Parallel Marginalization with Applications to Conditional Path Sampling

Jonathan Weare *

Department of Mathematics,

University of California and

Lawrence Berkeley National Laboratory,

Berkeley, CA 94720

September 11, 2007

Abstract

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

*E-Mail: weare@cims.nyu.edu

the bridge sampling and filtering/smoothing problems for a stochastic differential equation are presented.

1 Introduction

In spite of substantial effort to improve the efficiency of Markov chain Monte Carlo (MCMC) methods, spatial correlations remain a major impediment. These correlations can severely restrict the possible configurations of a system by imposing complicated relationships between variables. It is well known that judicious elimination of variables by renormalization can reduce long range correlations (see [1, 2]). The remaining variables are distributed according to the marginal distribution,

$$\overline{\pi}(x) = \int \pi(x, y)\, dy,$$

where $\pi(x, y)$ is the full distribution. Given the values of the x variables and the marginal distribution $\overline{\pi}$ the y variables are distributed according to the conditional distribution

$$\pi(y|x) = \frac{\pi(x, y)}{\overline{\pi}(x)}.$$

For systems exhibiting critical phenomena, the path through the space of distributions taken by marginal distributions under repeated renormalization can yield essential information about critical indices and the location of critical points (see [1, 2]). More generally, because these marginal distributions exhibit shorter correlation lengths and weaker local correlations, they are useful in the acceleration of Markov chain Monte Carlo methods. As explained in the next

section, parallel marginalization takes advantage of the shorter correlation lengths present in marginal distributions of the target density.

The use of Monte Carlo updates on lower dimensional spaces is not a new concept. In fact this is a necessary procedure in high dimensions. One simply constructs a chain with steps that preserve the conditional probability density of the full measure. This is usually accomplished by perturbing a few components of the chain while holding all other components of the chain constant. In other words the chain takes steps of the form

$$Y^{n+1} = (x_1, \ldots, x_{i-1}, x_i + \epsilon, x_{i+1}, \ldots, x_d)$$

where

$$Y^n = (x_1, \ldots, x_d)$$

and the move preserves $\pi(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$. There have been many important attempts to use proposals in more general sets of projected coordinates. The multi-grid Monte Carlo method presented in [3, 4] is one such method. These techniques do not incorporate marginal densities.

In [5], Brandt and Ron propose a multi-grid method which approximates successive marginal distributions of the Ising model and then uses these approximations to generate large scale movements of the Markov chain sampling the full joint distribution of all variables. Their method, while demonstrating the efficacy of incorporating information from successive marginal distributions, suffers from two limitations. First, the method used to approximate the marginal distributions is specific to a small class of problems. For example, it cannot be easily generalized to systems in continuous spaces. Second, infor-

mation from the approximate marginal distributions is adopted by the Markov chain in a way which does not preserve the target distribution of all variables.

The design of a generally applicable method which approximates the marginal distributions was addressed in [6, 7] by Chorin, and in [8] by Stinis. Both authors approximate the renormalized Hamiltonian of the system given by the formula,

$$\overline{\mathcal{H}}(x) = -\log \int \pi(x, y)\, dy.$$

Thus $\exp\left(-\overline{\mathcal{H}}(x)\right)$ is the marginal distribution of the x variables. Chorin determines the coefficients in an expansion of $\overline{\mathcal{H}}(x)$ by first expanding the derivatives $\frac{\partial \overline{\mathcal{H}}(x)}{\partial x}$, which can be expressed as conditional expectations with respect to the full distribution. Stinis shows that a maximum likelihood approximation to the renormalized Hamiltonian can be found by minimizing the error in the expectations of the basis functions in an expansion of $\overline{\mathcal{H}}(x)$. For applications of related ideas to MCMC simulations see [9] and [10].

Two Parallel marginalization algorithms are developed in the next section along with propositions that guarantee that the resulting Markov chains satisfy the detailed balance condition. In the final section the conditional path sampling problem is described and numerical results are presented for the bridge sampling and smoothing/filtering problems. A brief introduction to parallel marginalization can be found in [11].

2 Parallel marginalization

In this section, it is assumed that appropriate approximate marginal distributions are available. How to find these marginal distributions depends on the application and will be discussed here only in the context of the examples presented in this paper. A new Markov chain Monte Carlo method is introduced which uses approximate marginal distributions of the target distribution to accelerate sampling. 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.

Assume that the system of interest has a probability density, $\pi_0(x_0)$, where x_0 lies in some space E. Suppose further that, by the Metropolis-Hastings or any other method (see [12]), one can construct a Markov chain, $Y_0^n \in E$, which has π_0 as its stationary measure. That is, for two points $x_0, y_0 \in E$

$$\int \tau_0(y_0|x_0)\pi_0(x_0)\ dx_0 = \pi_0(y_0)$$

where $\tau_0(y_0|x_0)$ 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.

In order to take advantage of the shorter spatial correlations exhibited by marginal distributions of π_0, a collection of lower dimensional Markov chains which approximately sample marginal distributions of π_0 is considered. Sup-

pose the random variable X_0 has d_0 components. Divide these into two subsets,

$$X_0 = \left(\widehat{X}_0, \widetilde{X}_0\right),$$

where $\widehat{X}_0$ has d_1 components and $\widetilde{X}_0$ has $d_0 - d_1$ components. Recall that the $\widehat{X}_0$ variables are distributed according to the marginal density,

$$\overline{\pi}_0(\hat{x}_0) = \int \pi_0(\hat{x}_0, \tilde{x}_0) d\tilde{x}_0 \tag{1}$$

and that given the value of the $\widehat{X}_0$ variables, the $\widetilde{X}_0$ variables are distributed according to the conditional density,

$$\pi(\tilde{x}_0|\hat{x}_0) = \frac{\pi_0(\hat{x}_0, \tilde{x}_0)}{\overline{\pi}_0(\hat{x}_0)} \tag{2}$$

Label the domain of the $\widehat{X}_0$ variables E_1. Suppose further that an approximation to the marginal distribution of the $\widehat{X}_0$ variables,

$$\pi_1\left(\hat{x}_0\right) \approx \overline{\pi}_0\left(\hat{x}_0\right)$$

is available. The sense in which π_1 approximates $\overline{\pi}_0$ is intentionally left vague. In applications of parallel marginalization the accuracy of the approximation manifests itself through an acceptance rate.

Now let $X_1 \in E_1$ be independent of the X_0 random variables and drawn from $\pi_1\left(\widehat{x}_0\right)$. Notice that X_1 represents the same physical variables as $\widehat{X}_0$ though its probability density is not the exact marginal density. Continue in this way to remove variables from the system by decomposing $X_l \in E_l$ into proper subsets as

$$X_l = \left(\widehat{X}_l, \widetilde{X}_l\right)$$

and defining $X_{l+1} \in E_{l+1}$ to be independent of the $\{X_0, \ldots, X_l\}$ random variables and drawn from an approximation π_{l+1} to $\overline{\pi}_l(\hat{x}_l)$. Clearly each X_{l+1} represents fewer physical variables than X_l.

Just as one can construct a Markov chain $Y_0^n \in E_0$ to sample X_0, one can also construct Markov chains $Y_l^n \in E_l$ to sample π_l. In other words, for each Y_l^n choose a transition probability density τ_l, such that

$$\int \tau_l(y_l|x_l)\pi_l(x_l)\,dx_l = \pi_l(y_l)$$

for all i.

The chains Y_l^n can be arranged in parallel to yield a larger Markov chain,

$$Y^n = (Y_0^n, \ldots, Y_L^n) \in E_0 \times \cdots \times E_L.$$

The probability density of a move to $\{Y^{n+1} = y\}$ given that $\{Y^n = x\}$ for $x, y \in E_0 \times \cdots \times E_L$ is given by

$$\tau(y|x) = \prod_{l=0}^{L} \tau_l(y_l|x_l). \tag{3}$$

Since

$$\int \left(\tau(y|x) \prod_{l=0}^{L} \pi_l(x_l) \right) dx_0 \ldots dx_L = \prod_{l=0}^{L} \pi_l(y_l)$$

the stationary distribution of Y^n is

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

The next step in the construction is to allow interactions between the chains $\{Y_l^n\}$ and to thereby pass information from the rapidly equilibrating chains on the lower dimensional spaces (large l) down to the chain on the original

space ($l = 0$). This is accomplished by swap moves. In a swap move between levels l and $l+1$, a subset, $\hat{x}_l \in E_{l+1}$, of the x_l variables is exchanged with the $x_{l+1} \in E_{l+1}$ variables. The remaining $\tilde{x}_l$ variables are resampled from the conditional distribution $\pi_l(\tilde{x}_l|x_{l+1})$. For the full chain, this swap takes the form of a move from $\{Y^n = x\}$ to $\{Y^{n+1} = y\}$ where

$$x = (\dots, \hat{x}_l, \tilde{x}_l, x_{l+1}, \dots)$$

and

$$y = (\dots, x_{l+1}, \tilde{y}_l, \hat{x}_l, \dots).$$

The $\tilde{y}_l$ variables are drawn from $\pi_l(\tilde{x}_l|x_{l+1})$ and the ellipses represent components of Y^n that remain unchanged in the transition.

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

$$A_l = \min\left\{1,\ \frac{\overline{\pi}_l(x_{l+1})\pi_{l+1}(\hat{x}_l)}{\overline{\pi}_l(\hat{x}_l)\pi_{l+1}(x_{l+1})}\right\} \tag{4}$$

is introduced. Recall that $\overline{\pi}_l$ is the function resulting from the integration of π_l over the $\tilde{x}_l$ variables as in equation (1). Given that $\{Y^n = x\}$, the probability density of $\{Y^{n+1} = y\}$, after the proposal and either acceptance with probability A_l or rejection with probability $1 - A_l$, of a swap move, is given by

$$\begin{aligned}\psi_l(y|x) = (1 - A_l) \prod \delta_{\{y_j = x_j\}} \\ + A_l\ \pi_l(\tilde{y}_l|x_{l+1})\ \delta_{\{(\hat{y}_l, y_{l+1}) = (x_{l+1}, \hat{x}_l)\}} \prod_{j \notin \{l, l+1\}} \delta_{\{y_j = x_j\}}\end{aligned}$$

for $x, y \in E_0 \times \cdots \times E_L$. δ is the Dirac delta function.

We have the following proposition.

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

$$\Pi(x)\ \psi_l\left(y|x\right) = \Pi(y)\ \psi_l\left(x|y\right)$$

where $x, y \in E_0 \times \cdots \times E_L$.

Proof. Fix $x, y \in E_0 \times \cdots \times E_L$ such that $x \neq y$.

$$\begin{aligned}\Pi(x)\ \psi_l\left(y|x\right) = &\left(\prod_{j \notin \{l, l+1\}} \pi_j\left(x_j\right) \delta_{\{y_j = x_j\}}\right) \pi_l\left(x_l\right) \pi_{l+1}\left(x_{l+1}\right) \\ &\times \left((1 - A_l)\ \delta_{\{(y_l, y_{l+1}) = (x_l, x_{l+1})\}} + A_l\ \pi_l(\tilde{y}_l|x_{l+1})\ \delta_{\{(\hat{y}_l, y_{l+1}) = (x_{l+1}, \hat{x}_l)\}}\right)\end{aligned}$$

When $x \neq y$ $(\Pi(x)\psi_l\left(y|x\right))$ and $(\Pi(y)\psi_l\left(x|y\right))$ are both zero unless $x_j = y_j$ for all j except l and $l+1$ and $(\hat{y}_l, y_{l+1}) = (x_{l+1}, \hat{x}_l)$. Therefore it is enough to check that the function

$$R\left(\left(x_l, x_{l+1}\right), \left(y_l, y_{l+1}\right)\right) = \pi_l\left(x_l\right) \pi_{l+1}\left(x_{l+1}\right) \pi_l\left(\tilde{y}_l|x_{l+1}\right)\ A_l$$

is symmetric in (x_l, x_{l+1}) and (y_l, y_{l+1}) when $(\hat{y}_l, y_{l+1}) = (x_{l+1}, \hat{x}_l)$. Plugging in the definition of A_l,

$$R = \pi_l\left(x_l\right) \pi_{l+1}\left(x_{l+1}\right) \pi_l\left(\tilde{y}_l|x_{l+1}\right)\ \min\left\{1,\ \frac{\overline{\pi}_l(x_{l+1})\pi_{l+1}(\hat{x}_l)}{\overline{\pi}_l(\hat{x}_l)\pi_{l+1}(x_{l+1})}\right\}$$

Rearranging terms gives,

$$\begin{aligned}R = &\pi_l\left(x_l\right) \pi_{l+1}\left(x_{l+1}\right) \pi_l\left(\tilde{y}_l|x_{l+1}\right) \overline{\pi}_l\left(x_{l+1}\right) \pi_{l+1}\left(y_{l+1}\right) \\ &\times \min\left\{\frac{1}{\overline{\pi}_l(x_{l+1})\pi_{l+1}(y_{l+1})},\ \frac{1}{\overline{\pi}_l(y_{l+1})\pi_{l+1}(x_{l+1})}\right\}\end{aligned}$$

Recall from (2), that $\pi_l(\tilde{y}_l|x_{l+1})\overline{\pi}_l(x_{l+1}) = \pi_l(x_{l+1}, \tilde{y}_l)$. Therefore, since $x_{l+1} = \hat{y}_l$,

$$R = \pi_l\left(x_l\right)\pi_{l+1}\left(x_{l+1}\right)\pi_l\left(y_l\right)\pi_{l+1}\left(y_{l+1}\right) \times \min\left\{\frac{1}{\overline{\pi}_l(x_{l+1})\pi_{l+1}(y_{l+1})},\ \frac{1}{\overline{\pi}_l(y_{l+1})\pi_{l+1}(x_{l+1})}\right\}$$

The final formula is clearly symmetric in (x_l, x_{l+1}) and (y_l, y_{l+1}). □

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. If the chain is also Harris recurrent then averages over a trajectory of $\{Y^n\}$ will converge to averages over Π. In fact, chains generated by swaps as described above cannot be recurrent and must be combined with another transition rule to generate a convergent Markov chain. Since

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

if $\{Y^n\}$ is Harris recurrent with invariant distribution Π, averages over π_0 can be calculated by taking averages over the trajectories of the first d_0 components of $\{Y^n\}$.

2.1 Approximation of acceptance probabilities

Notice that the formula (4) for A_l requires the evaluation of $\overline{\pi}_l$ at the points $\hat{x}_l, x_{l+1} \in E_{l+1}$. While the approximation of $\overline{\pi}_l$ by functions on E_{l+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 Chapter 3, the $\widehat{X}_l$ variables can be chosen so that the remaining $\widetilde{X}_l$ variables are conditionally independent given $\widehat{X}_l$.

Despite this mitigating factor, the requirement that $\overline{\pi}_l$ be evaluated before acceptance of any swap is inconvenient. 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 also preserve the target measure. Thus even a poor estimate of the ratio appearing in (4) can give rise to a method that is exact in the sense that the resulting Markov chain will asymptotically sample the target measure.

Before moving on to the description of the resulting Markov chain Monte Carlo algorithms consider briefly the general problem of evaluating marginal densities. Let $p_1(x, y)$ and $p_2(x, y)$ be the densities of two equivalent measures with marginal densities,

$$\overline{p}_1(x) = \int p_1(x, y)dy$$

and

$$\overline{p}_2(x) = \int p_2(x, y)dy$$

respectively. For any integrable function $\gamma(x, y)$,

$$\begin{aligned}
\mathbf{E}_{p_1}\left[\gamma\left(X, Y\right) p_2\left(X, Y\right) | \left\{X = x\right\}\right] &= \int \gamma(x, y) p_2(x, y) p_1(y|x)dy \\
&= \frac{\overline{p}_2(x)}{\overline{p}_1(x)} \int \gamma(x, y) p_2(y|x) p_1(x, y)dy \\
&= \frac{\overline{p}_2(x)}{\overline{p}_1(x)} \mathbf{E}_{p_2}\left[\gamma\left(X, Y\right) p_1\left(X, Y\right) | \left\{X = x\right\}\right]
\end{aligned}$$

Thus given $\overline{p}_2(x)$, the value of $\overline{p}_1$ at x can be obtained through the formula,

$$\overline{p}_1(x) = \overline{p}_2(x) \frac{\mathbf{E}_{p_2}\left[\gamma\left(X,Y\right) p_1\left(X,Y\right) \mid \{X = x\}\right]}{\mathbf{E}_{p_1}\left[\gamma\left(X,Y\right) p_2\left(X,Y\right) \mid \{X = x\}\right]} \tag{5}$$

Of course, the usual importance sampling concerns apply here. In particular, the approximation of the conditional expectations in (5) will be much easier when Y lives in a lower dimensional space.

Similar approximations can be inserted into our acceptance probabilities A_l in place of the ratio $\frac{\overline{\pi}_l(x_{l+1})}{\overline{\pi}_l(\hat{x}_l)}$. For example, if $p_l(\tilde{x}_l|\hat{x}_l)$ is a reference density approximating $\pi_l(\tilde{x}_l|\hat{x}_l)$, then the choice

$$\gamma(\hat{x}_l, \tilde{x}_l) = \frac{1}{p_l(\hat{x}_l, \tilde{x}_l)}$$

yields

$$\overline{\pi}_l(\hat{x}) \approx \overline{p}_l(\hat{x}) \frac{1}{M} \sum \frac{\pi_l\left(\hat{x}_l, V^j\right)}{p_l\left(\hat{x}, V^j\right)} = \frac{1}{M} \sum \frac{\pi_l\left(\hat{x}_l, V^j\right)}{p_l\left(V^j|\hat{x}_l\right)} \tag{6}$$

where the $\{V^j\}$ are samples from $p_l(\tilde{x}_l|\hat{x}_l)$. Thus if $\{U^j\}$ are samples from $p_l(\tilde{x}_l|x_{l+1})$, then

$$\frac{\frac{1}{M}\sum_{j=1}^{M} \frac{\pi_l\left(x_{l+1}, U^j\right)}{p_l(U^j|x_{l+1})}}{\frac{1}{M}\sum_{j=1}^{M} \frac{\pi_l(\hat{x}_l, V^j)}{p_l(V^j|\hat{x}_l)}} \xrightarrow[M\to\infty]{a.s.} \frac{\mathbf{E}_{p_l}\left[\frac{\pi_l\left(x_{l+1}, \widetilde{X}_l\right)}{p_l\left(\widetilde{X}_l|x_{l+1}\right)} \mid \left\{\widehat{X}_l = x_{l+1}\right\}\right]}{\mathbf{E}_{p_l}\left[\frac{\pi_l\left(\hat{x}_l, \widetilde{X}_l\right)}{p_l\left(\widetilde{X}_l|\hat{x}_l\right)} \mid \left\{\widehat{X}_l = \hat{x}_l\right\}\right]} = \frac{\overline{\pi}_l(x_{l+1})}{\overline{\pi}_l(\hat{x}_l)}$$

In the numerical examples presented here, $p_l(\,\cdot\,|\hat{x}_l)$ is a Gaussian approximation of $\pi_l(\tilde{x}_l|\hat{x}_l)$. How p_l is chosen depends on the problem at hand (see numerical examples below). In general $p_l(\,\cdot\,|\hat{x}_l)$ should be easily evaluated and independently sampled, and it should "cover" $\pi_l(\,\cdot\,|\hat{x}_l)$ in the sense that regions where $\pi_l(\,\cdot\,|\hat{x}_l)$ is not negligible should be contained in regions where $p_l(\,\cdot\,|\hat{x}_l)$ is not negligible. In the case mentioned above that the $\widehat{X}_l$ variables

can be chosen so that the remaining $\widetilde{X}_l$ variables are conditionally independent given $\widehat{X}_l$ the conditional density $\pi_l(\tilde{x}_l|\hat{x}_l)$ can be written as a product of many low dimensional densities. As mentioned above, the problem of finding a reference density for importance sampling is much simpler in low dimensional spaces.

The following algorithm results from replacing A_l in (4) with approximation of the form (6). Assume that the current position of the chain is $\{Y^n = x\}$ where

$$x = (\dots, \hat{x}_l, \tilde{x}_l, x_{l+1}, \dots).$$

Algorithm 1 will result in either $\{Y^{n+1} = x\}$ or $\{Y^{n+1} = y\}$ where

$$y = (\dots, x_{l+1}, \tilde{y}_l, \hat{x}_l, \dots)$$

and $\tilde{y}_l$ is approximately drawn from $\pi_l(\tilde{x}_l|x_{l+1})$.

Algorithm 1 (Parallel Marginalization 1)**.** *The chain moves from Y^n to Y^{n+1} as follows:*

1. *Let U^j for $j = 1, \dots, M$ be independent random variables sampled from $p_l(\,\cdot\,|x_{l+1})$ (recall that the swap is between $\hat{x}_l$ and x_{l+1} which are both in E_{l+1}).*

2. *Evaluate the weights*

$$W_U^j = \frac{\pi_l(x_{l+1}, U^j)}{p_l(U^j|\ x_{l+1})}.$$

The choice of p_l made above affects the variance of these weights, and therefore the variance of the acceptance probability below.

3. *Draw the random index* $J \in \{1, \dots, M\}$ *according to the probabilities*

$$\mathbf{P}\left[J = j\right] = \frac{W_U^j}{\sum_{m=1}^M W_U^m}.$$

Set $\widetilde{Y}' = U^J$. *Notice that* $\widetilde{Y}'$ *is an approximate sample from* $\pi_l(\,\cdot\,|x_{l+1})$.

4. *Let* $V^J = \tilde{x}_l$ *and draw* V^j *for* $j \neq J$ *independently from* $p_l(\,\cdot\,|\hat{x}_l)$, *Notice that the* $\{U^j\}$ *variables depend on* x_{l+1} *while the* $\{V^j\}$ *variables depend on* $\hat{x}_l$.

5. *Define the weights*

$$W_V^j = \frac{\pi_l\left(\hat{x}_l, V^j\right)}{p_l\left(V^j|\hat{x}_l\right)}$$

6. *Set*

$$Y^{n+1} = \Big(\dots, x_{l+1}, \widetilde{Y}', \hat{x}_l, \dots\Big)$$

with probability

$$A_l^M = \min\left\{1,\ \frac{\pi_{l+1}(\hat{x}_l)\sum_{j=1}^M W_U^j}{\pi_{l+1}(x_{l+1})\sum_{j=1}^M W_V^j}\right\} \tag{7}$$

and

$$Y^{n+1} = Y^n = (\dots, \hat{x}_l, \tilde{x}_l, x_{l+1}, \dots)$$

with probability $1 - A_l^M$.

The transition probability density for the above swap move from x to y for $x, y \in E_0 \times \cdots \times E_L$is given by

$$\begin{aligned}\psi_l^M(y|x) = \mathbf{P}\left[\{\text{Swap is rejected}\}\right] \prod \delta_{\{y_j = x_j\}} \\ + \mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\widetilde{Y}' = \tilde{y}_l\right\}\right] \\ \times \delta_{\{(\hat{y}_l, y_{l+1}) = (x_{l+1}, \hat{x}_l)\}} \prod_{j \notin \{i, i+1\}} \delta_{\{y_j = x_j\}},\end{aligned}$$

where δ is again the Dirac delta function. Notice that to find the probability density $\mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\widetilde{Y}' = \tilde{y}_l\right\}\right]$ one must integrate over the possible values of the $\{U^j\}$ and $\{V^j\}$ variables. Since π_l appears in the integrand it is not possible, in general, to evaluate the integral. However, as indicated in the proof of the next proposition, it is not necessary to evaluate this density to show that the method converges.

While the preceding swap move corresponds to a method for approximating the ratio

$$\frac{\overline{\pi}_l(x_{l+1})}{\overline{\pi}_l(\hat{x}_l)}$$

appearing in formula (4) for A_l, it also has similarities with the multiple-try Metropolis method, presented in [13, 14], that uses multiple suggestion samples to improve acceptance rates of standard MCMC methods. In fact the proof of the following proposition is motivated by the proof of the detailed balance condition for the multiple try method.

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

Proof. For $x, y \in E_0 \times \cdots \times E_L$ such that $x \neq y$,

$$\Pi(x)\ \psi_l^M(y|x) = \Pi(x)\ \mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\widetilde{Y}' = \tilde{y}_l\right\}\right] \times \delta_{\{(\hat{y}_l, y_{l+1}) = (x_{l+1}, \hat{x}_l)\}} \prod_{j \notin \{i, i+1\}} \delta_{\{y_j = x_j\}},$$

As in the previous proof it can be assumed that $x_j = y_j$ for all j except l and $l+1$ and $(\hat{y}_l, y_{l+1}) = (x_{l+1}, \hat{x}_l)$. Since in this case $\pi(x_j) = \pi(y_j)$ for all

$j \notin \{l, l+1\}$, it remains to show that if $(\hat{y}_l, y_{l+1}) = (x_{l+1}, \hat{x}_l)$ then

$$R\left((x_l, x_{l+1}), (y_l, y_{l+1})\right) = \pi_l(x_l)\,\pi_{l+1}(x_{l+1}) \times \mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\widetilde{Y}_l^{n+1} = \tilde{y}_l\right\}\right]$$

is symmetric in (x_l, x_{l+1}) and (y_l, y_{l+1}). Define a random index $J \in \{1, \ldots, M\}$ by the relation $\tilde{y}_l = U^J$. Then, since the U^j are *i.i.d.*,

$$\begin{aligned}\mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\widetilde{Y}' = \tilde{y}_l\right\}\right] &= \\ &\sum_{j=1}^{M} \mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\tilde{y}_l = U^j\right\} \cap \{J = j\}\right] \\ &= M\ \mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\tilde{y}_l = U^1\right\} \cap \{J = 1\}\right]\end{aligned}$$

Thus,

$$R\left((x_l, x_{l+1}), (y_l, y_{l+1})\right) = M\ \pi_l(x_l)\,\pi_{l+1}(x_{l+1}) \times \mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\tilde{y}_l = U^1\right\} \cap \{J = 1\}\right]$$

Writing out the density on the right of this relation gives,

$$\begin{aligned}R = M\ \pi_l(x_l)\,\pi_{l+1}(x_{l+1}) \int \min\left\{1,\ \frac{\pi_{l+1}(\hat{x}_l)\sum_{j=1}^{M} W_U^j}{\pi_{l+1}(x_{l+1})\sum_{j=1}^{M} W_V^j}\right\} \frac{\frac{\pi_l\left(x_{l+1}, u^1\right)}{p(u^1|x_{l+1})}}{\sum_{j=1}^{M} W_U^j} \\ \times\, p\left(u^1|x_{l+1}\right) \prod_{j>1} p\left(u^j|x_{l+1}\right) p\left(v^j|\hat{x}_l\right) du^j dv^j\end{aligned}$$

Replacing u^1 by $\tilde{y}_l$ and rearranging gives,

$$\begin{aligned}R = M\ \pi_l(x_l)\,\pi_{l+1}(x_{l+1})\,\pi_l(x_{l+1}, \tilde{y}_l)\,\pi_{l+1}(\hat{x}_l) \\ \times \int \min\left\{\frac{1}{\pi_{l+1}(\hat{x}_l)\sum_{j=1}^{M} W_U^j},\ \frac{1}{\pi_{l+1}(x_{l+1})\sum_{j=1}^{M} W_V^j}\right\} \\ \times \prod_{j>1} p\left(u^j|x_{l+1}\right) p\left(v^j|\hat{x}_l\right) du^j dv^j.\end{aligned}$$

Since $x_{l+1} = \hat{y}_l$, $\pi_l(x_{l+1}, \tilde{y}_l) = \pi_l(y_l)$. Therefore, after replacing $\hat{x}_l$ by y_{l+1},

$$R = M\ \pi_l\left(x_l\right)\pi_{l+1}\left(x_{l+1}\right)\pi_l\left(y_l\right)\pi_{l+1}\left(y_{l+1}\right)$$
$$\times \int \min\left\{\frac{1}{\pi_{l+1}\left(y_{l+1}\right)\sum_{j=1}^{M} W_U^j},\ \frac{1}{\pi_{l+1}(x_{l+1})\sum_{j=1}^{M} W_V^j}\right\}$$
$$\times \prod_{j>1} p\left(u^j|x_{l+1}\right) p\left(v^j|y_{l+1}\right) du^j dv^j.$$

which is symmetric in (x_l, x_{l+1}) and (y_l, y_{l+1}). □

For small values of M in (13), calculation of the swap acceptance probabilities is very cheap. However, higher values of M may improve the acceptance rates. For example, if the $\{\pi_l\}_{i>0}$ are exact marginals of π_0, then $A_l \equiv 1$ while $A_l^M \leq 1$. In practice one has to balance the speed of evaluating A_l^M for small M with the possible higher acceptance rates for M large.

In analogy again with the multiple-try method, the above algorithm can be generalized to include correlated samples $\{U^j\}$ and $\{V^j\}$. This generalization is useful because it allows reference densities that cannot be independently sampled. Again consider a transition from $\{Y^n = x\}$ where

$$x = (\ldots, \hat{x}_l, \tilde{x}_l, x_{l+1}, \ldots)$$

to either $\{Y^{n+1} = x\}$ or $\{Y^{n+1} = y\}$ where

$$y = (\ldots, x_{l+1}, \tilde{y}_l, \hat{x}_l, \ldots).$$

First choose some reference transition densities $p_l^j\left(u^j|\left(u^0, \ldots, u^{j-1}\right),\ \hat{x}_l\right)$ that sample a variable U^j given the previous $j-1$ samples and the value of

the $\hat{x}_l$ variables. Let

$$p_l^j\left((u^{k+1},\ldots,u^j)|(u^0,\ldots,u^k),\ \hat{x}_l\right)=\prod_{k<m\leq j}p_l^m\left(u^m|(u^0,\ldots,u^{m-1}),\ \hat{x}_l\right). \quad (8)$$

For example, one might choose the $\left\{p_l^j\right\}$ to be a Markov transition kernel associated with some Markov chain Monte Carlo method with stationary measure $\pi_l(\tilde{x}_l|\hat{x}_l)$. Also let $\lambda^j\left((u^0,\ldots,u^j),\ \hat{x}_l,x_{l+1}\right)$ be any function satisfying the relation

$$\lambda^j\left((u^0,\ldots,u^j),\ \hat{x}_l,x_{l+1}\right)=\lambda^j\left((u^j,\ldots,u^0),\ x_{l+1},\hat{x}_l\right)$$

Algorithm 2 (Parallel Marginalization 2)**.** *We move the chain from Y^n to Y^{n+1} as follows:*

1. *For $j=1,\ldots,M$ sample U^j from $p_l^j\left(\ \cdot\ |(\tilde{x}_l,U^1,\ldots,U^{j-1}),\ x_{l+1}\right)$. Notice the conditioning on the value $\widehat{X}_l=x_{l+1}$.*

2. *Define the weights*

$$W_U^j=\pi_l\left(U^j,x_{l+1}\right)p_l^j\left(\left(U^{j-1},\ldots,U^1,\tilde{x}_l\right)|U^j,\ \hat{x}_l\right)$$
$$\times\lambda^j\left(\left(\tilde{x}_l,U^1,\ldots,U^j\right),\ \hat{x}_l,x_{l+1}\right).$$

   Notice the reversal in the ordering of the $\{U^j\}$ and the conditioning on $\widehat{X}_l=\hat{x}_l$.

3. *Choose the random index $J\in\{1,\ldots,M\}$ according to the probabilities*

$$\mathbf{P}\left[J=j\right]=\frac{W_U^j}{\sum_{m=1}^M W_U^m}.$$

   Set $\widetilde{Y}'=U^J$.

4. *Let* $V^J = \tilde{x}_l$ *and for* $j = 1, \ldots, J-1$ *let* $V^j = U^{J-j}$. *For* $j = J+1, \ldots, M$ *sample* V^j *from* $p_l^j\left(\ \cdot\ |(\widetilde{Y}', \ldots, V^{j-1}),\ \hat{x}_l\right)$. *Notice the conditioning on the value* $\widehat{X}_l = \hat{x}_l$.

5. *Define the weights*

$$W_V^j = \pi_l\left(V^j, \hat{x}_l\right) p_l^j\left((V^{j-1}, \ldots, V^1, \widetilde{Y}')|V^j,\ x_{l+1}\right) \times \lambda^j\left((\widetilde{Y}', V^1, \ldots, V^j),\ x_{l+1}, \hat{x}_l\right).$$

6. *Set*

$$Y^{n+1} = \left(\ldots, x_{l+1}, \widetilde{Y}', \hat{x}_l, \ldots\right)$$

with probability

$$A_l^M = \min\left\{1,\ \frac{\pi_{l+1}(\hat{x}_l)\sum_{j=1}^M W_U^j}{\pi_{l+1}(x_{l+1})\sum_{j=1}^M W_V^j}\right\} \tag{9}$$

and

$$Y^{n+1} = Y^n = (\ldots, \hat{x}_l, \tilde{x}_l, x_{l+1}, \ldots)$$

with probability $1 - A_l^M$.

The transition probability density for the above swap move from x to y for $x, y \in E_0 \times \cdots \times E_L$is again given by

$$\begin{aligned}\psi_l^M(y|x) = \mathbf{P}\left[\{\text{Swap is rejected}\}\right] \prod \delta_{\{y_j = x_j\}} \\ + \mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\widetilde{Y}' = \tilde{y}_l\right\}\right] \\ \times \delta_{\{(\hat{y}_l, y_{l+1}) = (x_{l+1}, \hat{x}_l)\}} \prod_{j \notin \{i, i+1\}} \delta_{\{y_j = x_j\}}\end{aligned}$$

where and δ is again the Dirac delta function. Again, the density $\mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\widetilde{Y}' = \tilde{y}_l\right\}\right]$ cannot and need not be evaluated.

Algorithm 1 can be derived from Algorithm 2 by setting

$$p_l^j\left(u^j|(u^0,\ldots,u^{j-1}),\ \hat{x}_l\right) = p_l\left(u^j|\hat{x}_l\right)$$

and

$$\lambda^j\left((u^0,\ldots,u^j),\ \hat{x}_l, x_{l+1}\right) = \frac{1}{p_l\left(u^j|\hat{x}_l\right) p_l\left(u^0|x_{l+1}\right)}.$$

Notice also that if

$$\lambda^j\left((u^0,\ldots,u^j),\hat{x}_l,x_{l+1}\right) = \frac{q^j\left((u^1,\ldots,u^{j-1})|\hat{x}_l,x_{l+1}\right)}{p_l^j\left((u^{j-1},\ldots,u^0)|u^j,\ \hat{x}_l\right) p_l^j\left((u^1,\ldots,u^j)|u^0,\ x_{l+1}\right)},$$

where, for each j, q^j is a conditional density satisfying $q^j\left((u^1,\ldots,u^{j-1})|\hat{x}_l,x_{l+1}\right) = q^j\left((u^{j-1},\ldots,u^1)|x_{l+1},\hat{x}_l\right)$ then

$$\mathbf{E}_{p^j}\left[W_U^j\right] = \int \pi_l(u^j,x_{l+1}) q^j\left((u_1,\ldots,u_{j-1})|\hat{x}_l,x_{l+1}\right)\prod_{i=1}^{j} du^i = \overline{\pi}_l(x_{l+1}).$$

Thus, if the $\{p^j\}$ generate an ergodic sequence, then $\frac{1}{M}\sum W_U^j \rightarrow \overline{\pi}_l(x_{i+1})$. The same holds for the $\left\{W_V^j\right\}$ so that

$$A_l^M \rightarrow \min\left\{1,\ \frac{\overline{\pi}_l(x_{l+1})\pi_{l+1}(\hat{x}_l)}{\overline{\pi}_l(\hat{x}_l)\pi_{l+1}(x_{l+1})}\right\} = A_l.$$

More general choices of $\{\lambda^j\}$ lead to A_l^M which converge to correpondingly more general acceptance probabilities than A_l.

Of course, expression (5) points the way to even more general algorithms. Algorithms 1 and 2 correspond to choices of γ in (5) that make the conditional expectation on the bottom of (5) equal to one. Other choices of γ may improve the variance of the resulting weights.

Proposition 3. *The transition probabilities ψ_l^M satisfy the detailed balance condition for the measure* Π.

Proof. Fix $x, y \in E_0 \times \cdots \times E_L$ such that $x \neq y$. For $x, y \in E_0 \times \cdots \times E_L$ such that $x \neq y$,

$$\Pi(x)\ \psi_l^M(y|x) = \Pi(x)\ \mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\widetilde{Y}' = \tilde{y}_l\right\}\right] \times \delta_{\{(\hat{y}_l, y_{l+1}) = (x_{l+1}, \hat{x}_l)\}} \prod_{j \notin \{i, i+1\}} \delta_{\{y_j = x_j\}},$$

As in the previous two proofs it can be assumed that $x_j = y_j$ for all j except l and $l+1$ and $(\hat{y}_l, y_{l+1}) = (x_{l+1}, \hat{x}_l)$. Since in this case $\pi(x_j) = \pi(y_j)$ for all $j \notin \{l, l+1\}$, it remains to show that if $(\hat{y}_l, y_{l+1}) = (x_{l+1}, \hat{x}_l)$ then

$$R\left((x_l, x_{l+1}), (y_l, y_{l+1})\right) = \pi_l(x_l)\, \pi_{l+1}(x_{l+1})\ \mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\widetilde{Y}' = \tilde{y}_l\right\}\right]$$

is symmetric in (x_l, x_{l+1}) and (y_l, y_{l+1}). Summing over disjoint events,

$$\mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\widetilde{Y}' = \tilde{y}_l\right\}\right] = \sum_{j=1}^{M} \mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\tilde{y}_l = U^j\right\} \cap \{J = j\}\right]$$

Thus R will be symmetric if for each j the function

$$R_j\left((x_l, x_{l+1}), (y_l, y_{l+1})\right) = \pi_l(x_l)\, \pi_{l+1}(x_{l+1})\ \mathbf{P}\left[\{\text{Swap is accepted}\} \cap \left\{\widetilde{Y}' = \tilde{y}_l\right\} \cap \{J = j\}\right]$$

is symmetric.

$$
\begin{aligned}
&R_j\left((x_l, x_{l+1}), (y_l, y_{l+1})\right) = \\
&\qquad \pi_l\left(x_l\right) \pi_{l+1}\left(x_{l+1}\right) \int \min\left\{1,\ \frac{\pi_{l+1}(\hat{x}_l) \sum_{k=1}^M W_U^k}{\pi_{l+1}(x_{l+1}) \sum_{k=1}^M W_V^k}\right\} \frac{W_U^k}{\sum_{k=1}^M W_U^k} \\
&\quad \times p_l^M\left((v^{j+1}, \ldots, v^M) | (\tilde{y}_l, v^1, \ldots, v^j),\ \hat{x}_l\right) p_l^M\left((u^1, \ldots, u^M) | \tilde{x}_l,\ x_{l+1}\right) \\
&\qquad \times \delta(u^j - \tilde{y}_l) \delta(v^j - \tilde{x}_l) \left(\prod_{1 \le k < j} \delta(u^{j-k} - v^k)\right) \left(\prod_{k>1,\ k \neq j} du^k dv^k\right)
\end{aligned}
$$

Recall the definition of the weights and the fact that $U^J = \tilde{y}_l$, $V^J = \tilde{x}_l$, and $V^j = U^{J-j}$ for $j = 1, \ldots, J-1$

$$
\begin{aligned}
W_U^j &= \pi_l\left(\tilde{y}_l, x_{l+1}\right) p_l^j\left(\left(u^{j-1}, \ldots, u^1, \tilde{x}_l\right) | \tilde{y}_l,\ \hat{x}_l\right) \lambda^j\left(\left(\tilde{x}_l, u^1, \ldots, u^{j-1}, \tilde{y}_l\right),\ \hat{x}_l, x_{l+1}\right) \\
&= \pi_l\left(\tilde{y}_l, x_{l+1}\right) p_l^j\left(\left(v^1, \ldots, v^{j-1}\right) | \tilde{y}_l,\ \hat{x}_l\right) \lambda^j\left(\left(\tilde{x}_l, u^1, \ldots, u^{j-1}, \tilde{y}_l\right),\ \hat{x}_l, x_{l+1}\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&R_j\left((x_l, x_{l+1}), (y_l, y_{l+1})\right) = \\
&\qquad \pi_l\left(x_l\right) \pi_{l+1}\left(x_{l+1}\right) \int \min\left\{1,\ \frac{\pi_{l+1}(\hat{x}_l) \sum_{k=1}^M W_U^k}{\pi_{l+1}(x_{l+1}) \sum_{k=1}^M W_V^k}\right\} \frac{1}{\sum_{k=1}^M W_U^k} \\
&\quad \times \pi_l\left(\tilde{y}_l, x_{l+1}\right) p_l^j\left(\left(v^1, \ldots, v^{j-1}\right) | \tilde{y}_l,\ \hat{x}_l\right) \lambda^j\left(\left(\tilde{x}_l, u^1, \ldots, u^{j-1}, \tilde{y}_l\right),\ \hat{x}_l, x_{l+1}\right) \\
&\quad \times p_l^M\left((v^{j+1}, \ldots, v^M) | (\tilde{y}_l, v^1, \ldots, v^j),\ \hat{x}_l\right) p_l^M\left((u^1, \ldots, u^M) | \tilde{x}_l,\ x_{l+1}\right) \\
&\qquad \times \delta(u^j - \tilde{y}_l) \delta(v^j - \tilde{x}_l) \left(\prod_{1 \le k < j} \delta(u^{j-k} - v^k)\right) \left(\prod_{k>1,\ k \neq j} du^k dv^k\right)
\end{aligned}
$$

Definition (8) implies that for all j,

$$
\begin{aligned}
&p_l^j\left(\left(v^1, \ldots, v^{j-1}\right) | \tilde{y}_l,\ \hat{x}_l\right) p_l^M\left((v^{j+1}, \ldots, v^M) | (\tilde{y}_l, v^1, \ldots, v^{j-1}, \tilde{x}_l),\ \hat{x}_l\right) \\
&\qquad\qquad = p_l^M\left(\left(v^1, \ldots, v^M\right) | \tilde{y}_l,\ \hat{x}_l\right),
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&R_j\left((x_l, x_{l+1}), (y_l, y_{l+1})\right) = \\
&\qquad \pi_l\left(x_l\right)\pi_{l+1}\left(x_{l+1}\right)\int \min\left\{1,\ \frac{\pi_{l+1}(\hat{x}_l)\sum_{k=1}^{M} W_U^k}{\pi_{l+1}(x_{l+1})\sum_{k=1}^{M} W_V^k}\right\} \\
&\qquad\qquad \times \frac{\pi_l\left(\tilde{y}_l, x_{l+1}\right)\lambda^j\left((\tilde{x}_l, u^1, \ldots, u^{j-1}, \tilde{y}_l),\ \hat{x}_l, x_{l+1}\right)}{\sum_{k=1}^{M} W_U^k} \\
&\qquad \times p_l^M\left((v^1, \ldots, v^M)|\tilde{y}_l,\ \hat{x}_l\right) p_l^M\left((u^1, \ldots, u^M)|\tilde{x}_l,\ x_{l+1}\right) \\
&\qquad \times \delta(u^j - \tilde{y}_l)\delta(v^j - \tilde{x}_l)\left(\prod_{1\le k<j} \delta(u^{j-k} - v^k)\right)\left(\prod_{k>1,\ k\neq j} du^k dv^k\right)
\end{aligned}
$$

which can be rewritten,

$$
\begin{aligned}
&R_j\left((x_l, x_{l+1}), (y_l, y_{l+1})\right) = \pi_l\left(x_l\right)\pi_{l+1}\left(x_{l+1}\right)\pi_l\left(\tilde{y}_l, x_{l+1}\right)\pi_{l+1}\left(\hat{x}_l\right) \\
&\qquad \times \int \min\left\{\frac{1}{\pi_{l+1}(\hat{x}_l)\sum_{k=1}^{M} W_U^k},\ \frac{1}{\pi_{l+1}(x_{l+1})\sum_{k=1}^{M} W_V^k}\right\} \\
&\qquad\qquad \times \lambda^j\left(\left(\tilde{x}_l, u^1, \ldots, u^{j-1}, \tilde{y}_l\right),\ \hat{x}_l, x_{l+1}\right) \\
&\qquad \times p_l^M\left((v^1, \ldots, v^M)|\tilde{y}_l,\ \hat{x}_l\right) p_l^M\left((u^1, \ldots, u^M)|\tilde{x}_l,\ x_{l+1}\right) \\
&\qquad \times \delta(u^j - \tilde{y}_l)\delta(v^j - \tilde{x}_l)\left(\prod_{1\le k<j} \delta(u^{j-k} - v^k)\right)\left(\prod_{k>1,\ k\neq j} du^k dv^k\right)
\end{aligned}
$$

Plugging $y_{l+1} = \hat{x}_l$ and $\hat{y}_l = x_{l+1}$, into this expression yields,

$$\begin{aligned} R_j\left((x_l, x_{l+1}), (y_l, y_{l+1})\right) &= \pi_l\left(x_l\right)\pi_{l+1}\left(x_{l+1}\right)\pi_l\left(y_l\right)\pi_{l+1}\left(y_{l+1}\right) \\ &\times \int \min\left\{ \frac{1}{\pi_{l+1}(y_{l+1})\sum_{k=1}^{M} W_U^k},\ \frac{1}{\pi_{l+1}(x_{l+1})\sum_{k=1}^{M} W_V^k} \right\} \\ &\times \lambda^j\left(\left(\tilde{x}_l, u^1, \ldots, u^{j-1}, \tilde{y}_l\right),\ y_{l+1}, x_{l+1}\right) \\ &\times p_l^M\left((v^1, \ldots, v^M)|\tilde{y}_l,\ y_{l+1}\right) p_l^M\left((u^1, \ldots, u^M)|\tilde{x}_l,\ x_{l+1}\right) \\ &\times \delta(u^j - \tilde{y}_l)\delta(v^j - \tilde{x}_l)\left(\prod_{1\le k<j} \delta(u^{j-k} - v^k)\right)\left(\prod_{k>1,\ k\neq j} du^k dv^k\right) \end{aligned}$$

By the symmetry property of λ_j this expression is symmetric in (x_l, x_{l+1}) and (y_l, y_{l+1}). □

Clearly a Markov chain that evolves only by swap moves cannot sample all configurations, ie. the chain generated by ψ is not ϕ-irreducible for any non trivial measure ϕ. These swap moves must therefore be used in conjunction with a transition rule that can reach any region of space. More precisely, let τ from expression (3) be Harris recurrent with stationary distribution Π (see [15]). The the transition rule for parallel marginalization is

$$\tau_{pm}(y|x) = (1-\alpha)\ \tau(y|x) + \alpha \int \tau(z|x)\psi\left(y|z\right) dz$$

where

$$\psi(y|x) = \sum_{k=0}^{L-1} \frac{1}{L} \psi_l^M\left(y|x\right)$$

and $\alpha \in [0, 1)$ is the probability that a swap move occurs. τ_{pm} 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 $\{0, \ldots, L-1\}$. Next, the

chain evolves according to τ. With probability $1-\alpha$ the chain moves only according to τ and does not attempt a swap. The next result guarantees the invariance of Π under evolution by τ_{pm}.

It is not difficult to verify that the chain generated by τ_{pm} has invariant measure Π and is Harris recurrent if the chain generated by τ has these properties. Thus by combining standard MCMC steps on each component, governed by the transition probability τ, with swap steps between the components governed by ψ, an MCMC method results that not only uses information from rapidly equilibrating lower dimensional chains, but is also convergent.

3 Numerical Examples

In this section I consider applications of parallel marginalization to two conditional path sampling problems for a one dimensional stochastic differential equation,

$$dZ(t) = f\left(Z(t)\right) dt + \sigma\left(Z(t)\right) dW(t), \tag{10}$$

where f and σ are real valued functions of $\mathbb{R}$. One must first approximate $\{Z(t)\}$ by a discrete process for which the path density is readily available. Let $t_0 = 0, t_1 = \frac{T}{N}, \ldots, t_N = T$ be a mesh on which one wishes to calculate path averages. One such approximate process is given by the linearly implicit

Euler scheme (a balanced implicit method, see [16]),

$$\begin{aligned} X(n+1) &= X(n) + f\left(X(n)\right) \triangle \\ &\quad + \left(X(n+1) - X(n)\right) f^{'}\left(X(n)\right) \triangle + \sigma\left(X(n)\right) \sqrt{\triangle}\, \xi(n), \qquad (11) \\ X(0) &= Z(0). \end{aligned}$$

Here $X(n)$ is an approximation to Z at time t_n. The reader should note that the rate of convergence of the above scheme to the solution of (10) would not be effected by the insertion in (11) of a non-negative constant in front of the $f^{'}$ term. The choice of 1 made here seemed to improve the stability of the resulting scheme for large values of $\triangle$. The $\{\xi(n)\}$ are independent Gaussian random variables with mean 0 and variance 1, and $\triangle = \frac{T}{N}$. N is assumed to be a power of 2. The choice of this scheme over the Euler scheme (see [17]) is due to its favorable stability properties as explained later. It is henceforth assumed that $X(t)$ instead of $Z(t)$ is the process of interest.

The first of the conditional sampling problems discussed here is the bridge sampling problem in which one generates samples of transition paths between two states. This problem arises, for example, in financial volatility estimation where, given a sequence of observations, $(z(s_0), \ldots, z(s_K))$ with $\{s_j\} \subset \{t_l\}$, the goal is to estimate the diffusion term σ (assumed here to be constant) appearing in the stochastic differential equation. Since in general one cannot easily evaluate the transition probability between times s_j and s_{j+1} (and thus the likelihood of the observations) it is necessary to generate samples between the observations,

$$V(j) = (X(j,1), \ldots, X(j,N_j))$$

where $N_j = N\,(s_{j+1} - s_j) - 1$ (assumed to be an integer) and $X(j,n)$ denotes the value of the process at time $s_j + \frac{n}{N_j}$. It is then easy to evaluate the likelihood of a path

$$X(s_0), V(0), \ldots, X(s_K), V(K)$$

given a particular value of the volatility, σ.

The filtering/smoothing problem is similar to the financial volatility example of the previous paragraph except that now it is assumed that the observations are noisy functions of the underlying process. For example, one may wish to sample possible trajectories taken by a rocket given somewhat unreliable GPS observations of its position. If the conditional density of the observations given the position of the rocket is known, it is possible to generate conditional samples of the trajectories.

3.1 Bridge path sampling

In the bridge path sampling problem one seeks to approximate conditional expectations of the form

$$\mathbf{E}\left[g\left(Z(t_1), \ldots, Z(t_{N-1})\right) \;|\{Z(0) = z^-\}, \{Z(T) = z^+\}\right]$$

where g is a real valued function, and $\{Z(t)\}$ is solution to (10).

Without the condition $Z(T) = z^+$ above, generating an approximate sample $(X(0), \ldots, X(N))$ path is a relatively straitforward endeavor. One simply generates a sample of $Z(0)$, then evolves (11) with this initial condition. However, the presence of information about $\{Z(t)\}_{t>0}$ complicates the task. In

general, some sampling method which requires only knowlege of a function proportional to conditional density of $(X(1), \ldots, X(N-1))$ must be applied. The approximate path density associated with discretization (11) is

$$\pi_0\left(x_0(1), \ldots, x_0(N-1) \mid x_0(0), x_0(N)\right) \propto \exp\left(-\sum_{k=0}^{N-1} \mathcal{V}\left(x_0(n), x_0(n+1), \triangle\right)\right) \tag{12}$$

where

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

At this point the parallel marginalization sampling procedure is applied 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 [6, 8]. These calculations can be carried out before implementation of parallel marginalization, or they can be integrated into the sampling procedure.

In some cases, computer generation of the $\{\pi_l\}_{l>0}$ can be completely avoided. The examples presented here are two such cases. Let $S_l = \left\{0, 2^l, 2(2^l), 3(2^l), \ldots, N\right\}$ (recall N is a power of 2). Decompose S_l as $\widehat{S}_l \sqcup \widetilde{S}_l$ where

$$\widehat{S}_l = \left\{0, 2(2^l), 4(2^l), 6(2^l), \ldots, N\right\}$$

and

$$\widetilde{S}_l = \left\{2^l, 3(2^l), 5(2^l), 7(2^l), \ldots, N - 2^l\right\}.$$

In the notation of the previous sections, $x_l = (\hat{x}_l, \tilde{x}_l)$ where $\hat{x}_l = \{x_l(n)\}_{n \in \widehat{S}_l \setminus \{0,N\}}$ and $\tilde{x}_l = \{x_l(n)\}_{n \in \widetilde{S}_l}$. In words, the hat and tilde variables represent alternat-

ing time slices of the path. For all l fix $x_l(0) = z^-$ and $x_l(N) = z^+$. We choose the approximate marginal densities

$$\pi_l\left(\{x_l(n)\}_{n\in S_l\setminus\{0,N\}} \;|x_l(0), x_l(N)\right) \propto q_l\left(\{x_l(n)\}_{n\in S_l}\right)$$

where for each l, q_l is defined by successive coarsenings of (11). That is,

$$q_l\left(\{x_l(n)\}_{n\in S_l}\right) = \exp\left(-\sum_{k=0}^{N/2^l-1} \mathcal{V}\left(x_l(2^l k), x_l(2^l(k+1)), 2^l\triangle\right)\right).$$

Since π_l will be sampled using a Metropolis-Hastings method with $x(0)$ and $x(N)$ fixed, knowlege of the normalization constants

$$\mathcal{Z}_l\left(x_l(0), x_l(N)\right) = \int q_l \prod_{n\in S_l\setminus\{0,N\}} dx_l(n)$$

is unnecessary.

Notice from (12) that, conditioned on the values of $X(n-1)$ and $X(n+1)$, the variance of $X(n)$ is of order $\triangle$. Thus any perturbation of $X(n)$ which leaves $X(m)$ fixed for $m \neq n$ and which is compatible with joint distribution (12) must be of the order $\sqrt{\triangle}$. This suggests that distributions defined by coarser discretizations of (12) 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 $z^- = z^+ = 0$. $\triangle = 2^{-10}$. $Y_l^n \in \mathbb{R}^{10/\left(2^l \triangle\right)+1}$ is the l^{th} level of the parallel marginalization Markov chain at algorithmic time n. There are 10 chains ($L = 9$). The observed swap acceptance rates are reported in table (1). Notice that the swap rates are highest at the lower levels but seems to stabilize at the higher levels.

Table 1: Swap acceptance rates for bridge sampling problem

Levels[1]	0/1	1/2	2/3	3/4	4/5	5/6	6/7	7/8	8/9
	0.86	0.83	0.75	0.69	0.54	0.45	0.30	0.22	0.26

[1] Swaps between levels l and $l + 1$.

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 Figure 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 using 1 dimensional Gaussian random walk proposals. 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.

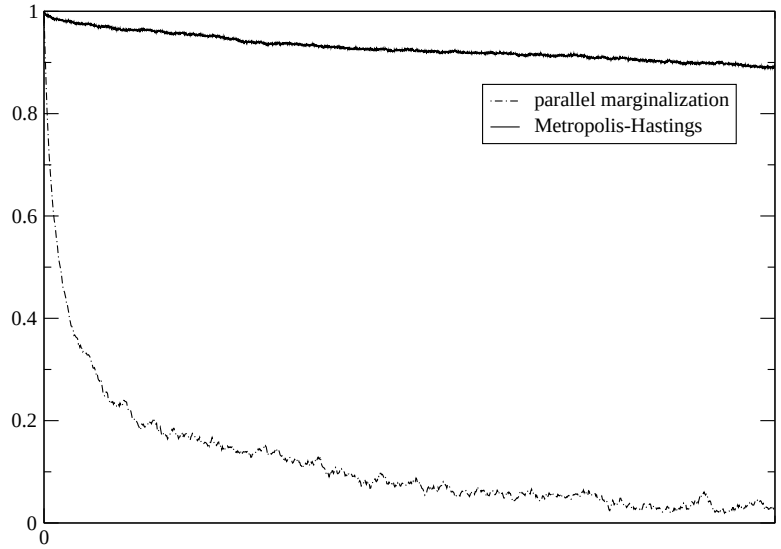

Figure 1: Autocorrelation of Y^n_{mid} for Metropolis-Hastings method with 1-d Gaussian random walk proposals (solid) and parallel marginalization (dotted). The x-axis runs from 0 to 10000 iterations of the Metropolis-Hastings method and from 0 to 1000 iterations of parallel marginalization. This rescaling more than compensates for the extra work for parallel marginalization per iteration.

In these numerical examples, parallel marginalization is applied with a slight simplification as detailed in the following algorithm.

Algorithm 3. *The chain moves from Y^n to Y^{n+1} as follows:*

1. *Generate M independent Gaussian random paths $\{\zeta^m(n)\}_{n\in\widetilde{S}_l}$ with independent components $\zeta^m(n)$ of mean 0 and variance $2^{l-1}\triangle$.*

2. *For each j and $n \in \widetilde{S}_l$ let*

$$U^m(n) = \zeta^m(n) + 0.5\left(x_{l+1}(n-1) + x_{l+1}(n+1)\right).$$

3. *Define the weights*

$$W_U^m = \frac{\pi_l\left(x_{l+1}, U^m\right)}{p_l\left(U^m | x_{l+1}\right)},$$

where p_l is defined by the choice in step 1 as

$$p_l\left(\tilde{x}_l | \hat{x}_l\right) \propto \exp\left(\sum_{n \in \widetilde{S}_l} -\frac{\left(\tilde{x}_l(n) - 0.5\left(\hat{x}_l(n-1) + \hat{x}_l(n+1)\right)\right)^2}{2^l \triangle}\right).$$

4. *Choose $J \in \{1, \ldots, M\}$ according to the probabilities*

$$\mathbf{P}\left[J = j\right] = \frac{W_U^j}{\sum_{k=1}^M W_U^k}.$$

Set $\widetilde{Y}' = U^J$.

5. *Set $V^J = \tilde{x}_l$ and for $j \neq J$ set*

$$V^j(n) = \zeta^j(n) + 0.5\left(\hat{x}_l(n-1) + \hat{x}_l(n+1)\right).$$

6. *Define the weights*

$$W_V^m = \frac{\pi_l\left(\hat{x}_l, V^m\right)}{p_l\left(V^m | \hat{x}_l\right)}.$$

7. *Set*

$$Y^{n+1} = \left(\ldots, x_{l+1}, \tilde{y}_l, \hat{x}_l, \ldots\right)$$

with probability

$$A_l^M = \min\left\{1, \ \frac{\pi_{l+1}(\hat{x}_l) \sum_{m=1}^M W_U^m}{\pi_{l+1}(x_{l+1}) \sum_{m=1}^M W_V^m}\right\} \tag{13}$$

and

$$Y^{n+1} = Y^n = \left(\ldots, \hat{x}_l, \tilde{x}_l, x_{l+1}, \ldots\right)$$

with probability $1 - A_l^M$.

This simplification reduces by half the number of gaussian random variables needed to evaluate the acceptance probability but may not be appropriate in all settings. For this problem, the choice of M in (13), the number of samples $\{U^m\}$ and $\{V^m\}$, seems to have little effect on the swap acceptance rates. In the numerical experiment $M = l + 1$ for swaps between levels l and $l + 1$.

The results of the Metropols-Hastings and parallel marginalization methods applied to the above bridge sampling problem after a run time of 10 minutes on a standard workstation are presented in Figures 2 and 3. Apparently the sample generated by parallel marginalization is a reasonable bridgepath while the Metropolis-Hastings method has clearly not converged.

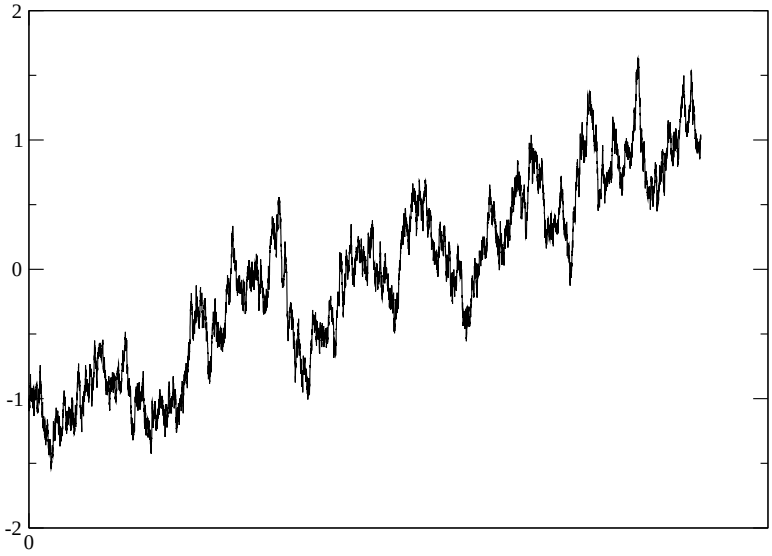

Figure 2: Metropolis generated bridge path from Section 3.1 after a 10 minute run on a standard desktop workstation. Clearly the method has not converged.

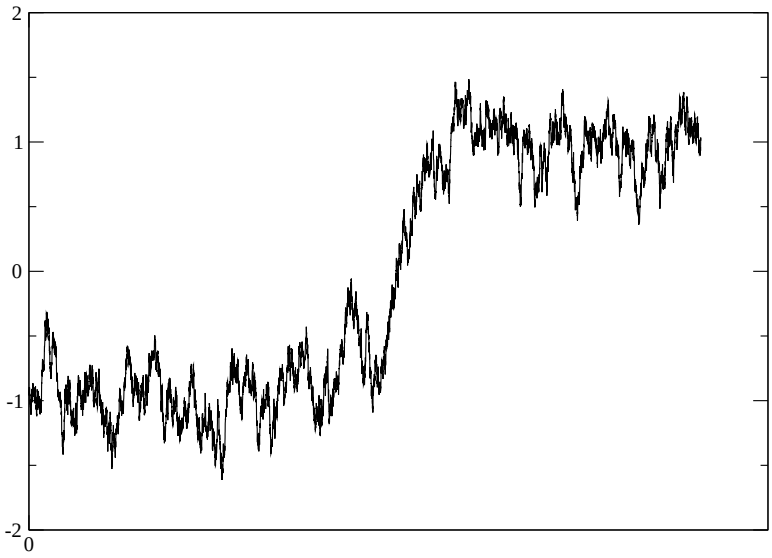

Figure 3: Parallel marginalization generated bridge path from Section 3.1 after a 10 minute run on a standard desktop workstation. Apparently the method has converged.

3.2 Non-linear smoothing/filtering

In the non-linear smoothing and filtering problem one seeks to approximate conditional expectations of the form

$$\mathbf{E}\left[g\left(Z(0), Z(t_1), \ldots, Z(T)\right) \mid \{H(j) = h(j)\}_0^K\right]$$

where the real valued processes $\{Z(t)\}$ and $\{H(j)\}$ are given by the system

$$\begin{aligned} dZ(t) &= f\left(Z(t)\right) dt + \sigma\left(Z(t)\right) dW(t), \\ H(j) &= r\left(Z(s_j)\right) + \chi(j), \\ Z(0) &\sim \rho, \qquad \chi(j) \sim i.i.d.\ \mu. \end{aligned} \tag{14}$$

g, f, σ, and r are real valued functions of $\mathbb{R}$. The $\{\chi(j)\}$ are real valued independent random variable drawn from the density μ and are independent of the Brownian motion $\{W(t)\}$. $\{s_j\} \subset \{t_j\}$, and $0 = s_0 < s_1 < ... < s_K = T$. The process $Z(t)$ is a hidden signal and the $\{H(j)\}$ are noisy observations. The idea of computing the above conditional expectation by conditional path sampling has been suggested in [18, 19]. Popular alternatives include particle filters (see [20]) and ensemble Kalman filters (see [21]).

Again, begin by discretizing the system. Assume that $N_j = N(s_{j+1} - s_j) - 1$ is an integer and let $\triangle = \frac{T}{N}$. The linearly implicit Euler scheme gives

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

where $X(j,n)$ represents the discrete time approximation to $Z(s_j + n\triangle)$, for $0 \leq n \leq N_j$. The $\{\xi(n)\}$ are independent Gaussian random variables with mean 0 and variance 1. The $\{\xi(n)\}$ are independent of the $\{\chi^m\}$. N is again assumed to be a power of 2.

The approximate path measure for this problem is

$$\begin{aligned}
\pi_0(x_0(0), \ldots, x_0(N) \,|h(0), \ldots, h(K)) \propto \exp\left(-\sum_{k=0}^{N-1} \mathcal{V}(x_0(n), x_0(n+1), \triangle)\right) \\
\times \rho(x_0(0)) \prod_{n=0}^{K} \mu(x_0(j) - r(h(j)))
\end{aligned}$$

The approximate marginals are chosen as

$$\pi_l\left(\{x_l(n)\}_{n\in S_l}\ \middle|h(0),\ldots,h(K)\right)\propto$$
$$q_l\left(\{x_l(n)\}_{n\in S_l}\right)\rho\left(x_l(0)\right)\prod_{n=0}^{K}\mu\left(x_l(j)-r\left(h(j)\right)\right)$$

where V, q_l and S_l 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 function r in (14) is the identity function. 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 $j = 6, \ldots, 10$. $\triangle = 2^{-10}$. There are 8 chains ($L = 7$). The observed swap acceptance rates are reported in table (2). Notice that the swap rates are again highest at the lower levels but, for this problem, become unreasonably small at the highest level.

Table 2: Swap acceptance rates for filtering/smoothing problems

Levels[1]	0/1	1/2	2/3	3/4	4/5	5/6	6/7
	0.86	0.83	0.74	0.65	0.46	0.23	0.04

[1] Swaps between levels l and $l+1$.

Again, $Y^n_{mid} \in \mathbb{R}$ denotes the midpoint of the path defined by Y^n_0 (i.e. an approximate sample of the path at time 5). In Figure 4 the autocorrelation of

Y^n_{mid} 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.

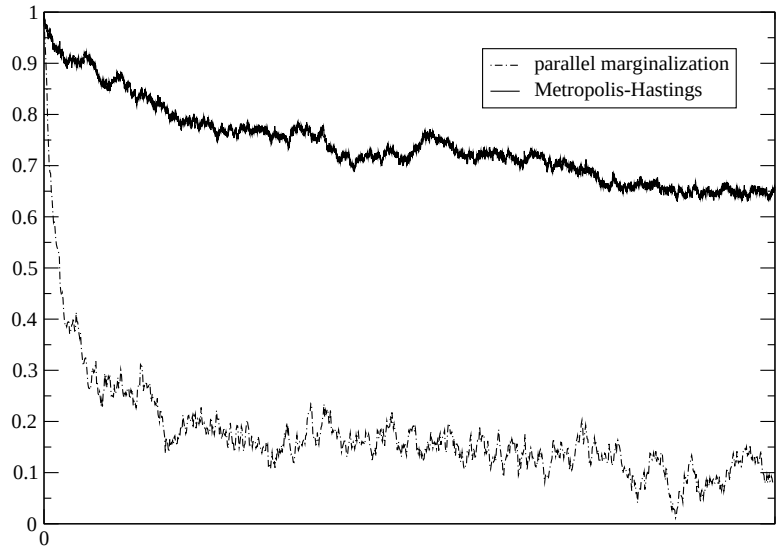

Figure 4: Autocorrelation of Y^n_{mid} for Metropolis-Hastings method with 1-d Gaussian random walk proposals (solid) and parallel marginalization (dotted). The x-axis runs from 0 to 10000 iterations of the Metropolis-Hastings method and from 0 to 1000 iterations of parallel marginalization. This rescaling more than compensates for the extra work for parallel marginalization per iteration.

The algorithm is modified as in the previous example. For this problem, acceptable swap rates require a higher choice of M in (13) than needed in the bridge sampling problem. In this numerical experiment $M = 2^l$ for swaps between levels l and $l + 1$.

The results of the Metropols-Hastings and parallel marginalization meth-

ods applied to the smoothing problem above after a run time of 10 minutes on a standard workstation are presented in figure 5 and 6. Apparently the sample generated by parallel marginalization is a reasonable bridgepath while the Metropolis-Hastings method has clearly not converged.

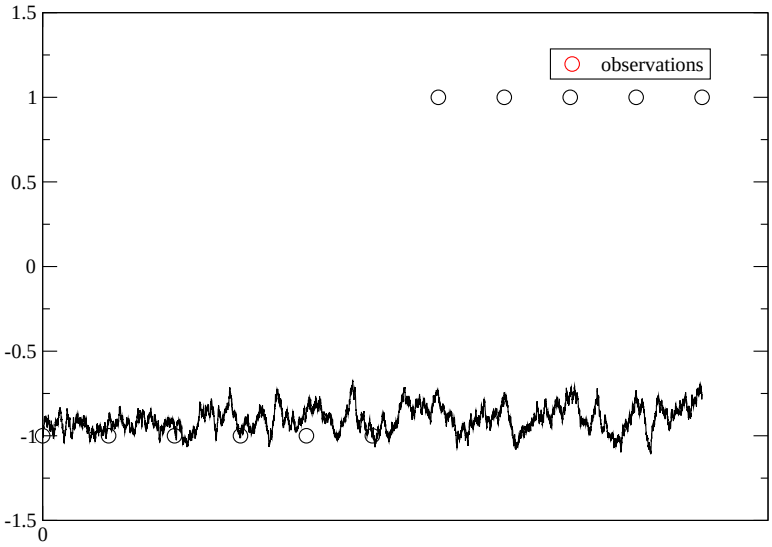

Figure 5: Metropolis-Hastings generated smoothed path from Section 3.2 after a 10 minute run on a standard desktop workstation. Clearly the method has not converged.

4 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

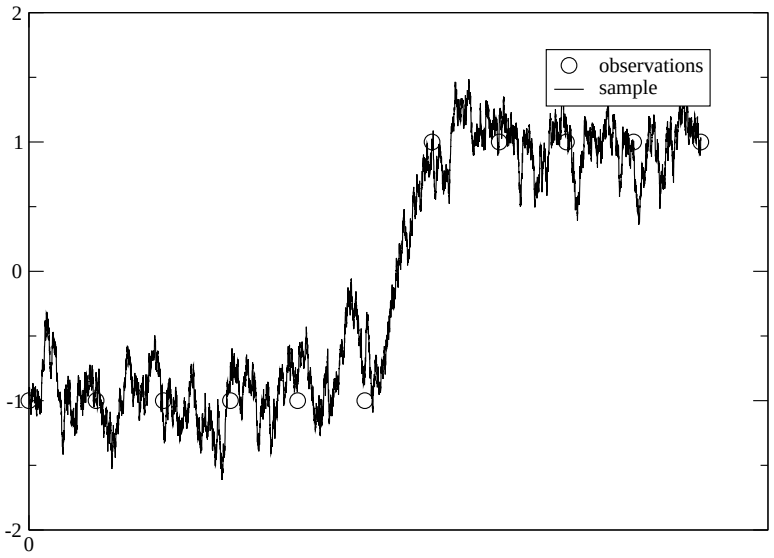

Figure 6: Parallel marginalization generated smoothed path from Section 3.2 after a 10 minute run on a standard desktop workstation. Apparently the method has converged.

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 [6] and [8].

5 Acknowledgments

I would like to thank Prof. A. Chorin for his guidance during this research, which was carried out while I was a Ph.D. student at U. C. Berkeley. I would also like to thank Prof. O. Hald, Dr. P. Okunev, Dr. P. Stinis, and Dr. Xuemin Tu, for their very helpful 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 and National Science Foundation grant DMS0410110.

References

[1] Leo P. Kadanoff. *Statistical physics*. World Scientific Publishing Co. Inc., River Edge, NJ, 2000. Statics, dynamics and renormalization.

[2] J.J. Binney, N.J.Dowrick, A.J. Fisher, and M.E.J. Newman. *The Theory of Critical Phenomena: An Introduction to the Renormalization Group*. Oxford University Press, USA, 1992.

[3] J. Goodman and A. Sokal. Multigrid Monte Carlo methods for lattice field theories. *Phys. Rev. Lett.*, 56:1015–1018, 1986.

[4] J. Goodman and A. Sokal. Multigrid Monte Carlo method: Conceptual foundations. *Phys. Rev. D*, 40(6):2035–2071, 1989.

[5] A. Brandt and D. Ron. Renormalization multigrid (rmg): coarse-to-fine Monte Carlo acceleration and optimal derivation of macroscopic descriptions. *J. Stat. Phys.*, 102(1/2):163–186, 2001.

[6] Alexandre J. Chorin. Conditional expectations and renormalization. *Multiscale Model. Simul.*, 1(1):105–118 (electronic), 2003.

[7] Alexandre J. Chorin and Ole H. Hald. *Stochastic tools in mathematics and science*, volume 1 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer, New York, 2006.

[8] Panagiotis Stinis. A maximum likelihood algorithm for the estimation and renormalization of exponential densities. *J. Comput. Phys.*, 208(2):691–703, 2005.

[9] Pavel Okunev. *Renormalization methods with applications to spin physics*. U. C. Berkeley Math. Dept., 2005.

[10] Alexandre J. Chorin. Monte Carlo without chains, with applications to spin glasses. *in preparation*.

[11] Jonathan Weare. Efficient conditional path sampling of stochastic differential equations by parallel marginalization. *submitted to Proc. Nat. Acad. Sci. USA*.

[12] Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2002.

[13] Jun S. Liu, Faming Liang, and Wing Hung Wong. The multiple-try method and local optimization in metropolis sampling. *J. Amer. Statist. Assoc.*, 95(449):121–134, 2000.

[14] Z. Qin and J. S. Liu. Multi-point metropolis method with application to hybrid Monte Carlo. *J. Comp. Phys.*, 172:827–840, 2001.

[15] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993.

[16] G. N. Milstein, E. Platen, and H. Schurz. Balanced implicit methods for stiff stochastic systems. *SIAM J. Numer. Anal.*, 35(3):1010–1019 (electronic), 1998.

[17] Peter E. Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1992.

[18] Francis Alexander, Greg Eyink, and Juan Restrepo. Accelerated Monte Carlo for optimal estimation of time series. *J. Stat. Phys.*, 119:1331–1345, 2004.

[19] Andrew M. Stuart, Jochen Voss, and Petter Wiberg. Fast communication conditional path sampling of SDEs and the Langevin MCMC method. *Commun. Math. Sci.*, 2(4):685–697, 2004.

[20] Nando De Freitas, Arnaud Doucet, and Neil Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2005.

[21] G. Evensen. The ensemble kalman filter: theoretical formulation and practical impementation. *Ocean Dynamics*, 53:343–367, 2003.