Chapter 1

Approximate inference for continuous-time Markov processes

Cédric Archambeau[1] *and Manfred Opper*[2]

1.1 Introduction

Markov processes are probabilistic models for describing data with a sequential structure. Probably the most common example is a dynamical system, of which the state evolves over time. For modelling purposes it is often convenient to assume that the system states are not directly observed: each observation is a possibly incomplete, nonlinear and noisy measurement (or transformation) of the underlying hidden state. In general, observations of the system occur only at discrete times, while the underlying system is inherently continuous in time. Continuous-time Markov processes arise in a variety of scientific areas such as physics, environmental modeling, finance, engineering and systems biology.

The continuous-time evolution of the system imposes strong constraints on the model dynamics. For example, the individual trajectories of a diffusion process are rough, but the mean trajectory is a smooth function of time. Unfortunately, this information is often under- or unexploited when devicing practical systems. The main reason is that inferring the state trajectories and the model parameters is a difficult problem as trajectories are infinite dimensional objects. Hence, a practical approach usually requires some sort of approximation. For example, Markov Chain Monte Carlo (MCMC) methods usually discretise time (Shephard, 2001; Eraker, 2001; Roberts and Stramer, 2001; Alexander et al., 2005; Golightly and Wilkinson, 2006), while particle filters approximate continuous densities by a finite number of point masses (Crisan and Lyons, 1999; Del Moral and Jacod, 2001; Del Moral et al., 2002). More recently, approaches using perfect simulation have been proposed (Beskos et al., 2006, 2008; Fearnhead et al., 2008). The main advantage of these MCMC techniques is that they do not require approximations of the transition density using time discretisations. Finally, a variety of approaches like extensions to the Kalman filter/smoother (Särkkä, 2006) and moment closure methods (Eyink et al., 2004) express the statistics of state variables by a finite set of moments, for example based on Gaussian assumptions.

In this work we discuss a promising variational approach to the inference problem for continuous-time Markov processes, which was introduced by Archambeau et al. (2007, 2008). We will focus on diffusion processes, where the system state is a

[1] Centre for Computational Statistics and Machine Learning, University College London, Gower Street, London WC1E 6BT, United Kingdom. (c.archambeau@cs.ucl.ac.uk)

[2] Technische Universität Berlin, Fakultät IV – Elektrotechnik und Informatik, Franklinstr. 28/29, D-10587 Berlin, Germany. (opperm@cs.tu-berlin.de)

continuous variable subject to a deterministic forcing, called *drift*, and a stochastic noise process, called *diffusion*. However, the approach can also be applied to other processes, such as Markov jump processes (MJPs) (Opper and Sanguinetti, 2008; I.Cohn et al., 2009; Sanguinetti et al., 2009). In MJPs the state trajectories are still functions of continuous-time, but the system state can only take discrete values. For diffusions, the approach is based on a Gaussian approximation, but as in perfect simulation MCMC, it is not based on a discrete-time approximation of the transition density. The approximate statistics are made not *ad hoc* as in the case of the Kalman filter/smoother, but introduced in such a way that the true intractable probability measure is optimally approximated.

This chapter is organised as follows. In Section 1.2 we will define partly observed diffusion processes and state the inference problem. Next we will characterise the probability measure over state trajectories given the data and show that the resulting posterior process is a non-stationary Markov process. In Section 1.4 we introduce the variational approximation and show how this approach popular in Machine Learning can be applied to Markov processes and in particular to diffusions. Note, however, that unlike in most variational approaches we will not assume any form of factorised approximation. In Section 1.5, we will consider a practical smoothing algorithm based on the Gaussian variational approximation and discuss the form of the solution in more detail. Finally, we will draw conclusions in Section 1.8.

1.2 Partly observed diffusion processes

We will be concerned with (Itô) stochastic differential equations, where the dynamics of a state variable $x(t) \in R^d$ is given by

$$dx(t) = f(x(t))dt + D^{1/2}(x(t))\, dW(t)\ . \tag{1.1}$$

The vector function f is called the drift. The second term describes a (in general state dependent) white noise process defined through a positive semi-definite matrix D, called diffusion matrix, and a Wiener process $W(t)$. We can think of this process as the limit of the discrete-time process

$$x(t+\Delta t) - x(t) = f(x(t))\Delta t + D^{1/2}(x(t))\sqrt{\Delta t}\, \epsilon_t\ , \tag{1.2}$$

where ϵ_t is now a vector of i.i.d. Gaussian random variables. The specific scaling of the white noise with $\sqrt{\Delta t}$ gives rise to the nondifferentiable trajectories of *sample paths* characteristic for a diffusion process (Karatzas and Schreve, 1998; Kloeden and Platen, 1999; Øksendal, 2005). The form (1.2) is known as the Euler-Maruyama approximation of (1.1).

We assume the diffusion process is stationary, i.e. f and D are not explicit functions of time, although this is not required. We have only access to a finite set of noisy observations $Y \equiv \{y_i\}_{i=1}^N$ of the unobserved process $x(t)$ at times t_i for $i = 1, \ldots, N$. Conditioned on the state x we assume that observations are independent with an observation likelihood $p(y|x)$. We are interested in the problem where f and D are known only up to some unknown parameters θ. It is usually necessary to add the initial state $x(0) = x_0$ as an unknown to the parameters to infer.

Our goals are then to learn as much as possible from the observations in order to infer the system parameters θ, the initial state x_0 and to estimate the unknown sample path $x(t)$ over some interval $0 \leq t \leq T$. The latter task (when all observations during this time are used) is called *smoothing*.

In a maximum likelihood approach (or more precisely type II maximum likelihood (Berger, 1985) or evidence maximisation (MacKay, 1992; Bishop, 1995)) one would solve the first two problems by integrating out the latent process $x(t)$ and then maximising the marginal likelihood $p(Y|x_0, \theta)$ with respect to θ and x_0. In a fully Bayesian approach, one would encode prior knowledge in a prior density $p(x_0, \theta)$ and would obtain the information about the unknowns in the posterior density $p(x_0, \theta|Y) \propto p(Y|x_0, \theta)\ p(x_0, \theta)$. In both cases, the computation of $p(Y|x_0, \theta)$ is essential, but in general analytically intractable.

1.3 Hidden Markov characterisation

Let us first assume the parameters are known. To deal with the reconstruction of a sample path we compute $p_t(x|Y, x_0, \theta)$, which is the marginal posterior of the state at time t, i.e. $x(t) = x$. This marginal can be computed in the same way as the marginals of standard discrete-time hidden Markov models (Rabiner, 1989). The only difference is that we have to deal with continuous-time and continuous states.

Using the Markov nature of the process[3] and Bayes' rule it is not hard to show that we can represent this posterior as a product of two factors

$$p_t(x|Y, x_0, \theta) \propto \underbrace{p(x(t)|Y_{<t}, x_0, \theta)}_{\doteq p_t^{(F)}(x)}\ \underbrace{p(Y_{\geq t}|x(t))}_{\doteq \psi_t(x)} \ .$$

where $p_t^{(F)}(x)$ is the posterior density of $x(t) = x$ based on the observations $Y_{<t} \equiv \{y_i\}_{t_i<t}$ before time t and $\psi_t(x)$ is the likelihood of the future observations $Y_{\geq t} = \{y_i\}_{t_i \geq t}$ conditioned on $x(t) = x$.

For times smaller than the next observation time $p_t^{(F)}(x)$ fulfils the *Fokker-Planck* (or *Kolmogorov forward*) *equation* (see for example Karatzas and Schreve, 1998, for a detailed discussion) corresponding to the SDE defined in (1.1):

$$\left\{\frac{\partial}{\partial t} + \nabla^\top f - \frac{1}{2}\mathrm{Tr}(\nabla\nabla^\top)D\right\} p_t^{(F)}(x) = 0 \ , \tag{1.3}$$

where ∇ is the vector differential operator. The Fokker-Planck equation describes the time evolution of the density $p_t^{(F)}(x)$ given some earlier density, e.g. at the most recent observation time.

The second factor is found to obey the *Kolmogorov backward equation* corresponding to the SDE defined in (1.1), that is

$$\left\{\frac{\partial}{\partial t} + f^\top\nabla + \frac{1}{2}\mathrm{Tr}(D\nabla\nabla^\top)\right\} \psi_t(x) = 0 \ . \tag{1.4}$$

This equation describes the time evolution of $\psi_t(x)$, i.e. the likelihood of future observations. The knowledge of $\psi_t(x)$ also gives us the desired marginal likelihood as

$$p(Y|x_0, \theta) = \psi_0(x) \ .$$

[3] More specifically, we need the *Chapman-Kolmogorov equation* to compute $p_t(x|Y_{<t}, x_0, \theta)$. By the Markov property we have $p(x(t)|x(s), x(r)) = p(x(t)|x(s))$, such that

$$\int dx(s)\ p(x(t), x(s)|x(r)) = \int dx(s)\ p(x(t)|x(s))\ p(x(s)|x(r)) = p(x(t)|x(r)) \ ,$$

for all $r \leq s \leq t$. Hence, using this result recursively and then applying Bayes' rule leads to

$$p(x(t)|Y_{<t}, x_0, \theta) \propto p(x(t)|x_0, \theta)\ p(Y_{<t}|x(t)) \ .$$

The equations (1.3) and (1.4) hold for times between observations. The information about the observations enters the formalism through a set of jump conditions for $p_t^{(F)}$ and $\psi_t(x)$ at the observation times. This result is known as the so-called *KSP equations* (Kushner, 1962; Stratonovich, 1960; Pardoux, 1982).

Intuitively, the occurrence of jumps can be understood as follows. Assume we are moving forward in time up to time t, where we encounter the observation $y(t)$. The information associated to y_t is removed from $\psi_t(x)$ and incorporated into $p_t^{(F)}(x)$. Mathematically, the "prior" $p_t^{(F)}(x)$ is updated using the likelihood factor $p(y(t)|x(t))$ causing jumps in $p_t^{(F)}(x)$ and $\psi_t(x)$ at time t:

$$p_t^{(F)}(x) \leftarrow \frac{1}{Z} p_t^{(F)}(x)\; p(y(t)|x(t)) \;, \tag{1.5}$$

$$\psi_t(x) \leftarrow \frac{\psi_t(x)}{p(y(t)|x(t))} \;, \tag{1.6}$$

where Z is a normalising constant.

Moreover, by direct differentiation of $p_t(x|Y, x_0, \theta)$ with respect to time and using (1.3) and (1.4), we find after some calculations that the posterior also fulfils the Fokker-Planck equation:

$$\left\{ \frac{\partial}{\partial t} + \nabla^\top g - \frac{1}{2}\mathrm{Tr}(\nabla\nabla^T)D \right\} p_t(x|Y, x_0, \theta) = 0 \;, \tag{1.7}$$

with a new drift defined as

$$g(x,t) = f(x) + D(x)\nabla \ln \psi_t(x) \;. \tag{1.8}$$

This shouldn't be too surprising because conditioning on the observations does not change the causal structure of the process $x(t)$. It is still a Markov process, but a non-stationary one due to the observations. Note that there are no jumps for $p_t(x|Y, x_0, \theta)$ as it found as the product of (1.5) and (1.6).

Hence, the process of exact inference boils down to solving the linear partial differential equation (1.4) backwards in time starting with the final condition $\psi_T(x)$ and taking the jumps $\psi_{t_i^-}(x) = \psi_{t_i}(x) p(y_i|x_i)$ into account to get the function $\psi_t(x)$ from which both the likelihood $p(Y|x_0, \theta)$ and the posterior drift (1.8) are obtained. Finally, the posterior marginals are computed by solving the linear partial differential equation (1.7) forwards in time for some initial condition $p_0^{(F)}(x)$.

1.3.1 Example

As an analytically tractable one dimensional example we consider the simple Wiener process $dx(t) = dW(t)$ starting at $x(0) = 0$ together with a single, noise free observation at $t = T$, i.e. $x(T) = y$.

The forward equation

$$\frac{\partial p_t^{(F)}(x)}{\partial t} - \frac{1}{2}\frac{\partial^2 p_t^{(F)}(x)}{\partial x^2} = 0$$

with initial condition $p_0^{(F)}(x) = \delta(x)$ is solved by $p_t^{(F)}(x) = \mathcal{N}(0, t)$, while the backward equation

$$\frac{\partial \psi_t(x)}{\partial t} + \frac{1}{2}\frac{\partial^2 \psi_t(x)}{\partial x^2} = 0$$

with end condition $\psi_T(x) = \delta(x-y)$ is solved by $\psi_t(x) = \mathcal{N}(y, T-t)$. The posterior density and the posterior drift are then respectively given by

$$p_t(x|x(T)=y, x(0)=0) \propto p_t^{(F)}(x)\ \psi_t(x) = \mathcal{N}(ty/T, t(T-t)/T)\ , \tag{1.9}$$

$$g(x,t) = \frac{\partial \ln \psi_t(x)}{\partial x} = \frac{y-x}{T-t}\ , \tag{1.10}$$

for $0 < t < T$. A process with drift (1.10) is known as a *Brownian bridge* (see Figure 1.1). Inspection of (1.9) shows that any path of the process starting at the origin and diffusing away will eventually go to the noise free observation y at time T.

In general, especially in higher dimensions, the solution of the partial differential equations (PDEs) will not be analytically tractable. Also numerical methods for PDE solving (Kloeden and Platen, 1999) might become time consuming. Hence, we may have to consider other types of approximations. One such possibility will be discussed next.

1.4 The variational approximation

A different idea for solving the inference problem might be to attempt a direct computation of the marginal likelihood or *partition function* $Z(x_0, \theta) \doteq p(Y|x_0, \theta)$. Using the Markov property of the process $x(t)$ we obtain

$$Z(x_0, \theta) = \int \prod_{i=1}^{N} \{dx_i\ p\left(x_i|x_{i-1}, \theta\right) p(y_i|x_i)\}\ ,$$

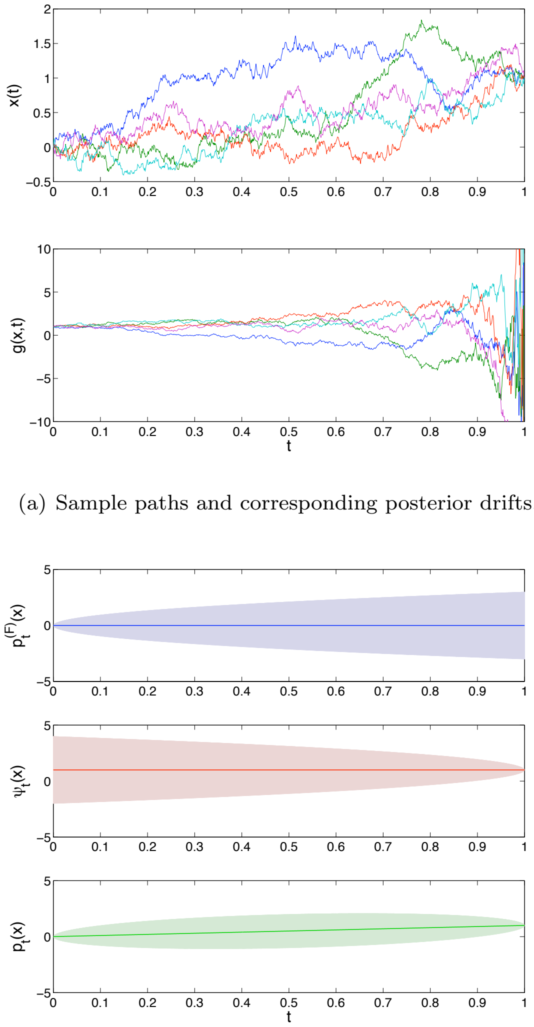
where x_i is a shorthand notation for $x(t_i)$ and x_0 is fixed. Unfortunately, except for simple linear SDEs, the transition density $p(x_i|x_{i-1}, \theta)$ is not known analytically. In fact it would have to be computed by solving the Fokker-Planck equation (1.3).

Nevertheless, at least formally we can write Z as an *infinite dimensional* or functional integral over paths $x(t)$ starting at x_0 using a proper weighting of the paths. Using the Girsanov change of measure formula from stochastic calculus (Øksendal, 2005) one could write such a path integral as:

$$Z = \int d\mu \exp\left(-\frac{1}{2}\int_0^T \{f^\top D^{-1} f dt - 2 f^\top D^{-1} dx\}\right) \prod_{i=1}^{N} p(y_i|x_i)\ ,$$

where $d\mu$ denotes a Gaussian measure over paths starting at $x(0) = x_0$ induced by the simple *linear* SDE $dx(t) = D^{1/2}(x(t))\ dW(t)$ without drift. Note, that in the case of a diagonal diffusion matrix and a drift derived from a potential, the Itô integral $\int f^\top D^{-1} dx$ can be transformed into an ordinary integral. These types of functional integrals play an important role in quantum statistical physics (usually written in a slightly different notation). Most functional integrals cannot be solved exactly, but the variational approach of statistical physics pioneered by Feynman, Peierls, Bogolubov and Kleinert (Feynman and Hibbs, 1965; Kleinert, 2006) gives us an idea how to approximate Z.

Consider some configuration χ of the system of interest. In our application χ is identified with the path $x(t)$ in the time window $[0, T]$. We can represent the probabilities over configurations in the form $dp(\chi) = \frac{1}{Z}\ d\mu(\chi)\ e^{-H(\chi)}$, where $H(\chi) = \frac{1}{2}\int_0^T \{f^\top D^{-1} f dt - 2 f^\top D^{-1} dx\} - \sum_{i=1}^N \ln p(y_i|x_i)$ is the *Hamiltonian*, which in statistical physics corresponds to the energy associated to the configuration. To compute an approximation to the partition function $Z = \int d\mu(\chi)\ e^{-H(\chi)}$,



(a) Sample paths and corresponding posterior drifts.

(b) Prior, likelihood and posterior densities.

Figure 1.1: Illustration of a one dimensional diffusion process without drift and unit diffusion coefficient, starting at the origin and with a noise free observation $y = 1$ at $t = 1$. The posterior process is a Brownian bridge. Note how the drift increases drastically when getting close to the final time. (a) shows 5 sample paths with their corresponding posterior drift functions. (b) shows the mean and variance (shaded region) of the prior, the likelihood and the posterior marginals. Observe how the variance of the posterior $p_t(x)$ is largest in the middle of the time interval and eventually decreases to 0 at $t = 1$.

we first approximate $dp(\chi)$ by a simpler distribution $dq(\chi) = \frac{1}{Z_0}\, d\mu(\chi)\, e^{-H_0(\chi)}$, which is defined by a simpler Hamiltonian $H_0(\chi)$ and for which Z_0 is tractable. Using a simple convexity argument and *Jensen's inequality*, we get an approximation to the log partition function or *free energy* by the bound

$$-\ln Z \leq -\ln Z_0 + \langle H \rangle - \langle H_0 \rangle \ , \tag{1.11}$$

where the brackets denote expectations with respect to the measure q. Usually H_0 contains free parameters, which can be adjusted in such a way that the inequality becomes as tight as possible by minimising the upper bound on the right hand side.

To define a tractable variational approximation (1.11) for our inference problem, we would use an H_0 which is quadratic functional in the process $x(t)$. This would lead to a Gaussian measure over paths. While this is indeed possible we prefer a different, but equivalent formulation of the variational method, which neither needs the definition of a Hamiltonian, nor the application of stochastic calculus. The variational method in this formulation has been extensively applied in recent years in Machine Learning to problems involving finite dimensional latent variables (Jordan, 1998; Opper and Saad, 2001; Bishop, 2006).

1.4.1 The variational approximation in Machine Learning

Let us denote the observations by Y and assume a finite dimensional latent variable X. Consider some prior distribution $p(X|\theta)$ parametrised by θ and some likelihood function $p(Y|X)$. To approximate the intractable posterior $p(X|Y, \theta) \propto p(Y|X)\ p(X|\theta)$ we directly choose a simpler trial distribution $q(X)$. The optimal q is chosen to minimise the *Kullback-Leibler* (KL) *divergence* or relative entropy (Cover and Thomas, 1991)

$$\mathrm{KL}[q\|p] = \left\langle \ln \frac{q(X)}{p(X|Y,\theta)} \right\rangle \geq 0 \ . \tag{1.12}$$

This inequality directly leads to the bound

$$-\ln Z(\theta) \leq -\langle \ln p(Y|X) \rangle + \mathrm{KL}[q(X)\|p(X|\theta)] \doteq \mathcal{F}(q,\theta) \ . \tag{1.13}$$

The right hand side of (1.13) defines the so-called *variational free energy* which is an upper bound to the marginal likelihood of the data. Hence, minimising such a bound with respect to the parameters θ can be viewed as an approximation to the (type II) maximum likelihood method.

One can also apply the variational method in a Bayesian setting, where we have a prior distribution $p(\theta)$ over model parameters (Lappalainen and Miskin, 2000). To approximate the posterior $p(\theta|Y)$, we set $p(X, \theta|Y) \approx q(X|\theta)q(\theta)$ and apply the variational method to the joint space of variables X and θ. Let $q(X|\theta)$ be the distribution which minimises the variational free energy $\mathcal{F}(q, \theta)$ of (1.13). We then get

$$q(\theta) = \frac{e^{-\mathcal{F}(q,\theta)}\ p(\theta)}{\int e^{-\mathcal{F}(q,\theta)}\ p(\theta)\ d\theta} \tag{1.14}$$

as the best variational approximation to $p(\theta|Y)$.

1.4.2 The variational approximation for Markov processes

In the case of partly observed diffusion processes we are interested in the posterior measure over latent paths, which are infinite dimensional objects. The prior measure

$p(\chi|x_0,\theta)$ is derived from an SDE of the form (1.1) and the posterior measure $p(\chi|Y,x_0,\theta)$ is computed from Bayes' rule:

$$\frac{p(\chi|Y,x_0,\theta)}{p(\chi|x_0,\theta)} = \frac{\prod_{i=1}^N p(y_i|x_i)}{Z(x_0,\theta)} , \qquad\qquad 0 \le t \le T .$$

When the exact posterior is analytically intractable, we consider a trial posterior $q(\chi)$ that we would like to match to the true posterior by applying the variational principle. All we need is an expression for the KL divergence. From Section 1.3, we already know that the posterior process is also Markovian and that it obeys an SDE with the time-dependent drift (1.8):

$$dx(t) = g(x(t),t)dt + D^{1/2}(x(t))\, dW(t) . \tag{1.15}$$

Consider two continuous-time diffusion processes having the same[4] diffusion matrix $D(x)$, but different drift functions $f(x)$ and $g(x)$. We call the probability measures induced over the corresponding sample paths respectively $p(\chi)$ and $q(\chi)$. Although we could prove the following rigorously using Girsanov's change of measure theorem (Karatzas and Schreve, 1998; Øksendal, 2005), we will use a simpler, more intuitive heuristic in this paper which can also be applied to Markov jump processes.

Let us discretise time into small intervals of length Δt and consider discretised sample paths $X = \{x_k = x(t_k = k\Delta t)\}_{k=1}^K$ with their corresponding multivariate probabilities $p(X|x_0)$ and $q(X|x_0)$. We then aim to compute the KL divergence between the measures dp and dq over some interval $[0,T]$ as the limit of

$$\begin{aligned}\mathrm{KL}\left[q(X)\|p(X)\right] &= \int dX\ q(X|x_0) \ln \frac{q(X|x_0)}{p(X|x_0)} \qquad (1.16)\\ &= \sum_{k=1}^K \int dx_{k-1}\ q(x_{k-1}) \int dx_k\ q(x_k|x_{k-1}) \ln \frac{q(x_k|x_{k-1})}{p(x_k|x_{k-1})} ,\end{aligned}$$

where we have used the Markov property to represent $p(X|x_0)$ and $q(X|x_0)$ respectively as $\prod_k p(x_k|x_{k-1})$ and $\prod_k q(x_k|x_{k-1})$. Next, we plug in the specific short term behaviour (i.e. $\Delta t \to 0$) of the transition probabilities. Since we are dealing with diffusions we obtain the Gaussian forms

$$\begin{aligned} p(x_k|x_{k-1}) &\propto \exp\left(-\frac{1}{2\Delta t}\left\|x_k - x_{k-1} - f(x_{k-1})\Delta t\right\|^2_{D(x_{k-1})}\right) ,\\ q(x_k|x_{k-1}) &\propto \exp\left(-\frac{1}{2\Delta t}\left\|x_k - x_{k-1} - g(x_{k-1})\Delta t\right\|^2_{D(x_{k-1})}\right) ,\end{aligned}$$

where $\|f\|^2_D = f^\top D^{-1} f$. Following Archambeau et al. (2008), a direct computation taking the limit $\Delta t \to 0$ yields

$$\mathrm{KL}\left[q(X)\|p(X)\right] = \frac{1}{2}\int_0^T dt \left\{\int dq_t(x)\ \|g(x) - f(x)\|^2_{D(x)}\right\} ,$$

where $q_t(x)$ is the posterior marginal at time t. Note that this result is still valid if the drift function and the diffusion matrix are time-dependent.

Hence, the variational free energy in the context of diffusion processes can be written as

$$\mathcal{F}(q,\theta) = \mathrm{KL}[q(\chi)\|p(\chi|\theta)] - \sum_i \langle \ln p(y_i|x_i)\rangle_{q_{t_i}} , \tag{1.17}$$

[4] It can be shown that the KL divergence diverges for different diffusions matrices.

where χ is a continuous sample path in the interval $[0,T]$. The bound (1.11) is equivalent to the bound (1.17) for appropriate definitions of Hamiltonians $H(\chi)$ and $H_0(\chi)$. The advantage of (1.17) is that we can directly compute the KL divergence for Markov processes, without defining $H(\chi)$ and $H_0(\chi)$ explicitly. The results can also be applied to MJPs as proposed by Opper and Sanguinetti (2008).

1.4.3 The variational problem revisited

Before discussing approximations, we will show that total minimisation of the free energy yields our previous result (1.8). For the corresponding derivation in the case of MJPs see Ruttor et al. (2009). The free energy can be written as

$$\mathcal{F}(q,\theta) = \int_0^T dt \int dx\; q_t(x) \left\{ \frac{1}{2} \left\| g(x,t) - f(x) \right\|^2_{D(x)} + u(x,t) \right\} ,$$

where the observations are included in the term

$$u(x,t) = -\sum_i \ln p(y_i|x)\; \delta(t-t_i) \; .$$

The drift g and the marginal q_t are connected by the Fokker-Planck equation

$$\frac{\partial q_t}{\partial t} = \left\{ -\nabla^\top g + \frac{1}{2}\mathrm{Tr}(\nabla\nabla^T)D \right\} q_t \doteq L_g q_t$$

as a constraint in the optimisation of q_t. We can deal with this constraint by introducing a Langrange multiplier function $\lambda(x,t)$ to obtain the following *Lagrange functional*:

$$\mathcal{L} \doteq \mathcal{F}(q,\theta) - \int_0^T dt \int dx\; \lambda(x,t) \left(\frac{\partial q_t(x)}{\partial t} - (L_g q_t)(x) \right) \; .$$

Performing independent variations of q_t and g leads respectively to the following *Euler-Lagrange* equations:

$$\frac{1}{2}\left\| g - f \right\|^2_D + u + \left\{ g^\top \nabla + \frac{1}{2}\mathrm{Tr}(D\nabla\nabla^\top) \right\} \lambda + \frac{\partial \lambda}{\partial t} = 0 \; ,$$
$$D^{-1}(g-f) + \nabla\lambda = 0 \; ,$$

where we have used integration by parts when appropriate. Defining the logarithmic transformation $\lambda(x,t) = -\ln \psi_t(x)$ and rearranging yields then the conditions

$$\left\{ \frac{\partial}{\partial t} - u(x,t) \right\} \psi_t(x) = \left\{ -f^\top(x)\nabla - \frac{1}{2}\mathrm{Tr}(D(x)\nabla\nabla^\top) \right\} \psi_t(x) \; , \qquad (1.18)$$

$$g(x,t) = f(x) + D(x)\nabla \ln \psi_t(x) \; , \qquad (1.19)$$

for all $t \in [0,T]$. By noting that $u(x,t) = 0$ except at the observation times, we find that these results are equivalent to (1.4) and (1.8); the Dirac δ functions yield the proper jump conditions when there are observations. Note that this derivation still holds if f and D are time-dependent.

1.5 The Gaussian variational approximation

In practice, rather than assuming the correct functional form (1.19), we will view g as a variational function with a simplified form. The function g can then be optimised to minimise the free energy.

Gaussian distributions are a natural choice for approximations. For example, they have been used frequently in statistical physics applications. For previous (finite dimensional) applications in machine learning see Barber and Bishop (1998); Seeger (2000); Honkela and Valpola (2005). In the present inference case, a Gaussian approximating measure over paths, that is a Gaussian process, is considered. In this case the drift must be a *linear* function of the state x. We consider a drift of the form $g(x,t) = -A(t)x + b(t)$, where $A(t)$ and $b(t)$ are functions to be optimised. In addition, we limit ourselves to the special case of a *constant* diffusion matrix D (Archambeau et al., 2007, 2008). The approximation equally holds in the case of time-dependent diffusions. The more general case of multiplicative noise processes, that is with state dependent diffusion matrices, will be discussed in Section 1.6.

Since we are dealing with a Gaussian process, the marginals $q_t(x)$ are Gaussian densities. This result represents a significant simplification of the calculations. First, $q_t(x)$ are fully specified by their marginal means $m(t)$ and their marginal covariances $S(t)$. Second, we don't need to solve PDEs, but are left with simpler ordinary differential equations (ODEs). Since (1.15) is a linear SDE, we have

$$\begin{aligned} dm &\doteq \langle dx \rangle = (-Am + b)dt \ , \\ dS &\doteq \langle d((x-m)(x-m)^\top) \rangle = (-AS - SA^\top)dt + Ddt + \mathcal{O}(dt^2) \ , \end{aligned}$$

where the term Ddt is obtained by applying the stochastic chain rule.[5] Hence, the evolution of $m(t)$ and of $S(t)$ is described by the following set of ODEs:

$$\frac{dm(t)}{dt} = -A(t)m(t) + b(t) \ , \tag{1.20}$$

$$\frac{dS(t)}{dt} = -A(t)S(t) - S(t)A^\top(t) + D \ . \tag{1.21}$$

We can follow a similar approach as in Section 1.4.3 to optimise the Gaussian variational approximation. More specifically, we use these ODEs as a contstraint during the optimisation. Let us define $e(x,t) = \frac{1}{2}\|g(x,t) - f(x)\|_D^2$. The Lagrangian functional is now defined as

$$\begin{aligned} \mathcal{L} = \int_0^T dt \, \langle e(x,t) + u(x,t) \rangle_{q_t} - \int_0^T dt \, \lambda^\top(t) \left(\frac{dm(t)}{dt} + A(t)m(t) - b(t) \right) \\ - \int_0^T dt \, \mathrm{Tr} \left(\Psi(t) \left(\frac{dS(t)}{dt} + A(t)S(t) + S(t)A^\top(t) - D \right) \right) \ , \end{aligned} \tag{1.22}$$

where $\lambda(t)$ and $\Psi(t)$ are vector and matrix Lagrange parameter functions which depend on time only. Performing independent variations of $m(t)$ and $S(t)$ (which is equivalent to performing an independent variation of q_t) yields an additional set of ODEs:

$$\frac{d\lambda(t)}{dt} = -\nabla_m \langle e(x,t) \rangle_{q_t} + A^\top(t)\lambda(t) \ , \tag{1.23}$$

$$\frac{d\Psi(t)}{dt} = -\nabla_S \langle e(x,t) \rangle_{q_t} + \Psi(t)A(t) + A^\top(t)\Psi(t) \ , \tag{1.24}$$

along with jump conditions at observation times

$$\lambda_i = \lambda_i^- - \nabla_m \langle u(x,t) \rangle_{q_t} \big|_{t=t_i} \ , \qquad \lambda_i^- = \lim_{t \uparrow t_i} \lambda(t) \ , \tag{1.25}$$

$$\Psi_i = \Psi_i^- - \nabla_S \langle u(x,t) \rangle_{q_t} \big|_{t=t_i} \ , \qquad \Psi_i^- = \lim_{t \uparrow t_i} \Psi(t) \ . \tag{1.26}$$

[5] This result can also be obtained by an informal derivation not relying on stochastic calculus but only using properties of the Wiener process (see Archambeau et al., 2007, Appendix).

Hence, the Fokker-Planck equation is replaced by (1.20) and (1.21) in the Gaussian variational approximation, while the Kolmogorov backward equation is replaced by (1.23) and (1.24). Based on (1.23–1.26) we can devise a smoothing algorithm as described in Archambeau et al. (2007, 2008). Also, a procedure to infer of θ (which parametrises f and D) is discussed in detail in Archambeau et al. (2008).

One important advantage of the Gaussian variational approach is that representations can be based on a discretisation of ODEs instead of a direct discretisation of the SDE. The loss of accuracy is expected to be less severe because of the smoothness of the paths (Kloeden and Platen, 1999). Also the approximation holds in continuous-time and is thus independent of the chosen representations unlike most MCMC schemes (Alexander et al., 2005; Golightly and Wilkinson, 2006). In contrast to these discrete-time MCMC schemes, perfect simulation MCMC for continuous-time systems was recently proposed (Beskos et al., 2006, 2008; Fearnhead et al., 2008). This method is sofar restricted to problems where drift terms are derived as gradients of a potential function, which is not required in the Gaussian variational approximation. The main similarity between the Gaussian variational approximation and these advanced MCMC approaches is that they do not depend on a discrete-time approximation of the transition density. However, the Gaussian variational approximation differs from perfect simulation in its approximation of the non-Gaussian transition density by a (time dependent) Gaussian one.

Thorough experimental comparisons are still required to assess the advantages and disadvantages of the different methods, but the Gaussian variational approximation is likelily to be computationally faster as it is not based on sampling; it only cares about the marginal means and covariances, which can be computed efficiently by forward integration (1.20) and (1.21). On the other hand, perfect sampling MCMC will capture the posterior measure more accurately if run for a sufficiently long period of time.

1.5.1 Interpretation of the solution

In this subsection we discuss the form of the Gaussian variational solution in more detail. Let us perform the independent variation of $A(t)$ and $b(t)$, which can be viewed as performing the independent variation of g as in Section 1.4.3. This leads to the following conditions

$$A(t) = -\left\langle \nabla(f^\top)(x,t)\right\rangle_{q_t} + 2D\Psi(t), \tag{1.27}$$

$$b(t) = \langle f(x,t)\rangle_{q_t} + A(t)m(t) - D\lambda(t), \tag{1.28}$$

for all t. In order to obtain (1.27) we used the identity $\langle f(x-m)^\top\rangle = \langle\nabla(f^\top)\rangle S$, which holds for any nonlinear function $f(\cdot)$ applied to a Gaussian random variable x. The solution (1.27–1.28) is closely related to a solution known as *statistical linearisation* (Roberts and Spanos, 2003).

Consider a nonlinear function f, which is applied to a continuous random variable x with density q. We are interested in the best linear approximation $-Ax+b$ to f. Instead of directly truncating the Taylor series of f to obtain a linear approximation, we would like to take into account the fact that x is a random variable. Statistical linearisation takes this information into account by taking A and b such that the linear approximation is optimal in the mean squared sense:

$$A, b \leftarrow \min_{A,b} \left\langle \|f(x) + Ax - b\|^2\right\rangle_q \ .$$

When x is a Gaussian random variable it is easy to show that the solution to this problem is given by $A = -\left\langle\nabla(f^\top)(x)\right\rangle_q$ and $b = \langle f(x)\rangle_q + Am$. Comparing these expressions to (1.27) and (1.28), it can be observed that the variational solution reduces to the statistical linearisation solution when the Lagrange multipliers are zero. Recalling that the Lagrange mulitpliers account for the constraints and the observations, one can see that the solution (1.27–1.28) is biased compared to the standard statistical linearisation solution. This bias corresponds to a correction based on future information and it is weighted by the diffusion matrix. The weighting by D makes sense as the magnitude of the correction should depend on the amount of stochasticity in the system.

1.5.2 Example

Applications of the Gaussian variational approximation to statistical inference for nonlinear SDEs can be found in Archambeau et al. (2007, 2008). We will illustrate the basic idea of the calculation only for the simple, analytically tractable case of Section 1.3.1. For later use, we introduce an extra parameter σ in the model which controls the diffusion coefficient, i.e. we set $dx(t) = \sigma dW(t)$.

We have $g(x,t) = -a(t)x(t) + b(t)$. The evolution of the mean $m(t)$ and variance $s(t)$ simplify to

$$\frac{dm}{dt} = -a(t)m(t) + b(t) \ , \tag{1.29}$$

$$\frac{ds}{dt} = -2a(t)s(t) + \sigma^2 \ . \tag{1.30}$$

We model the noise free observation as the limit of a Gaussian observation centred at y and with variance $\sigma_0^2 \to 0$. Hence, we have

$$\begin{aligned} \mathcal{L} = &\int_0^T dt \left\{ \frac{1}{2\sigma^2} a^2(s+m^2) + \frac{1}{2\sigma^2} b^2 - \frac{1}{\sigma^2} amb \right\} \\ &+ \int_0^T dt \ \frac{1}{2\sigma_0^2}(y^2 + s + m^2 - 2my)\delta(t-T) \\ &- \int_0^T dt \ \lambda \left(\frac{dm}{dt} + am - b \right) - \int_0^T dt \ \psi \left(\frac{ds}{dt} + 2as - \sigma^2 \right) \ . \end{aligned} \tag{1.31}$$

The Euler-Lagrange equations (1.23–1.28) are then given by

$$\frac{d\lambda(t)}{dt} = -\frac{a^2(t)m(t)}{\sigma^2} + \frac{a(t)b(t)}{\sigma^2} + a(t)\lambda(t) \ , \tag{1.32}$$

$$\frac{d\psi(t)}{dt} = -\frac{a^2(t)}{2\sigma^2} + 2\psi(t)a(t) \ , \tag{1.33}$$

$$a(t) = 2\sigma^2\psi(t) \ , \tag{1.34}$$

$$b(t) = a(t)m(t) - \sigma^2\lambda(t) \ , \tag{1.35}$$

along with the jump conditions

$$\lambda(T) = \lambda(T^-) - \frac{m(T) - y}{\sigma_0^2} \ , \qquad \psi(T) = \psi(T^-) - \frac{1}{2\sigma_0^2} \ .$$

Substitution of (1.34) into (1.33) leads to $\frac{d\psi}{dt} = 2\sigma^2\psi^2$ with the end condition $\frac{1}{2\sigma_0^2}$. It follows that the solution to this ODE is given by $\psi(t) = \frac{1}{2\sigma^2(T-t)+2\sigma_0^2}$. Second, substitution of (1.35) into (1.32) implies λ is a constant. The end condition yields

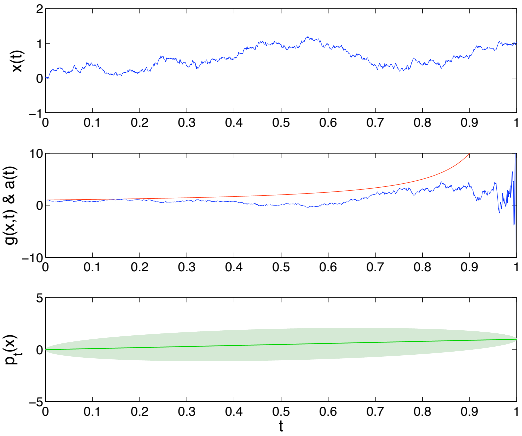


Figure 1.2: Illustration of a one dimensional diffusion process without drift and unit diffusion coefficient, starting at the origin and with a noise free observation $y = 1$ at $t = 1$. The top panel shows a sample path and the corresponding drift $g(x, t) = -a(t)x(t) + b(t)$ is shown in the middle panel. The middle panel also shows the variational parameter $a(t)$. The bottom panel shows the posterior process, which corresponds to the one obtained in Figure 1.1.

$\lambda = \frac{m(T)-y}{\sigma_0^2}$. Next, substitution of (1.35) into (1.29) leads to $m(t) = -\sigma^2\lambda t + c$ with $c = 0$ as $m(0) = 0$, such that $m(T) = \frac{yT}{\sigma_0^2/\sigma^2+T}$. Hence, we obtain:

$$a(t) = \frac{\sigma^2}{\sigma^2(T-t)+\sigma_0^2} \, ,$$

$$b(t) = \frac{\sigma^2 y}{\sigma^2(T-t)+\sigma_0^2} \, .$$

This leads to the same result for $g(x, t)$ as (1.10) when $\sigma_0^2 \to 0$. The solution is illustrated in Figure 1.2.

1.6 Diffusions with multiplicative noise

In Section 1.5 we discussed the Gaussian variational approximation of diffusion processes with a constant or a time-dependent diffusion matrix. However, the methodology can still be applied to diffusion processes with multiplicative noise.

In some cases one can apply an explicit transformation to transform the original diffusion process with multiplicative noise into a diffusion process with a unit diffusion matrix (e.g., Ait-Sahalia, 2008). The resulting drift is then expressed in terms of f and D via Itô's formula. Although this *add-hoc* approach is always possible when the state space is one dimensional, such a transformation typically does not exist in the multivariate case.

In the general case of a state dependent diffusion, the ODEs describing the evolution of the mean $m(t)$ and the covariance $S(t)$ are defined by

$$\frac{dm(t)}{dt} = -A(t)m(t) + b(t) \, , \tag{1.36}$$

$$\frac{dS(t)}{dt} = -A(t)S(t) - S(t)A^\top(t) + \langle D(x(t), t)\rangle_{q_t} \, . \tag{1.37}$$

The only difference with (1.20) and (1.21) is that in (1.37) the expectation of the diffusion matrix appears. The variational energy is still given by (1.17). Hence, we can construct a valid Gaussian process approximation of the posterior process using the constraints (1.36) and (1.37). Note, however, that $A(t)$ and $b(t)$ are no longer given by (1.27) and (1.28), but have a more complicated form.

1.7 Parameter inference

The formulation of the variational approach in terms of (1.22) is especially useful when we would like to estimate model parameters by an approximate type II maximum likelihood method. In this approximation, we use the free energy $\mathcal{F}(q^*, \theta)$ evaluated at the optimal variational Gaussian measure q^* for given parameters θ as a proxy for the negative log-marginal likelihood $-\ln Z(\theta)$.

The optimal parameters θ^* are obtained by minimising $\mathcal{F}$, which requires the computation of the gradients $\nabla_\theta \mathcal{F}(q^*, \theta)$. Although q^* is a function of θ, this optimisation problem is facilitated by the following argument. For each θ, we have $\mathcal{L} = \mathcal{F}(q^*, \theta)$ at the stationary solution, which is also stationary with respect to marginal moments, variational parameters and Lagrange parameters. Hence, to compute the gradients $\nabla_\theta \mathcal{F}(q^*, \theta)$, we just have to take the *explicit* gradients of $\mathcal{L}$ with respect to θ, while keeping all other quantities fixed.

1.7.1 Example

This idea is illustrated for the simple diffusion example of Section 1.3.1, where we have introduced a parameter σ to control the diffusion variance: $dx(t) = \sigma dW(t)$. We are interested in computing the derivative of the negative log-marginal likelihood of the single observation y (at time T) with respect to σ^2.

For a direct computation, let us first note that $p_t^{(F)}(x) = \mathcal{N}(0, \sigma^2 t)$. The marginal likelihood for y is given by

$$p(y|, \sigma^2) = \int \delta(y - x(T)) p_T^{(F)}(x) dx(T) = \mathcal{N}(0, \sigma^2 T) \ ,$$

which yields

$$-\frac{\partial \ln p(y|\sigma^2)}{\partial \sigma^2} = \frac{1}{2\sigma^2} - \frac{y^2}{2\sigma^4 T}.$$

On the other hand, differentiating (1.31) with respect to σ^2 leads to

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = -\frac{1}{2\sigma^4} \int_0^T dt \ \left(a^2 s + \sigma^4 \lambda^2\right) + \int_0^T dt \ \psi(t) = \frac{1}{2\sigma^2} - \frac{y^2}{2\sigma^4 T}.$$

The first equality is obtained by differentiating (1.31) and using (1.35). To get the final result we have inserted the explicit results for $a(t)$, $\lambda(t)$ and $\psi(t)$ obtained in Section 1.5.2 for $\sigma_0^2 \to 0$, as well as the corresponding solution to (1.30): $s(t) = \frac{\sigma^2 t}{T}(T - t)$.

1.8 Discussion and outlook

Continuous-time Markov processes, such as diffusion processes and Markov jump processes, play an important role in the modelling of dynamical systems. In a variety of applications, the state of the system is a (time-dependent) random variable of which the realisation is not directly observed. One has only access to noisy observations taken at a discrete set of times. The problem is then to infer from data the unknown state trajectory and the model parameters, which define the dynamics. While it is fairly straightforward to present a theoretical solution to these estimation problems, a practical solution in terms of PDEs or by MCMC sampling can be time consuming. One is thus interested in efficient approximations.

In this work we described a method to fit a Gaussian process to a non-Gaussian process induced by a SDE. The method is based on the variational principle originally developed in statistical physics and now extensively used in Machine Learning. It provides a practical alternative to exact methods and MCMC. Unlike previous variational approaches (Wang and Titterington, 2004) it is not required to discretise the sample paths, nor to factorise the posterior across time. Although this might lead to good results when the number of observations is large compared to the speed of the dynamics, this approach leads in general to poor results. For a systematic discussion of the effect of factorisation in discrete-time dynamical systems we refer the interested reader to Chapter **??**. By contrast our approximation does not assume any form of factorisation of the posterior. Rather, we choose a posterior process within a tractable family, namely the Gaussian family, which explicitly preserves the time dependency. Moreover, the approximation holds in continuous-time such that discretisation is only required for representation purposes.

The Gaussian variational approximation is attractive as it replaces the problem of directly solving a SDE (or equivalently a set of PDEs) by the simpler problem of solving a set of ODEs. The variational parameters are optimised to obtain the best possible approximation. This optimisation is done concurrently with the estimation of the model parameters, which enable us to learn the dynamics of the system. However, the proposed approach might be too time consuming in high dimensional applications, such as numerical weather prediction. The main reason is that the dynamics of the marginal covariance S scales with d^2, d being the state space dimension. Hence, one could envisage suboptimal schemes in which the variational parameters are reparametrised by a small number of auxiliary quantities. Another potential issue is the estimation of the multivariate Gaussian expectations, which appear in the computation of A and b, as well as the computation of free energy $\mathcal{F}$. In low dimensional state spaces they can be estimated naively using sampling. Alternatively, one can use quadrature methods, but most existing approaches break down or are too slow in higher dimensional spaces and/or for highly nonlinear dynamics.

As mentioned earlier there are ongoing efforts to develop computationally efficient algorithms for fully Bayesian inference in diffusion processes. A very promising direction is to combine the Gaussian variational method and MCMC. One could for example develop a MCMC algorithm which uses the variational approximating process as a proposal process (Shen et al., 2008). Sample paths could then be simulated using the optimal non-stationary linear diffusion and flexible blocking strategies would be used to further improve the mixing.

Acknowledgements

This work has been partly funded by the EPSRC as part of the VISDEM project (EP/C005848/1), as well as by the PASCAL 2 European network of excellence.

Bibliography

Ait-Sahalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusions. *Annals of Statistics*, 36:906–937.

Alexander, F. J., Eyink, G. L., and Restrepo, J. M. (2005). Accelerated Monte Carlo for optimal estimation of time series. *Journal of Statistical Physics*, 119:1331–1345.

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. (2007). Gaussian process approximation of stochastic differential equations. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 1:1–16.

Archambeau, C., Opper, M., Shen, Y., Cornford, D., and Shawe-Taylor, J. (2008). Variational inference for diffusion processes. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 17–24. The MIT Press.

Barber, D. and Bishop, C. M. (1998). Ensemble learning for Multi-Layer Networks. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems 10 (NIPS)*. The MIT Press.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer, New York.

Beskos, A., Papaspiliopoulos, O., Roberts, G., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society B*, 68(3):333–382.

Beskos, A., Roberts, G., Stuart, A., and Voss, J. (2008). MCMC methods for diffusion bridges. *Stochastics and Dynamics*, 8(3):319–350.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press, New York.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer, New York.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* John Wiley and Sons.

Crisan, D. and Lyons, T. (1999). A particle approximation of the solution of the Kushner-Stratonovitch equation. *Probability Theory and Related Fields*, 115(4):549–578.

Del Moral, P. and Jacod, J. (2001). Interacting particle filtering with discrete observations. In Doucet, A., de Freitas, N., and Gordon, N., editors, *Sequential Monte Carlo Methods in Practice*, pages 43–76. The MIT press.

Del Moral, P., Jacod, J., and Protter, P. (2002). The Monte Carlo method for filtering with discrete-time observations. *Probability Theory and Related Fields*, 120:346–368.

Eraker, B. (2001). MCMC analysis of diffusion models with application to finance. *Journal of Business and Economic Statistics*, 19:177–191.

Eyink, G. L., Restrepo, J. L., and Alexander, F. J. (2004). A mean field approximation in data assimilation for nonlinear dynamics. *Physica D*, 194:347–368.

Fearnhead, P., Papaspiliopoulos, O., and Roberts, G. O. (2008). Particle filters for partially-observed diffusions. *Journal of the Royal Statistical Society B*, 70:755–777.

Feynman, R. P. and Hibbs, A. R. (1965). *Quantum Mechanics and Path integrals.* McGraw - Hill Book Company.

Golightly, A. and Wilkinson, D. J. (2006). Bayesian sequential inference for nonlinear multivariate diffusions. *Statistics and Computing*, 16:323–338.

Honkela, A. and Valpola, H. (2005). Unsupervised variational Bayesian learning of nonlinear models. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17 (NIPS)*, pages 593 – 600. The MIT Press.

I.Cohn, El-hay, T., Friedman, N., and Kupferman, R. (2009). Mean field variational approximation for continuous-time Bayesian networks. In *25th International Conference on Uncertainty in Artificial Intelligence (UAI).*

Jordan, M. I., editor (1998). *Learning in Graphical Models.* The MIT Press.

Karatzas, I. and Schreve, S. E. (1998). *Brownian Motion and Stochastic Calculus.* Springer, New York.

Kleinert, H. (2006). *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets.* World Scientific, Singapore.

Kloeden, P. E. and Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations.* Springer, Berlin.

Kushner, H. J. (1962). On the differential equations satisfied by conditional probability densities of Markov processes with applications. *Journal of SIAM, Series A: Control*, 2:106–119.

Lappalainen, H. and Miskin, J. W. (2000). Ensemble learning. In Girolami, M., editor, *Advances in Independent Component Analysis*, pages 76–92. Springer-Verlag.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, 4(3):415–447.

Øksendal, B. (2005). *Stochastic Differential Equations.* Springer-Verlag, Berlin.

Opper, M. and Saad, D., editors (2001). *Advanced Mean Field Methods: Theory and Practice*. The MIT Press.

Opper, M. and Sanguinetti, G. (2008). Variational inference for Markov jump processes. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20 (NIPS)*. The MIT Press.

Pardoux, E. (1982). Equations du filtrage non linéaire, de la prédiction et du lissage. *Stochastics*, 6:193–231.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 2(77):257–286.

Roberts, G. and Stramer, O. (2001). On inference for partially observed non-linear diffusion models using the Metropolis-Hastings algorithm. *Biometirka*, 88(3):603–621.

Roberts, J. B. and Spanos, P. D. (2003). *Random Vibration and Statistical Linearization*. Dover Publications.

Ruttor, A., Sanguinetti, G., and Opper, M. (2009). Approximate inference for stochastic reaction processes. In *Learning and Inference in Computational Systems Biology*, pages 189–205. The MIT Press.

Sanguinetti, G., Ruttor, A., Opper, M., and Archambeau, C. (2009). Switching regulatory models of cellular stress response. *Bioinformatics*, 25(10):1280–1286.

Särkkä, S. (2006). *Recursive Bayesian Inference on Stochastic Differential Equations*. PhD thesis, Helsinki University of Technology, Finland.

Seeger, M. (2000). Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 12 (NIPS)*, pages 603–609. The MIT Press.

Shen, Y., Archambeau, C., Cornford, D., and Opper, M. (2008). Markov Chain Monte Carlo for inference in partially observed nonlinear diffusions. In *Proceedings Newton Institute for Mathematical Sciences workshop on Inference and Estimation in Probabilistic Time-Series Models*, pages 67–78.

Shephard, O. E. S. C. N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrika*, 69(4):959–993.

Stratonovich, R. L. (1960). Conditional Markov processes. *Theory of Probability and its Applications*, 5:156–178.

Wang, B. and Titterington, D. M. (2004). Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Processing Letters*, 20(3):151–170.