ISBA 2000, Proceedings, pp. 000–000

Exploring Hybrid Monte Carlo in Bayesian Computation

Lingyu Chen *Stanford University*
Zhaohui Qin *Harvard University*
Jun S. Liu *Harvard University*

SUMMARY

Hybrid Monte Carlo (HMC) has been successfully applied to molecular simulation problems since its introduction in the late 1980s. Its use in Bayesian computation, however, is relatively recent and rare (Neal 1996). In this article, we investigate statistical models in which HMC shows an edge over the more standard Monte Carlo techniques such as the Metropolis algorithm and the Gibbs sampler. The models under investigation include the indirect observation model, nonlinear state-space model and non-linear random-effects model. We also propose two methods, the multi-point method and parallel tempering, for improving HMC's efficiency.

Keywords: HYBRID MONTE CARLO; MARKOV CHAIN MONTE CARLO; MULTIPLE-POINT; PARALLEL TEMPERING; INDIRECT OBSERVATION; STOCHASTIC VOLATILITY; NONLINEAR RANDOM-EFFECTS MODEL; BAYESIAN COMPUTATION.

1. MOTIVATING PROBLEM

Hybrid Monte Carlo (HMC) as first introduced by Duane *et al.* (1987) is a Markov chain Monte Carlo (MCMC) technique built upon the basic principle of Hamiltonian mechanics. Its applications in molecular simulation have attracted much interest from researchers. Its potential in Bayesian computation, however, has not been fully explored. We show in this article that HMC can be very effective means for exploring complex posterior distributions.

To motivate our investigation, consider the following indirect observation model. Let θ be a parameter vector and let X_θ be a vector of random variables whose distribution is completely known given θ. Suppose we observe only Y, where

$$Y = g(X_\theta, \theta), \tag{1}$$

and the functional form of $g(\cdot)$ is known, whereas X_θ is not directly observable. Of interest is the Bayesian inference on the parameter vector θ. Since the analytical computation of the likelihood function of θ is generally infeasible (when g is complex), the standard maximum likelihood estimation method cannot be applied. We overcome this difficulty by formulating a new model which can be viewed as a "contaminated version" of (1):

$$Y = g(X_\theta, \theta) + \epsilon \tag{2}$$

where $\epsilon \sim N(0, \sigma^2 I)$, I is the identity matrix and σ is a tuning parameter controlled by the user. Treating the problem as a usual missing-data problem, we write the "pseudo-posterior distribution" of X and θ as follows:

$$\pi_\sigma(X, \theta \mid Y) \propto g_\sigma(Y \mid X, \theta) f(X \mid \theta) \pi_0(\theta) \tag{3}$$

where g_σ represents the density function of the model (2) and π_0 is the prior for θ. It can be shown that under mild conditions, the "pseudo-posterior" of θ converges to its true posterior almost surely as $\sigma^2 \to 0$.

However, producing satisfactory Monte Carlo samples from (3) is not easy to achieve either. Although a MCMC procedure such as the Metropolis algorithm (Metropolis *et al.* 1953) might be applicable, the random walk nature of the algorithm makes it very inefficient to explore the posterior distribution defined by (3). For instance, in the following trivial example

$$y = \theta x, \qquad x \sim N(\theta, 1),$$

the samples from the "pseudo-posterior" (3), when σ is small, lie in the vicinity of the curve $x\theta = y$, as displayed in Figure 1. When $\sigma = 0.05$, for example, it needs many iterations for a Metropolis sampler to traverse the entire banana-shaped valley depicted in the figure and the situation becomes worse as σ decreases. In contrast, HMC can follow the dynamics of this distribution rather well.

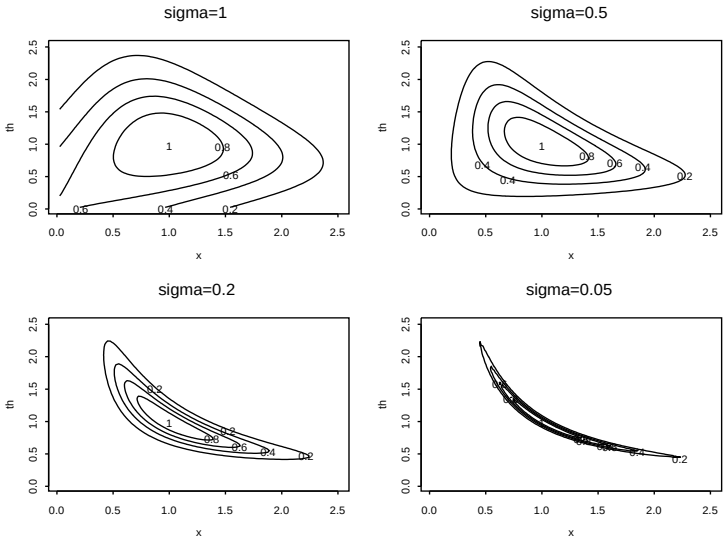

Figure 1. *The contour plots of density $\pi_\sigma(x, \theta)$ with a flat prior $\pi_0(\theta) \equiv c$.*

The remainder of this paper is organized as follows. Section 2 reviews briefly the general HMC procedure. Section 3 describes two improvement methods. Section 4 studies some examples and compares HMC with some other MCMC approaches. Section 5 concludes with a brief discussion.

2. HYBRID MONTE CARLO

Suppose we wish to draw Monte Carlo samples from $\pi(\mathbf{x}) \propto \exp\{-U(\mathbf{x})\}$, where $\mathbf{x} = (x_1, \cdots, x_d)$. In physics contexts, $\mathbf{x}$ can be regarded as a position vector and $U(\mathbf{x})$ the potential energy function. We introduce a fictitious "momentum vector" $\mathbf{p} = (p_1, \cdots, p_d)$ and the corresponding kinetic energy $K(\mathbf{p}) = \frac{1}{2}\sum_{i=1}^{d} p_i^2/m_i$, where m_i represents the "mass" of component i and we write $\mathbf{m} = (m_1, \ldots, m_d)$. The total energy is then

$$H(\mathbf{x}, \mathbf{p}) = U(\mathbf{x}) + K(\mathbf{p}). \tag{4}$$

Clearly, if we can sample $(\mathbf{x}, \mathbf{p})$ from the distribution $\pi(\mathbf{x}, \mathbf{p}) \propto \exp\{-H(\mathbf{x}, \mathbf{p})\}$, then the marginal distribution of $\mathbf{x}$ is exactly the target distribution $\pi(\mathbf{x})$.

On the other hand, if a physical system under consideration conserves the total energy (i.e. H remains as a constant), then its evolution dynamics can be described by the *Hamiltonian equations* which are derived by differentiating (4) with respect to $\mathbf{x}$ and $\mathbf{p}$:

$$\frac{\partial H}{\partial \mathbf{x}} = -\dot{\mathbf{p}}, \qquad \frac{\partial H}{\partial \mathbf{p}} = \dot{\mathbf{x}}. \tag{5}$$

Because the Hamiltonian dynamics are time-reversible, volume-preserving, and energy-preserving, the resulting moves leave $\pi(\mathbf{x}, \mathbf{p})$ invariant. That is, if $(\mathbf{x}^{(0)}, \mathbf{p}^{(0)}) \sim \pi$, then after the conserved system evolves for time t, the new configuration at time t, $(\mathbf{x}^{(t)}, \mathbf{p}^{(t)})$, also follows distribution π.

In practice, the Hamiltonian dynamics is often approximated by a discretized version, called the *the leapfrog algorithm*, with a small time step-size δ:

$$\begin{aligned} \mathbf{x}(t+\delta) &= \mathbf{x}(t) + \delta \frac{\mathbf{p}(t+\delta/2)}{\mathbf{m}} \\ \mathbf{p}(t+\delta/2) &= \mathbf{p}(t-\delta/2) - \delta \left.\frac{\partial U}{\partial \mathbf{x}}\right|_{\mathbf{x}(t)} \end{aligned}$$

where the ratio between two vectors is operated component-wise. Although each leapfrog move remains time-reversible and volume-preserving, it no longer keeps H constant. Duane et al. (1987) suggested to use the Metropolis rule to correct this discrepancy. Suppose the configuration at the n-th iteration of HMC is $(\mathbf{x}_n, \mathbf{p}_n)$. The next state is obtained as follows:

1. Generate a new momentum vector $\mathbf{p}$ from the Gaussian distribution $\pi(\mathbf{p}) \propto \exp\{-K(\mathbf{p})\}$;
2. Run the leapfrog algorithm (or any time-reversible and volume-preserving algorithm) for L steps to reach a new configuration in the phase space, $(\mathbf{x}', \mathbf{p}')$;
3. Let $(\mathbf{x}_{n+1}, \mathbf{p}_{n+1}) = (\mathbf{x}', -\mathbf{p}')$ with probability
$$\min\big[1, \exp\{-H(\mathbf{x}', -\mathbf{p}') + H(\mathbf{x}_n, \mathbf{p})\}\big] \tag{6}$$
and let $(\mathbf{x}_{n+1}, \mathbf{p}_{n+1}) = (\mathbf{x}_n, \mathbf{p})$ with the remaining probability.

The success of the method stems from the fact that the exploration of the phase space is driven by basic physics laws. See Neal (1996) for a detailed review of HMC and its application to neural network training.

3. IMPROVEMENTS ON HMC

3.1. *Multiple-point Method*

The basic HMC considers only the ending state of an L-step leapfrog trajectory as a candidate configuration. This makes the acceptance probability very low when the step-size δ is large. Neal (1994) presented a window method to increase the acceptance rate by considering windows of states at both ends of a trajectory. We propose a multi-point method which approaches the problem from a different angle.

Suppose at iteration n the configuration is $\varphi^{(0)} = (\mathbf{x}_n, \mathbf{p}_n)$ in the phase space. Starting from state $\varphi^{(0)}$, we run L leapfrog steps to obtain $\varphi^{(0)}, \varphi^{(1)}, \cdots, \varphi^{(L)}$. For a fixed W (between 1 and $L+1$), a candidate state within the window $(\varphi^{(L-W+1)}, \cdots, \varphi^{(L)})$ is chosen according to the probability distribution

$$P(\varphi^{(k)}) = \frac{w_k \exp(-H(\varphi^{(k)}))}{\sum_{k'=L-W+1}^{L} w_{k'} \exp(-H(\varphi^{(k')}))}, \tag{7}$$

where w_k is a weighting factor given in advance by the user. Reasonable choices of w_k include $\sqrt{k}$ and $\log k$, both giving higher weights to states closer to the end of a trajectory. Suppose the chosen state is $\varphi^{(L-K)}$. We then run K backward leapfrog steps from the current state $\varphi^{(0)}$, producing states $\varphi^{(-K)}, \cdots, \varphi^{(-1)}$. Using the "generalized" Metropolis-Hastings rule by Liu *et al.* (2000), we accept $\varphi^{(L-K)}$ with probability

$$p = \min\left\{1, \frac{\sum_{k=L-W+1}^{L} w_k \exp\{-H(\varphi^{(k)})\}}{\sum_{k'=L-W+1}^{L} w_{k'} \exp\{-H(\varphi^{(L-K-k')})\}}\right\} \tag{8}$$

and accept $\varphi^{(0)}$ with probability $1-p$. The multi-point method is valid in that the above dynamical transitions satisfy the detailed balance condition. A graphical illustration of the method is given in Figure 2.

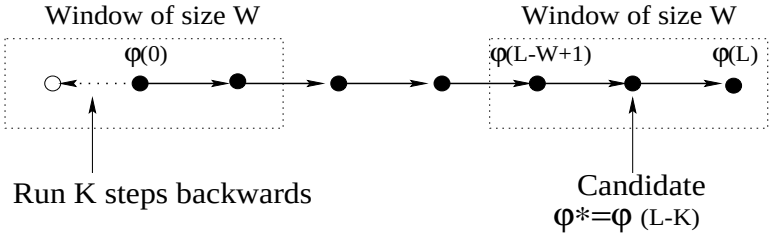

Figure 2. *A graphical view of the multiple-point HMC method.*

3.2. *Parallel Tempering*

We consider the indirect observation model mentioned in Section 1. When the control parameter σ is sufficiently small, the Markov chain might oscillate within a local region. To address this problem, we incorporate parallel tempering (Geyer 1991) in HMC. The basic idea of parallel tempering is to allow the system to "exchange" configurations corresponding to differently "tempered" distributions, enabling the sampler to explore the phase space in a more flexible way.

In the indirect observation model, we run M HMC chains in parallel with $\sigma_1 > \sigma_2 > \cdots > \sigma_M$. After a fixed number ($k_0$, for instance) of HMC transitions within each chain, we choose two chains, i and j (corresponding to σ_i and σ_j), say, at random. Suppose φ_i and φ_j are the current states of these two HMC chains, respectively. We exchange them with probability

$$\min\left\{1, \frac{\exp\{-H(\varphi_i;\sigma_j^2) - H(\varphi_j;\sigma_i^2)\}}{\exp\{-H(\varphi_i;\sigma_i^2) - H(\varphi_j;\sigma_j^2)\}}\right\}. \tag{9}$$

It is easy to show that the joint distribution $\pi_{\sigma_1}(\varphi_1) \times \cdots \times \pi_{\sigma_M}(\varphi_M)$ is invariant under this exchange operation. Thus, at the end of the parallel tempering simulation, we obtain M estimates (may be posterior means or modes), $\hat{\theta}(\sigma_1), \hat{\theta}(\sigma_2), \cdots, \hat{\theta}(\sigma_M)$. A quadratic function

$$\hat{\theta}(\sigma) = \beta_0 + \beta_1\sigma + \beta_2\sigma^2$$

can then be fitted and the parameter θ is estimated by the estimated intercept $\hat{\beta}_0$.

4. NUMERICAL EXAMPLES

4.1 *Uncoupled Oscillators*

Consider a system of N uncoupled oscillators (Neal 1994) with the potential energy function

$$U(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{N}\omega_i^2 x_i^2. \tag{10}$$

Four HMC methods, i.e. the basic HMC, the window HMC of Neal (1996), the unweighted, and the weighted multi-point methods, were applied to a 1600-dimensional system. The step-sizes were sampled uniformly from interval $(0, 2c/\omega_{\max})$. We compared the *integrated autocorrelation time (IAT)*, defined as the sum of all autocorrelations, and the CPU time (in seconds) per effective sample (where the effective sample size (ESS) is defined as total sample size/IAT).

Table 1. *Comparison of four methods: (A) basic HMC, (B) window HMC, (C) unweighted multi-point, and (D) weighted multi-point.*

c=1	A	B	C	D	c=2	A	B	C	D
IAT	7.35	8.32	5.04	1.11		11.19	7.31	7.61	3.25
$\frac{\text{CPU}}{\text{ESS}}$	0.82	1.11	0.71	0.15		1.30	0.99	1.11	0.44

From the two realizations of the algorithms reported in Table 1, one can clearly see that the two multi-point HMC methods are superior to the other two HMC methods.

4.2. *Competing Risk Model*

Suppose $X = (X_1, X_2)^T$ follows a bivariate Gaussian distribution with unknown mean $\mu = (\mu_1, \mu_2)^T$ and unknown covariance matrix Σ. We observe $Y = \max(X_1, X_2)$. To draw inference on μ, we introduce an artificial Gaussian noise with mean 0 and variance σ^2 into the model:

$$Y = \max(X_1, X_2) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2). \tag{11}$$

We then sample (X, μ, Σ) from their joint pseudo-posterior distribution by using HMC.

For illustration, we simulated 100 independent observations from $N(\mu, \Sigma)$ with

$$\mu = \begin{pmatrix} 9 \\ 12 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}.$$

The prior distribution for μ was chosen as $N(\mu, 10I)$ and that for Σ, a Wishart distribution. Our σ^2 takes four values: $1, .5, .2$, and $.1$, respectively. We constructed an independent HMC chain for each σ^2. The parameters of HMC were tuned according to the specific noise level. A large step-size is always preferred unless it makes the acceptance rate too low. A rule of thumb is to maintain an acceptance rate of $\sim$70%. The number of leapfrog steps in each dynamic transition is usually chosen to be reasonably large so that the trajectory is long enough to avoid a random walk; on the other hand, an excessively large number of leapfrog steps might be wasteful and also requires more evaluations of the derivatives of the Hamiltonian.

Table 2 displays the numbers of leapfrog steps L and the corresponding step-sizes δ for different σ^2. As can be seen from the table, a larger step-size is often followed by a larger σ^2. We also observed that for small σ^2, the autocorrelations were very high even when a relatively large number of leapfrog transitions were carried out. This slow-mixing problem can be alleviated by parallel tempering.

Table 2. *Tuning parameters for HMC in 4.2.*

σ^2	1	.5	.2	.1
L	40	60	70	80
δ	.14	.12	.10	.08

Figure 3 plots the posterior density estimates and autocorrelations for μ_1 and μ_2, respectively. It can be seen from the figure that the posterior distribution for μ_1 has a high mode near 12 and a low and flat mode near 8. This is in fact due to the nature of the problem: with the information at hand one cannot obtain a consistent estimator of μ_1 even with infinite number of observations. The middle two plots of Figure 3 show that the autocorrelations were still rather high even with the aid of parallel tempering. These autocorrelations can be further reduced by using the multi-point method, as shown by the bottom two plots of Figure 3.

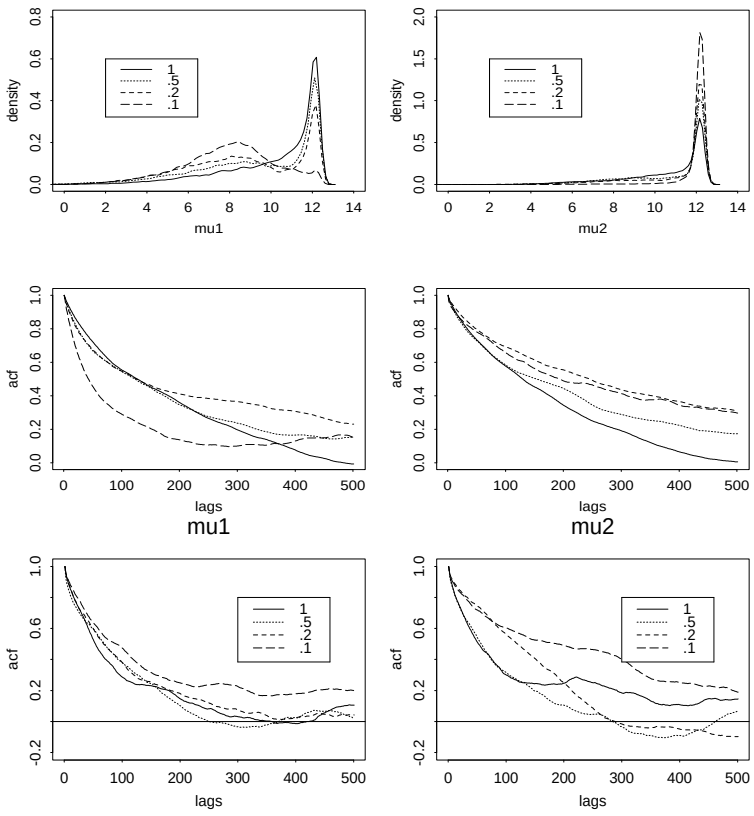

Figure 3. *Top plots from left to right: the posterior density estimates for μ_1 and μ_2; middle plots: their respective autocorrelations using HMC; bottom plots: their respective autocorrelations using the multiple-point method.*

4.3. *Stochastic Volatility Model*

The stochastic volatility model is a nonlinear state-space model and can be considered as a generalization of the celebrated Black-Scholes formula (Hull and White 1987). A simple stochastic volatility model has the form:

$$y_t = \epsilon_t \beta \exp(x_t/2), \qquad x_{t+1} = \phi x_t + \eta_t, \;\; t = 1, \cdots, T \tag{12}$$

where $\epsilon_t \sim N(0,1)$ and $\eta_t \sim N(0, \sigma^2)$. One can see that $\log\{\text{var}(y_t)\}$ follows an AR(1) process. Due to its nonlinear nature, the usual Gibbs sampler converges extremely slowly. Shephard and Pitt (1997) provided an improved MCMC algorithm which employs a "grouping" technique based on a Gaussian approximation to the log-likelihood.

We now report some promising results by using HMC to impute the state variables $x_t, t = 1, \cdots, T$. Our dataset consists of daily exchange rates of Pound/Dollar from

10/1/1981 to 6/28/1985 (i.e. 946 observations). Let r_t denote the daily exchange rate and let $dr_t = \log r_{t+1} - \log r_t$. Define

$$y_t = 100 \cdot (dr_t - \sum dr_t/T) \tag{13}$$

for $t = 1, \cdots, T$. We employed the following strategy in our implementation:

1. Given the states, sample β, σ and ϕ from their conditional distributions.

2. Given β, σ and ϕ, impute the states $x_t, t = 1, \cdots, T$ by HMC.

Table 3 summarizes the Bayesian estimates of β, σ and ϕ obtained from $10,000$ iterations in the equilibrium stage. The posterior density estimate and the autocorrelations for ϕ, which measures the persistence of volatility over periods, are displayed in Figure 4. These results indicate that the efficiency of HMC is comparable to that of the multiple-move simulation in Shephard and Pitt (1997). Since the HMC algorithm is applicable to all the systems where the derivatives of the log-likelihood functions are available, it should be useful for the Bayesian analysis of many other nonlinear and non-Gaussian state-space models.

Table 3. *Bayesian estimates of β, σ and ϕ in the stochastic volatility model.*

Parameter	Mean	Std Err	Covariance		
β	.6647	.1237	1.5306e-02	-5.3229e-04	2.9903e-04
σ	.1428	.0262	-5.3229e-04	6.8651e-04	-1.5714e-04
ϕ	.9815	.0092	2.9903e-04	-1.5714e-04	8.4321e-05

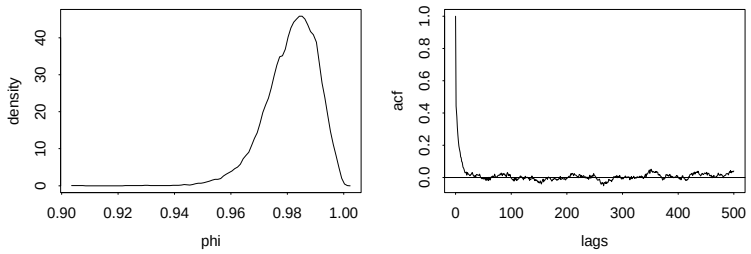

Figure 4. *The posterior density estimate and the autocorrelations for ϕ.*

4.4. *Nonlinear Random-Effects Model*

Consider the following model

$$y_{ij} = \log\left(\frac{200 k_{ai} k_{ei}}{Cl_i(k_{ai} - k_{ei})}(e^{-k_{ei}t_{ij}} - e^{-k_{ai}t_{ij}})\right) + \epsilon_{ij} \tag{14}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$. Here y_{ij} stands for the j-th observation of subject i. Let

$$\theta_i = \begin{pmatrix} \log(Cl_i) \\ \log(k_{ai}) \\ \log(k_{ei}) \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + e_i$$

and $e_i \sim N(0, \Sigma)$ where $\beta = (-3, .2, -2)^T, \sigma^2 = .01$, Σ is a diagonal matrix with diagonal elements $.04, .04$ and $.01$. A data set of 50 subjects is simulated from the model with $t_{ij} \sim \text{Unif}(0, 12)$. Only one observation ($j = 1$) is collected for each subject. Assume σ^2 is known. We wish to estimate the mean β and the covariance matrix Σ.

Shih (1999) applied the rejection Gibbs, the independent Metropolis-Hastings and the random-walk Metropolis algorithms on this model. For comparison, we used the same settings as those in Shih (1999). As with Section 4.3, we iterate the following two steps: (a) draw β and Σ from their posterior distributions conditional on the state variables $\theta_i, i = 1, \cdots, 50$; and (b) draw the state variables θ_i by HMC conditional on β and Σ. Table 4 gives the IAT, ESS, and the CPU time (in seconds) per effective sample for the HMC method. The CPU time per effective sample for the rejection Gibbs, whose performance was the best among the three MCMC methods in Shih (1999), is also included in Table 4 for comparison.

Table 4. *Simulation results for the random-effects model.*

Parameter	IAT (HMC)	ESS (HMC)	CPU time / ESS (HMC)	CPU time / ESS (rejection Gibbs)
β_1	20	2150	.12	.14
β_2	32	1344	.20	.35
β_3	31	1387	.19	.24

5. DISCUSSION

This paper presents some experimental results for using HMC in Bayesian computation and two methods for improving the performance of a standard HMC. Our experimentation with different HMC methods for a system of uncoupled oscillators shows that the multi-point method can significantly improve the efficiency of a standard HMC. The numerical analyses of the indirect observation model, the nonlinear state-space model, and the non-linear random-effects model demonstrate that HMC can be more efficient than the standard MCMC methods in these very nonlinear situations.

Although HMC has been found useful for Bayesian computations, many important issues remain open. For example, how to choose tuning parameters in HMC, e.g., the step-size and the number of the leapfrog iterations, is still a difficult problem. A rule of thumb is to maintain an acceptance rate of $\sim$70%. But there seems to be no clear theoretical basis for this rule. From our numerical studies, we also find that the efficiency of HMC can often be improved significantly by adjusting the fictitious mass variable m_i for $i = 1, \ldots, d$. This is equivalent to adopting different leapfrog step-sizes along different directions. It is intuitive to choose the m_i inversely proportional to the marginal standard deviation of π along that direction. But this may not be desirable when a strong correlation between two components of $\mathbf{x}$ is present. The multi-point method requires even more tuning: the window size W and the weighting factor w_k can both be adjusted freely. How to tune these new parameters to result in an efficient multi-point HMC warrants further investigation.

REFERENCES

Duane, S., Kennedy, A.D., Pendleton, B., and Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195**, 216–222.

Geyer, C. (1991). Markov Chain Monte Carlo Maximum Likelihood. *In* E. Keramigas (ed.), *Computing Science and Statistics: the 23rd symposium on the interface*, 156–163.

Hukushima, K. and Nemoto, K. (1996). Exchange Monte Carlo Method and Application to Spin Glass Simulation. *J. Phys. Soc. Jpn.* **65**, 1604–1608.

Hull, J. and White, A. (1987). The Pricing of Options on Assets with Stochastic Volatility. *J. Finance* **42**, 281–300.

Liu, J. S., Liang, F., and Wong, W. H. (2000). The Multiple-Try Method and Local Optimization in Metropolis Sampling. *J. Amer. Statist. Assoc.* **95**, 121–134.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21**, 1087–1091.

Neal, R. (1994). An Improved Acceptance Procedure for the Hybrid Monte Carlo Algorithm. *J. Comput. Phys.* **111**, 194–203.

Neal, R. (1996). *Bayesian Learning for Neural Networks*. Berlin: Springer.

Shephard, N. and Pitt, M. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**, 653–667.

Shih, M. (1999). *Estimation in Nonlinear Mixed Effects Models: Parametric and Nonparametric Approaches*. Ph.D. Thesis, Stanford University.