(x) \; \mathbf{k}_s^\top(x) dx \right\} \mathbf{K}_s^{-1} \tag{74}$$

and

$$\mathbf{y}_s = \mathbf{K}_s^{-1} \int \mathbf{D}(x)^{-1} \mathbf{k}_s(x) \; b(x) \; dx. \tag{75}$$

With these results, the approximate MAP estimate is

$$\bar{f}_s(x) = \mathbf{k}_s^\top(x) (\mathbf{I} + \mathbf{\Lambda}_s \mathbf{K}_s)^{-1} \mathbf{y}_s. \tag{76}$$

The integrals over x in Eqs. (74) and (75) can be computed analytically for many kernels of interest such as polynomial and RBF ones. However, we found it more efficient to treat the time integration in Eqs. (64) and (65) as well as the x integrals by sampling, where time points t are drawn uniformly at random and x points from the multivariate Gaussian $q_t(x)$. A related expression for the variance,

$$\hat{V}_s(x) = K(x,x) - \mathbf{k}_s^\top(x) (\mathbf{I} + \mathbf{\Lambda} \mathbf{K}^s)^{-1} \mathbf{\Lambda}_s \mathbf{k}_s(x), \tag{77}$$

can only be viewed as a crude estimate, because it does not include the impact of the GP fluctuations on the path probabilities.

Finally, a possible approximate evidence for our model is given by the product of the local Ornstein-Uhlenbeck transition probabilities:

$$p(\mathbf{z}) \approx p_{\mathrm{ou}}(\mathbf{z}|\hat{\mathbf{f}}) = p(x_1) \prod_{j=1}^{n-1} q_\tau^{(k)}(z_{k+1}|z_k). \tag{78}$$

The expression is a product of Gaussian transition densities and therefore of analytical form. Note that in addition to the Ornstein-Uhlenbeck linearization, this approximation also neglects the uncertainty of $\mathbf{f}$, since the GP in the M step only uses the expectation.

Nevertheless, in our experiments we found that the use of the approximate evidence is a reasonable choice for the optimization of the diffusion $D(x)$, see Sec. VI D. However, the optimization of the kernel hyperparameters is more problematic, since the approximate evidence depends on the drift estimate $\hat{\mathbf{f}}$, which itself depends on the choice of the hyperparameters through the application of the GP. Since we assume that prior knowledge of a suitable kernel hyperparameters is often available, we did not pursue this problem further.

B. Experiments

We created the synthetic data sets in this section by first using the Euler method from the corresponding SDE with grid size $\Delta_{\mathrm{dense}} = 0.002$. Then for a data set of N observations separated by $\Delta t \gg \Delta_{\mathrm{dense}}$, we keep every $k = (\Delta t / \Delta_{\mathrm{dense}})$th path sample value as observation, until the desired observation number N is reached.

The EM algorithm is initialized with the sparse direct GP estimator, which works well in practice as a reasonable first approximation to the true system dynamics. Although the monotonicity property of the EM algorithm is no longer satisfied due to the approximation in the E-step, convergence will be assumed, once $\mathcal{L}$ stabilizes up to some minor fluctuations. In our experiments convergence was typically attained after a few (<10) iterations.

1. Performance comparison

First, we compare the estimation accuracy of the direct GP and the EM algorithm on the double-well model with constant known diffusion,

$$dX = 4(X - X^3) dt + dW_t, \tag{79}$$

for different time discretization Δt. For each time step, we generated 20 data sets, each of size $n = 4000$, and computed the MSE on a test set of size $n = 2000$ for each data set and for both algorithms using the RBF kernel. As benchmark reference, we include the estimation results of a Monte Carlo sampler (see Appendix B). The latter one is represented only for one data set at small and medium time intervals, respectively, due to its long computation time. To improve comparability, we fixed the length scale of the RBF kernel to $l = 0.62$ for all data sets.

The results are given in Fig. 14. The MSE of the direct GP grows quite rapidly for smaller intervals until it reaches an upper bound roughly equivalent with randomly guessing the drift function. On the other hand, the MSE for the EM algorithm increases at a much slower rate, giving good results even for

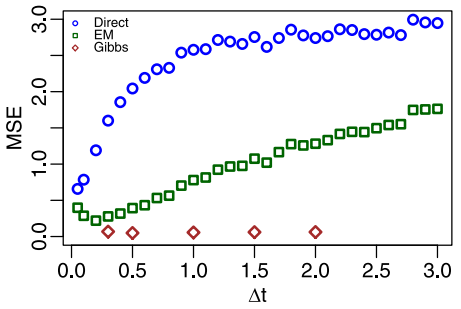

FIG. 14. Comparison of the MSE for different methods over different time intervals.

data sets with bigger time distances. The estimation results for the Gibbs sampler are independent of the discretization rate, but take considerable time to compute: while the EM algorithm runs for a couple of minutes, the sampler takes up to two days.

2. Double-well model with known state-dependent diffusion

As our next example we examine the double-well model with state-dependent diffusion and larger time discretization. Here we assume that the diffusion function $D(x)$ is known. Specifically, we sample $n = 4000$ observation at $\Delta t = 0.5$ and run the EM algorithm with a polynomial kernel of order $p = 4$. The direct GP and the EM result are given in Figs. 13 and 15, respectively. One can clearly see, that an application of EM leads to a significantly better estimator of the drift function, compared to the direct GP method.

3. Two-dimensional synthetic model

We now turn to a two-dimensional process with the following dynamics:

$$dX = [X(1 - X^2 - Y^2) - Y]dt + dW_t^{(1)}, \tag{80}$$

$$dY = [Y(1 - X^2 - Y^2) + X]dt + dW_t^{(2)}, \tag{81}$$

where the component indices are denoted by superscripts. For this model we generated $n = 10\,000$ observations with step size $\Delta t = 0.2$ shown in Fig. 16. The estimation in Fig. 17 uses a polynomial kernel of order $p = 4$ and shows a good fit to the true drift especially in the regions where the observations are concentrated. Note that this is a nonequilibrium model, where

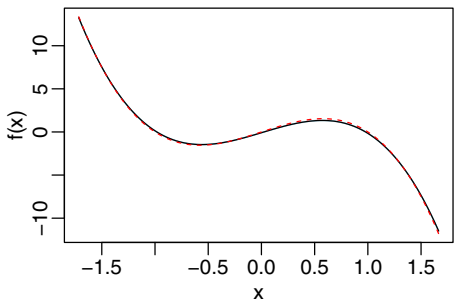

FIG. 15. GP estimation after one iteration of the EM algorithm. Again, the solid black and red dashed lines denote estimator and true drift function, respectively.

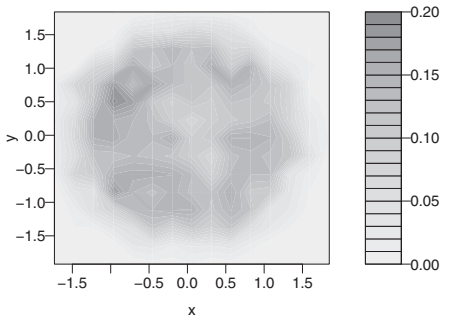

FIG. 16. Empirical density of the data set generated from the two-dimensional synthetic model.

the drift cannot be expressed as the gradient of a potential. Hence, the density based method of Ref. [29] cannot be applied here.

4. Lorenz'63 model

We next analyze a stochastic version of the three-dimensional Lorenz'63 model. It consists of the following system of nonlinear coupled stochastic differential equations:

$$dX = \sigma(Y - X)dt + dW_t^{(1)}, \tag{82}$$

$$dY = (\rho X - X - XZ)dt + dW_t^{(2)}, \tag{83}$$

$$dZ = (XY - \beta Z)dt + dW_t^{(3)}. \tag{84}$$

Lorenz'63 is a chaotic system which was developed as a simplified model of thermal convection in the atmosphere [30]. The parameters $\boldsymbol{\theta} = (\sigma, \rho, \beta)$ are set to the commonly used $\boldsymbol{\theta} = (10, 28, 8/3)$ known to induce chaotic behavior in the system. To analyze the model we simulate $n = 3000$ data points with time discretization $\Delta t = 0.2$. In the inference, we used a polynomial kernel of order $p = 2$ and assume that the constant diffusion is known.

To visualize the quality of the drift estimation, we computed it with both the direct GP and the EM algorithm. Figures 18 and 19 show simulated paths using the resulting mean estimators as drift functions. Here, the application of EM leads to a vastly superior estimation result compared to the direct method. As shown in Fig. 20, the direct GP estimator path collapses to a small region of the function space, whereas the EM trajectory

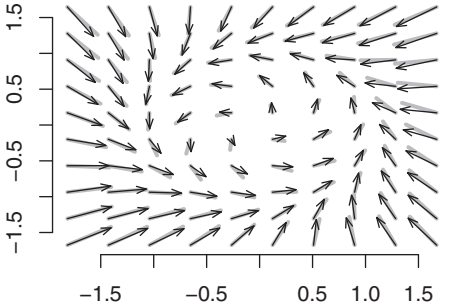

FIG. 17. Vector fields of the true drift depicted as gray lines and the estimated drift as black arrows for the two-dimensional synthetic model.

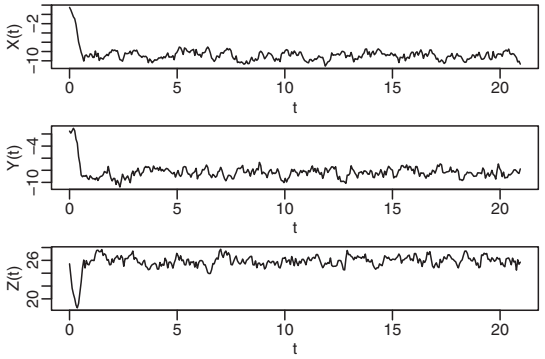

FIG. 18. Simulated sample path of the Lorenz'63 model learned by the direct GP algorithm.

of Fig. 21 nicely captures the true dynamics of the Lorenz'63 model, faithfully recreating the famous butterfly pattern in the X-Z plane.

5. Cart and pole model

Next, we consider an example from the class of mechanical systems. Our model describes the dynamics of a pole attached to a cart moving randomly along an one-dimensional axis. Formally, we get a system of two-dimensional differential equations with x denoting the angle of the pendulum, and v the angular velocity. We define the upright position of the pendulum as $X = 0$. This particular *cart and pole* model is frequently studied in the context of learning control policies [31], where the goal is to move the cart in such a way as to stabilize the pendulum in the upright position. The complete system looks as follows:

$$dX = V dt, \tag{85}$$

$$dV = \frac{-\gamma V + mgl\sin(X)}{ml^2}dt + d^{1/2}dW_t, \tag{86}$$

where $\gamma = 0.05$ is the friction coefficient, $l = 1$ m and $m = 1$ kg are the length and mass of the pendulum, respectively, and $g = 9.81$ m s^{-2} denotes the gravitational constant. For our experiment, we generated $N = 4000$ data points (x, v) on a grid with $\Delta t = 0.3$ and known diffusion constant $d = 1$. Here, the

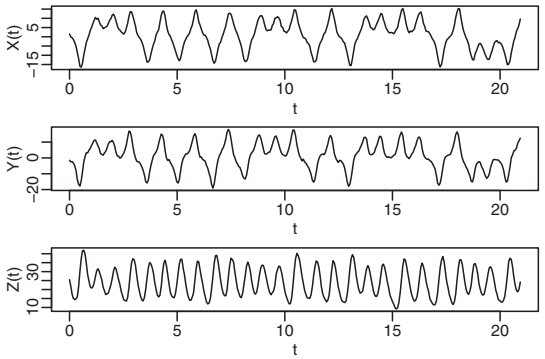

FIG. 19. Simulated sample path of the Lorenz'63 model learned by the EM algorithm.

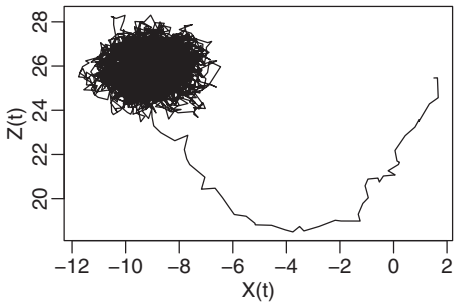

FIG. 20. Simulated path in the X-Z plane from the Lorenz'63 model learned by the direct GP algorithm.

full diffusion matrix

$$D = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \tag{87}$$

for both X and V is rank deficient due to its noiseless first equation. However, we note that our EM algorithm is also applicable to models with deterministic components, since the E-Step in the EM algorithm remains well defined. In the kernel function we incorporate our prior knowledge that the pendulum angle is periodic and the velocity acts as a linear friction term inside the system. Specifically, we define the following multiplicative kernel for the dV equation:

$$K[(x,v),(x',v')] = K_{\text{Per}}(x,x')K_{\text{Poly}}(v,v'), \tag{88}$$

where K_{Per} denotes the periodic kernel over the state x with hyperparameters $l = 1.21$ and K_{Poly} the polynomial kernel of order $p = 1$ over the velocity V. The multiplicative kernel structure allows for interactions between its components. Since in this model the components are independent, we could also use an additive kernel, which neglects interactions terms, but we have chosen the more generally applicable variant here. For the dX equation, we use a polynomial kernel of order $p = 1$, which captures the linear relationship between X and V. If we adapt our choice of the kernel to the specific form of the system, we get an accurate estimate even for data points separated by a wider time spacing (see Figs. 22 and 23).

C. External forces

We can expect a reasonably good estimation of $f(x)$ only in regions of x where we have enough observations. This is of clear importance, when the system is multistable and the noise

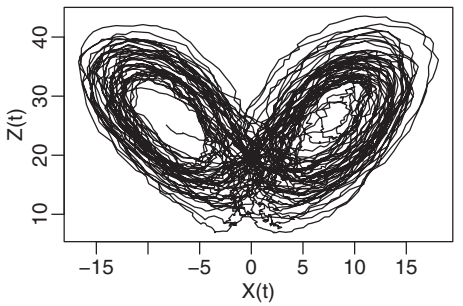

FIG. 21. Simulated path in the X-Z plane from the Lorenz'63 model learned by the EM algorithm.

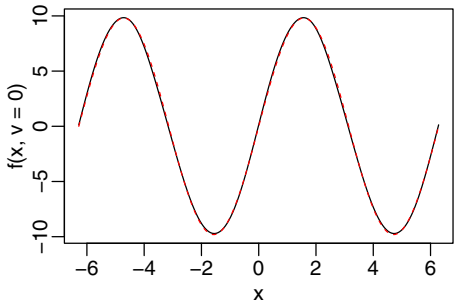

FIG. 22. One-dimensional drift estimation of the second (dV) SDE of the cart and pole model. This figure shows the drift as a function of the position X for the fixed velocity $V = 0$. The solid black line denotes the estimation and the red dashed line the true function.

is too small to allow for a sufficient exploration of space. An alternative method for exploration would be to add a known external *deterministic control force* $u(t)$ to the dynamics which is designed to drive the system from one locally stable region to another one. Hence, we assume a SDE,

$$dX_t = [f(X_t) + u(t)]dt + D^{1/2}dW_t. \tag{89}$$

This situation is easily incorporated into our formalism. In all likelihood terms, we replace $f(X_t)$ by $f(X_t) + u(t)$ but keep the zero mean GP prior over functions. The changes for the corresponding transition probabilities of the approximating time-dependent Ornstein-Uhlenbeck bridge are given in Appendix D.

We demonstrate the concept by applying it to the double-well model. We get

$$dX = [4(X - X^3) + u(t)]dt + \sigma dW_t. \tag{90}$$

As external force we choose a periodic control function of the form $u(t) = a\sin(\omega t)$ with parameters $a = 1$ and $\omega = 3$. We generated a data set of $n = 2000$ observations on a regular grid with distance $\Delta t = 0.2$ from the model with known diffusion $D^{1/2} = 0.5$. The addition of $u(t)$ leads to observations from both of the wells, whereas in the uncontrolled case only one part of the underlying state space is explored. Hence, the drift estimation in the latter case leads to an accurate result solely around the well at $X = 1$, as opposed to the controlled case,

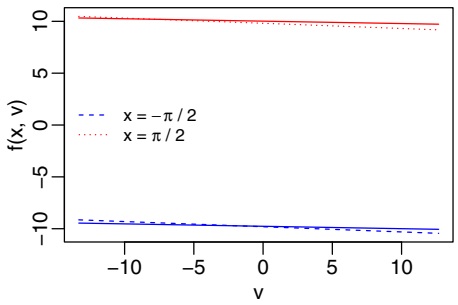

FIG. 23. One-dimensional drift estimation of the second (dV) SDE of the cart and pole model: Drift as a function of V for the two horizontal positions with the top pointing to the left $X = -\pi/2$ and to the right side $X = \pi/2$. Full lines denote the drift estimation and dashed and dotted lines the true values.

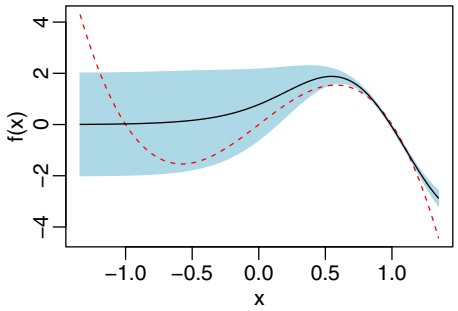

FIG. 24. EM algorithm predictions for the uncontrolled double-well path with the solid black line denoting the estimation and the dashed red line the true drift. Here, the estimation of the well around $X = -1$ basically equals the GP prior, since there are no observations in this region. The shaded area can be interpreted as the 95%-confidence bound.

where both modes are truthfully recovered (Figs. 24 and 25). In both cases, we used a RBF kernel with $\tau = 1$. The length scales was set to $l = 0.74$ in the controlled and $l = 0.53$ in the uncontrolled case.

D. Diffusion estimation

As in the dense data scenario, we look at constant and state-dependent diffusions in turn. If D does not depend on the state, we can proceed in analogy to the dense data case and maximize the approximate evidence Eq. (78) with respect to the diffusion values.

For the state-dependent case $D(x)$ we assume a parametric function $D(x;\boldsymbol{\theta})$, which is specified by its parameter vector $\boldsymbol{\theta}$. Here, we again maximize the likelihood with respect to the corresponding $\boldsymbol{\theta}$.

For an illustration, we do not show the constant diffusion case and instead restrict ourself to the more interesting case of a state-dependent $D(x)$. We sampled $n = 8000$ observations at $\Delta t = 0.3$ from the following process:

$$dX = 0.4(4 - X)dt + \max[2 - (X - 4)^2, 0.25]dW_t. \tag{91}$$

The diffusion function was modelled as $D(x,\boldsymbol{\theta}) = \theta_1 x^2 + \theta_2 x + \theta_3$. As kernel function for the drift, we used a polynomial kernel of order $p = 1$. Optimizing the evidence with respect to $\boldsymbol{\theta}$ leads to the results shown in Fig. 26. One can see that the

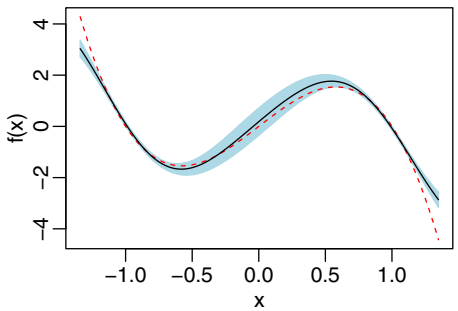

FIG. 25. EM algorithm predictions for the controlled double-well path. The solid black line is the estimated drift and the dashed red line the true function.

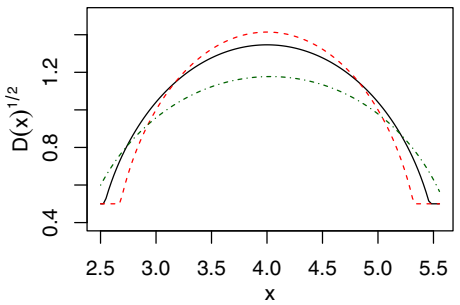

FIG. 26. Comparison of the diffusion estimation for data generated from Eq. (91). The dashed red line is the true square root $D(x)^{1/2}$ of the diffusion and the solid black line the parametric estimation based on the EM algorithm. For comparison, we include the estimate based on the direct GP denoted by the green dashed-dotted line.

estimation gives a reasonably good fit to the true diffusion function even with the bigger time discretization. We note however, that the diffusion estimate is of a lower quality than the drift estimate, since in this case the evidence is less accurate.

VII. DISCUSSION

It would be interesting to replace the ad hoc local linear approximation of the posterior drift by a more flexible time-dependent Gaussian model. This could be optimized in a variational EM approximation by minimizing a free energy in the E-step, which contains the Kullback-Leibler divergence between the linear and true processes [20,32]. Such a method could be extended to noisy observations and the case, where some components of the state vector are not observed. Also, this method could be turned into a variational Bayesian approximation, where one optimizes posteriors over both drifts and over state paths. The path probabilities are then influenced by the uncertainties in the drift estimation, which would lead to more realistic predictions of error bars.

Finally, nonparametric diffusion estimation deserves further attention. Incorporating a fully nonparametric model of the diffusion function $D(x)$ in our scheme would be infeasible in practice, since this would involve the joint estimation of n diffusion matrices. In our preliminary experiments, we tried a (quasi-)nonparametric approach, where we represented the diffusion function by its value at a few supporting points and took these as inputs for a GP regression, which we then used as function approximation. However, our experiments have shown that to achieve a reasonable estimation quality we need supporting points on a relatively dense grid. The corresponding optimization over the vector of grid points turned out to be too inefficient, which makes the approach impractical. Furthermore, the evidence over which we optimize is often too inaccurate to lead to a reasonable quality.

If performance time is not at all critical, one can resort to a Markov chain Monte Carlo (MCMC) algorithm, which generates exact samples from the corresponding drift and diffusion functions. In contrast to the EM algorithm, the sampler evaluates the diffusion function on a dense grid and also does not use the assumption of constant diffusion between adjacent observations, thereby overcoming the significant estimation errors for larger time distances. We plan to report on this in a future publication.

ACKNOWLEDGMENT

This work was supported by the European Community's Seventh Framework Programme (FP7, 2007-2013) under Grant Agreement No. 270327 (CompLACS).

APPENDIX A: POLYNOMIAL KERNEL

A *kernel* is defined as function k, such that for all $\mathbf{x},\mathbf{x}' \in \mathcal{X}$,

$$k(\mathbf{x},\mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle$$

holds, where $\langle \cdot\,,\cdot \rangle$ defines an inner product and $\boldsymbol{\phi}(\mathbf{x})$ denotes a map from $\mathcal{X}$ to a feature space $\mathcal{F}$,

$$\boldsymbol{\phi} : \mathbf{x} \to \boldsymbol{\phi}(\mathbf{x}) \in \mathcal{F}.$$

To show that the function Eq. (15) defines a valid kernel, we assume input vectors $\mathbf{x},\mathbf{x}' \in \mathcal{R}^D$ and incorporate the constant by concatenating it with the input. Then

$$\begin{aligned} k(\mathbf{x},\mathbf{x}') &= \left(\sum_{d=1}^{D+1} x_d x'_d \right)^p \\ &= \left(\sum_{d_1=1}^{D+1} x_{d_1} x'_{d_1} \right) \cdots \left(\sum_{d_p=1}^{D+1} x_{d_p} x'_{d_p} \right) \\ &= \sum_{d=1}^{D+1} \cdots \sum_{d=1}^{D+1} \left(x_{d_1} \cdots x_{d_p} \right) \left(x'_{d_1} \cdots x'_{d_p} \right) \\ &\doteq \langle \boldsymbol{\phi}_P(\mathbf{x}) \boldsymbol{\phi}_P(\mathbf{x}') \rangle . \end{aligned}$$

Hence, the polynomial kernel can be written as an inner product in the feature space induced by the map $\boldsymbol{\phi}_P$. Further details can be found in Refs. [18,33].

APPENDIX B: MCMC SAMPLER

We briefly describe the Markov chain Monte Carlo (MCMC) algorithm, which generates samples from the drift function of a system of SDEs with known diffusion. Similar to the EM algorithm in the main text, the drift is modeled in a nonparametric way.

As before, our data will be a set of N observations $\mathbf{Y} = (y_1, \ldots, y_N)$, where $y_k = X_{k\tau}$. Since the time distance between adjacent observations is taken to be large, we impute the process between observations in interval $I_k = [k\,\tau; (k+1)\tau]$ on a fine grid of step size $\Delta = \tau/M$ for some suitable integer M. The imputed path of the kth subinterval will be denoted by $\mathbf{X}_k = \{X_{k\tau}, X_{k\tau+\Delta}, \ldots, X_{k\tau+M\Delta}\}$.

If we write the complete imputed path of length MN as

$$\begin{aligned} \mathbf{X} = (&y_0, X_\Delta, \ldots, X_{(M-1)\Delta}, \ldots, y_1, \ldots, \\ &X_{(k-1)\tau+(M-1)\Delta}, y_k, X_{k\tau+\Delta}, \ldots, y_N), \end{aligned}$$

then the joint posterior distribution of the data and the drift and diffusion function for a given set of observations is given by

$$p(\mathbf{X}, f | \mathbf{Y}, D) \propto p_0(f) \prod_{l=1}^{NM} p(X^{l+1} | X^l, f, D).$$

Here, the density $p(\mathbf{X}, f|\mathbf{Y}, D)$ is approximately normally distributed [see Eq. (5)] on the fine grid with mean and variance given by Eqs. (29) and (30), respectively. A straightforward way to sample from this posterior is given by the following Gibbs sampler:

Algorithm 1. Gibbs Sampler.

1: Initialize $f^{(0)}$ with the direct GP solution
2: **for** $i = 1, \ldots, N$ **do**
3: Sample $\mathbf{X}^{(i)} \sim p(\mathbf{X}|\mathbf{Y}, f^{(i-1)}, D)$
4: Sample $f^{(i)} \sim p(f|\mathbf{X}^{(i)}, \mathbf{Y})$

Here, the superscripts denote the iteration. The number of iterations for a particular model is determined by the usual MCMC convergence diagnostics; see, for example, Ref. [34]. Since an analytic form for the imputed path distribution $p(\mathbf{X}|\mathbf{Y}, f, D)$ does not exist, we have to resort to a Metropolis-Hasting (MH) step. As proposal distribution q, we use the so-called *modified diffusion bridge* (MDB) of Ref. [35]. Here, for each interval I_k the density of a grid point X_k^{j+1} from $\mathbf{X}_k$ is normally distributed, conditioned on X_k^j and the interval endpoint y_{k+1}:

$$q\big(X_k^{j+1}|X_k^j, y_{k+1}, f_q, D_q\big) = \mathcal{N}\big[X_k^{j+1}|X_k^j + f_q\big(X_k^j\big)\Delta, D_q\big(X_k^j\big)\big], \tag{B1}$$

with drift and diffusion

$$f_q\big(X_k^j\big) = \frac{y_{k+1} - X^j}{\tau - j\Delta}, \quad D_q\big(X_k^j\big) = \frac{\tau - (j+1)\Delta}{\tau - j\Delta} D\big(X_k^j\big).$$

Now, since for each subinterval I_k the bridge proposal starts in observation y_k and terminates in y_{k+1}, we can generate a sample of the complete path $p(\mathbf{X}|\mathbf{Y}, f, D)$ by sampling a MDB proposal separately for each the N subintervals. Specifically, for subinterval I_k we simulate a path $\mathbf{X}_k^*$ on the dense grid by recursively sampling from Eq. (B1) and move from current state $\mathbf{X}_k$ to $\mathbf{X}_k^*$ with probability

$$\alpha(\mathbf{X}_k, \mathbf{X}_k^*) = \min\left\{1, \left[\prod_{j=1}^{M-1} \frac{p\big(X_k^{*(j+1)}|X_k^{*j}, f, D\big)}{p\big(X_k^{j+1}|X_k^j, f, D\big)}\right] \times \left[\prod_{j=1}^{M-2} \frac{q\big(X_k^{j+1}|X_k^j, y_{k+1}, f_q, D_q\big)}{q\big(X_k^{*(j+1)}|X_k^{*j}, y_{k+1}, f_q, D_q\big)}\right]\right\},$$

with probability $[1 - \alpha(\mathbf{X}_k, \mathbf{X}_k^*)]$ we retain the current path $\mathbf{X}_k$.

The sampling from the drift $p(f|\mathbf{X}, \mathbf{Y})$ is easier to accomplish, since under a GP prior $P_0 \sim \mathcal{GP}$ assumption, the distribution $p(f|\mathbf{Y}, \mathbf{X})$ of the SDE drift corresponds to a GP posterior and is therefore of analytic form. Since the number of dense path observations is usually quite substantial, we resort to the sparse version of the GP with mean and variance given by Eqs. (54) and (55), respectively. In each iteration of the Gibbs sampler, we simulate a new f on a fine grid over the (slightly extended) range of the path observations $\mathbf{X}$ and then interpolate these points by nonparametric regression to arrive at an approximate drift function. The interpolation step, for which we again resort to a sparse GP, is motivated by computational considerations, since this way evaluating the function values for the path can be can be done very efficiently, while also being accurate due to the smoothness of the underlying drift.

APPENDIX C: CONDITIONAL DRIFT

Here, we give the derivation of the conditional drift term $g_t(x)$, which occurs in the E-step of the EM algorithm:

$$\begin{aligned}
g_t(x) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{E}[X_{t+\Delta t} - X_t | X_t = x, X_\tau = y] \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\int (x' - x)\; p_{\tau - t - \Delta t}(y|x') p_{\Delta t}(x'|x)\, dx'}{\int p_{\tau - t - \Delta t}(y|x') p_{\Delta t}(x'|x)\, dx'} \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{f(x)\Delta t + \mathrm{E}_u[p_{\tau - t - \Delta t}(y|x + f(x)\Delta t + u)u]}{\mathrm{E}_u[p_{\tau - t - \Delta t}(y|x + f(x)\Delta t + u)]} \\
&= f(x) + D \lim_{\Delta t \to 0} \frac{\nabla_x \mathrm{E}_u[p_{\tau - t - \Delta t}(y|x + f(x)\Delta t + u)]}{\mathrm{E}_u[p_{\tau - t - \Delta t}(y|x + f(x)\Delta t + u)]} \\
&= f(x) + D \lim_{\Delta t \to 0} \nabla_x \ln\{\mathrm{E}_u[p_{\tau - t - \Delta t}(y|x + f(x)\Delta t + u)]\} \\
&= f(x) + D \nabla_x \ln\{p_{\tau - t}(y|x)\}.
\end{aligned}$$

The second line follows from the definition of the conditional density, the third line from the fact that $p_{\Delta t}(x'|x) = \mathcal{N}(x + f(x)\Delta t; D\Delta t)$ and $u \sim \mathcal{N}(0; \sigma^2 \Delta t)$. The fourth line is based on the fact that for zero mean Gaussian random vectors with covariance S, we have $\mathrm{E}[ug(u)] = S\mathrm{E}[\nabla_u g(u)]$. Finally, the last line is obtained by noting that the covariance of u vanishes for $\Delta t \to 0$.

APPENDIX D: ORNSTEIN-UHLENBECK BRIDGE WITH EXTERNAL FORCES

If there is an additional time-dependent and known drift term $u(t)$, e.g., a control force, in the Ornstein-Uhlenbeck model, i.e.,

$$dX_t = [f(y_k) - \Gamma_k(X_t - y_k) + u(t)]dt + D_k^{1/2} dW,$$

with $\Gamma_k = -\nabla f(y_k)$ and $D_k = D(y_k)$, the mean of the marginal posterior is changed to

$$\begin{aligned}
m(t) = C(t) e^{-\Gamma_k^\top (\tau - u)} S_{\tau - u}^{-1} \left[x_{k+1} - \alpha_k + e^{-\Gamma_k(\tau - u)}\alpha_k - \int_u^\tau e^{-\Gamma_k(\tau - v)} u(t - u + v) dv\right] \\
+ C(t) S_u^{-1} \left[\alpha_k + e^{-\Gamma_k u}(x_k - \alpha_k) + \int_0^u e^{-\Gamma_k(u - v)} u(t - u + v) dv\right],
\end{aligned}$$

but the covariance matrix stays the same. For the posterior drift, we get in this case

$$g_t(x) \approx f(x_k) - \Gamma_k(x - x_k) + D_k e^{-\Gamma_k^\top(\tau-u)} S^{-1}_{\tau-u}\left[x_{k+1} - \alpha_k - e^{-\Gamma_k(\tau-u)}(x-\alpha_k) - \int_u^\tau e^{-\Gamma_k(\tau-v)} u(t-u+v)dv\right].$$

For $u(t) = a\sin(\omega t)$:

$$\begin{aligned} m(t) = C(t)e^{-\gamma_k(t_{k+1}-t)} S^{-1}_{t_{k+1}-t}\Bigg(x_{k+1} - \alpha_k + e^{-\gamma_k(t_{k+1}-t)}\alpha_k - \frac{a}{\gamma_k^2+\omega^2}\{[\gamma_k \sin(\omega t_{k+1}) - \omega\cos(\omega t_{k+1})] \\ - e^{-\gamma_k(t_{k+1}-t)}[\gamma_k\sin(\omega t) - \omega\cos(\omega t)]\}\Bigg) + C(t)S^{-1}_{t-t_k}\Bigg(\alpha_k + e^{-\gamma_k(t-t_k)}(x_k - \alpha_k) \\ + \frac{a}{\gamma_k^2+\omega^2}\{[\gamma_k\sin(\omega t) - \omega\cos(\omega t)] - e^{-\gamma_k(t-t_k)}[\gamma_k\sin(\omega t_k) - \omega\cos(\omega t_k)]\}\Bigg), \end{aligned}$$

$$\begin{aligned} g_t(x) \approx f(x_k) + a\sin(\omega t) - \gamma_k(x-x_k) + De^{-\gamma_k(t_{k+1}-t)} S^{-1}_{t_{k+1}-t}\Bigg(x_{k+1} - \alpha_k - e^{-\gamma_k(t_{k+1}-t)}(x-\alpha_k) \\ - \frac{a}{\gamma_k^2+\omega^2}\{[\gamma_k\sin(\omega t_{k+1}) - \omega\cos(\omega t_{k+1})] - e^{-\gamma_k(t_{k+1}-t)}[\gamma_k\sin(\omega t) - \omega\cos(\omega t)]\}\Bigg). \end{aligned}$$

[1] C. W. Gardiner, *Handbook of Stochastic Methods*, 2nd ed. (Springer, Berlin, 1996).

[2] S. M. Iacus, *Simulation and Inference for Stochastic Differential Equations: With R Examples (Springer Series in Statistics)*, 1st ed. (Springer, Berlin, 2008).

[3] H. J. Kappen, Linear Theory for Control of Nonlinear Stochastic Systems, Phys. Rev. Lett. **95**, 200201 (2005).

[4] M. Johannes and N. Polson, MCMC methods for continuous-time financial econometrics, in *Handbook of Financial Econometrics: Applications*, edited by Y. Ait-Sahalia and L. P. Hansen, Handbooks in Finance, Vol. 2 (Elsevier, San Diego, 2010), pp. 1–72.

[5] H. Wu and F. Noe, Bayesian framework for modeling diffusion processes with nonlinear drift based on nonlinear and incomplete observations, Phys. Rev. E **83**, 036705 (2011).

[6] A. Golightly and D. J. Wilkinson, Markov chain Monte Carlo algorithms for SDE parameter estimation, *Learning and Inference for Computational Systems Biology* (MIT Press, Cambridge, MA, 2010), pp. 253–276.

[7] C. K. Wikle and M. B. Hooten, A general science-based framework for dynamical spatiotemporal models, Test **19**, 417 (2010).

[8] I. Sgouralis and S. Pressé, Icon: An adaptation of infinite HMMS for time traces with drift, Biophys. J. **112**, 2117 (2017).

[9] I. Sgouralis and S. Pressé, An introduction to infinite HMMS for single-molecule data analysis, Biophys. J. **112**, 2021 (2017).

[10] F. van der Meulen, M. Schauer, and H. van Zanten, Reversible jump MCMC for nonparametric drift estimation for diffusion processes, Comput. Stat. Data Anal. **71**, 615 (2014).

[11] For non-Bayesian nonparametric approaches to SDE models, see, e.g., Refs. [36,37], which are effectively restricted to one-dimensional SDE models. See also Ref. [29] for the special case, where the drift can be expressed as gradient of a potential, and where one can utilize the relationship between stationary density and potential.

[12] D. S. Oliver, L. B. Cunha, and A. C. Reynolds, Markov chain Monte Carlo methods for conditioning a permeability field to pressure data, Math. Geol. **29**, 61 (1997).

[13] O. Papaspiliopoulos, Y. Pokern, G. O. Roberts, and A. M. Stuart, Nonparametric estimation of diffusions: A differential equations approach, Biometrika **99**, 511 (2012).

[14] Y. Pokern, A. M. Stuart, and J. H. van Zanten, Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDEs, Stoch. Process. Appl. **123**, 603 (2013).

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. Ser. B (Methodological) **39**, 1 (1977).

[16] A. Ruttor, P. Batz, and M. Opper, Approximate gaussian process inference for the drift function in stochastic differential equations, in *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc., Red Hook, NY, 2013), pp. 2040–2048.

[17] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations* (Springer, New York, 2011).

[18] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2006).

[19] M. Lázaro-Gredilla and M. K. Titsias, Variational heteroscedastic Gaussian process regression, in *Proceedings of the 28th International Conference on Machine Learning (ICML'11), Bellevue, WA* (ACM, New York, NY, 2011), pp. 841–848.

[20] C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Taylor, Variational inference for diffusion processes, in *Advances in Neural Information Processing Systems 20*, edited by J. C. Platt, D. Koller, Y. Singer, and S. Roweis, (MIT Press, Cambridge, MA, 2008), pp. 17–24.

[21] K. K. Andersen, N. Azuma, J-M. Barnola, M. Bigler, P. Biscaye, N. Caillon, J. Chappellaz, H. B. Clausen, D. Dahl-Jensen, H. Fischer *et al.*, High-resolution record of northern hemisphere climate extending into the last interglacial period, Nature **431**, 147 (2004).

[22] F. Kwasniok and G. Lohmann, Deriving dynamical models from paleoclimatic records: Application to glacial millenial-scale climate variability, Phys. Rev. E **80**, 066104 (2009).

[23] F. Kwasniok, Analysis and modeling of glacial climate transitions using simple dynamical systems, Philos. Trans. Roy. Soc. A: Math. Phys. Eng. Sci. **371**, 20110472 (2013).

[24] W. Dansgaard, S. J. Johnsen, H. B. Clausen, D. Dahl-Jensen, N. S. Gundestrup, C. U. Hammer, C. S. Hvidberg, J. P. Steffensen, A. E. Sveinbjörnsdottir, J. Jouzel *et al.*, Evidence for general instability of past climate from a 250-kyr ice-core record, Nature **364**, 218 (1993).

[25] M. Titsias, Variational learning of inducing variables in sparse Gaussian processes, in *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 5, edited by D. van Dyk and M. Welling (PMLR, 2009), pp. 567–574.

[26] L. Csató, M. Opper, and O. Winther, TAP Gibbs free energy, belief propagation and sparsity, in *Advances in Neural Information Processing Systems 14*, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani (MIT Press, Cambridge, MA, 2002), pp. 657–663.

[27] A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York, 1965).

[28] H. A. Sturges, The choice of a class interval, J. Am. Stat. Assoc. **21**, 65 (1926).

[29] P. Batz, A. Ruttor, and M. Opper, Variational estimation of the drift for stochastic differential equations from the empirical density, J. Stat. Mech. (2016) 083404.

[30] E. N. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci. **20**, 130 (1963).

[31] M. P. Deisenroth, C. E. Rasmussen, and J. Peters, Gaussian process dynamic programming, Neurocomputing **72**, 1508 (2009).

[32] M. D. Vrettas, D. Cornford, and M. Opper, Variational mean-field algorithm for efficient inference in large systems of stochastic differential equations, Phys. Rev. E **91**, 012148 (2015).

[33] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis* (Cambridge University Press, Cambridge, 2004).

[34] C. Robert and G. Casella, *Monte Carlo Statistical Methods* (Springer Science & Business Media, Berlin, 2013).

[35] G. B. Durham and A. R. Gallant, Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes, J. Bus. Econ. Stat. **20**, 297 (2002).

[36] S. J. Lade, Finite sampling interval effects in Kramers-Moyal analysis, Phys. Lett. A **373**, 3705 (2009).

[37] F. M. Bandi and P. C. B. Phillips, Fully nonparametric estimation of scalar diffusion models, Econometrica **71**, 241 (2003).