Smoothing Estimates of Diffusion Processes

H.C. Ruiz Euler*
Donders Institute
Radboud University
Nijmegen, NL
hruiz@science.ru.nl

H.J. Kappen
Donders Institute
Radboud University
Nijmegen, NL
bkappen@science.ru.nl

Abstract

We present a novel method to sample efficiently from the posterior distribution over hidden diffusion processes given a time series of noisy observations. We use high-performance computing to increase iteratively the effective sample size (ESS) of the posterior, or "smoothing" distribution. This is done using an adaptive importance sampler based on the path integral cross entropy (PICE) control theory [16]. We call this method PICE Smoother (PICES). The novelty of this method vis-à-vis [16] and [24] is the implementation of the learning rule (5) to deep neural networks, its comparison to the bootstrap particle filter-smoother and its application to fMRI data.

Introduction

Given a time series of J observations $y_{0:T} = (y_{t_1}, y_{t_2}, \dots, y_{t_J})$ with $0 \leq t_1 < \cdots < t_J \leq T \in \mathbb{R}$ we are interested in the posterior distribution over continuous time hidden processes $x_{[0,T]} = \{x_t\}_{t\in[0,T]}$ on the interval $[0, T] \subset \mathbb{R}$. This posterior distribution is conditioned on two aspects of the model. First, a stochastic differential equation (SDE) describes the dynamics of the hidden state,

$$dx_t = F(x_t, t)dt + \sigma_{dyn}(x_t, t)dW_t \tag{1}$$

where $dW_t \sim \mathcal{N}(0, dt)$ is Gaussian distributed with variance dt. Second, the observation model $g(y|x)$ gives the probability to observe y conditioned on the latent state x.

When a time discretization dt is chosen, this process corresponds to a first order Markov process with a transition probability $f(x_{t+dt}|x_t)$ given by a Gaussian with mean $x_t + F(x_t, t)dt$ and a covariance $\sigma^2_{dyn}(x_t, t)dt$. This defines a prior distribution $p(x_{[0,T]}) := \prod_{s=0}^{T} f(x_{s+dt}|x_s)p(x_0)$ over all paths following (1), where $s \in [0, T]$ and $p(x_0)$ is the prior distribution over initial conditions. We call a "particle" an entire path generated by this type of processes.

Additionally, the observation model defines the likelihood $p(y_{0:T}|x_{[0,T]}) := \prod_{j=1}^{J} g(y_{t_j}|x_{t_j})$. Thus, the smoothing distribution can be written as

$$p(x_{[0,T]}|y_{0:T}) \propto p(x_{[0,T]})\exp\left(\sum_{j=1}^{J} \log\left[g(y_{t_j}|x_{t_j})\right]\right). \tag{2}$$

The estimation of the statistics or marginals of the smoothing distribution is in general intractable when (1) is non-linear or $g(y|x)$ non-Gaussian. One may use sequential Monte Carlo (SMC) sampling

*The work of H.-Ch. Ruiz Euler is supported by the European Commission through the FP7 Marie Curie Initial Training Network 289146, NETT: Neural Engineering Transformative Technologies. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

or particle methods to overcome this difficulty, e.g. [18, 9, 7, 12]. Nevertheless, a major challenge for sampling from the smoothing distribution is the weight degeneracy of most smoothing procedures. The consequence is a poor representation of the smoothing distribution and a small ESS, which is defined as $N_{eff} = N/\mathbb{E}[w^2]$, where w are the normalized importance weights [19].

In general, weight degeneracy appears whenever the high-density region of the prior has little overlap with the likelihood of the data. Hence, only a few of the importance weights are sufficiently dominant to have a contribution in the estimation of the target statistics. In the case of the smoothing distribution, this effect increases exponentially with the number of observations [4]. The problem becomes acuter whenever the generative model $p(x_{[0,T]})$ inaccurately approximates the posterior dynamics. For this reason, much work as been devoted to reducing the weight degeneracy and improve smoothing estimates, e.g. [13, 25, 2, 1, 6, 8, 21].

The method proposed here significantly reduces the degeneracy of the particles by adapting the process with a feedback controller. Hence, the problem of the smoothing estimate translates to a problem of stochastic optimal control. The relation between filtering of continuous-time hidden processes and control is not new. In [20] it was shown that the solution to the Kushner-Stratonovich equation for the normalized probability density conditioned on noisy measurements is given in terms of a normalized Feynman-Kac formula similar to the solution of the PI control theory [15]. More recently, [25] shows that for a general non-linear non-Gaussian problem, the optimal Kalman gain can be computed at each time as an Euler-Lagrange boundary value problem. However, this is restricted to one-dimensional diffusion processes only. In [23] the authors propose an approach that describes particles whose density evolves according to a Fokker-Planck equation controlled by a Hamilton-Jacobi-Bellman equation. Unfortunately, it is not recognized that the solution to the optimal cost-to-go function can be given formally.

Method

In [16] the cross-entropy method [5] is applied to the Path Integral control problem [15]. This gives a gradient descent method (PICE) to minimize the KL divergence between the target distribution and the distribution given by the controlled process

$$dx_t = F(x_t, t)dt + u_\theta(x_t, t)dt + \sigma_{dyn}(x_t, t)dW_t \tag{3}$$

where $u_\theta(x_t, t)$ is the feedback controller with an arbitrary parametrization. This method gives an adaptive importance sampler with particles generated according to (3). At each iteration, starting with some controller, usually $u_\theta(x, t) = 0$, we sample $i = 1, \ldots, N$ particles from (3) and compute their corresponding importance weight $w_i = \exp(-S_i) / \sum_{j=1}^{N} \exp(-S_j)$ defined by the cost of each particle,

$$S_i := -\sum_{k=1}^{J} \log\left[g(y_{t_k}|x^i_{t_k})\right] + \frac{1}{2}\int_0^T u^\dagger_\theta(x^i_s, s)u_\theta(x^i_s, s)ds + \int u^\dagger(x^i_s, s)dW^i_s \tag{4}$$

where $u^\dagger$ denotes the transposed and dW^i_s is the noise realization of particle i at time s. The first term comes from the likelihood of the data, the second and third term come from the importance sampling correction for diffusion processes given by Grisanov's theorem [11]. The form of the Grisanov correction can be easily understood by considering the infinitesimal transition density $f(x_{s+dt}|x_s, u)$ given the controlled process. This is proportional to the product of the transition density of the uncontrolled process and the correction term for an infinitesimal time step $f(x_{s+dt}|x_s, u = 0)\exp(\frac{1}{2}u^\dagger_\theta(x_s, s)u_\theta(x_s, s)ds + u^\dagger(x_s, s)dW_s)$.

The normalized weights $\{w_i\}_{i=1}^N$ in the $l - th$ iteration are then used to compute the update rule for the parameters $\hat{\theta}_l$ of the feedback control function $\hat{u}_{\hat{\theta}}(x, t)$,

$$\hat{\theta}_{l+1} = \hat{\theta}_l + \eta \left\langle \int_0^T \left[\hat{u}_{\hat{\theta}}(x_s, s) - \left(u_\theta(x_s, s) + \frac{dW_s}{ds}\right)\right] \frac{\partial \hat{u}_{\hat{\theta}}(x_s, s)}{\partial \hat{\theta}_l} ds \right\rangle_u \tag{5}$$

where $\langle h(x)\rangle_u := \sum_{i=1}^N w_i h(x^i)$ is the weighted average with weights (4) and η is a learning rate [16]. Notice the dependency of S_i on the feedback controller $u_\theta(x, t)$ used to sample the particles.

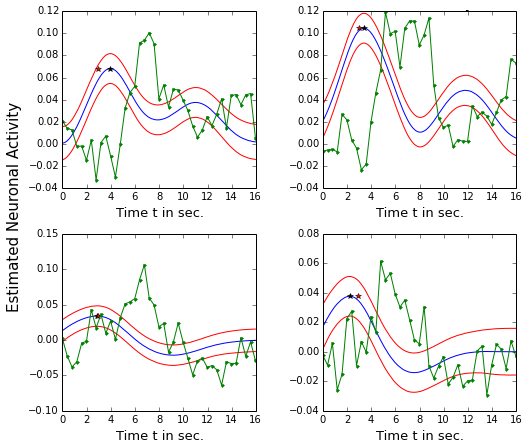

Figure 1: Examples of smoothing estimates of the hidden neuronal activity in the motor cortex (blue: mean; red: standard deviation). Black star: maximum of the mean estimate; red star: measured response time.

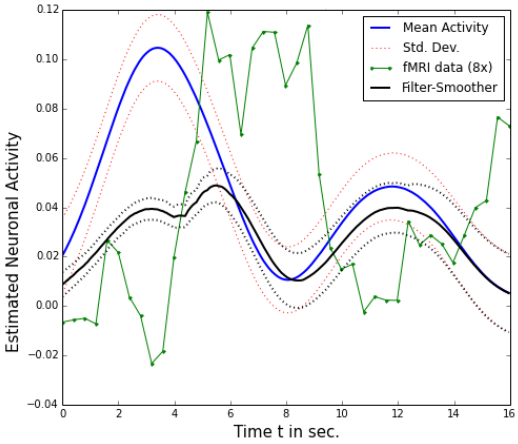

Figure 2: Comparison of estimates: standard Bootstrap Filter-Smother (black) and PICES estimate (blue). The values of the fMRI time-series are scaled with a factor of 8 for comparison (green). The CPU time for both methods is about 40 minutes.

A good parametrization of the controller is crucial because better approximations to the optimal control result in more efficient importance samplers. However, the design of a "good" controller is a non-trivial task. Thus, we need a flexible parametrization to approximate a complex function without specifying the concrete features needed for good performance. In this regard, deep neural networks are a natural candidate. Since PICE places no restriction on the parametric form of $\hat{u}_{\hat{\theta}}(x,t)$, we choose a deep neural network.

Notice that in 5 we distinguish between the controller used to sample particles $u_\theta(x,t)$, and the controller $\hat{u}_{\hat{\theta}}(x,t)$ that is being updated. The expression provides an unbiased estimate of the gradient for any choice of $u_\theta(x,t)$, but the variance of the estimate depends on the controller. A particular choice that we take here is $u = \hat{u}$, i.e. sample particles with the most recently constructed sampler $\hat{u}$.

The initialization of the particles at time $t = 0$ requires special care. A priori, the particles are distributed according to $p(x_0)$, but this distribution changes to the posterior marginal $p(x_0|y_{0:T})$. In general, it is hard to sample from this distribution. For this purpose, we introduce a Gaussian as importance sampler at the initial time which is adapted at each iteration using the statistics of the marginal obtained in the previous iteration [24].

Results

An important example of hidden state estimation from indirect sparse measurements comes from neuroscience. Non-invasive methods to measure brain activity, e.g. fMRI, are important to understand cognitive processes in the human brain. However, the measurement of the brain activity through fMRI is indirect and delayed due to the hemodynamics.

We use PICES to estimate the neuronal activity of a subject during an experiment. In it, visual or auditory stimuli were presented and the subject had to press a button as soon as the stimulus was perceived [22]. For the estimations, we consider the motor cortex as the region of interest (ROI) modeled by a 5-dimensional hidden state where the neuronal activity z follows the one-dimensional stochastic dynamics,

$$dz = -zdt + \sigma_z dW_z. \tag{6}$$

The way in which the neural process affects the BOLD response is modeled in a standard way involving four coupled nonlinear differential equations, two for the Hemodynamic equations [10] and two for the Balloon model. The observation model is a Gaussian with mean given by a non-linear function of the states of the Balloon model. This function represents the Bold signal change [3].

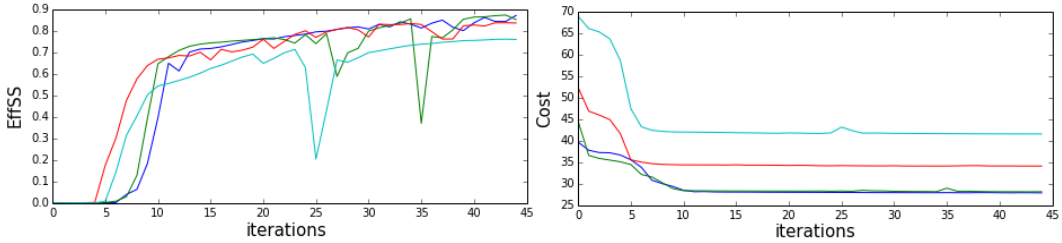

Figure 3: Effective Sampling Size vs iterations (left) for four fMRI time series: The initial fraction of ESS is around 3×10^{-6}, i.e. only one particle out of the $N = 3 \times 10^5$ has a significant contribution to the estimations. Right: The mean cost $\mathbb{E}_u[S]$ under the control at each iteration.

The objective is to reconstruct the neuronal signal z from the observed Bold signal. Notice that we disregard any inputs to the ROI (from visual or auditory stimuli) and consider them as unknown.

Each of the 30 stimuli presented to a subject gives an fMRI time series consisting of 41 observations at an interval of $\Delta t = 0.399\ s$. We estimate the hidden neuronal activity by iterating the learning procedure 45 times. Per learning iteration we use 500 CPUs (or "workers") of the Dutch high-performance computing platform Cartesius with 600 particles each. As a controller, we use a neural network with 8 hidden layers and 50 nodes each. All nodes have a rectified linear transfer function. In figure 1 we present four examples of the estimation using PICES. Notice that although no input is assumed, the estimated neuronal response has a clear peak close to the experimentally measured response time.

For comparison, we estimate the smoothing distribution over the neuronal activity with a standard Bootstrap Filter-Smoother (BFS [17]). We use 500 workers with 10^4 particles per CPU. In each worker, we estimate the mean and variance of the smoothing distribution. Since the effective sampling size of the BFS deteriorates for early times, the variance of the estimates at these times is large. For this reason, the estimations are done using 90 forward passes on each CPU to get better estimates.

Figure 2 compares the estimates from PICES and BFS for one of the 30 fMRI time courses that we analyzed. The result shows that the BFS has problems estimating a larger amplitude of the neuronal activity. This is due to the fact that the BFS cannot reconstruct a posterior which significantly deviates from the prior due to the lack of input in the model. In contrast, the control drift $u(x_t, t)$ in PICES accounts for the lack of inputs in our model.

Finally, in figure 3 left we show how the fraction of the ESS (EffSS) increases several orders of magnitude, from around 3×10^{-6} to $0.7 - 0.85$. On the right, the mean cost of the particles in each iteration. The results show how reducing the mean cost S is accompanied by an increase in the ESS of the particles.

Conclusion

In this paper, we propose a stochastic optimal control method to estimate posterior processes, as an alternative to particle smoothing. In these problems, it is expected that the prior dynamics give a poor representation of the posterior process, which explains the poor performance of the Bootstrap Filter-Smoother. Our results on fMRI show that the adaptive importance sampling using a controlled diffusion process improves the efficiency of the sampler several orders of magnitude. We can learn an arbitrary parametrized controller. Here, we used a deep neural network. We found that a shallow network or a simpler linear feedback controller performed significantly worse in this case (results not shown).

Both the sampling and gradient computations are easy to parallelize and can be implemented efficiently in a distributed manner. This is in strong contrast with the standard SMC methods where the resampling is a bottleneck when considering a distributed implementation and extra considerations are needed [14].

The proposed method can be applied to a large range of problems where the dynamics of the hidden state can be described by the general SDE (1). In addition, there is no constraint imposed on the

observation model. Thus, PICES is applicable in many inference problems with different modalities of data, e.g. spike inference from Ca-Imaging recordings or model based decoding of spike trains. This makes PICES a promising alternative to the current particle smoothing methods.

References

[1] Mark Briers, Arnaud Doucet, and Simon Maskell. Smoothing algorithms for state–space models. *Annals of the Institute of Statistical Mathematics*, 62(1):61–89, 2010. (document)

[2] Pete Bunch and Simon Godsill. Improved particle approximations to the joint smoothing distribution using markov chain monte carlo. *Signal Processing, IEEE Transactions on*, 61(4):956–963, 2013. (document)

[3] Richard B Buxton, Eric C Wong, and Lawrence R Frank. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic resonance in medicine*, 39(6):855–864, 1998. (document)

[4] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer New York, 2006. (document)

[5] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005. (document)

[6] Randal Douc, Aurélien Garivier, Eric Moulines, and Jimmy Olsson. Sequential monte carlo smoothing for general state space hidden markov models. *The Annals of Applied Probability*, 21(6):2109–2145, Dec 2011. (document)

[7] Arnaud Doucet and Adam M Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12:656–704, 2009. (document)

[8] Cyrille Dubarry and Randal Douc. Improving particle approximations of the joint smoothing distribution with linear computational cost. In *Statistical Signal Processing Workshop (SSP), 2011 IEEE*, pages 209–212. IEEE, 2011. (document)

[9] Paul Fearnhead, David Wyncoll, and Jonathan Tawn. A sequential smoothing algorithm with linear computational cost. *Biometrika*, 97(2):447–464, 2010. (document)

[10] Karl J Friston, Andrea Mechelli, Robert Turner, and Cathy J Price. Nonlinear responses in fmri: the balloon model, volterra kernels, and other hemodynamics. *NeuroImage*, 12(4):466–477, 2000. (document)

[11] Igor Vladimirovich Girsanov. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability & Its Applications*, 5(3):285–301, 1960. (document)

[12] Simon J Godsill, Arnaud Doucet, and Mike West. Monte carlo smoothing for nonlinear time series. *J. Amer. Statist. Assoc.*, 99(465):156–168, Mar 2004. (document)

[13] Shixiang Gu, Zoubin Ghahramani, and Richard E Turner. Neural adaptive sequential monte carlo. In *Advances in Neural Information Processing Systems*, pages 2629–2637, 2015. (document)

[14] Soren Henriksen, Adrian Wills, Thomas B Schön, and Brett Ninness. Parallel implementation of particle mcmc methods on a gpu. *IFAC Proceedings Volumes*, 45(16):1143–1148, 2012. (document)

[15] Hilbert J Kappen. Linear theory for control of nonlinear stochastic systems. *Physical review letters*, 95(20):200201, 2005. (document)

[16] H.J. Kappen and H.C. Ruiz. Adaptive importance sampling for control and inference. *Journal of Statistical Physics*, 2016. (document)

[17] Genshiro Kitagawa. Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of computational and graphical statistics*, 5(1):1–25, 1996. (document)

[18] Fredrik Lindsten and Thomas B Schön. Backward simulation methods for monte carlo statistical inference. *Foundations and Trends in Machine Learning*, 6(1):1–143, 2013. (document)

[19] Jun S Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119, 1996. (document)

[20] Sanjoy K. Mitter. *Nonlinear filtering of diffusion processes a guided tour*, volume 42 of *Lecture Notes in Control and Information Sciences*, chapter chapter 23, pages 256–266. Springer-Verlag, 1982. (document)

[21] Lawrence Murray and Amos Storkey. Particle smoothing in continuous time: A fast approach via density estimation. *Signal Processing, IEEE Transactions on*, 59(3):1017–1026, 2011. (document)

[22] Mayur Narsude, Daniel Gallichan, Wietske Van Der Zwaag, Rolf Gruetter, and José P Marques. Three-dimensional echo planar imaging with controlled aliasing: A sequence for high temporal resolution functional mri. *Magnetic resonance in medicine*, 2015. (document)

[23] Sergio Pequito, Pedro Aguiar, Bruno Sinopoli, and Diogo Gomes. Nonlinear estimation using mean field games. In *NetGCOOP 2011: International conference on NETwork Games, COntrol and OPtimization*. IEEE, 2011. (document)

[24] H-Ch Ruiz and HJ Kappen. Particle smoothing for hidden diffusion processes: Adaptive path integral smoother. *arXiv preprint arXiv:1605.00278*, 2016. (document)

[25] Tao Yang, Prashant G Mehta, and Sean P Meyn. Feedback particle filter. *Automatic Control, IEEE Transactions on*, 58(10):2465–2480, 2013. (document)