Free energy approximations for inference in continuous time Markov processes

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

We propose a new representation of the free energy for continuous time Markov processes where the approximating process is parametrised in terms of *marginal* probabilities and a set of 'conjugate' functions. It allows for the use of Gaussian marginal approximations for both diffusion processes with state dependent diffusions and for Markov jump processes.

1 Introduction

Recently there has been growing interest in inference of stochastic dynamical systems within the machine learning community [1, 7, 2]. A specific machine learning approach to inference, the *variational approximation* has been applied to diffusion processes, Markov jump processes and dynamical belief networks. For this method the exact posterior probability measure over paths is approximated by a tractable one, e.g. a Gaussian for the case of diffusions. The approximation is optimised by the minimisation of a certain *free energy* functional based on the Kullback-Leibler (KL) divergence between these measures. While this approach has been applied successfully to a variety of models, we find that it also has a *severe drawback*. This comes from the fact that in quite natural cases the approximating probability distribution and the exact one may not be absolutely continuous with respect to each others (ie a set which has nonzero probability for one measure, may have zero probability for the other). In this case, the KL divergence is *infinite* ! This can happen for diffusion processes where the diffusion terms for both processes must be equal in order to have a finite KL divergence. Hence, it is not possible to approximate a process with *state dependent noise* by a Gaussian measure for which the noise must be state independent. A similar problem appears for processes driven by coloured noise, where the noise itself is modeled by another diffusion process. The singular structure of the diffusion matrix for the joint processes makes the KL divergence finite only when the drift terms of the two driven process are *identical*. This leaves no room for approximations. Finally, it is unclear how to use the variational approach, when measures over discrete variables have to approximated by Gaussians. This seems a quite natural idea e.g. in a chemical reaction processes when the the number of molecules is large enough to be described by a fluctuating continuous random variable.

In this paper we propose a new formulation of the free energy for continuous time Markov processes which is expressed in terms of certain *marginal densities*. Rather than working with approximations of the complex infinite dimensional probabilities over paths we have to approximate only these marginal functions. This allows for a much more flexible application of eg Gaussian approximations. However, one nice feature of the old variational approximation is lost: Since the new variational formulation is of the minimax type, the approximate free energy is no longer an upper bound on the exact negative log–likelihood. As a proof of concept, we demonstrate our method on a few simpler problems for which a 'standard' Gaussian variational approximation would be not be applicable.

2 Continuous time Markov processes

We assume that the evolution of a d–dimensional *state vector* X_t in continuous time t follows a Markov process (for a detailed presentation see [5]). An important example is the class of *diffusion processes* defined by a *stochastic differential equation*

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

with the *drift* vector $f(x)$ and *diffusion matrix* $D(x)$ which is driven by a Wiener process W. The *marginal probability* $p_t(x)$ of X_t obeys a *forward Kolmogorov* equation of the form $\dot{p_t}= \mathcal{L}p_t$ where the dot denotes time derivative and where the operator $\mathcal{L}$ is a second order differential operator acting as $(\mathcal{L}p)(x) = \{-\nabla(fp) + \frac{1}{2}\mathrm{Tr}(\nabla^T\nabla D)\}p(x)$ with $\nabla = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_d})$.

A second case is given by *Markov jump processes* defined through the short time behaviour of the transition probability $p_{t+\Delta t,t}\,(x',|x) \simeq \delta_{x'x} + \Delta t\;\; f\,(x'|x)$ as $\Delta t \to 0$ where f is the rate function. In this case we have a *forward Kolmogorov* equation with $(\mathcal{L}p)(x) = \sum_{x'}\,[p(x')f(x|x') - p(x)f(x'|x)]$.

Another important operator is **the generator** of the Markov process $\mathcal{L}^\dagger$ which is related to the change of functions $G(X_t)$ over short times Δt via

$$E[G(X_{t+\Delta t})|X_t = x] \simeq G(x) + \Delta t\,(\mathcal{L}^\dagger G)(x)\ . \tag{2}$$

For diffusion processes one has $(\mathcal{L}^\dagger G)(x) = \{f^\top(x)\nabla + \frac{1}{2}\mathrm{Tr}(D(x)\nabla^\top\nabla)\}G(x)$ and in case of jump processes $(\mathcal{L}^\dagger G)(x) = \sum_{x'\neq x} f(x'|x)G(x')$.

We will specialise on jump processes of the *reaction model* type, where $f(x'|x)\sum_\alpha h_\alpha(x)\delta_{x'-x,R_\alpha}$ where for each reaction α a jump of the state of size $x' - x = R_\alpha$ occurs with rate $h_\alpha(x)$.

We find it therefore convenient to work with a notation which unifies both reaction and diffusion models by writing

$$(\mathcal{L}_p^\dagger\psi)(x) = \int W(z, x)\psi(x + z)\; dz \tag{3}$$

i.e. we have jumps of size z from the state x each occuring with a weight $W(z, x)$. For reaction systems we have $W(z, x) = \sum_\alpha h_\alpha(x)\delta(z - R_\alpha)$ where δ denotes the Dirac function. For diffusions we can use $W(z, x) = \left\{f^\top(x)\nabla_z + \frac{1}{2}\mathrm{Tr}(D(x)\nabla_z^\top\nabla_z)\right\}\delta(z)$

3 The variational approach to inference

Let us assume that we observe data $D = (y_1, \ldots, y_n)$ at certain discrete times t_i, $i = 1, \ldots, N$ over a time window $0 \le t \le T$ which are probabilistically related to the *entire path* $X_{0:T}$ of the state via a likelihood $p(D|X_{0:T})$. We will assume that this is of the form

$$p(D|X_{0:T}) = e^{-\int_0^T U(X_t,x)\;dt} \tag{4}$$

which includes noisy observations of the state given by a likelihood $p(y|x)$ as

$$U(x, t) \doteq -\sum_i \ln p(y_i|x)\delta(t - t_i) \tag{5}$$

This definition will also apply to the case where the *observation process* is more complicated, e.g. when we identify the y_i with arrival times of events in an inhomogeneous Poisson process with a rate function $\lambda(X_t)$ that depends on the state X_t. For the latter we would have $U(x, t) = -\sum_i \ln\lambda(x)\delta(t - t_i) + \lambda(x)$ which gives a contribution even when no event arrives in $[0 : T]$.

Usually, we are interested in computing the marginal posterior $p_t(x|D)$ of the state X_t based on *all* observations in $[0:T]$ and the marginal probability of the data $p(D) = E[p(D|X_{0:T})]$ where E denotes expectation wrt the prior measure of the process X.

The variational approach to inference is based on the variational formulation of a posterior measure as the optimisation problem $p(\cdot|D) = \arg\min_q \mathcal{F}[q]$ where the *variational free energy* is defined as

$$\mathcal{F} = KL(q\|p_{prior}) - E_q[\ln p(D|\cdot)] \tag{6}$$

where KL denotes the Kullback–Leibler divergence and p_{prior} is the prior measure over state variables.

3.1 The KL divergence

We apply the variational approach to probabilities over paths $q(x_{0:T})$ following [1]. Similar to their work we do not attempt a rigorous measure theoretic approach based on stochastic analysis but rather discretise time in steps of size Δt and work on the finite representation of paths $X_{0:T}$ by $\{X_{k\Delta t}\}_{k=0,\dots,K}$ with $K\Delta t = T$. Both prior and posterior probabilities over paths are Markovian with transition probabilities $p_{t+\Delta t,t}(x'|x)$ and $q_{t+\Delta t,t}(x'|x)$. Using the representation of the joint density in term of conditionals and the Markov property we get (assuming $q_0(x) = p_0(x)$)

$$\begin{aligned} KL\,[q\|p] &= \int dx_{0:T}\, q(x_{0:T}) \ln \frac{q(x_{0:T})}{p(x_{0:T})} \\ &= \sum_{k=0}^{K-1} \int dx\, q_{t_k}(x) \int dx'\, q_{t_{k+1},t_k}(x'|x) \ln \frac{q_{t_{k+1},t_k}(x'|x)}{p_{t_{k+1},t_k}(x'|x)} \\ &= \sum_{k=0}^{K-1} \int dx\, q_{t_k}(x) KL\left[q_{t_{k+1},t_k}(\cdot|x)\|p_{t_{k+1},t_k}(\cdot|x)\right] \end{aligned}$$

in terms of transition and marginal probabilities. Note that we have written expectation in terms of integrals. For discrete random variables these should be replaced by sums.

At this step we deviate from the derivations of [1] Rather than evaluating the KL divergence between conditional probabilities directly for the different types of Markov processes in the limit $\Delta t \to 0$, we will first introduce Lagrange parameters for the dynamics of the marginals which will take the dependency of marginals and transition probabilities into account.

$$q_{t+\Delta t}(x') = \int q_{t+\Delta t,t}(x'|x)\; q_t(x)\; dx\ . \tag{7}$$

Using Lagrange parameter functions $\phi_t(x)$ (one per time slice, with $\phi_T(x) \equiv 0$) we get a new form of the free energy functional which depends on the set of transition probabilities, the marginals and the Lagrange parameters

$$\begin{aligned} \mathcal{F} &= \sum_t \int dx\, q_t(x) \left\{ \int dx'\, q_{t+\Delta t,t}(x'|x) \ln \frac{q_{t+\Delta t,t}(x'|x)}{p_{t+\Delta t,t}(x'|x)} \right\} \\ &+ \sum_t \int dx'\, \phi_t(x') \left\{ q_{t+1}(x') - \int dx\, q_{t+\Delta t,t}(x'|x)\, q_t(x) \right\} + \Delta t \sum_t \int dx\, q_t(x) U(x,t) \end{aligned}$$

where we have used the specific form of the likelihood (4). Now, we can optimise wrt $q_{t+\Delta t,t}(x'|x)$ independently from the marginals and get

$$\frac{q_{t+\Delta t,t}(x'|x)}{p_{t+\Delta t,t}(x'|x)} = \frac{e^{\phi_t(x')}}{E_p[e^{\phi_t(X_{t+\Delta t})}|X_t = x]} \tag{8}$$

where $E_p[e^{\phi_t(X_{t+\Delta t})}|X_t = x] = \int dx' e^{\phi_t(x')} p_{t+\Delta t,t}(x'|x)$. Inserting this result for the optimal transition probabilities into the free energy yields

$$\begin{aligned} \mathcal{F} = \sum_t \int dx'\, \phi_t(x') q_{t+\Delta t}(x') - \sum_t \int dx\, q_t(x) \left\{ \ln E_p[e^{\phi_t(X_{t+\Delta t})}|X_t = x] \right\} \\ + \Delta t \sum_t \int dx\, q_t(x) U(x,t) \end{aligned} \tag{9}$$

3.2 Transition to continuous times

In order to take the limit $\Delta t \to 0$ we use the (2) assuming that the functions ϕ_t can be extended to continuous functions of time. We get

$$\mathcal{F} = \int_0^T dt \left\{ \int dx\, \phi(x,t) \frac{\partial q_t(x)}{\partial t} - \mathcal{H}[q,\phi] \right\} \tag{10}$$

with the functional

$$\mathcal{H}[q,\phi] = \int dx\, q_t(x) \left(e^{-\phi_t(x)} (\mathcal{L}^\dagger e^{\phi_t})(x) - U(x,t) \right) \tag{11}$$

Note, that we have not made any approximations so far. The most important point is that the free energy functional is no longer in terms of measures over entire paths, but over *marginal* probabilities and a set of 'conjugate' functions.

The present formalism resembles somewhat Eyink's approach [3, 4] who developed a method for approximating marginals for Markov processes which has been also used to approximate conditional distributions. However to our knowledge it does not give a method for computing the likelihood $p(D)$.

3.3 Specific forms

Using the form of the generator (3) we have

$$\mathcal{H} = \int dx q_t(x) \left\{ \int W(z,x) e^{\phi_t(x+z) - \phi_t(x)}\, dz - U(x,t) \right\}$$

which for diffusions becomes

$$\mathcal{H} = \int dx q_t(x) \left(f^\top(x) \nabla\phi + \frac{1}{2}\mathrm{Tr}(D(x)\nabla^\top\nabla)\phi_t(x) + \frac{1}{2}\mathrm{Tr}(D(x)(\nabla\phi_t(x))^\top(\nabla\phi_t(x))) - U(x,t) \right) \tag{12}$$

3.4 Variational equations

We next perform a variation of the free energy with respect to q and ϕ. Note, that this is not a joint minimum, because the Lagrange multipliers make this rather into a saddle - point (minimax) problem (this can also be seen when we use an alternative derivation of the result (9) using the variational representation of the KL divergence $KL[q\|p] = \sup_\phi \left\{ E_q[\phi] - \ln E_p[e^\phi] \right\}$). Taking functional derivatives gives

$$\frac{\delta \mathcal{F}}{\delta \phi} = 0 \quad \to \quad \frac{\partial q_t(x)}{\partial t} + q_t(x) e^{-\phi_t(x)} (\mathcal{L}^\dagger e^{\phi_t})(x) - e^{\phi_t(x)} (\mathcal{L} q_t e^{-\phi_t})(x) = 0 \tag{13}$$

$$\frac{\delta \mathcal{F}}{\delta q} = 0 \quad \to \quad -\frac{\partial \phi_t}{\partial t} - e^{-\phi_t(x)} (\mathcal{L}^\dagger e^{\phi_t})(x) + U(x,t) = 0 \tag{14}$$

The meaning of the function ψ_t becomes clear when we do an integration by parts with respect to time in (10) and use equation (14) we arrive at $\mathcal{F} = -\phi_0(x_0)$ assuming deterministic intial conditions of the process. On the other hand we know that at the fixed points, the free energy equals the negative log–likelihood of the data and thus $\phi_0(x) = \ln E\left[e^{-\int_0^T U(X_s,s)\, ds} | X_0 = x_0 \right]$ We can apply the same argument to a process which starts at an arbitrary time t at a state x with the result that $e^{\phi(x,t)} = E\left[e^{-\int_t^T U(X_s,s)\, ds} | X_t = x \right] = \psi_t(x)$.

3.5 Equivalence to forward–backward equations

The variational equations can be simplified by introducing the definitions $\psi_t(x) = e^{\phi_t(x)}$ and $p_t = \frac{q_t}{\psi_t}$. Inserting into (13) and (14) gives after some calculations

$$\dot{p_t} = \mathcal{L} p_t - U p_t \qquad \dot{\psi_t} + \mathcal{L}_p^\dagger \psi_t - U\psi_t = 0 \tag{15}$$

Our results can be understood as a generalisation of the well known forward-backward approach to inference in hidden Markov models [8]. Here the the marginal posterior is expressed as $p_t(x|D) \propto p_t(x)\psi_t(x)$ where $p_t(x) = p_t(x|D_{<t})$ is the filtering probability (i.e. the conditioning is on the past data) which follows the filtering equation (15, right-hand side) and where $\psi_t(x) = E[D_{>t}||X_t = x] = E\left[e^{-\int_t^T U(X_s,s)\,ds}|X_t = x\right]$ is the likelihood of future observation conditioned on the present state being x which fulfils the *backward equation* (15) (to be solved backward in time with $\psi_T(x) = 1$).

4 Gaussian approximations

Approximate inference methods can be obtained by restricting marginal probabilities q_t and the ϕ_t to simpler *parametric families* of functions. A natural choice is to approximate q_t by a Gaussian density $q_t \approx \mathcal{N}(m(t), S(t))$. A corresponding approximation

$$\phi_t(x) \approx b^\top x + \frac{1}{2}(x - m(t))^\top A(t)(x - m(t)) \tag{16}$$

seems also reasonable. It is motivated by the fact that it is *exact for a Gaussian posterior process* such as the *Ornstein Uhlenbeck model* which is a diffusion process defined by $f = -\gamma x$ and $D = \sigma^2$ (we also have to assume that the 'data' term $U(x)$ is a quadratic form). The functions b and A can be understood as as Lagrange multipliers which enforce consistency only for the first two moments rather than for the full marginal distribution (7). Note, that we are no longer having the guarantee that the variational free energy provides an upper bound on the true negative log–likelihood.

Inserting these approximations into (10) we arrive at

$$\mathcal{F} = \int_0^T dt \left\{ b(t)^\top \dot{m}(t) + \frac{1}{2}\mathrm{Tr}(A(t)\dot{S}(t)) - \mathcal{H}(m, S, b, A) \right\}$$

where $\mathcal{H}(m, S, b, A)$ is obtained by inserting (16) into (11). In contrast to the variational approximation of [1] our approach does not lead to a consistent approximation *for the measure over paths* ! This is because we are using a Gaussian approximation *only for the marginals* q_t. The transition probabilities (8) are still not Gaussian. Hence, we do not run into trouble with infinite KL divergencies when applying our method to state dependent diffusions.

Using the general form (3) we have

$$\mathcal{H}(m, S, b, A) = \int \exp\left(b^\top z + \frac{1}{2}z^\top A z\right) E\left[W(z, m + u)e^{z^\top u}\right]\,dz - E[U(m + u, t)] \tag{17}$$

where we have set $x = m + u$ with $u \sim \mathcal{N}(0, S)$.

The variational equations are (note that $S_{ij} = S_{ji}$ there is an extra factor 1/2 for the diagonal elements of A and S.

$$\dot{m}_i = \frac{\partial \mathcal{H}}{\partial b_i} \quad \dot{b}_i = -\frac{\partial \mathcal{H}}{\partial m_i} \qquad \dot{S}_{ij} = \frac{\partial \mathcal{H}}{\partial A_{ij}} \quad \dot{A}_{ij} = -\frac{\partial \mathcal{H}}{\partial S_{ij}}\; j \neq i \quad \dot{S}_{ii} = 2\frac{\partial \mathcal{H}}{\partial A_{ii}} \quad \dot{A}_{ii} = -2\frac{\partial H}{\partial S_{ii}} \tag{18}$$

It is interesting to note, that these equations are of the type of *Hamilton's equations* well known from classical mechanics (and also used extensively in the hybrid Monte Carlo method by machine learners), where m and S play the role of generalised coordinates and b_i and S those of generalised momenta (forgetting a trivial factor of 2 for the diagonal elements) and where $\mathcal{H}$ is the Hamilton function. This fact could play an important role in efficient and stable methods for solving these equation using symplectic ODE solvers.

4.1 Weak noise limit

It can be shown for the diffusion case with additive noise treated by [1], that their variational approximation does not coincide with ours. However, our method is plagued by similar computational problems. We have to solve a boundary value problem, where typically the values for m, S are known at $t = 0$, but those for b and A are known at the end time, ie $b(T) = A(T) = 0$. Hence, the solution of the variational equations requires an iterative procedure which for the matrix ODEs can be time consuming when the dimensionality of the problem is large. To avoid that problem, we introduce a further approximation which decouples the fluctuation terms from the means. Assuming that fluctuations are not large we ignore them completely in equation (17) replacing $\mathcal{H}(m, S, b, A)$ by

$$\mathcal{H}_0(m, b) = \int e^{b^\top z}\, W(z, x)dz - U(m, t) \tag{19}$$

which for diffusion process gives $\mathcal{H}_0(m, b) = f^\top(m)b + \frac{1}{2}b^\top D(m)b - U(m, t)$. We then solve the variational equations for $m(t)$ and $b(t)$. In the second step, we fix $m(t)$ and $b(t)$ in and expand $\mathcal{H}$ in (17) around $u = 0$ up to second order in u and also up to second order in A (we are guided by the fact that this gives the exact result for a Gaussian model). The resulting powers in z, which are generated from these expansions can also be generated by derivatives of $\mathcal{H}_0(m, b)$ with respect to b. Also Gaussian expectations over polynomials in u can be expressed by similar derivatives wrt m. After a somewhat lengthy and tedious calculation, we arrive at the result

$$\mathcal{H} \approx \mathcal{H}_0(m, b) + \mathrm{Tr}\left(FSA + \frac{1}{2}\mathcal{D}A + \frac{1}{2}\mathcal{D}ASA\right) - \frac{1}{2}\mathrm{Tr}(SV(m, t))$$

where

$$\mathcal{D}_{ij} = \frac{\partial^2 \mathcal{H}_0}{\partial b_i \partial b_j} \qquad F_{ij} = \frac{\partial^2 \mathcal{H}_0}{\partial b_i \partial m_j} \qquad V_{ij} = \frac{\partial^2 U(m, t)}{\partial m_i \partial m_j} \tag{20}$$

and where we assume that the m and b in the matrices $\mathcal{D}$ and F and V are taken as fixed when we perform variation wrt A and S. We get

$$\dot{A} = -F^\top A - AF - A\mathcal{D}A + V \qquad \dot{S} = (F + \mathcal{D}A)S + S(F^\top + A\mathcal{D}) + \mathcal{D}$$

For the case of simple noisy observations (5) with Gaussian noise, $p(y|x) = \mathcal{N}(0, \sigma_o^2)$, the data terms U and V lead to the simple jump condition $b(t_i^+) = b(t_i^-) + \frac{m(t_i) - y_i}{\sigma_o^2}$ and $A(t_i^+) = A(t_i^-) + I\frac{1}{\sigma_o^2}$ where I is the unit matrix and $t_i^\pm$ are times just before and after the jump. Using the variational equations of motion again we can show that within the weak noise expansion, the free energy is

$$\mathcal{F}_{wn} = \int_0^T dt \left\{ b(t)^\top \dot{m}(t) - \mathcal{H}_0(m(t), b(t)) - \frac{1}{2}\mathrm{Tr}(\mathcal{D}A)dt \right\}$$

It should be noted that the present weak noise expansion does not coincide with the approximations presented by [8]. The latter ones assumes essentially that the system performs small fluctuations around a *fixed deterministic motion* which can be treated within a Gaussian approximation. This method would however fail to compute he likelihood of a rare event (when e.g. a data point is highly unlikely for the model and is far away from the deterministic path) in a proper way. The present approximation is expected to work also in such cases of large deviations as we will see next for a toy example of a Poisson process.

5 Examples

5.1 A simple Poisson process

This is described by single reaction with constant rate $h(x) = \lambda$ and unit jump $R = 1$. If we assume a single noise free integer observation y at $t = T$, i.e. $X_T = y$, and $X_0 = 0$ one can show that the marginal posterior is *binomial* $p_t(x|D) = \frac{y!}{x!(y-x)!}\left(\frac{t}{T}\right)^x \left(1 - \frac{t}{T}\right)^{y-x}$ with mean yt/T and variance $yt/T(1 - t/T)$. Surprisingly, the posterior does not depend on the original rate of the process. The likelihood is $p(y) = \frac{\lambda^y}{y!}e^{-\lambda T}$. In our weak noise approximation, we compute a Gaussian approximation with the same mean and variance. The free energy is $\mathcal{F} = -\ln p(y) = y \ln\left(\frac{y}{\lambda T e}\right) + \lambda T + \frac{1}{2}\ln(2\pi y)$ which would correspond to

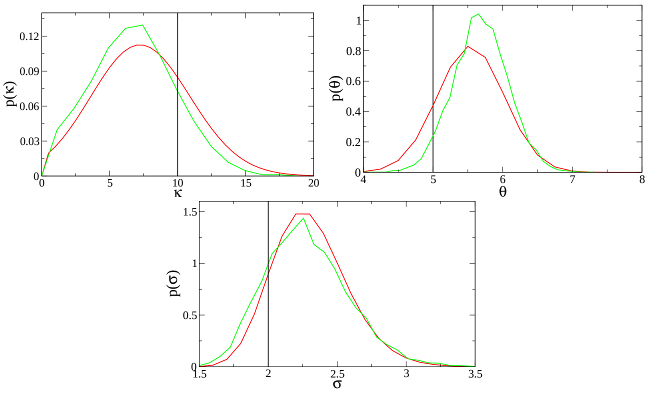

Figure 1: Comparison of parameter prediction using MCMC (green) and Free energy methods (red).

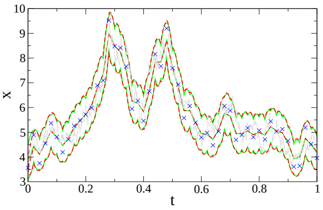

Figure 2: . Comparison of path prediction using MCMC (green) and Free energy methods (red). The middle line denotes the mean prediction, and the upper and lower lines denote the 2 standard deviation confidence bounds.

a simple Stirling-type of approximation of the factorial in the exact result. Repeating the calculation in an 'ordinary' small noise approximation [8] leads to a posterior of the form $p_t(x|D) \sim \mathcal{N}(yt/T, \lambda t(1-t/T))$ i.e. a wrong variance. Both agree only at the maximum likelihood estimate $\lambda_{ML} = y/T$, i.e. when the data point is typical wrt the model.

5.2 Cox-Ingersoll-Ross model

This is defined by the SDE [**?**]

$$dX_t = -\kappa(X_t - \theta)dt + \sigma\sqrt{X_t}dW_t$$

which is used for modeling interest rates. Here the Hamilton function is

$$\mathcal{H}_0(m, b) = \kappa(\theta - m)b + \frac{1}{2}\sigma^2 b^2 m.$$

We have simulated 50 data point using $\kappa = 10, \theta = 5, \sigma = 2$ and observation noise with standard deviatiation $\sigma_o = 0.4$. We have then computed the likelihood $p(D|\theta, \kappa, \sigma) \approx e^{\mathcal{F}}$ using the approximate free energy. Keeping two parameters fixed, we have obtained an approximation to the posterior of the 3rd parameter using a flat prior in large enough interval from this likelihood. Figure presents a comparison with an MCMC sampler for SDE inference based on [6]. Fig shows a comparison with the mean prediction of the path and two standard deviations based on the approximation and on the MCMC sampler. Both show excellent agreements.

5.3 Bicoid reaction model

This is a stochastic *reaction* system which has been suggested for the *Bicoid* protein evolution in *Drosophila*. If we assume a *compartment* model [**?**], where space is discretized into small cells of size Δx. (assuming a single space dimension for simplicity) The state is described by the vector $X = (x_1, \ldots, x_N)$ with x_i denoting the number of molecules in cell i. The model consists of the following processes

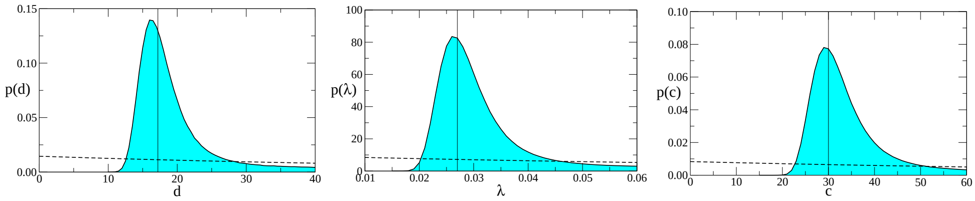

Figure 3: Results of the Laplace approximation for the Bicoid model. The posterior density is plotted as solid line, while the vertical line marks the value used to generate the observations and the prior is shown as dashed line.

- *Particle diffusion*, say from cell i into the neighboring cell $i+1$, is modelled by a transition $x'_i = x_i - 1$ and $x'_{i+1} = x_{i+1} + 1$ with a rate $dx_i/\Delta x^2$, where d is the diffusion constant of molecules the system. The proportionality of the rate to the number n_i of molecules accounts for the fact that each of the molecules can perform the jump to the neighboring cell.
- *Molecule degradation* $x'_i = x_i - 1$ which occurs with a rate $\lambda\, x_i$ in all cells.
- *Creation of molecules* in the system occurs by injection with a fixed rate c in a single cell only.

We have used $N = 8$ cells and have taken the parameters $d = 17.2$, $\lambda = 0.027$, $c = 30$. We generated data $8 -$ *dimensional* data samples at 11 equidistant time points using the Gillespie algorithm [], which were corrupted with Gaussian noise of variance $\sigma_o^2 = 0.08$.

Approximate inference for this model has been discussed within a variational mean field approach [**?**] and a weak noise approximation [9]. Here we have used the Hamiltonian approach in order to compute the likelihood and from there the marginal posterior distributions for each of the three parameters d, λ, c by using the Laplace approximation. As prior for each parameter we have assumed a exponential distribution with mean equal to four times the value used to generate the observations.

6 Discussion

The free energy approximation based on the weak noise approach has shown promising performance for simple models or models of medium complexity. One could easily extend this to other models, such as coloured noise problems. However, one may see our free energy approach as a more general framework for approximations. It would be interesting to find other, problem specific parametric choices for q_t and ϕ_t which on the one hand should be chosen rich enough to give good approximations to the exact forms but on the other hand should be simple enough to lead to tractable computations.

References

[1] Cedric Archambeau, Dan Cornford, Manfred Opper, and John Shawe-Taylor. Gaussian process approximations of stochastic differential equations. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 1(1):1–16, 2007.

[2] Ido Cohn, Tal El-hay, Nir Friedman, and Raz Kupferman. Mean field variational approximation for continuous-time bayesian networks. *Journal of Machine Learning Research*, 11:2745–2783, 2010.

[3] Gregory L. Eyink. Action principle in nonequilibrium statistical dynamics. *Phys. Rev. E*, 54(4):3419–3435, 1996.

[4] Gregory L. Eyink, Julian M. Restrepo, and Francis J. Alexander. A mean field approximation in data assimilation for nonlinear dynamics. *Physica D*, 195:347–368, 2004.

[5] C. W. Gardiner. *Handbook of Stochastic Methods*. Springer, Berlin, second edition, 1996.

[6] A. Golightly and D. J. Wilkinson. Markov chain monte carlo algorithms for sde parameter estimation. chapter 9, pages 253–275. MIT Press, 2010.

[7] Manfred Opper and Guido Sanguinetti. Variational inference for Markov jump processes. In *Advances in Neural Information Processing Systems 20*, 2007.

[8] Andreas Ruttor and Manfred Opper. Efficient statistical inference for stochastic reaction processes. *Phys. Rev. Lett.*, 103(23):230601, 2009.

[9] Andreas Ruttor and Manfred Opper. Approximate inference in reaction-diffusion processes. In *Proceedings of the Thirteenth Interhantional Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.