Variational estimation of the drift for stochastic differential equations from the empirical density

Philipp Batz, Andreas Ruttor and Manfred Opper

Artificial Intelligence Group, Technische Universität Berlin, Marchstraße 23, Berlin 10587, Germany
E-mail: philipp.batz@tu-berlin.de, andreas.ruttor@tu-berlin.de and manfred.opper@tu-berlin.de

Received 3 March 2016, revised 20 May 2016
Accepted for publication 24 July 2016
Published 22 August 2016

Online at stacks.iop.org/JSTAT/2016/083404
doi:10.1088/1742-5468/2016/08/083404

Abstract. We present a method for the nonparametric estimation of the drift function of certain types of stochastic differential equations from the empirical density. It is based on a variational formulation of the Fokker–Planck equation. The minimization of an empirical estimate of the variational functional using kernel based regularization can be performed in closed form. We demonstrate the performance of the method on second order, Langevin-type equations and show how the method can be generalized to other noise models.

Keywords: dynamical processes, statistical inference, stochastic processes
1. Introduction

An important problem in modeling a random process by a stochastic differential equation (SDE) is the fitting of the model to observed data. An SDE is determined by its drift function and the diffusion. For models of thermal equilibrium, where the diffusion is proportional to the unit matrix and the drift is the gradient of a potential, a rather simple and well known approach for estimating the drift from data is available (see e.g. [1]): one can use the fact that the potential is proportional to the logarithm of the stationary density. Hence, from an estimator of the density, provided e.g. by a kernel density estimator (KDE) one can get an explicit estimator for the drift. This estimator is based on the empirical distribution of data alone which completely ignores the temporal ordering of observations and the time lag between them. One only needs an ergodic sample of the process. The KDE estimator is also nonparametric i.e. it does not assume a specific parametric functional form of the drift. For non-equilibrium models such an explicit expression of the density in terms of the drift is, in general, not known. Also in higher dimensions the convergence of the KDE to the true density with increasing data sample size may be slow [2].

Other parametric and nonparametric approaches to drift estimation have to deal with the problem of low data sampling rates [3]. E.g. a nonparametric method based
on Kramers–Moyal coefficients [4] (conditional moments) requires numerical solutions of the Kolmogorov backward equation over a time interval given by the time lag. Bayesian estimators using a Gaussian process prior over drift functions provide an elegant solution to estimation [5] when a complete path of dense observations is available. But in general it requires the imputation of unobserved diffusion paths as hidden random variables between neighboring observations. This can lead to time consuming computations or requires further approximations. Papaspiliopoulos et al [5] introduced a Monte Carlo Gibbs sampler, which switches between sampling hidden paths of the process and sampling drift functions. An alternative approach was given in [6], where the latent path was treated by an expectation maximization approach with the hidden process approximated by a linear stochastic differential equation. This seems to work faster, but the quality of the linear approximation deteriorates for larger time lags leading to an asymptotic bias in the inference of the drift.

The goal of this paper is to construct classes of nontrivial SDE models for which a computationally efficient nonparametric estimation of the drift is possible using the empirical distribution alone. Our method is based on a variational formulation of the stationary Fokker–Planck equation which gives a unique solution to the drift under certain conditions. These generalize the potential condition of thermal equilibrium. The functional to be minimized is an expectation over the stationary density. By replacing this density with the empirical one, i.e. with an unordered data sample and by minimizing the empirical functional (e.g. in a parametric family of potentials), one can get an estimator of the drift. The method can be generalized to a nonparametric estimate if the empirical functional is regularized with a kernel based penalty term. Our approach is not based on an explicit representation of the drift in terms of the density. Hence, it does not use a direct estimator of the density such as the KDE. Thus it is not expected to suffer from the bad convergence properties of the KDE in higher dimensions.

The paper is organized as follows. The second chapter introduces the variational formulation of the Fokker–Planck equation from which the drift can be derived by minimization of a functional. The third chapter shows how a regularized empirical approximation of the functional leads to a nonparametric estimate. The fourth chapter presents examples of this estimator for the class of Langevin equations for which the extra conditions lead to only mild restrictions. The fifth chapter explains how the method can be extended to other types of noise, such as jump processes. We conclude with a discussion and possible extensions of the method in chapter six.

2. A variational formulation for the Fokker–Planck equation

We consider stochastic differential equations for the dynamics of a $d$-dimensional diffusion process $Z_t \in \mathbb{R}^d$ given by

$$
\text{d}Z_t = g(Z_t)\text{d}t + \sigma(Z_t)\text{d}W_t. 
$$

The drift function $g(\cdot) \in \mathbb{R}^d$ represents the deterministic part of the driving force and $W$ is a $k$-dimensional ($k \leq d$) vector of independent Wiener processes acting as a white
noise source. The strength of the noise is determined by the state dependent \( d \times k \) dimensional noise matrix \( \sigma(Z) \).

Suppose that we are given the stationary density \( p(z) \) of the process. How can we determine the drift \( g \) which corresponds to this density? To give a partial answer to this question, we assume that \( \sigma(z) \) is known and the drift splits into two parts \( g(z) = r(z) + f(z) \), where \( r(z) \) is a known part and we try to compute \( f \). Of course, in the multivariate case there is not enough information to reconstruct \( f \) uniquely. However, we may search for a minimal solution which minimizes a quadratic functional

\[
\frac{1}{2} \int p(z) f(z) \cdot A^{-1}(z)f(z) \, dz
\]

for a given positive definite matrix \( A(z) \). Introducing a Lagrange multiplier function \( \psi(z) \) for the condition that the density \( p \) fulfills the stationary Fokker–Planck equation with drift \( g \), we can derive the minimal \( f \) from the variation of the Lagrange-functional

\[
\frac{1}{2} \int f(z) \cdot A^{-1}(z)f(z) \, dz - \int \psi(z) \{\mathcal{L} p(z) - \nabla \cdot (f(z)p(z))\} \, dz
\]

where the Fokker–Planck operator \( \mathcal{L} \) corresponding to the known drift \( r(z) \) is given by

\[
\mathcal{L} p(z) = -\nabla \cdot (r(z)p(z)) + \frac{1}{2} \text{tr}[\nabla \nabla^\top (D(z)p(z))]
\]

with \( D(z) \doteq \sigma(z)\sigma(z)^\top \). Variation of (3) with respect to \( f \) yields \( f(z) = A(z)\nabla \psi(z) \). Inserting this solution back into (3) shows that the unknown potential \( \psi \) can be derived from the minimization of the functional

\[
\varepsilon[\psi] = \int \left\{ \frac{1}{2} \nabla \psi(z) \cdot A(z) \nabla \psi(z) + \mathcal{L}^* \psi(z) \right\} p(z) \, dz
\]

where \( \mathcal{L}^* \) is the adjoint operator of \( \mathcal{L} \), (4) which fulfills

\[
\int \psi(z) \mathcal{L} p(z) \, dz = \int p(z) \mathcal{L}^* \psi(z) \, dz
\]

and is given by

\[
\mathcal{L}^* \psi(z) = r(z) \cdot \nabla \psi(z) + \frac{1}{2} \text{tr}[D(z)\nabla \nabla^\top \psi(z)].
\]

In fact, a direct minimization of (5) with respect to \( \psi \) yields

\[
\mathcal{L}[\psi] p(z) \doteq \mathcal{L} p(z) - \nabla \cdot (A(z)\nabla \psi(z)p(z)) = 0,
\]

which is the stationary Fokker–Planck equation corresponding to the density \( p(z) \) and the drift \( g(z) = r(z) + A(z)\nabla \psi \). Hence, if the drift is actually of this form, then the minimization of (5) will give us the desired unique result. For the special case \( D = A = I \) and \( r = 0 \) the functional (5) was introduced in the field of machine learning as a score-function for estimating \( \ln p(x) \) up to a normalization constant [7]. This case corresponds to an SDE for thermal equilibrium, where the drift \( f(z) = \nabla \psi(z) \) is the gradient of a potential \( \psi \) and the stationary density fulfills \( p(z) \propto e^{2\psi(z)} \).
The matrix $A$ introduces an extra degree of freedom which could be chosen using prior knowledge of the SDE model. Of special interest are models with $A = D$. As we show in appendix A, for such models we have asymptotically $\varepsilon[\psi] \approx \frac{1}{T} \varepsilon_{\text{ML}}[\psi]$, where $\varepsilon_{\text{ML}}$ is the negative log-likelihood, when the process $Z_t$ was sampled continuously in time over a large time $T$. Hence if observations are dense in time a minimization of $\varepsilon[\psi]$ using the empirical distribution should become asymptotically equivalent to maximum likelihood estimation. The discussion in appendix A also gives another interpretation of the cost function (5) for $A = D$. The drift given by the minima of (5) leads to the process with path measure that is closest in relative entropy (Kullback–Leibler divergence) rate $[^8, ^9]$ to the path measure of the process with drift $r(z)$, when $p(z)$ is given.

3. Minimizing the empirical functional

Our goal is to estimate $\psi$ from data by replacing the average over the stationary density $p(z)$ in the functional (5) by the empirical distribution

$$\hat{p}(z) = \frac{1}{n} \sum_{i=1}^{n} \delta(z - z_i)$$

(9)

where $z_1, \ldots, z_n$ is a random, ergodic sample drawn from this density. An obvious possibility to construct estimators is to work with a parametric representation

$$\psi_w(z) = \sum_{k=1}^{K} w_k \phi_k(x)$$

(10)

where the $\phi_k$ are a set of given basis functions. The weights $w_k$ could be determined by minimization of the empirical version of the functional (5)

$$\varepsilon_{\text{emp}}[\psi_w] = \frac{1}{2} \sum_{i=1}^{n} \left\{ \nabla \psi_w(z_i) \cdot A(z_i) \nabla \psi_w(z_i) + \mathcal{L}^* \psi_w(z_i) \right\}$$

(11)

which is a quadratic form in the $w_k$ and can thus be performed in closed form. We are however interested in the case where a representation in terms of a finite set of basis functions is not rich enough to represent $\psi$. Thus we will resort to a more general, non-parametric representation allowing for an infinite set of functions $\phi_k$. Since one has only a finite number of data $z_i$ for estimation, the estimator needs to be regularized by introducing an extra penalty term. This will be chosen as a quadratic form $\frac{1}{2} \sum_k w_k^2 / \lambda_k$, where the $\lambda_k$ are hyper-parameters. This penalty can also be viewed from a pseudo-Bayesian perspective where $\exp\{-C \varepsilon_{\text{emp}}[\psi_w]\}$ is interpreted as a likelihood and $\exp\{-\frac{1}{2} \sum_k w_k^2 / \lambda_k\}$ as a Gaussian prior distribution over parameters $w_k$. $C$ can be chosen to give different weight to the data and to the penalty. In this interpretation, (10) could be understood as a Gaussian process model $[^10]$ for the function $\psi$. As shown in appendix A, the likelihood interpretation becomes correct asymptotically for densely sampled observations if $A = D$ for which we would set $C \approx T/n$ being the time between observations.

Motivated by the Gaussian process point of view we will introduce the kernel trick into our formalism avoiding an explicit specification of $\phi_k$ and $\lambda_k$ and assume instead
that these are defined implicitly as orthonormal eigenfunctions and eigenvalues of a positive definite kernel function $K(z, z')$ via

$$K(z, z') = \sum_k \lambda_k \phi_k(z) \phi_k(z').$$

This can be viewed as the covariance kernel of a Gaussian process prior distribution for functions $\psi(z)$ [10]. Kernels can be adapted to the prior knowledge which is available about the function $\psi$. This might include a known periodicity of the function, the length scale of its typical variation, or the fact that $\psi$ is a polynomial of a given order. Appendix B gives a short summary of the kernels used in our experiments.

In the kernel approach the regularized functional can be written as

$$C \sum_{i=1}^n \left\{ \frac{1}{2} \nabla \psi(z_i) \cdot A(z_i) \nabla \psi(z_i) + \mathcal{L}^* \psi(z_i) \right\}$$

$$+ \frac{1}{2} \int \int \psi(z) K^{-1}(z, z') \psi(z') \, dz \, dz', \quad (12)$$

where $K^{-1}(z, z')$ is the formal inverse of the kernel operator. One can also show that the penalty term on the right hand side equals the so-called reproducing kernel Hilbert space (RKHS) norm of $\psi$ defined by the kernel $K$. Using this formalism a nonparametric extension of the score function approach for estimating $\ln p(x)$ was introduced in [2].

Our discussion shows that there are two ways for computing the estimator of $\psi(z)$ explicitly, both leading to the same result. The first one is based on setting the variational derivative of (12) equal to zero and the second uses the formalism of Gaussian process inference [10] with $\varepsilon_{\text{emp}}[\psi]$ interpreted as negative log-likelihood. Due to the non-standard form of the functional defined in (11) the regression algorithm has to be adapted as shown in appendix C. We will next give a derivation of the estimator using the first method. Performing the variation with respect to $\psi$ yields

$$Cn \mathcal{L}[\psi] \hat{p}(z) + \int K^{-1}(z, z') \psi(z') \, dz' = 0,$$

where $\mathcal{L}[\psi]$ was defined in (8). Multiplying both sides of this equation with the operator $K$ we get

$$\psi(z) + C \sum_{j=1}^n \mathcal{L}_{z'j}^*[\psi] K(z, z')_{z'=z_j} = 0 \quad (13)$$

where the adjoint operator acts on functions $h$ as

$$\mathcal{L}_{z'}^*[\psi] h(z') = \langle r(z') + A(z') \nabla \psi(z') \rangle \nabla h(z') + \frac{1}{2} \text{tr}[D(z') \nabla \nabla^\top h(z')]. \quad (14)$$

We can understand (13) as a regularized version of the equation

$$\int p(z') \mathcal{L}_{z'}^*[\psi] h(z') \, dz' = \int h(z') \mathcal{L}_{z}^*[\psi] p(z') \, dz' = 0 \quad (15)$$

applied to the family of kernel functions $h_{\delta}(z') = K(z, z')$ when the stationary density $p$ is replaced by its empirical approximation $\hat{p}$ (9).
Equations (13) and (14) show that if $\nabla \psi(z)$ is known at all sample points $z = z_i$, we can evaluate the second term and get then the function $\psi(z)$ for all $z$. The gradient of $\psi$ at the data points is computed by taking the gradient of (13) and setting $z = z_i$. This yields the set of linear equations

$$\nabla \psi(z_i) + \sum_{j=1}^{n} C_j [\psi] \nabla z K(z, z')_{z = z_i, z' = z_j} = 0$$

for the $d \times n$ unknowns $\nabla \psi(z_i)$ which can be plugged into (13) to obtain the explicit result for the estimator as shown in appendix D.

### 4. Application: Langevin dynamics

To show that the condition $f(z) = A(z) \nabla \psi(z)$ includes classes of non-trivial non-equilibrium models, we will specialize to second order (Langevin-type) SDEs which appear naturally when systems of classical mechanics are driven by deterministic and random forces. The time evolution is described in terms of (generalized) coordinates and velocities $X, V \in R^d$ as

$$dX_t = V_t dt, \quad dV_t = g_v(X_t, V_t) dt + \sigma_v(X_t, V_t) dW_t.$$  

(17)

The noise acts only on the acceleration and the drift for this model is of the form $(g_x, g_v)$ where $g_v = v$ is known. Hence, we may choose the matrix elements of $A$ in (5) to be zero except for the sub-matrix $A_{vv}$ which we will set to the unit matrix $A_{vv} = I$ for simplicity. Thus the reduced drift vector for the velocity is assumed to be of the form $g_v(x, v) = r_v(x, y) + \nabla_v \psi(x, v)$ and the functional (5) becomes

$$\varepsilon[\psi] = \int p(x, v) \left\{ \mathcal{L}^* \psi(x, v) + \frac{1}{2} (\nabla_v \psi(x, v))^2 \right\} dx \, dv$$  

(18)

where

$$\mathcal{L}^* \psi(x, v) = \left( v \cdot \nabla_x + r_v(x, v) \cdot \nabla_v + \frac{1}{2} \text{tr}(D_v(x, v) \nabla_v^\top \nabla_v) \right) \psi(x, v),$$  

(19)

with $D_v = \sigma_v \sigma_v^\top$. The integrability condition on the unknown part of the drift $f_v(x, v) = \nabla_v \psi(x, v)$ restricts the velocity dependency rather than its coordinate dependency. We will specialize on dynamical systems with a position dependent external force $f(x)$ and a friction term which is linear in the velocity. This is given by

$$f_v(x, v) = f(x) - \Lambda v = \nabla_v \left\{ v \cdot f(x) - \frac{1}{2} v \cdot \Lambda v \right\},$$  

(20)

with a positive diagonal matrix $\Lambda$ and an arbitrary (e.g. non-conservative) vector field $f(x)$. In order to learn $f_v(x, v)$ from $n$ pairs of observations $(x_i, v_i)$, a suitable kernel function has

\[\text{doi:10.1088/1742-5468/2016/08/083404}\]
to be chosen. If no prior assumptions about the drift function are made, we use a RBF kernel for $\psi(x, v)$. Alternatively, if the specific form of (20) as an additive combination of a coordinate term and a (parametric) velocity term is taken into account, we use a kernel for $f_v(x, v)$ which is a product of a first order polynomial kernel in $v$ and a periodic (B.2) or polynomial (B.3) kernel in $x$. The latter two kernels will be referred to as product kernels. In order to optimize hyperparameters of the kernel functions (length scale parameter of RBF (B.1) and periodic (B.2) kernels, respectively), we use 2-fold cross-validation. This method consists of randomly dividing the observations into two subsets of equal size, where on each subset an estimator is trained and then tested on the remaining subset. As cost function we use (18) and the minimization is done by a conjugate gradient method. We found that the constant $C$ in (13) did not have a strong influence on the accuracy of the estimator and have used $C = 1$ throughout the experiments.

We illustrate this method for the case of a bistable system with two locally stable equilibria (double well model) which corresponds to the drift $f(x) = -4(x - x^3)$. We simulated the process with diffusion $\sigma = 1.3$ and friction constant $\lambda = 1.1$ and generated a data set of size $n = 3000$ observations with time lag $\tau = 0.5$. The estimator uses a RBF kernel for $\psi(x, v)$ with a common length scale parameter $l = 1.29$ determined by cross validation.

The left of figure 1 shows the estimate $\hat{f}(x, v)$ as a function of the coordinate $x$ for three different values of the velocity $v$. We have also studied how fast our estimator $\hat{f}(x, v)$ converges to the true drift. For that purpose we have generated several training sets with different number of observations and a test set from the double well model with diffusion constant $\sigma = 1$ and friction constant $\lambda = 1.1$. Learning of $f_v(x, v)$ then used a RBF kernel for $\psi(x, v)$ with a fixed length scale $l = 1.25$ for all data sets in order to improve comparability. Specifically, we applied our method to each of the 370 training sets (10 training sets each on a range of 37 different sizes ranging from $n = 600$ to $n = 14\,000$) and computed the approximate mean squared error (MSE).
Variational estimation of the drift for stochastic differential equations from the empirical density

\[ \int p(z)(\hat{f}(z) - f(z))^2 dz \approx \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}}(\hat{f}(z_i) - f(z_i))^2 \]  

(21)

of the corresponding estimator. Here \( \hat{f}(z) \) denotes the estimated drift and \( f(z) \) the true drift value, on a fixed test set consisting of \( n_{\text{test}} = 1000 \) observations with time lag \( \tau = 4 \). We repeated this procedure three times, where each time we varied the time lag \( \tau \) between the observations in the training sets to \( \tau = 0.4, 1 \) and \( \tau = 2 \), respectively, but kept the same test set. The results are shown as a log–log plot in the right figure 1. Here, each MSE value for a given observation number \( n \) represents the mean MSE over the 10 corresponding training sets. For the data sets with smaller \( (\tau = 0.4) \) and medium \( (\tau = 1) \) time lag, the learning curves turn out to be slightly flatter at the beginning. This is due to the fact that smaller training sets sometimes do not cover the whole range of the double well potential, which in turn leads to higher MSE values on the test sets. With bigger training sets these effects vanish. We therefore decided to compute the regression lines for each of the learning curves only on the observation sets of size \( n = 3500 \) and bigger. Comparing the regression lines, one can see that all decays lie between \( \propto n^{-1/2} \) and \( \propto n^{-1} \), which is consistent with the results in [2] reported for score matching.

So far we have assumed that the diffusion \( D_v \) is known. If on the other hand, the friction parameter \( \Lambda \) is known, i.e. \( r_v = -\Lambda v \), then \( \mathcal{L} \psi(x, v) = \mathcal{L}(v \cdot f(x)) \) is independent of the diffusion term \( D_v(x, v) \) and we can estimate \( f(x) \) without knowing the diffusion. We demonstrate this estimate on a two dimensional Langevin model with a nonconservative drift with components \( f^{(1)}(x) = x^{(1)}(1 - (x^{(1)})^2 - (x^{(2)})^2) - x^{(1)} \) and \( f^{(2)}(x) = x^{(2)}(1 - (x^{(1)})^2 - (x^{(2)})^2) - x^{(1)} \), and friction constants \( \lambda^{(1,2)} = 0.8 \) (component indices denoted by superscripts). The \( n = 2000 \) four-dimensional data (position and velocity) observations were generated with constant diffusion constants \( D_v = 9 I \) and time lag \( \tau = 0.5 \). The drift vector was estimated by penalizing each component of \( f(x) \) independently using a product polynomial kernel (B.3) of order \( p = 4 \), assuming that the true drift is at most a polynomial of order 4. The results are shown in figure 2. Further discussion on the influence of kernel function and hyperparameters on the estimation results for both this model as well as the double well model can be found in appendix E.
To show that this approach also works for state dependent diffusion (multiplicative noise), we consider a model of a pendulum on which gravitation and friction act as drift terms given by $f(x) = a \sin x$ and $-\lambda v$ in (20). Here $x$ is the angle relative to the upward position and $v$ the angular velocity. The pendulum is mounted on a cart that is accelerated in the horizontal direction by a white noise force. This leads to an additional stochastic angular acceleration with diffusion $D_v \cos v^2 \sigma$. For the simulation, we used a data set of $n = 2000$ observations with time lag $\tau = 0.25$, diffusion $\sigma = 1$, $a = 9.81$ and $\lambda = 0.05$. As kernel function we chose a product kernel with periodic kernel (B.2) in $x$ and length scale $l_{\text{per}} = 1.2$ determined by cross-validation. One can clearly see from the left of figure 3 that most of the time the pendulum stays in the downward position and only occasionally crosses the upright position (corresponding to $x = 0$). Nevertheless, the right panel of figure 3 shows that regularization with the periodic kernel leads to an excellent estimation of the drift for all values of $x$.

One might wonder if the estimation of the drift could have also be achieved by a kernel density approach. While for a general model (1) with drift $f(z) = r(z) + A(z)\nabla \psi(z)$ there does not seem to be a way of expressing $f(z)$ in terms of $p(z)$ in closed form, a somewhat complicated expression can be given for Langevin equations with drifts of the form (20). Multiplying the Fokker–Planck equation for the process (17) with a component $v^i$ of the velocity vector and integrating over $v$, one obtains the following explicit representation for the drift

$$f^{(i)}(x) = \frac{d}{j=1} \frac{\partial E[v^{(i)} v^{(j)}]}{\partial x^{(j)}} + \frac{d}{j=1} E[v^{(i)} v^{(j)}] \frac{\partial \ln p(x)}{\partial x^{(j)}} - E[r^{(i)}|x],$$

where components of vectors are denoted by superscripts and $E[\cdot|x]$ denote conditional expectations. This shows that, in general, one would need not only a KDE for estimating $p(x)$ but also a nonparametric regression method for estimating the conditional expectations as a function of $x$. Of course, for the equilibrium case where $f(x) = \nabla \phi$ and $D_v \propto \Sigma$, $r = -\Lambda v$ where $\Lambda$ and $\Sigma$ are diagonal matrices which satisfy $2\lambda_\Sigma \sigma^2 = \beta$, which is the inverse temperature, (22) simplifies because one has $E[v^{(i)} v^{(j)}] = \frac{1}{2\beta} \delta_{ij}$ and $E[r^{(i)}|x] = 0$. For this case, the velocity samples $v_i$ are not needed.

1 This toy model is known as cart and pole, frequently used to test control methods for stabilizing the pendulum in the upright position [11].
5. Generalization to other noise processes

For applications where noise is used as a part of an external control signal acting on a dynamical system, the assumption of white noise is not realistic, because its non-decaying high frequency components are, in practice, filtered out in the control circuit. Hence, we would like to include other processes, e.g. colored noise or a noise source with a finite state space. To adapt our method to this situation, we replace the white noise $\sigma_t(X,V)\,dW$ in (17) by $U_t\,dt$ where $U_t$ is a Markov process, which is included in the state variable $Z = (X, V, U)$ and observed at the same times as $X$ and $V$. Our formalism does not change when $U_t$ is a diffusion process itself, because the entire system can be described by a Fokker–Planck equation for the density $p(x, v, u)$. But it is also possible to include other Markov processes such as jump processes as noise sources. We just have to replace the Fokker–Planck equation by the appropriate Master equation in the definition (4). We will illustrate this idea for $U_t$ being a random telegraph process [12] that switches with equal rates $\gamma$ between $u = \pm 1$. We study a one-dimensional system $x, v, u \in \mathbb{R}$ with drift given by $g_\nu(x, v) = -\lambda v + f(x)$ with a known friction constant $\lambda$. The Master equation [12] for the stationary density is given by

$$
- \partial_x v p(x, v, u) - \partial_v [f(x) - \lambda v + u] p(x, v, u) + \gamma (p(x, v, -u) - p(x, v, u)) = 0.
$$

The drift $f$ can be estimated by the minimization of the functional (18), when we use the adjoint operator given by

$$
\mathcal{L}^* \psi(x, v, u) = \{ v \partial_x + (u - \lambda v) \partial_v \} \psi(x, v, u) + \gamma (\psi(x, v, -1) - \psi(x, v, 1)).
$$

The parameterization $\psi(x, v) = uf(x)$ leads to the functional

$$
\varepsilon[f] = \frac{1}{2} \sum_{u=\pm 1} \int p(x, v, u) \{ f^2(x) + 2f'(x)v^2 + 2f(x)(u - \lambda v) \} \, dx \, dv
$$

(25)

to be minimized with respect to $f$. Experiments (not included here) for the cart and pole model show that this method achieves similar performance as the one shown in figure 3.

6. Discussion and outlook

We have presented a method for a nonparametric estimation of the drift of certain types of stochastic differential equations from the empirical density alone. The method is not related to kernel density estimation and can be applied to cases where the use of a KDE would not be simple or impossible. The method should be of interest for situations where external noise is used to explore the state space of a mechanical system in order to learn the deterministic part of the forces which can be used to later control the system. On the other hand, one might use our variational approach for solving a specific type of stochastic control problem [9]: we would be able to compute a state dependent control $f(z)$ which has to be added to the known drift $r(z)$ of a system such that a new desired stationary density $p(z)$ will be reached.
In future work we will explore different possibilities to increase the applicability of our methods. We will investigate carefully the role of the matrix valued model parameter $A$ in (4) but also try to generalize the functional (5) by including other types of operators. E.g. a second derivative of a convex potential $\psi$ could be used to estimate the diffusion for known drift.

It will be important to develop a theoretical approach for studying the convergence properties of the drift estimator. For certain special cases preliminary results are already available. If we restrict ourselves to parametric representations (10), the resulting system of linear equations can be understood as sample approximations to equation (15) applied to functions $h(z) = \phi_k(z)$. The resulting parametric estimator belongs to the class of so-called generalized methods of moments (GMM) estimators which have been analyzed by Hansen in [13] for data sampled from stationary ergodic processes. If the true drift can be expressed in the same parametric form (10) with a finite dimensional vector of parameters, the results in [13] guarantee almost sure convergence of estimated parameters to the true ones. Furthermore, the fluctuations of estimated parameters are asymptotically Gaussian with a covariance matrix which is of the order $O(1 / n)$ as $n \to \infty$. We expect that these asymptotic results could be used to test the hypothesis of a finite dimensional model (10) by applying sample approximations of the expectations in (15) to other appropriate functions $h(z)$. These should be close to zero if the model is correct. Unfortunately, the work [13] cannot be applied to the nonparametric kernel based estimators (13). Although the nonparametric estimators are determined by a finite set of linear equations (16), the dimensionality of this set equals $n$ and increases with the number of observations. An exception are polynomial kernels, which could also be cast into a form (10) with a fixed number of parameters.

Two other theoretical investigations relate to special cases of the nonparametric kernel estimators. As mentioned before, the equilibrium case $A = D \propto I$ with $r = 0$ corresponds to the score matching approach to density estimation for which Sriperumbudur et al [2] have analyzed the convergence properties. Since these authors consider data points which are independently sampled from a density, their results would apply to the asymptotic limit where observations from an SDE are sampled with a very small sampling rate. Their results on convergence for the so-called Fisher divergence apply to the convergence of the mean squared error of the estimated drift. If the true potential lies in the reproducing kernel Hilbert space of the kernel used for estimation, an asymptotic power law decay of the error with a decay constant depending on the smoothness of the potential can be shown. Similar results on power law decays for estimation errors (for 1-dimensional SDE with state independent diffusion) have been obtained by Pokern et al [14] for the opposite case of densely sampled data where the nonparametric estimator corresponds to a proper Bayesian estimator with a Gaussian process prior. Such power law decays are also observed in our simulations as shown in figure 1.

Another question is how to lift the restriction that all coordinates of the random state vector $z$ need to be observed jointly. For the Langevin type equations with a drift of the form (20) it would be interesting to see if the method could be generalized to estimating a drift $f(x)$ based on coordinate observations $x_i$ alone. For potentials of the type $v \cdot f(x)$ one can integrate over the velocities in (18) to obtain a functional which depends on $f(x)$, $p(x)$ and conditional moments of the velocities (see (22)). For an equilibrium problem, these conditional moments are constant and known and the velocity
observations are not needed. For the general case one could use the temporal order of coordinate observations to obtain a preliminary approximation to the unobserved velocities. An initial estimate of the drift \( f(x) \) could then be derived by a minimization of the functional (18). This estimate could be used to create new velocity samples and estimates for conditional velocity moments by performing forward sampling of the SDE (17) and the method could be iterated. Preliminary experiments using this iterative procedure are promising, but we do not yet have conditions on the convergence of such a procedure.

Acknowledgments

This work was supported by the European Community’s Seventh Framework Programme (FP7, 2007–2013) under the grant agreement 270327 (CompLACS).

Appendix A. Likelihood for dense observations

We will derive the likelihood function for the drift \( g \) of an SDE assuming that we have access to a dense path \( Z_{0:T} \) of observations in a time window from \( t = 0 \) to \( t = T \). Discretizing time into small intervals of length \( \Delta t \) and using the fact that for \( \Delta t \to 0 \), the transition density of (1) becomes Gaussian, we find that the part of the negative log-likelihood (NLL) which depends on the drift function \( g \) can be approximated by

\[
- \ln p(Z_{0:T}|g) \approx \frac{1}{2} \sum_t \{||g(Z_t)||^2 \Delta t - 2\langle g(Z_t), (Z_{t+\Delta t} - Z_t) \rangle \} + \text{const.}
\]

where we have introduced the inner product \( \langle u, v \rangle \equiv u \cdot D^{-1}v \) and the corresponding squared norm \( ||u||^2 \equiv u \cdot D^{-1}u \). For \( \Delta t \to 0 \), the second sum becomes a Ito stochastic integral [12]. If the drift can be written as \( g = r + D\nabla \psi \) where \( r \) is a known function and we are only interested in estimating \( \psi \), we can transform the Ito integral into an ordinary time integral using Ito’s formula. The part of the NLL which contains \( \psi \) becomes

\[
\varepsilon_{\text{ML}}[\psi] = \frac{1}{2} \int_0^T \{\nabla \psi \cdot D \nabla \psi \, dt + 2r \cdot \nabla \psi \, dt - 2\nabla \psi \cdot \, dZ_t \}
\]

\[
= \frac{1}{2} \int_0^T \{\nabla \psi \cdot D \nabla \psi + 2r \cdot \nabla \psi + \text{tr}(D \nabla \nabla^\top \psi)\} dt - \psi(Z_T) + \psi(Z_0). \tag{A.1}
\]

We will now assume that for large \( T \), the process becomes stationary with density \( p(z) \). We can then replace the time integral by an integral over \( p \). For a mathematical rigorous treatment see e.g. [5]. Neglecting the contribution from the boundary terms in (A.1) for large \( T \) we arrive at

\[
\varepsilon_{\text{ML}}[\psi] \approx \frac{T}{2} \int \{\nabla \psi \cdot D \nabla \psi + 2r \cdot \nabla \psi + \text{tr}(D \nabla \nabla^\top \psi)\} p(z) \, dz. \tag{A.2}
\]

Using the Gaussian form of the transition density for short times \( \Delta t \) it can also be shown that the relative entropy or Kullback–Leibler (KL) divergence between the path probabilities for two diffusion processes with drifts \( g(z) \) and \( r(z) \), where \( g(z) = r(z) + f(z) \), is given by

\[
doi:10.1088/1742-5468/2016/08/083404\]
Variational estimation of the drift for stochastic differential equations from the empirical density

\[ D(p(Z_0, T | g) \| p(Z_0, T | r)) = \int_0^T dt \int p_t(z) f(z) \cdot D^{-1}(z) f(z) dz \]

assuming they have the same diffusion term \( D(z) \) and the same non-random initial state (see e.g. [8]). Here \( p_t(z) \) is the marginal density of the process with drift \( g \) at time \( t \). Hence, assuming that the process becomes stationary with density \( p(z) \), we get for the relative entropy rate

\[ \lim_{T \to \infty} \frac{1}{T} D(p(Z_0, T | g) \| p(Z_0, T | r)) = \int p(z) f(z) \cdot D^{-1}(z) f(z) dz. \quad (A.3) \]

A comparison with (3) shows that for \( A = D \) the minimization of (5) leads to a process with given stationary density that is closest to the process with drift \( r(z) \) in relative entropy. Hence, this may be understood as a generalized maximum entropy (minimum relative entropy) solution where the stationary density is given as a constraint.

**Appendix B. Kernel functions**

For the experiments we have used the following kernels:

- The radial basis function (RBF) kernel
  \[ K_{\text{RBF}}(x, y) = \exp\left( -\frac{(x - y)^\top (x - y)}{2l_{\text{RBF}}^2} \right). \quad (B.1) \]
  This has the length scale \( l_{\text{RBF}} \) as a hyper parameter and is used for estimating smooth functions.

- The (one-dimensional) periodic kernel
  \[ K_{\text{Per}}(x, y) = \exp\left( -\frac{2 \sin\left( \frac{x - y}{2} \right)^2}{l_{\text{Per}}^2} \right) \quad (B.2) \]
  is used for estimating smooth periodic functions.

- The polynomial kernel of degree \( p \)
  \[ K_{\text{Poly}}(x, y) = (1 + x^\top y)^p \quad (B.3) \]
  is used for estimating functions which are polynomials with degrees at most \( p \).

**Appendix C. Gaussian process inference**

Compared to standard Gaussian process (GP) regression tasks the form of the functional \( \varepsilon_{\text{emp}}[\psi] \) defined in (11) and used as log-likelihood is different in two important aspects:

\[ \text{doi:10.1088/1742-5468/2016/08/083404} \]
The (pseudo-)likelihood contains not only function values, but also derivatives of the unknown function $\psi$.

Both function values and derivatives can appear linearly in the log-likelihood without a corresponding quadratic term.

Here we show how to do Gaussian process inference efficiently in this case.

It is well known that GPs can be used to infer solutions of linear partial differential equations (PDE) and operator equations from discrete space-time observations $[15-17]$. It should be noted, however, that the present application of a GP to the Fokker–Planck PDE (8) is mathematically fairly different, and has, to our knowledge, so far not been discussed.

If the likelihood contains differential operators, a consistent prior over the function $\psi(\cdot)$ to estimate and its derivatives has to be used. The covariance between two function values can be specified as

$$\text{Cov}(\psi(x), \psi(y)) = K(x, y)$$  \hspace{1cm} (C.1)

by choosing a suitable kernel function $K(x, y)$. Then the correlations between derivatives of $\psi(\cdot)$ are already determined by the corresponding derivatives of $K(\cdot, \cdot)$:

$$\text{Cov}(\nabla \psi(x), \nabla \psi(y)) = \nabla_x K(x, y),$$

$$\text{Cov}(\nabla^2 \psi(x), \nabla^2 \psi(y)) = \nabla_x \nabla_y K(x, y),$$

$$\text{Cov}(\nabla \psi(x), \nabla^2 \psi(y)) = \nabla_x \nabla_y^2 K(x, y).$$

In order to simplify the notation we collect all variables of interest, function values and derivatives at observations and test points, in the vector $\Psi$ and denote the corresponding covariance matrix as $K$. Then the effective prior is, in fact, a multivariate Gaussian distribution

$$p(\Psi) \propto |K|^{-1/2} \exp \left(-\frac{1}{2} \Psi^T K^{-1} \Psi \right)$$

with mean zero, where $|K| = \det(K)$.

As the (pseudo-)log-likelihood in our approach contains some components of $\Psi$ only linearly, it cannot be written as the standard Gaussian likelihood

$$p(\Psi | \psi) \propto |S|^{-1/2} \exp \left(-\frac{1}{2} (P\Psi - a)^T S^{-1} (P\Psi - a) \right)$$

for direct observations $\Psi_o \approx a$, where the projector $P$ selects the observed variables and $S$ describes the observation noise. Instead we can represent it as

$$p(\Psi | \psi) \propto |S|^{-1/2} \exp \left(-\frac{1}{2} (P\Psi)^T S^{-1} (P\Psi) + b^T R \Psi \right)$$  \hspace{1cm} (C.2)

with two projector matrices $P$ and $R$ selecting the components of $\Psi$ occurring in quadratic and linear terms of the log-likelihood. If we additionally define $Q$ as the projector onto the test points $\Psi_*$ we want to estimate, the moments of the predictive distribution $p(\Psi_* | \Psi_o) = \mathcal{N}(\Theta; \mu, \Sigma)$ are given as

$$p(\Psi_* | \Psi_o) = \mathcal{N}(\Theta; \mu, \Sigma)$$

with $\Theta = Q \Psi_o$, $\mu = Q \mu$, and $\Sigma = Q \Sigma$. The moments are given by

$$\mu = Q \mu, \quad \Sigma = Q \Sigma.$$
Variational estimation of the drift for stochastic differential equations from the empirical density

\[ \Sigma = Q(K^{-1} + P^T S^{-1} P)^{-1} Q^T, \]  
(C.3)

\[ \mu = Q(K^{-1} + P^T S^{-1} P)^{-1} R^T b. \]  
(C.4)

Using the matrix inversion lemma, these equations can be simplified to

\[ \Sigma = QKQ^T - QKP^T (PKP^T + S)^{-1} PKQ^T \]  
(C.5)

\[ \mu = Q[K - KP^T (PKP^T + S)^{-1} PK] R^T \]  
(C.6)

\[ = QKR^T b - QKP^T (PKP^T + S)^{-1} PKR^T b, \]  
(C.7)

which shows that only some parts of the full kernel matrix \( K \) are needed in order to calculate the estimates.

Appendix D. Explicit form of the estimator

Here we show how to obtain the explicit form of the estimator using the equations (13) and (16). In order to simplify the notation we combine the gradients \( \nabla \psi(z_i) \) and the known drift vectors \( r(z) \) at all observations into \( n \cdot d \)-dimensional vectors

\[ g = (\nabla \psi(z_1), \nabla \psi(z_2), \ldots, \nabla \psi(z_n)), \]
\[ r = (r(z_1), r(z_2), \ldots, r(z_n)), \]

and define the corresponding kernel matrix \( G \) as well as the combined block-diagonal weight matrix \( A \) consisting of submatrices

\[ G_{ij} = \nabla z_i \nabla z_j' K(z, z')|_{z=z_i, z'=z_j}, \]
\[ A_{ij} = \delta_{ij} A(z_j). \]

Then (16) can be written as

\[ (I + CGA)g + C(Gr + \nabla y) = 0, \]  
(D.1)

where \( I \) is the identity matrix and

\[ \nabla y = (\nabla y(z_1), \nabla y(z_2), \ldots, \nabla y(z_n)) \] with

\[ y(z) = \frac{1}{2} \sum_{j=1}^{n} \text{tr}[D(z') \nabla z_j \nabla z_j' K(z, z')]|_{z'=z_j}. \]

Rewriting (13) in a similar way yields

\[ \psi(z) + C(k(z)Ag + k(z)r + y(z)) = 0 \]  
(D.2)

with

\[ k(z) = (\nabla z_j' K(z, z')|_{z'=z_1}, \nabla z_j' K(z, z')|_{z'=z_2}, \ldots, \nabla z_j' K(z, z')|_{z'=z_n}). \]

Solving (D.1) for \( g \) results in

\[ g = -(I + CGA)^{-1} C(Gr + \nabla y). \]  
(D.3)
Together with (D.2) finally the explicit form of the estimator
\[ \psi(z) = Ck(z)A(I + CGA)^{-1}C(Gr + \nabla y) - C(k(z)r + y(z)) \] (D.4)
is obtained. We can now compare this result with Gaussian process regression using
\[ \exp(-C\varepsilon_{\text{emp}}[\psi]) = \exp\left(-\frac{1}{2}Cg \cdot Ag - Cr \cdot g - \frac{1}{2}Cd \cdot \nabla g\right) \] (D.5)
as (pseudo-)likelihood and the approach described in appendix C. Here \( \nabla g \) denotes the vector of second derivatives of \( \psi \) at all observations and \( d \) the vector of corresponding elements in the diffusion matrix \( D \), so that
\[ d \cdot \nabla g = \sum_{j=1}^{n} \text{tr}[D(z')\nabla z'\nabla z'\psi(z')]|_{z'=z} \cdot \]
Comparing (D.5) and (C.2) we clearly identify the parts
\[ PP = g, \quad S^{-1} = CA, \quad RR = (g, \nabla g), \quad b = (-Cr, -C/2d), \]
of these equations and compute the terms
\[ PKP^T = G, \quad PKR^Tb = -C(Gr + \nabla y), \quad QKP^T = k(z), \quad QKR^Tb = -C(k(z)r + y(z)) \]
in (C.7). Then the posterior mean
\[ \mu = k(z)((CA)^{-1} + G)^{-1}C(Gr + \nabla y) - C(k(z)r + y(z)) \] (D.6)
of the Gaussian process is equal to our estimator \( \psi(z) \) as found in (D.4).

**Appendix E. Influence of kernel functions on estimation**

Here, we will look at how the choice of kernel function and its hyperparameters influence the estimation results. We discuss the effects on the double well model, which was already introduced in section 4 of the paper. The process was always simulated with diffusion \( \sigma = 1.3 \), friction constant \( \lambda = 1.1 \) and time lag \( \tau = 1 \).

In order to show the influence of prior knowledge on the estimation both for smaller and for larger data sets, we generated a path with \( n = 300 \) observations and a path with \( n = 4000 \) observations. For each data set, we trained an estimator with RBF kernel for \( \psi(x, v) \) and an estimator with a product kernel for \( f(x, v) \) consisting of a linear kernel in \( v \) and a polynomial kernel of order \( p = 4 \) in \( x \), respectively. Notice that the latter kernel exactly captures the structure of the model, since we in fact have a linear friction and a drift function \( f(x) \) of order \( p = 3 \). The results are shown in figure E1. One can see that for the smaller data set the inclusion of prior knowledge leads to a significant improvement in the estimation quality. For the bigger data set, both estimators largely agree in the range of the drift function where a large amount of observations are available. Differences occur at the low-data ranges on the edges of the drift function, where the product kernel follows the true drift, whereas the RBF kernel prior influence slowly pushes the estimation back to zero.

\[ \text{doi:10.1088/1742-5468/2016/08/083404} \]
Next, we take a look at the effects of hyperparameters of the kernel function. We generated a data set with \( n = 2000 \) observations from the double well model with the parameters given above and trained three estimators with a RBF kernel function, one with the optimal length scale \( l = 1.19 \) as determined by 2-fold cross-validation, and additionally one with \( l = 0.7 \) and one with \( l = 1.5 \). The results are shown in the left figure E2. One can see that the length scale does have a noticeable influence on the estimation results. Still, the estimator is relatively robust with respect to changes in the length scale, as the estimation still works reasonably well for a range of choices around the optimal value. In the case of the polynomial kernel, it can be seen from the right figure E2 that the estimator with a polynomial kernel of order \( p = 6 \) in \( x \) is able to infer the true order of the drift function in the high-data area of the data set. In regions where few observations are available, however, the estimation error increases significantly. While using a polynomial of a high order \( p \) yields a slightly less accurate estimator, one has to be careful to avoid using a kernel of insufficient complexity for

\[ f(x, v) \]

Next, we take a look at the effects of hyperparameters of the kernel function. We generated a data set with \( n = 2000 \) observations from the double well model with the parameters given above and trained three estimators with a RBF kernel function, one with the optimal length scale \( l = 1.19 \) as determined by 2-fold cross-validation, and additionally one with \( l = 0.7 \) and one with \( l = 1.5 \). The results are shown in the left figure E2. One can see that the length scale does have a noticeable influence on the estimation results. Still, the estimator is relatively robust with respect to changes in the length scale, as the estimation still works reasonably well for a range of choices around the optimal value. In the case of the polynomial kernel, it can be seen from the right figure E2 that the estimator with a polynomial kernel of order \( p = 6 \) in \( x \) is able to infer the true order of the drift function in the high-data area of the data set. In regions where few observations are available, however, the estimation error increases significantly. While using a polynomial of a high order \( p \) yields a slightly less accurate estimator, one has to be careful to avoid using a kernel of insufficient complexity for

\[ f(x, v) \]
the particular problem at hand. In our double well case, using a polynomial kernel of order \( p = 2 \) in \( x \) to learn a function of order \( p = 3 \) gives poor results, as the estimation denoted by the red line in the right figure E2 shows. The inadequacy of an estimator using a polynomial of order \( p = 2 \) can also be seen by comparing the value of the cost function (18) with the values for estimator with \( p = 3 \) and higher orders.

Finally, we discuss the influence of the kernel function for the model with nonconservative drift, which was also introduced in section 4. In addition to the product kernel function used in the main text we trained an estimator with a RBF kernel for \( \psi(x, v) \) and length scale \( l = 0.89 \) determined by 2-fold cross-validation. The results are shown in figure E3. Notice that the left figure is identical to figure 2 and is reproduced for better comparison. Both estimators show a good fit to the true drift in the region of high observation density, while on the outer regions with sparse observations the RBF kernel significantly underestimates the true function. This effect can be explained by the stronger influence of the RBF kernel prior in low-data neighborhoods, which drags the function back to zero. We already noticed this behavior in the discussion of the double well model in figure E1.

References

[1] Iacus S M 2008 Simulation and Inference for Stochastic Differential Equations (New York: Springer)
[2] Sriperumbudur B K, Fukumizu K, Kumar R, Gretton A and Hyvärinen A 2014 Density estimation in infinite dimensional exponential families (arXiv:1312.3516v3)
[3] Gottschall J and Peinke J 2008 New J. Phys. 10 083034
[4] Honisch C and Friedrich R 2011 Phys. Rev. E 83 066701
[5] Papaspiiliopoulos O, Pokern Y, Roberts G O and Stuart A M 2012 Biometrika 99 511–31
[6] Ruttor A, Batz P and Opper M 2013 Approximate Gaussian process inference for the drift function in stochastic differential equations Advances in Neural Information Processing Systems vol 26, ed C Burges et al (Red Hook, NY: Curran Associates, Inc.) pp 2040–8
[7] Hyvärinen A 2005 J. Mach. Learn. Res. 6 695–709
[8] Archambeau C, Opper M, Shen Y, Cornford D and Shawe-Taylor J 2008 Variational inference for diffusion processes Advances in Neural Information Processing Systems vol 20, ed J Platt et al (Cambridge, MA: MIT Press) pp 17–24
[9] Chernyak V Y, Chertkov M, Bierkens J and Kappen H J 2013 J. Phys. A: Math. Theor. 47 022001
[10] Rasmussen C E and Williams C K I 2006 Gaussian Processes for Machine Learning (Cambridge, MA: MIT Press)
[11] Deisenroth M P, Rasmussen C E and Peters J 2009 Neurocomputing 72 1508–24

do:10.1088/1742-5468/2016/08/083404
Variational estimation of the drift for stochastic differential equations from the empirical density

[12] Gardiner C W 1996 *Handbook of Stochastic Methods* 2nd edn (Berlin: Springer)
[13] Hansen L P 1982 *Econometrica* **50** 1029–54
[14] Pokern Y, Stuart A M and van Zanten J H 2013 *Stoch. Process. Appl.* **123** 603–28
[15] Graepel T 2003 Solving noisy linear operator equations by gaussian processes: application to ordinary and partial differential equations *ICML* pp 234–41
[16] Alvarez M A, Luengo D and Lawrence N D 2009 Latent force models *Int. Conf. on Artificial Intelligence and Statistics* pp 9–16
[17] Melkumyan A 2012 Operator induced multi-task gaussian processes for solving differential equations *Neural Information Processing Systems (NIPS) Workshop New Directions in Multiple Kernel Learning*