Efficient Monte Carlo sampling by parallel marginalization

Jonathan Weare[*] [†]

[†]Department of Mathematics, University of California, Berkeley CA 94720-3840

Markov chain Monte Carlo sampling methods often suffer from long correlation times. Consequently, these methods must be run for many steps to generate an independent sample. In this paper a method is proposed to overcome this difficulty. The method utilizes information from rapidly equilibrating coarse Markov chains that sample marginal distributions of the full system. This is accomplished through exchanges between the full chain and the auxiliary coarse chains. Results of numerical tests on the bridge sampling and filtering/smoothing problems for a stochastic differential equation are presented.

Markov chain Monte Carlo | renormalization | multi-grid | filtering | parameter estimation

In order to understand the behavior of a physical system it is often necessary to generate samples from complicated high dimensional distributions. The usual tools for sampling from these distributions are Markov chain Monte Carlo methods (MCMC) by which one constructs a Markov chain whose trajectory averages converge to averages with respect to the distribution of interest. For some simple systems it is possible to construct Markov chains with independent values at each step. In general, however, spatial correlations in the system of interest result in long correlation times in the Markov chain and hence slow convergence of the chain's trajectory averages. In this paper, a method is proposed to alleviate the difficulties caused by spatial correlations in high dimensional systems. The method, parallel marginalization, is tested on two stochastic differential equation conditional path sampling problems.

Parallel marginalization takes advantage of the shorter correlation lengths present in marginal distributions of the target density. Auxiliary Markov chains that sample approximate marginal distributions are evolved simultaneously with the Markov chain that samples the distribution of interest. By swapping their configurations, these auxiliary chains pass information between themselves and with the chain sampling the original distribution. As shown below, these swaps are made in a manner consistent with both the original distributions and the approximate marginal distributions. The numerical examples indicate that improvement in efficiency of parallel marginalization over standard MCMC techniques can be significant.

The design of efficient methods to approximate marginal distributions was addressed in [1] by Chorin and in [2] by Stinis. The use of Monte Carlo updates on coarse subsets of variables is not a new concept (see [3] and the references therein). The method presented in [3] does not use marginal distributions. However, attempts have been made previously to use marginal distributions to accelerate the convergence of MCMC (see [4, 5]). In contrast to parallel marginalization, the methods proposed in [4] and [5] do not preserve the distribution of the full system and therefore are not guaranteed to converge. The parallel construction used here is motivated by the parallel tempering method (see [6]), and allows efficient comparison of the auxiliary chains and the original chain. See references [6] and [7] for expositions of standard MCMC methods.

Parallel marginalization for problems in Euclidean state spaces is described in detail in the next two sections. In the final sections the conditional path sampling problem is described and numerical results are presented for the bridge sampling and smoothing/filtering problems.

Parallel Marginalization

For the purposes of the discussion in this section, we assume that appropriate approximate marginal distributions are available. As discussed in a later section, they may be provided by coarse models of the physical problem as in the examples below, or they may be calculated via the methods in [1] and [2].

Assume that the d_0 dimensional system of interest has a probability density, $\pi_0(x_0)$, where $x_0 \in \mathbb{R}^{d_0}$. Suppose further that, by the Metropolis-Hastings or any other method (see [6]), we can construct a Markov chain, $Y_0^n \in \mathbb{R}^{d_0}$, which has π_0 as its stationary measure. That is, for two points $x_0, y_0 \in \mathbb{R}^{d_0}$

$$\int T_0(x_0 \to y_0)\pi_0(x_0)\, dx_0 = \pi_0(y_0)$$

where $T_0(x_0 \to y_0)$ is the probability density of a move to $\left\{Y_0^{n+1} = y_0\right\}$ given that $\{Y_0^n = x_0\}$. Here, n is the algorithmic step. Under appropriate conditions (see [6]), averages over a trajectory of $\{Y_0^n\}$ will converge to averages over π_0, i.e. for an objective function $g(x_0)$

$$\frac{1}{N}\sum_{n=0}^{N-1} g\left(Y_0^n\right) \to \mathbf{E}\left[g\left(X_0\right)\right]$$

The size of the error in the above limit decreases as the rate of decay of the time autocorrelation

$$\mathbf{corr}\left[g\left(Y_0^n\right), g\left(Y_0^0\right)\right] = \frac{\mathbf{E}\left[\left(g\left(Y_0^n\right) - \mathbf{E}\left[g\left(X_0\right)\right]\right)\left(g\left(Y_0^0\right) - \mathbf{E}\left[g\left(X_0\right)\right]\right)\right]}{\mathbf{Var}\left[g\left(X_0\right)\right]}$$

increases. In this formula, Y_0^0 is assumed to be drawn from π_0.

It is well known that judicious elimination of variables by renormalization can reduce long range spatial correlations (see e.g. [8]). The variables are removed by averaging out their effects on the full distribution. If the original density is $\pi(\hat{x}, \tilde{x})$

Conflict of interest footnote placeholder

Insert 'This paper was submitted directly to the PNAS office.' when applicable.

[*] E-mail: weare@math.berkeley.edu

and we wish to remove the $\tilde{x}$ variables, the distribution of the remaining $\hat{x}$ variables is given by the marginal density (see [1, 6])

$$\overline{\pi}\left(\hat{x}\right) = \int \pi\left(\hat{x}, \tilde{x}\right) d\tilde{x} \qquad [\mathbf{1}]$$

The full distribution can be factored as

$$\pi(\hat{x}, \tilde{x}) = \overline{\pi}(\hat{x})\pi(\tilde{x}|\hat{x})$$

where $\pi(\tilde{x}|\hat{x})$ is the conditional density of $\tilde{x}$ given $\hat{x}$. Because they exhibit shorter correlation lengths, the marginal distributions are useful in the acceleration of Markov chain Monte Carlo methods.

With this in mind we consider a collection of lower dimensional Markov chains $Y_i^n \in \mathbb{R}^{d_i}$ which have stationary distributions $\pi_i(x_i)$ where $d_0 > \cdots > d_i$. For each $i \leq L$ let T_i be the transition probability density of Y_i^n, i.e. $T_i(x_i \to y_i)$ is the probability density of $\left\{Y_i^{n+1} = y_i\right\}$ given that $\{Y_i^n = x_i\}$. The $\{\pi_i\}$ are approximate marginal distributions. For example, divide the x_i variables into two subsets, $\hat{x}_i \in \mathbb{R}^{d_{i+1}}$ and $\tilde{x}_i \in \mathbb{R}^{d_i - d_{i+1}}$, so that $x_i = (\hat{x}_i, \tilde{x}_i)$. The $\tilde{x}_i$ variables represent the variables of x_i that are removed by marginalization, i.e.

$$\pi_{i+1}\left(\hat{x}_i\right) \approx \int \pi_i\left(\hat{x}_i, \tilde{x}_i\right)\ d\tilde{x}_i.$$

After arranging these chains in parallel we have the larger process

$$Y^n = (Y_0^n, \ldots, Y_L^n) \in \mathbb{R}^{d_0} \times \cdots \times \mathbb{R}^{d_L}.$$

The probability density of a move to $\left\{Y^{n+1} = y\right\}$ given that $\{Y^n = x\}$ for $x, y \in \mathbb{R}^{d_0} \times \cdots \times \mathbb{R}^{d_L}$ is given by

$$T(x \to y) = \prod_{i=0}^{L} T_i(x_i \to y_i). \qquad [\mathbf{2}]$$

Since

$$\int \left(T(x \to y)\prod_{i=0}^{L}\pi_i\left(x_i\right)\right)\ dx_0 \ldots dx_L = \prod_{i=0}^{L}\pi_i\left(y_i\right)$$

the stationary distribution of Y^n is

$$\Pi\left(x_0, \ldots, x_L\right) = \pi_0\left(x_0\right) \ldots \pi_L\left(x_L\right).$$

The next step in the construction is to allow interactions between the chains $\{Y_i^n\}$ and to thereby pass information from the rapidly equilibrating chains on the lower dimensional spaces (large i) down to the chain on the original space ($i = 0$). This is accomplished by swap moves. In a swap move between levels i and $i+1$, we take a d_{i+1} dimensional subset, $\hat{x}_i$, of the x_i variables and exchange them with the x_{i+1} variables. The remaining $d_i - d_{i+1}$ $\tilde{x}_i$ variables are resampled from the conditional distribution $\pi_i\left(\tilde{x}_i|x_{i+1}\right)$. For the full chain, this swap takes the form of a move from $\{Y^n = x\}$ to $\left\{Y^{n+1} = y\right\}$ where

$$x = \left(\ldots, \hat{x}_i, \tilde{x}_i, x_{i+1}, \ldots\right)$$

and

$$y = \left(\ldots, x_{i+1}, \tilde{y}_i, \hat{x}_i, \ldots\right).$$

The ellipses represent components of Y^n that remain unchanged in the transition and $\tilde{y}_i$ is drawn from $\pi_i\left(\tilde{x}_i|x_{i+1}\right)$.

If these swaps are undertaken unconditionally, the resulting chain with equilibrate rapidly, but will not, in general, preserve the product distribution Π. To remedy this we introduce the swap acceptance probability

$$A_i = \min\left\{1,\ \frac{\overline{\pi}_i(x_{i+1})\pi_{i+1}(\hat{x}_i)}{\overline{\pi}_i(\hat{x}_i)\pi_{i+1}(x_{i+1})}\right\}. \qquad [\mathbf{3}]$$

In this formula $\overline{\pi}_i$ is the function on $\mathbb{R}^{d_{i+1}}$ resulting from marginalization of π_i as in equation 1. Given that $\{Y^n = x\}$, the probability density of $\left\{Y^{n+1} = y\right\}$, after the proposal and either acceptance with probability A_i or rejection with probability $1 - A_i$, of a swap move, is given by

$$\begin{aligned} S_i\left(x \to y\right) &= \left(1 - A_i\right)\ \delta_{\{y=x\}} \\ &+ A_i\ \pi_i(\tilde{y}_i|x_{i+1})\ \delta_{\left\{\left(\hat{y}_i, y_{i+1}\right) = \left(x_{i+1}, \hat{x}_i\right)\right\}} \prod_{j \notin \{i, i+1\}} \delta_{\left\{y_j = x_j\right\}} \end{aligned}$$

for $x, y \in \mathbb{R}^{d_0} \times \cdots \times \mathbb{R}^{d_L}$. δ is the Dirac delta function.

We have the following lemma.

Lemma 1. *The transition probabilities S_i satisfy the detailed balance condition for the measure Π, i.e.*

$$\Pi(x)\ S_i\left(x \to y\right) = \Pi(y)\ S_i\left(y \to x\right)$$

where $x, y \in \mathbb{R}^{d_0} \times \cdots \times \mathbb{R}^{d_L}$.

The detailed balance condition stipulates that the probability of observing a transition $x \to y$ is equal to that of observing a transition $y \to x$ and guarantees that the resulting Markov Chain preserves the distribution Π. Therefore, under general conditions, averages over a trajectory of $\{Y^n\}$ will converge to averages over Π. Since

$$\pi_0(x_0) = \int \Pi(x_0, \ldots, x_L)\ dx_1 \ldots dx_L$$

we can calculate averages over π_0 by taking averages over the trajectories of the first d_0 components of Y^n.

"Exact" approximation of acceptance probability

Notice that the formula 3 for A_i requires the evaluation of $\overline{\pi}_i$ at the points $\hat{x}_i, x_{i+1} \in \mathbb{R}^{d_{i+1}}$. While the approximation of $\overline{\pi}_i$ by functions on $\mathbb{R}^{d_{i+1}}$ is in general a very difficult problem, its evaluation at a single point is often not terribly demanding. In fact, in many cases, including the examples in this paper, the $\hat{x}_i$ variables can be chosen so that the remaining $\tilde{x}_i$ variables are conditionally independent given $\hat{x}_i$.

Despite these mitigating factors, the requirement that we evaluate $\overline{\pi}_i$ before we accept any swap is a little onerous. Fortunately, and somewhat surprisingly, this requirement is not necessary. In fact, standard strategies for approximating the point values of the marginals yield Markov chains that themselves preserve the target measure. Thus even a poor estimate of the ratio appearing in 3 can give rise to a method that is exact in the sense that the resulting Markov chain will asymptotically sample the target measure.

To illustrate this point, we consider the following example of a swap move. Assume that the current position of the chain is $\{Y^n = x\}$ where

$$x = \left(\ldots, \hat{x}_i, \tilde{x}_i, x_{i+1}, \ldots\right)$$

The following steps will result in either $\left\{Y^{n+1} = x\right\}$ or $\left\{Y^{n+1} = y\right\}$ where

$$y = \left(\ldots, x_{i+1}, \tilde{y}_i, \hat{x}_i, \ldots\right)$$

and $\tilde{y}_i \in \mathbb{R}^{d_i - d_{i+1}}$.

1. Let $v^0 = \tilde{x}_i$ and let $v^j \in \mathbb{R}^{d_i - d_{i+1}}$ for $j = 1, \dots, M-1$ be independent samples from $p_i(\,\cdot\,|\hat{x}_i)$, where $p_i(\,\cdot\,|\hat{x}_i)$ is a reference density conditioned by $\hat{x}_i$. For example, $p_i(\,\cdot\,|\hat{x}_i)$ could be a Gaussian approximation of $\pi_i(\tilde{x}_i|\hat{x}_i)$. How p_i is chosen depends on the problem at hand (see numerical examples below). In general $p_i(\,\cdot\,|\hat{x}_i)$ should be easily evaluated and independently sampled, and it should "cover" $\pi_i(\,\cdot\,|\hat{x}_i)$ in the sense that areas of $\mathbb{R}^{d_i}$ where $\pi_i(\,\cdot\,|\hat{x}_i)$ is not negligible should be contained in areas where $p_i(\,\cdot\,|\hat{x}_i)$ is not negligible.

2. Let $u^j \in \mathbb{R}^{d_i - d_{i+1}}$ for $j = 0, \dots, M-1$ be independent random variables sampled from $p_i(\,\cdot\,|x_{i+1})$ (recall that we are considering a swap of $\hat{x}_i$ and x_{i+1} which live in the same space). Notice that the $\left\{u^j\right\}$ variables depend on x_{i+1} while the $\left\{v^j\right\}$ variables depend on $\hat{x}_i$.

3. Define the weights
$$w_v^j = \frac{\pi_i\left(\hat{x}_i, v^j\right)}{p_i\left(v^j|\hat{x}_i\right)} \quad \text{and} \quad w_u^j = \frac{\pi_i\left(x_{i+1}, u^j\right)}{p_i\left(u^j|x_{i+1}\right)}$$
The choice of p_i made above affects the variance of these weights, and therefore the variance of the acceptance probability below.

4. Choose $\tilde{y}_i$ from among the $\left\{u^j\right\}$ according to the multinomial distribution with probabilities
$$\mathbf{P}\left(\tilde{y}_i = u^j\right) = \frac{w_u^j}{\sum_{l=0}^{M-1} w_u^l}.$$
Notice that $\tilde{y}_i$ is an approximate sample from $\pi_i(\,\cdot\,|x_{i+1})$.

5. Set
$$Y^{n+1} = (\dots, x_{i+1}, \tilde{y}_i, \hat{x}_i, \dots)$$
with probability
$$A_i^M = \min\left\{1,\ \frac{\pi_{i+1}(\hat{x}_i)\sum_{j=0}^{M-1} w_u^j}{\pi_{i+1}(x_{i+1})\sum_{j=0}^{M-1} w_v^j}\right\} \qquad \textbf{[4]}$$
and
$$Y^{n+1} = Y^n = (\dots, \hat{x}_i, \tilde{x}_i, x_{i+1}, \dots)$$
with probability $1 - A_i^M$.

The transition probability density for the above swap move from $x \to y$ for $x, y \in \mathbb{R}^{d_0} \times \cdots \times \mathbb{R}^{d_L}$is given by
$$\begin{aligned} S_i^M(x \to y) = {} & (1-R)\ \delta_{\{y=x\}} \\ & + R\ \delta_{\left\{\left(\hat{y}_i, y_{i+1}\right)=\left(x_{i+1}, \hat{x}_i\right)\right\}} \prod_{j \notin \{i, i+1\}} \delta_{\{y_j = x_j\}} \end{aligned}$$
where
$$\begin{aligned} R = M \int & p_i(u^0|x_{i+1}) \frac{w_u^0}{\sum_{j=0}^{M-1} w_u^j}\ A_i^M \\ & \times \prod_{j=1}^{M-1} p_i(v^j|\hat{x}_i) p_i(u^j|x_{i+1}) dv^j du^j \end{aligned}$$
and δ is again the Dirac delta function. In other words, S_i^M dictates that the Markov chain accepts the swap with probability R and rejects it with probability $1 - R$.

While the preceding swap move corresponds to a method for approximating the ratio
$$\frac{\overline{\pi}_i(x_{i+1})}{\overline{\pi}_i(\hat{x}_i)}$$
appearing in the formula for A_i above, it also has some similarities with the multiple-try Metropolis method presented in [10] which uses multiple suggestion samples to improve acceptance rates of standard MCMC methods. The following lemma is suggested by results in [10].

Lemma 2. *The transition probabilities S_i^M satisfy the detailed balance condition for the measure* Π.

As before the detailed balance condition guarantees that averages over trajectories of the first d_0 dimensions of Y^n will converge to averages over π_0.

The A_i^M contain an approximation to the ratio of marginals in 3
$$\begin{aligned} \frac{\sum_{j=0}^{M-1} w_u^j}{\sum_{j=0}^{M-1} w_v^j} &= \frac{\frac{1}{M}\sum_{j=0}^{M-1} \frac{\pi_i\left(x_{i+1}, u^j\right)}{p_i\left(u^j|x_{i+1}\right)}}{\frac{1}{M}\sum_{j=0}^{M-1} \frac{\pi_i\left(\hat{x}_i, v^j\right)}{p_i\left(v^j|\hat{x}_i\right)}} \\ &\xrightarrow[M\to\infty]{a.s.} \frac{\mathbf{E}_{p_i}\left[\frac{\pi_i\left(x_{i+1}, \widetilde{X}_i\right)}{p_i\left(\widetilde{X}_i|x_{i+1}\right)} \mid \left\{\widehat{X}_i = x_{i+1}\right\}\right]}{\mathbf{E}_{p_i}\left[\frac{\pi_i\left(\hat{x}_i, \widetilde{X}_i\right)}{p_i\left(\widetilde{X}_i|\hat{x}_i\right)} \mid \left\{\widehat{X}_i = \hat{x}_i\right\}\right]} \\ &= \frac{\overline{\pi}_i(x_{i+1})}{\overline{\pi}_i(\hat{x}_i)} \end{aligned}$$
where $\mathbf{E}_{p_i}$ denotes expectation with respect to the density p_i. When $0 < \mathbf{E}_{p_i}\left[w_v^j \mid \left\{\widehat{X}_i = \hat{x}_i\right\}\right] < \infty$, the convergence above follows from the strong law of large numbers and the fact that
$$\begin{aligned} \mathbf{E}_{p_i}\left[\frac{\pi_i\left(\widehat{X}_i|\widetilde{X}_i\right)}{p_i\left(\widetilde{X}_i|\widehat{X}_i\right)} \mid \left\{\widehat{X}_i = \hat{x}_i\right\}\right] &= \int \frac{\pi_i(\hat{x}_i, \tilde{x}_i)}{p_i(\tilde{x}_i|\hat{x}_i)} p_i(\tilde{x}_i|\hat{x}_i)\, d\tilde{x}_i \\ &= \int \pi_i(\hat{x}_i, \tilde{x}_i)\ d\tilde{x}_i = \overline{\pi_i}(\hat{x}_i) \end{aligned}$$
For small values of M in 4, calculation of the swap acceptance probabilities is very cheap. However, higher values of M may improve the acceptance rates. For example, if the $\{\pi_i\}_{i>0}$ are exact marginals of π_0, then $A_i \equiv 1$ while $A_i^M \leq 1$. Results similar to Lemma 2 hold when more general approximations replace the one given above; for example when the $\left\{u^j\right\}$ and $\left\{v^j\right\}$ are generated by a Metropolis-Hastings rule. In practice one has to balance the speed of evaluating A_i^M for small M with the possible higher acceptance rates for M large.

It is easy to see that a Markov chain which evolves only by swap moves will only sample a finite number of configurations. These swap moves must therefore be used in conjunction with a transition rule that can reach any region of space, such as T from expression 2. More precisely, T should be Π-irreducible and aperiodic (see [11]). The the transition rule for parallel marginalization is
$$\begin{aligned} P(x \to y) = {} & (1-\alpha)\ T(x \to y) \\ & + \alpha \int T(x \to z) S\,(z \to y)\, dz \end{aligned}$$

where

$$S(x \to y) = \sum_{k=0}^{L-1} \frac{1}{L} S_i^M \left(x \to y\right)$$

and $\alpha \in [0,1)$ is the probability that a swap move occurs. P dictates that, with probability α, the chain attempts a swap move between levels I and $I+1$ where I is a random variable chosen uniformly from $\{0,\dots,L-1\}$. Next, each level of the chain evolves independently according to the $\{T_i\}$. With probability $1-\alpha$ the chain does not attempt a swap move, but does evolve each level. The next result follows trivially from Lemma 2 and guarantees the invariance of Π under evolution by P.

Theorem 1. *The transition probability P satisfies the detailed balance condition for the measure Π, i.e.*

$$\Pi(x)\ P\left(x \to y\right) = \Pi(y)\ P\left(y \to x\right)$$

where $x, y \in \mathbb{R}^{d_0} \times \cdots \times \mathbb{R}^{d_N}$.

Thus by combining standard MCMC steps on each component governed by the transition probability T, with swap steps between the components governed by S, an MCMC method results which not only uses information from rapidly equilibrating lower dimensional chains, but is also convergent.

Numerical example 1: bridge path sampling

In the bridge path sampling problem we wish to approximate conditional expectations of the form

$$\mathbf{E}\left[g\left(Z^s\right)\ |\{Z^0 = z^-\}, \{Z^T = z^+\}\right]$$

where $s \in (0,T)$ and $\left\{Z^t\right\}$ is the real valued processes given by the solution of the stochastic differential equation

$$dZ^t = f\left(Z^t\right)dt + \sigma\left(Z^t\right)dW^t. \qquad \mathbf{[5]}$$

g, f and σ are real valued functions of $\mathbb{R}$. Of course we can also consider functions g of more than one time. This problem arises, for example, in financial volatility estimation. Because in general we cannot sample paths of 5 we must first approximate $\left\{Z^t\right\}$ by a discrete process for which the path density is readily available. Let $t_0 = 0, t_1 = \frac{T}{K}, \dots, t_K = T$ be a mesh on which we wish to calculate path averages. One such approximate process is given by the linearly implicit Euler scheme (a balanced implicit method, see [12]),

$$\begin{aligned} X^{t_{k+1}} &= X^{t_k} + f\left(X^{t_k}\right)\triangle \\ &\quad + \left(X^{t_{k+1}} - X^{t_k}\right) f^{'}\left(X^{t_k}\right)\triangle + \sigma\left(X^{t_k}\right)\sqrt{\triangle}\ \xi^k, \\ X^0 &= Z^0 \qquad X^{t_K} = Z^T. \end{aligned} \qquad \mathbf{[6]}$$

The $\left\{\xi^k\right\}$ are independent Gaussian random variables with mean 0 and variance 1, and $\triangle = \frac{T}{K}$. K is assumed to be a power of 2. The choice of this scheme over the Euler scheme (see [13]) is due to its favorable stability properties as explained later. Without the condition $X^{t_K} = Z^T$ above, generating samples of $\left(X^0, \dots, X^{t_K}\right)$ is a relatively straitforward endeavor. One simply generates a sample of Z^0, then evolves the system with this initial condition. However, the presence of information about $\left\{Z^t\right\}_{t>0}$ complicates the task. In general, some sampling method which requires only knowlege of a function proportional to conditional density of $\left(X^{t_1}, \dots, X^{t_{K-1}}\right)$ must be applied. The approximate path density associated with discretization 6 is

$$\pi_0\left(x^{t_1}, \dots, x^{t_{K-1}}\ |x^0, x^{t_K}\right) \propto \exp\left(-\sum_{k=0}^{K-1} V\left(x^{t_k}, x^{t_{k+1}}, \triangle\right)\right) \qquad \mathbf{[7]}$$

where

$$V\left(x,y\right) = \frac{\left[\left(1 - \triangle f^{'}\left(x\right)\right)\left(y-x\right) + \triangle f\left(x\right)\right]^2}{2\sigma^2\left(x\right)\triangle}$$

At this point we wish to apply the parallel marginalization sampling procedure to the density π_0. However, as indicated above, a prerequisite for the use of parallel marginalization is the ability to estimate marginal densities. In some important problems homogeneities in the underlying system yield simplifications in the calculation of these densities by the methods in [1, 2]. These calculations can be carried out before implementation of parallel marginalization, or they can be integrated into the sampling procedure.

In some cases, the numerical estimation of the $\{\pi_i\}_{i>0}$ can be completely avoided. The examples presented here are two such cases. Let $S_i = \left\{0, 2^i, \dots, K\right\}$. Decompose S_i as $\widehat{S}_i \sqcup \widetilde{S}_i$ where

$$\widehat{S}_i = \left\{0, 2(2^i), 4(2^i), \dots, K\right\}$$

and

$$\widetilde{S}_i = \left\{2^i, 3(2^i), 5(2^i), \dots, K - 2^i\right\}.$$

In the notation of the previous sections, $x_i = (\hat{x}_i, \tilde{x}_i)$ where $\hat{x}_i = \left\{x_i^{t_k}\right\}_{k \in \widehat{S}_i \setminus \{0,K\}}$ and $\tilde{x}_i = \left\{x_i^{t_k}\right\}_{k \in \widetilde{S}_i}$. In words, the hat and tilde variables represent alternating time slices of the path. For all i fix $x_i^0 = z^-$ and $x_i^{t_K} = z^+$. We choose the approximate marginal densities

$$\pi_i\left(\left\{x_i^{t_k}\right\}_{k \in S_i \setminus \{0,K\}}\ |x_i^0, x_i^{t_K}\right) \propto q_i\left(\left\{x_i^{t_k}\right\}_{k \in S_i}\right)$$

where for each i, q_i is defined by successive coarsenings of 6. That is,

$$q_i\left(\left\{x_i^{t_k}\right\}_{k \in S_i}\right) = \exp\left(-\sum_{k=0}^{K/2^i - 1} V\left(x_i^{t_{2^i k}}, x_i^{t_{2^i(k+1)}}, 2^i\triangle\right)\right).$$

Since π_i will be sampled using a Metropolis-Hastings method with x^0 and x^{t_K} fixed, knowlege of the normalization constants

$$\mathcal{Z}_i\left(x_i^0, x_i^{t_K}\right) = \int q_i \prod_{k \in S_i \setminus \{0,K\}} dx_i^{t_k}$$

is unnecessary.

Notice from 7 that, conditioned on the values of $x^{t_{k-1}}$ and $x^{t_{k+1}}$, the variance of x^{t_k} is of order $\triangle$. Thus any perturbation of x^{t_k} which leaves x^{t_j} fixed for $j \neq k$ and which is compatible with joint distribution 7 must be of the order $\sqrt{\triangle}$. This suggests that distributions defined by coarser discretizations of 7 will allow larger perturbations, and consequently will be easier to sample. However, it is important to choose a discretization that remains stable for large values of $\triangle$. For example, while the linearly implicit Euler method performs well in the experiments below, similar tests using the Euler method were less successful due to limitations on the largest allowable values of $\triangle$.

In this numerical example bridge paths are sampled between time 0 and time 10 for a diffusion in a double well potential

$$f(x) = -4x\left(x^2-1\right) \quad \text{and} \quad \sigma(x) = 1$$

The left and right end points are chosen as $z^- = z^+ = 0$. $Y_i^n \in \mathbb{R}^{10/\left(2^i\triangle\right)+1}$ is the i^{th} level of the parallel marginalization Markov chain at algorithmic time n. There are 10 chains ($L = 9$ in expression 2). The observed swap acceptance rates are reported in Table 1. Let $Y_{mid}^n \in \mathbb{R}$ denote the midpoint of the path defined by Y_0^n (i.e. an approximate sample of the path at time 5). In Fig. 1 the autocorrelation of Y_{mid}^n

$$\mathbf{corr}\left[Y_{mid}^n, Y_{mid}^0\right]$$

is compared to that of a standard Metropolis-Hastings rule. In the figure, the time scale of the autocorrelation for the Metropolis-Hastings method has been scaled by a factor of 1/10 to more than account for the extra computational time required per iteration of parallel marginalization. The relaxation time of the parallel chain is clearly reduced. In these numerical examples, the algorithm in the previous section is applied with a slight simplification. First generate M independent Gaussian random paths $\left\{\zeta^j\left(t_k\right)\right\}_{k\in\widetilde{S}_i}$ with independent components $\zeta^j\left(t_k\right)$ of mean 0 and variance $2^{i-1}\triangle$. For each j and $k \in \widetilde{S}_i$ let

$$u^j\left(t_k\right) = \zeta^j\left(t_k\right) + 0.5\left(x_{i+1}^{t_{k-1}} + x_{i+1}^{t_{k+1}}\right)$$

If in step 4, $\tilde{y}_i = u^{j^*}$, then in step 1 we set $v^0 = \tilde{x}_i$ and for each $k \in \widetilde{S}_i$

$$\left\{v^j\left(t_k\right)\right\}_{j>0} = \left\{\zeta^j\left(t_k\right) + 0.5\left(\hat{x}_i^{t_{k-1}} + \hat{x}_i^{t_{k+1}}\right)\right\}_{j\neq j^*}.$$

All other steps remain the same. This change yields a slightly faster though less generally applicable swap step that also preserves the density Π. Notice that this modification implies that the reference density p_i is given by

$$p_i\left(\tilde{x}_i|\hat{x}_i\right) \propto \exp\left(\sum_{k\in\widetilde{S}_i} -\frac{\left(\tilde{x}_i^{t_k} - 0.5\left(\hat{x}_i^{t_{k-1}} + \hat{x}_i^{t_{k+1}}\right)\right)^2}{2^i\triangle}\right).$$

For this problem, the choice of M in 4, the number of samples of $\left\{u^j\right\}$ and $\left\{v^j\right\}$, seems to have little effect on the swap acceptance rates. In the numerical experiment $M = i+1$ for swaps between levels i and $i+1$.

Numerical example 2: non-linear smoothing/filtering

In the non-linear smoothing and filtering problem we wish to approximate conditional expectations of the form

$$\mathbf{E}\left[g\left(Z^s\right) \;|\left\{H^j = h^j\right\}_0^J\right]$$

where $s \in (0,T)$ and the real valued processes $\left\{Z^t\right\}$ and $\left\{H^j\right\}$ are given by the system

$$\begin{aligned} dZ^t &= f\left(Z^t\right)dt + \sigma\left(Z^t\right)dW^t, \\ H^j &= r\left(Z^{s_j}\right) + \chi^j, \\ Z^0 &\sim \rho, \qquad \chi^n \sim i.i.d.\ \mu. \end{aligned}$$

g, f, σ, and r are real valued functions of $\mathbb{R}$. The $\left\{\chi^j\right\}$ are real valued independent random variable drawn from the density μ and are independent of the Brownian motion $\left\{W^t\right\}$. $\left\{s_j\right\} \subset \left\{t_j\right\}$, and $0 = s_0 < s_1 < ... < s_J = T$. The process Z^t is a hidden signal and the $\left\{H^j\right\}$ are noisy observations.

Again, the system must first be discretized. The linearly implicit Euler scheme gives

$$\begin{aligned} X^{t_{k+1}} &= X^{t_k} + f\left(X^{t_k}\right)\triangle \\ &\quad + \left(X^{t_{k+1}} - X^{t_k}\right)f'\left(X^{t_k}\right)\triangle + \sigma\left(X^{t_k}\right)\sqrt{\triangle}\,\xi^k, \\ H^j &= r\left(X^{s_j}\right) + \chi^j, \\ X^0 &= Z^0 \qquad \chi^n \sim i.i.d.\ \mu. \end{aligned}$$

The $\left\{\xi^k\right\}$ are independent Gaussian random variables with mean 0 and variance 1, and $\triangle = \frac{T}{K}$. The $\left\{\xi^k\right\}$ are independent of the $\left\{\chi^j\right\}$. K is again assumed to be a power of 2.

The approximate path measure for this problem is

$$\begin{aligned} &\pi_0\left(x^{t_0},\ldots,x^{t_K} \;|h^0,\ldots,h^T\right) \propto \\ &\qquad \exp\left(-\sum_{k=0}^{K-1} V\left(x^{t_k}, x^{t_{k+1}}, \triangle\right)\right) \\ &\qquad\qquad \times \rho\left(x_i^{t_0}\right)\prod_{n=0}^{J}\mu\left(x_i^{s_j} - r\left(h^{s_j}\right)\right) \end{aligned}$$

The approximate marginals are chosen as

$$\begin{aligned} &\pi_i\left(\left\{x_i^{t_k}\right\}_{k\in S_i} \;|h^0,\ldots,h^T\right) \propto \\ &\qquad q_i\left(\left\{x_i^{t_k}\right\}_{k\in S_i}\right)\rho\left(x^{t_0}\right)\prod_{n=0}^{J}\mu\left(x^{s_j} - r\left(h^{s_j}\right)\right) \end{aligned}$$

where V, q_i and S_i are as defined in the previous section.

In this example, samples of the smoothed path are generated between time time 0 and time 10 for the same diffusion in a double well potential. The densities μ and ρ are chosen as

$$\mu = N(0, 0.01) \quad \text{and} \quad \rho(x) \propto \exp\left(-\left(x^2-1\right)^2\right)$$

The observation times are $s_0 = 0, s_1 = 1, \ldots, s_{10} = 10$ with $H^j = -1$ for $j = 0,\ldots,5$ and $H^j = 1$ for $j = 6,\ldots,10$. There are 8 chains ($L = 7$ in expression 2). The observed swap acceptance rates are reported in Table 1. Again, $Y_{mid}^n \in \mathbb{R}$ denotes the midpoint of the path defined by Y_0^n (i.e. an approximate sample of the path at time 5). In Fig. 2 the autocorrelation of Y_{mid}^n is compared to that of a standard Metropolis-Hastings rule. The figure has been adjusted as in the previous example. The relaxation time of the parallel chain is again clearly reduced. The algorithm is modified as in the previous example. For this problem, acceptable swap rates require a higher choice of M in 4 than needed in the bridge sampling problem. In this numerical experiment $M = 2^i$ for swaps between levels i and $i+1$.

Conclusion

A Markov chain Monte Carlo method has been proposed and applied to two conditional path sampling problems for stochastic differential equations. Numerical results indicate that this method, parallel marginalization, can have a dramatically reduced equilibration time when compared to standard MCMC methods.

Note that parallel marginalization should not be viewed as a stand alone method. Other acceleration techniques such as hybrid Monte Carlo can and should be implemented at each level within the parallel marginalization framework. As the smoothing problem indicates, the acceptance probabilities at coarser levels can become small. The remedy for this is the development of more accurate approximate marginal distributions by, for example, the methods in [1] and [2].

I would like to thank Prof. A. Chorin for his guidance during this research, which was carried out during my Ph.D. studies at U. C. Berkeley. I would also like to thank Dr. P. Okunev, and Dr. P. Stinis for very helpful discussions and comments. This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U. S. Department of Energy under Contract No. DE-AC03-76SF00098.

1. Chorin, A. (2003) *Multiscale Model. Simul.* **1**, 105–118.
2. Stinis, P. (2005) *J. Comput. Phys.* **208**, 691–703.
3. Goodman, J. and Sokal, A. (1989) *Physical Review D* **40**, 2035–2071.
4. Brandt, A. and Ron, D. (2001) *J. Stat. Phys.* **102**, 163-186.
5. Okunev, P. (2005) *Renormalization Methods with Applications to Spin Systems and to Finance Ph.D. Thesis* (U. C. Berkeley)
6. Liu, J. (2002) *Monte Carlo Strategies in Scientific Computing.* (Springer)
7. Binder, K. and Heermann, D. (2002) *Monte Carlo Simulation in Statistical Physics.* (Springer)
8. Binney, J., Dowrick, N., Fisher, A. and Newman, M. (1992) *The Theory of Critical Phenomena: An Introduction to the Renormalization Group.* (Oxford University Press, USA)
9. Kadanoff, L. (1966) *Physics* **2**, 263.
10. Liu, J., Liang, F. and Wong, W. (2000) *J. Amer. Statist. Assoc.* **95**, 121-134.
11. Tierney, L. (1994) *Annals of Statistics* **22**, 1701-1728.
12. Milstein, G., Platen, E. and Schurz, H. (1998) *SIAM J. Numer. Anal.* **35**, 1010-1019.
13. Kloeden, P. and Platen, E. (1992) *Numerical solution of stochastic differential equations.* (Springer)
14. Apte, A., Hairer, H., Stuart, A. and Voss, J. (2001) *Physica D* to appear.
15. Chorin, A. and Hald, O. (2005) *Stochastic Tools for Mathematics and Science.* (Springer).
16. Chorin, A., Hald, O. and Kupferman, R. (2002) *Phys. D* **166**, 239–257.
17. E, W., Liu, D. and Vanden-Eijnden, E. (2005) *Comm. Pure Appl. Math.* **11**, 1544–1585.
18. E, W., Ren, W. and Vanden-Eijnden, E. (2003) *J. Phys. Chem. B* **109**, 6688-6693.
19. Efendiev, Y. and Hou, T. and Luo, W. (2006) *SIAM J. Sci. Comput* **28**, 776–803.
20. Gear, C. and Kevrekidis, I. (2003) *SIAM J. Sci. Comput.* **24**, 1091–1106.
21. Goldenfeld, N. (1992) *Lectures on Phase Transitions and the Renormalization Group.* (Westview Press)
22. Kadanoff, L. (2000) *Statistical Physics. Statics, Dynamics and Renormalization.* (World Scientific)
23. Stinis, P. (2004) *Multiscale Model. Simul.* **2**, 580–612.
24. Stuart, A., Voss, J. and Wiberg, P. (2004) *Commun. Math. Sci.* **2**, 685–697.

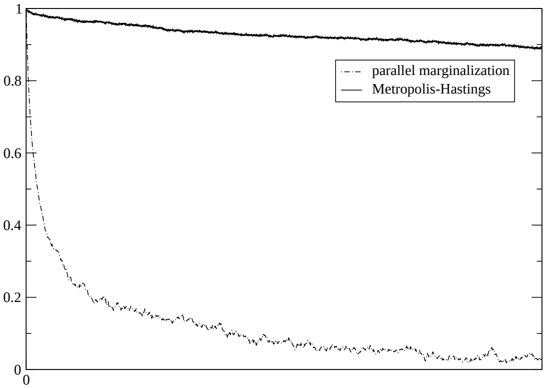

Fig. 1. Autocorrelations of parallel marginalization and standard Metropolis-Hastings methods for bridge sampling problem.

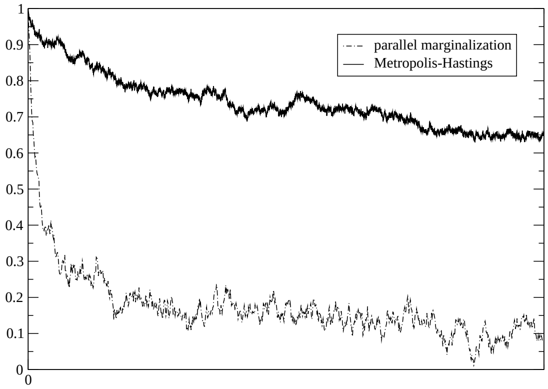

Fig. 2. Autocorrelations of parallel marginalization and standard Metropolis-Hastings methods for filtering and smoothing problem.

Table 1. Swap acceptance rates for bridge sampling and filtering/smoothing problems

Levels*	0/1	1/2	2/3	3/4	4/5	5/6	6/7	7/8	8/9
BS†	0.86	0.83	0.75	0.69	0.54	0.45	0.30	0.22	0.26
FS‡	0.86	0.83	0.74	0.65	0.46	0.23	0.04	NA	NA

*Swaps between levels i and $i+1$

†Bridge sampling problem

‡Filtering/smoothing problem