Resampling from the past to improve on MCMC algorithms

Yves F. Atchadé[*]

(Dec. 2007, First version May 2006)

Abstract: We explore two strategies that resample from previously sampled observations in a Markov Chain Monte Carlo algorithm. In one strategy the MCMC sampler reuses its own past. We show that in general this strategy generates a sampler with slower mixing. We propose another strategy based on multiple chains where some of the chains reuse past samples generated by other chains. This latter algorithm is related to the *Equi-Energy sampler* of [11]. We show by examples that this strategy yields a viable Monte Carlo methods with mixing properties similar to those of the *Equi-Energy sampler*.

AMS 2000 subject classifications: Primary 60C05, 60J27, 60J35, 65C40.

Keywords and phrases: Monte Carlo methods, Adaptive MCMC, Importance resampling, Stochastic volatility models.

1. Introduction

Markov Chain Monte Carlo (MCMC) methods have become the standard computational tool for bayesian inference. But the great flexibility of the method comes with a price. Namely, it is very difficult to determine whether a given MCMC sampler can mix or has mixed in a given computing time. Given this limitation, there is a lot of interest in developing new algorithms with improved mixing and convergence properties.

In this paper, we explore the idea of reusing previously sampled observations in a MCMC sampler. We investigate two strategies. In the first, rather naive approach, we update the chain by resampling from its sample path. More precisely, suppose that at time n, we want to sample X_n. Instead of sampling X_n from $P(X_{n-1}, \cdot)$ for some transition kernel P, we obtain X_n by resampling independently from $\{X_B, \ldots, X_{n-1}\}$, where $B \geq 0$ is some burn-in period. This *resampling from*

[*]Department of Statistics, University of Michigan, email: yvesa@umich.edu

the past step is then repeated during the simulation at some predetermined times $\tau_1 < \tau_2 < \ldots$. Heuristically, the idea is to look at $\{X_B, \ldots, X_{n-1}\}$ as an empirical measure approximation of π. Therefore resampling from the past allows the sampler to move more easily and according to a distribution that is close to π. But interestingly, it turns out that this strategy results in an algorithm with slower convergence rate. We show this with an example and also by a rigorous convergence analysis. This analysis reinforces the general wisdom that relying too much on the past in an adaptive Monte Carlo simulation is a bad idea.

Our second strategy uses multiple chains with different target distributions. In this approach, we run $K+1$ parallel chains where the l-th chain $\{X_n^{(l)}\}$ has target distribution $\pi^{(l)}$. During the simulation, the l-th chain is allowed to use the sample path of the $(l-1)$-th chain to built its kernel. The ideal situation for this sampler is when the target distribution $\pi^{(l-1)}$ of the $(l-1)$-th chain is close to $\pi^{(l)}$ but mixes more rapidly. In order to borrow samples from the $(l-1)$-th chain to the l-th chain, we rely on importance resampling. We call the algorithm, *importance-resampling MCMC* (IR-MCMC). This idea of resampling from an auxiliary process is not new and is the idea behind the *equi-energy sampler* recently proposed by [11]. IR-MCMC replies on the importance function $\pi^{(l)}/\pi^{(l-1)}$ to move samples from the $(l-1)$-th chain to the l-th chain; while the equi-energy sampler relies on *equi-energy rings*. We compare the two approach in a simulation study. We find that with good equi-energy rings, the equi-energy sampler in general performs better than IR-MCMC. But in practice, IR-MCMC is more easy to use since the construction of the equi-energy rings can be time-consuming. The idea of using empirical measure to build transition kernels in Monte Carlo similations has also been explored by [6] although in a much less efficient framework. On the theoretical side, we refer the reader to [4] where it is shown that under appropriate conditions IR-MCMC is ergodic and satisfies a strong law of large number and a central limit theorem. Similar results are also available on the equi-energy sampler ([2, 4]).

The paper has two main Sections. In Section 2, we present the idea of resampling from previously sampled observations in a MCMC algorithm and investigate its theoretical properties. The Importance-Resampling MCMC algorithm is discussed 3. Proofs are postponed to Section 5.

2. Resampling from the past

2.1. The algorithm

Let $\{X_n\}$ be a Markov chain with state space $(\mathcal{X}, \mathcal{B})$, transition kernel P, invariant distribution π and initial distribution μ. Throughout the paper, all random objects are defined on a fixed probability triplet $(\Omega, \mathcal{F}, \Pr)$ and we write $\mathbb{E}$ for the expectation with respect to Pr. If $\{X_n\}$ is ergodic, the distribution of X_n converges to π as $n \to \infty$. In which case, we can estimate integrals of the form $\pi(f) = \int f(x)\pi(dx)$ by the corresponding empirical average

$$\pi_n(f) = \frac{1}{n}\sum_{k=0}^{n-1} f(X_k).$$

It is well known that if the autocorrelation function of $\{f(X_n)\}$ decays fast enough then

$$n\mathbb{E}\left[(\pi_n(f) - \pi(f))^2\right] = \sigma^2_{\pi,P}(f) + O\left(\frac{1}{n}\right), \tag{1}$$

with $\sigma^2_{\pi,P}(f)$ given by:

$$\sigma^2_{\pi,P}(f) = \pi(\bar{f}^2) + 2\sum_{i=1}^{\infty} \pi\left[\bar{f}P^i\bar{f}\right],$$

where $\bar{f} = f - \pi(f)$, $P^i f(x) = \int P^i(x, dy)f(y)$ and $P^i(x, A) = \int P^{i-1}(x, dy)P(y, A)$ is the i-th power of P.

We see from (1) that the precision of $\pi_n(f)$ in approximating $\pi(f)$ is roughly $\sigma_{\pi,P}(f)/\sqrt{n}$. Intuitively, it seems that for a Markov chain in stationarity, we can reduce its autocorrelation by resampling from its sample path. This suggestions the following algorithm.

Let T be a transition kernel with invariant distribution π. For example $T = I$ the identity kernel or $T = P$. Suppose that after a burn-in period B, we have $X_B \approx \pi$. Let $B = \tau_0 < \tau_1 < \cdots$ be a deterministic sequence of resampling times, $\tau_i \in \mathbb{Z}^+ = \{0, 1, \ldots\}$. At any time $n > B$ and given $\{X_B, X_{B+1}, \ldots, X_{n-1}\}$, we do the following to generate X_n. If n is a resampling time, that is if $n \in \{\tau_1, \ldots\}$, we generate Y by resampling from $\{X_B, X_{B+1}, \ldots, X_{n-1}\}$ and then generate X_n from $T(Y, \cdot)$. If n is not a resampling time, then we generate X_n by sampling from $P(X_{n-1}, \cdot)$. In the sequel, and for simplicity, we will assume that the inter-resampling time is constant equal to m: $\tau_k = \tau_{k-1} + m$, $k \geq 1$.

Algorithm 2.1. *Resampling from the past*

Assume B and $B < \tau_1 < \cdots$ be given.

At some time $n > B$, given $\{X_B, X_{B+1}, \ldots, X_{n-1}\}$:

(i) *If $n \in \{\tau_1, \ldots\}$ generate Y by resampling from $\{X_B, X_{B+1}, \ldots, X_{n-1}\}$ and generate $X_n \sim T(Y, \cdot)$.*

(ii) *If $n \notin \{\tau_1, \ldots\}$, then we generate X_n by sampling from $P(X_{n-1}, \cdot)$.*

Suppose that $\mu = \pi$, that is $X_0 \sim \pi$ and define the empirical measure

$$\pi_n(A) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_A(X_k).$$

For n large enough, we expect π_n to be close to π. Therefore, each time we resampling from the past, the conditional distribution of X_n given $(X_0, \ldots, X_{n-1})$ is $\pi_n T \approx \pi$. In the general case where $X_0 \sim \mu$, the same argument carries through after a burn-in period B such that $X_B \approx \pi$. So the heuristic seems to suggest that this algorithm will improve on mixing. Interestingly, this turns out not to be the case.

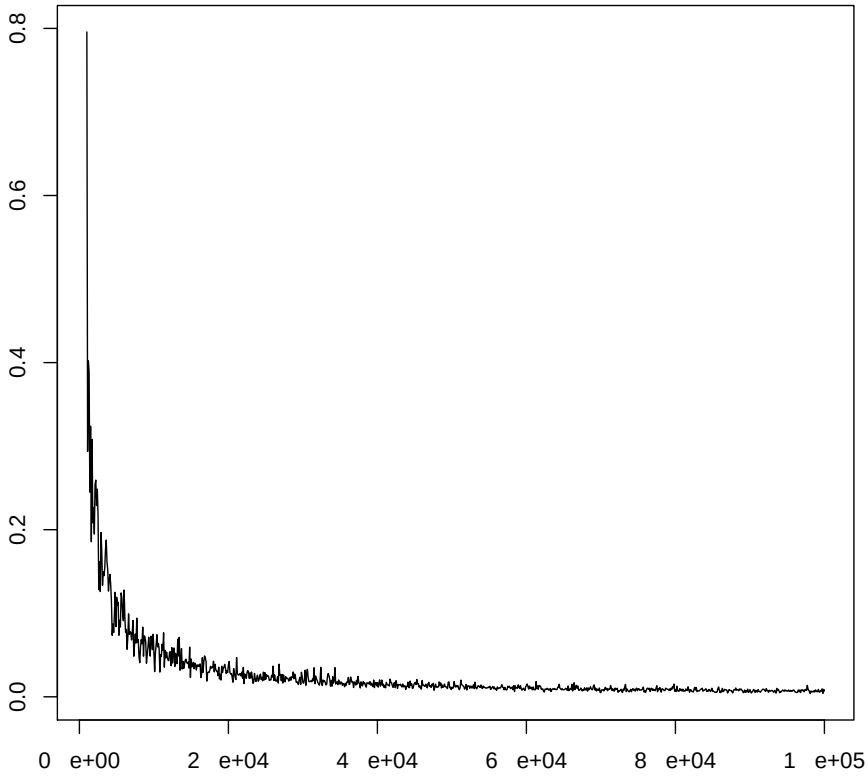
2.2. A toy example

Let $\mathcal{N}(x; \mu, \sigma^2)$ denotes the density of the normal distribution $N(\mu, \sigma^2)$ with mean μ and variance σ^2. Consider the Random Walk Metropolis (RWM) algorithm with target density $\pi(x) = \mathcal{N}(x; 0, 1)$ and proposal density $q(x, y) = \mathcal{N}(y - x; 0, \sigma^2)$, $\sigma = 0.1$. We compare the plain RWM with a RWM with resampling from the past (RFP). We set $T = P$ and let the initial distribution be π ($X_0 \sim \pi$). The question is: does RFP improves on the Monte Carlo precision.

We compare the RFP sampler with the plain Metropolis sampler on how well they estimate $\mu = \int_{-\infty}^{+\infty} x\pi(x) = 0$. The comparison is based on $\sigma^2(N)/\sigma_m^2(N)$ where $\sigma^2(N)$ (resp. $\sigma_m^2(N)$) is $\mathbb{E}\left[\left(\sum_{k=1}^{N} X_k/N\right)^2\right]$, $X_0 \sim \pi$, when $\{X_n\}$ is the RWM (resp. RFP with an inter-resampling time of m). For N fixed, we estimate these quantities from 100 replications of each sampler.

In Figure 1, we plot $\sigma^2(N)/\sigma_m^2(N)$ as a function of N for $m = 10$. We obtain the same pattern for different values of m. We see clearly, that the plain RWM performs better than RFP. Actually, the performance of RFP (compared to the plain RWM) worsens as $N \to \infty$. The curve

of $\sigma^2(N)/\sigma_m^2(N)$ as function of N clearly suggests that $\sigma_m^2(N) = O\left(N^{-\alpha}\right)$ for some $\alpha < 1$. This would mean that the rate of convergence to zero of Monte Carlo errors in RFP is actually slower than the classical rate of $1/\sqrt{N}$. We derive below some theoretical results that are consistent with this example.



Graph 1: $\sigma^2(N)/\sigma_m^2(N)$ as a function of N for $m = 10$.

2.3. Theoretical discussion

Resampling from the past should not disturb the target distribution of the algorithm. But contrary to the heuristic discussion that led to the algorithm, resampling from the past essentially slows down the simulation. Clearly, $\{X_n\}$ is no longer Markov and the conditional distribution of

X_n depends on the whole path $(X_0, \ldots, X_{n-1})$. For $f : \mathcal{X} \to \mathbb{R}$ and $V : \mathcal{X} \to [1, \infty)$, define $|f|_V := \sup_{x \in \mathcal{X}} |f(x)| / V(x)$. We work with the Banach space of V-bounded measurable functions $L_V^\infty := \{ \text{ meas. } f : (\mathcal{X}, \mathcal{B}) \to \mathbb{R} \text{ s.t. } |f|_V < \infty \}$. For mathematical simplicity, we assume that P is V-uniformly ergodic ([14]).

Assumption (A): *We assume that there exists* $V : (\mathcal{X}, \mathcal{B}) \to [1, \infty)$ *such that*

$$\sup_n \mathbb{E}\left[V(X_n)\right] < \infty, \tag{2}$$

and for any $\alpha \in (0, 1]$, *there exists* $C_\alpha < \infty$, $r_\alpha \in (0, 1)$ *such that:*

$$\left\| P^n - \pi \right\|_{V^\alpha} \leq C_\alpha r_\alpha^n, \tag{3}$$

where for any linear function operator Q,

$$\left\| Q \right\|_{V^\alpha} = \sup_{|f|_{V^\alpha} \leq 1} |Qf|_{V^\alpha} . \tag{4}$$

Assumption (A) can be checked using drift and minorization conditions ([14]). For $n \geq 0$, define $\mathcal{L}^{(n)}(A) = \Pr\left(X_n \in A\right)$ the distribution of X_n. Also, we will write $\bar{T} = T - \pi$. For a signed measure ν on $(\mathcal{X}, \mathcal{B})$, define $\|\nu\|_V = \sup_{|f|_V \leq 1} |\nu(f)|$.

Theorem 2.1. *1. Assume (A) and suppose that* $\left\| \bar{T} \right\|_V \leq 1$. *Define* $\rho = \left\| \bar{T} \right\|_V \sum_{k=0}^{m-1} r^k / m$, *where* $r = r_1$ *in (3) and* m *is the inter-resampling time. Then there exists* $C < \infty$ *such that*

$$\left\| \mathcal{L}^{(n)} - \pi \right\|_V \leq \frac{C}{n^{1-\rho}}, \quad n \geq 1. \tag{5}$$

2. Assume $B = 0$, $m = 1$ *and* $T = P$. *Take* f *such that* $\pi(f) = 0$ *and* $f^2 \in L_V^\infty$. *Then*

$$\mathbb{E}\left[\left(\frac{1}{n} \sum_{k=1}^{n} f(X_k) \right)^2 \right] \leq C \left| f \right|_{V^{1/2}}^2 \frac{1}{n^{\min(1, 2(1-\rho_1))}}, \tag{6}$$

for some finite constants $C \in [0, \infty)$ *and* $\rho_1 \in (0, 1)$.

Proof. See Section (5). □

Remark 2.1. We do not know whether these bounds are optimal or not. But Example 2.2 clearly suggests that the L^2 bound is tied.

3. Importance-Resampling MCMC

Let $\pi(dx) \propto h(x)\lambda(dx)$ be a probability measure of interest on a states space $(\mathcal{X}, \mathcal{B}, \lambda)$. Typically $\mathcal{X} \subseteq \mathbb{R}^d$, d large; λ is the Lebesgue measure and h is some complicated nonnegative measurable function. One of the most successful Monte Carlo strategy is the *annealing/tempering* method where the sampling algorithm moves progressively to π through a sequence of "easily sampled" distributions. This idea is behind many well known algorithms such as simulated annealing ([10]), simulated tempering ([8, 13]), parallel tempering ([7]), annealed importance sampling ([15]), equi-energy sampling ([11]). Let $\{\pi^{(l)},\ l = 0, \ldots, K\}$ be the sequence of *annealed* distributions, where $\pi^{(l)}(dx) \propto h^{(l)}(x)\lambda(dx)$ and $\pi^{(K)} = \pi$. In the spirit of the equi-energy sampler, we propose an algorithm that builds parallel chains where the entire sample path of the $(l-1)$-th chain can be used to move the l-th chain. This algorithm differs from the equi-energy sampler in that it relies on importance sampling.

3.1. The algorithm

Let $\{\pi^{(l)}\}, l = 0, \ldots, K$ be a sequence of distributions where $\pi^{(l)}(dx) \propto h^{(l)}(x)\lambda(dx)$ and $\pi^{(K)} = \pi$. To describe the IR-MCMC algorithm, we assume that a transition kernel $P^{(l)}$ is available that has invariant distribution $\pi^{(l)}$ and denote $\omega_l(x) = h^{(l)}(x)/h^{(l-1)}(x)$. There are many ways to contruct the tempered distributions. In the more traditional approach, $h^{(l)}(x) = h^{\gamma_l}(x)$ for some sequence of "inverse temperatures" $0 < \gamma_0 < \ldots < \gamma_K = 1$. In another approach, the initial distribution $\pi^{(0)} \propto h^{(0)}(x)\lambda(dx)$ is selected first, then $h^{(l)}(x) = [h^{(0)}(x)]^{1-\gamma_l}[h(x)]^{\gamma_l}$ or $h^{(l)}(x) = h^{(0)}(x)\left(h(\gamma x)/h^{(0)}(\gamma x)\right)$ for $0 = \gamma_0 < \ldots < \gamma_K = 1$. For example, in Bayesian inference, $\pi^{(0)}$ can be taken as the prior distribution.

Let $T^{(l)}$ be a transition kernel on $(\mathcal{X}, \mathcal{B})$ that also has invariant distribution $\pi^{(l)}$. The algorithm builds a parallel self-interacting chain $\{(X_n^{(0)}, \ldots, X_n^{(K)}),\ n \geq 0\}$ on $\mathcal{X}^{K+1}$ as follows. We start with some initial values $(X_0^{(0)}, \ldots, X_0^{(K)}) \in \mathcal{X}^{K+1}$. Then at time n, given the σ-algebra generated by $\{(X_k^{(0)}, \ldots, X_k^{(K)}),\ k \leq n\}$, we do the following. We move $\{X_n^{(0)}\}$ as a standard MCMC algorithm: $X_{n+1}^{(0)} \sim P^{(0)}(X_n^{(0)}, \cdot)$. For $l = 1, \ldots, K$, we generate a random variable $Y^{(l)}$ by resampling from $\{X_0^{(l-1)}, \ldots, X_n^{(l-1)}\}$ according to weights $\{\omega_0^{(l)}, \ldots, \omega_n^{(l)}\}$, where $\omega_k^{(l)} = \omega_l(X_k^{(l-1)})$. Then with probability θ_l, we sample $X_{n+1}^{(l)}$ from $P^{(l)}(X_n^{(l)}, \cdot)$ and with probability $1 - \theta_l$, we sample

$X_{n+1}^{(l)}$ from $T^{(l)}(Y^{(l)},\cdot)$.

Algorithm 3.1. *At time n, given $\{(X_k^{(0)},\ldots,X_k^{(K)}),\ k\le n\}$:*

1. *Sample $X_{n+1}^{(0)}\sim P^{(0)}(X_n^{(0)},\cdot)$.*
2. *For $l=1,\ldots,K$ and $\theta_l\in(0,1)$, Sample $X_{n+1}^{(l)}$ from $S_n^{(l)}(X_n^{(l)},\cdot)$, where*

$$S_n^{(l)}(x,A)=\theta_l P^{(l)}(x,A)+(1-\theta_l)\frac{\sum_{k=1}^n\omega_k^{(l)}T^{(l)}(X_k^{(l-1)},A)}{\sum_{k=1}^n\omega_k^{(l)}},$$

where $\omega_k^{(l)}=\omega_l(X_k^{(l-1)})$. In other words, with probability θ_l we sample $X_{n+1}^{(l)}$ from $P^{(l)}(X_n^{(l)},\cdot)$ and with probability $1-\theta_l$, we obtain $Y^{(l)}$ by resampling from $\{X_0^{(l-1)},\ldots,X_n^{(l-1)}\}$ with weights $\{\omega_k^{(l)},\ k\le n\}$ and then propose $X_{n+1}^{(l)}\sim T^{(l)}(Y^{(l)},\cdot)$.

3.2. Discussion of the method

Clearly $\{X_n^{(0)}\}$ is a Markov chain with stationary distribution $\pi^{(0)}$. Thus for n large enough, $\{X_k^{(0)},\ k\le n\}$ can be seen as a sample from $\pi^{(0)}$ and $\sum_{k=1}^n\omega_k^{(1)}T^{(1)}(X_k^{(0)},\cdot)/\sum_{k=1}^n\omega_k^{(1)}$ an empirical measure estimate of $\pi^{(1)}$. This means that for n large, $S_n^{(1)}(x,A)=\theta_1P^{(1)}(x,A)+(1-\theta_1)\sum_{k=1}^n\omega_k^{(1)}T^{(1)}(X_k^{(0)},A)/\sum_{k=1}^n\omega_k^{(0)}$ should behave like $S^{(1)}(x,A)=\theta_1P^{(1)}(x,A)+(1-\theta_1)\pi^{(1)}(A)$. This makes the algorithm particularly appealing: for n large, each importance resampling should give an almost perfect sampling from $\pi^{(1)}$. Now, once $X_n^{(1)}$ converges to $\pi^{(1)}$, the same heuristical argument just developed also applies and justified that $\{X_n^{(2)}\}$ should be sampling efficiently from $\pi^{(2)}$, and so on.

Typically, $P^{(l)}$ will be a Metropolis-Hastings kernel or a Gibbs kernel with invariant distribution $\pi^{(l)}$. $T^{(l)}$ can be any transition kernel (not necessarily ergodic) that is invariant with respect to $\pi^{(l)}$. The typical choice is $T^{(l)}=I$, the identity kernel. The choice of θ_l overall should depend on the quality of the interpolation. We give some guideline on how to choose θ through Example 3.3 below.

As with any *importance sampling* method, we can check the quality of the interpolation $\{\pi^{(l)}\}$ by computing for each $l=1,\ldots,K$, one plus the variance of the weight function ω_l:

$$\mathrm{eff}^{(l)}=1+\mathbb{V}\mathrm{ar}_{\pi^{(l-1)}}\left(\tilde{\omega}_l(X)\right),\tag{7}$$

where $\tilde{\omega}_l(x)=\omega_l(x)/\mathbb{E}_{\pi^{(l-1)}}(\omega_l(X))$. Large values of $\mathrm{eff}^{(l)}$ indicates a high variability in the importance weights which typically means a high discrepancy between $\pi^{(l-1)}$ and $\pi^{(l)}$. In these

cases, the resampling step will perform poorly. A consistent estimator of $\text{eff}^{(l)}$ is given by:

$$\text{eff}_n^{(l)} = \frac{n \sum_{k=1}^n \omega_l^2(X_k^{(l-1)})}{\left\{\sum_{k=1}^n \omega_l(X_k^{(l-1)})\right\}^2}. \tag{8}$$

The asymptotics of the IR-MCMC has been investigated in [4] where it is shown that a strong law of large numbers and a central limit theorem hold for a large class of functions under some verifiable conditions. We refer the reader to that work for more detail.

3.3. Illustrative Example: Sampling from a bivariate multimodal distribution

We apply IR-MCMC to a bivariate multimodal gaussian mixture example taken from [12]. The target distribution is given by:

$$\pi(x) = \frac{1}{2\pi\sigma^2} \sum_{i=1}^{20} \omega_i \exp\left\{-\frac{1}{2\sigma^2}(x-\mu_i)'(x-\mu_i)\right\}, \tag{9}$$

where $\sigma = 0.1$ and $\omega_i \equiv 0.05$. The μ_i are given by:

i	μ_{i1}	μ_{i2}	i	μ_{i1}	μ_{i2}	i	μ_{i1}	μ_{i2}	i	μ_{i1}	μ_{i2}
1	2.18	5.76	6	3.25	3.47	11	5.41	2.65	16	4.93	1.50
2	8.67	9.59	7	1.70	0.50	12	2.70	7.88	17	1.83	0.09
3	4.24	8.48	8	4.59	5.60	13	4.98	3.70	18	2.26	0.31
4	8.41	1.68	9	6.91	5.81	14	1.14	2.39	19	5.54	6.86
5	3.93	8.82	10	6.87	5.40	15	8.33	9.50	20	1.69	8.11

TABLE 1
Inefficiencies of the samplers for the Sterling dataset.

We define $\pi^{(l)} = \pi^{1/t_l}$ with $t_l \in \{50, 21.6, 13, 7.7, 4, 2.8, 1\}$. We take $T^{(l)} = I$, the identity kernel and set $\theta_l \equiv \theta$. We first investigate the choice of θ. For different values of θ, we estimate the Mean Square error of the sampler in estimating the first two moments of both components of π. The results are presented in Figure 2. We see that the precision of the algorithm improves very rapidly as θ moves away from zero. But there is almost no gain in efficiency after θ reaches 0.2. Overall, we recommend setting θ to a value between 0.1 and 0.4.

Setting $\theta = 0.33$, we compare IR-MCMC with the Equi-Energy sampler, again on how well both algorithms estimate the first two moments of both components of π. For the EE sampler, we use the parametrization given in [11] where the same example has also been used except the sequence of temperature which the same as above. Overall, the EE sampler has a better mixing and yields

smaller asymptotic variances. This is not surprising given the equi-energy moves. But IR-MCMC has the advantage of being simpler to implement. In practice setting up the equi-energy rings of the EE sampler can be tedious.

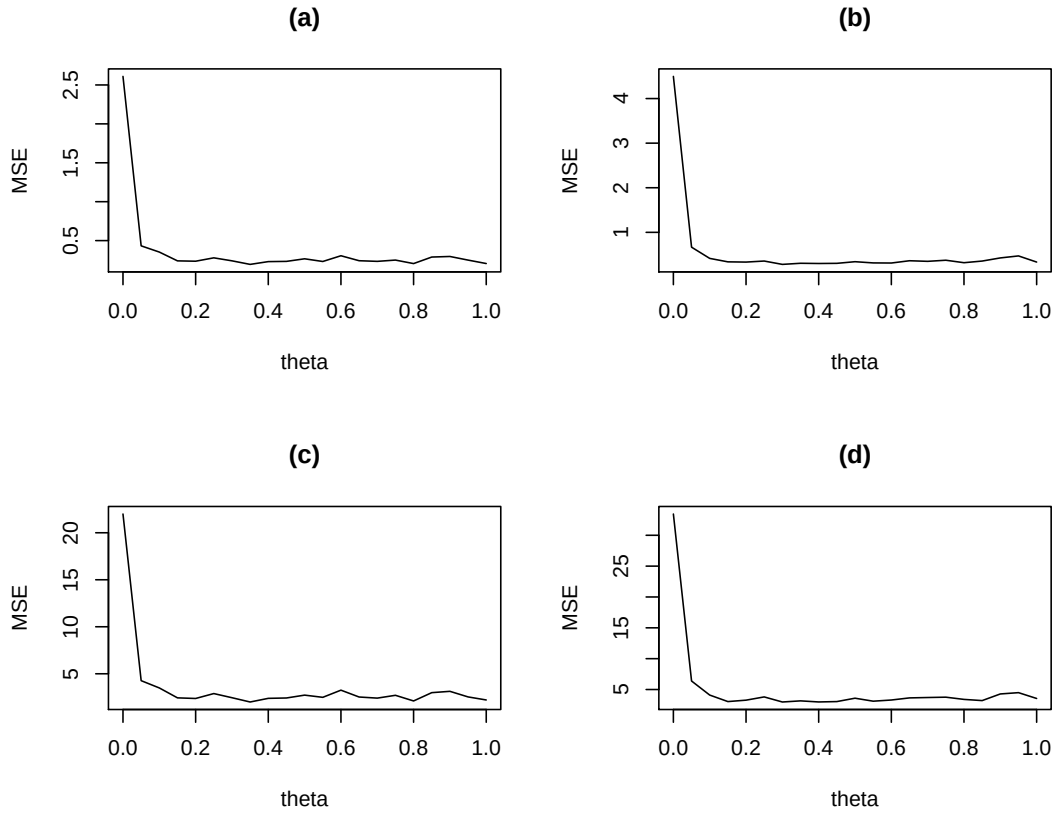


Figure 1: MSE of the Monte Carlo estimators for different value of θ. Based on 30 replications of the sampler each of length $N = 10,000$. (a): $\mathbb{E}(X_1)$; (b) $\mathbb{E}(X_1^2)$; (c) $\mathbb{E}(X_1^2)$; (d): $\mathbb{E}(X_2^2)$.

	$\mathbb{E}(X_1)$	$\mathbb{E}(X_2)$	$\mathbb{E}(X_1^2)$	$\mathbb{E}(X_2^2)$
$100 \times (MSE_{IR}/MSE_{EE} - 1)$	8.24	19.39	3.01	19.65

TABLE 2

Improvement of EE over IR-MCMC (in percentage). Based on 30 replications of $50,000$ *iterations of each sampler.*

4. Conclusion

Adaptive MCMC has become an important topic in the statistical Monte Carlo community. In its classical form, the problem is formulated as that of choosing the "best" kernel to sample from π among a class of potential transition kernels $\{P_\theta,\ \theta \in \Theta\}$. ([1, 3, 5, 9]). Typically Θ is a finite dimensional space and the problem is easily translated into a stochastic approximation problem. The resulting adaptive MCMC samplers are then designed so as to solve this stochastic approximation problem while also sampling from π. But yet, the more general question of adaptive MCMC is how to reuse efficiently samples from a Monte Carlo sampler in order to improve on its performances. The approach taken in this work can be seen as building nonparametric approximations of the target distribution. The lessons learned is that there are some good ways and some bad ways of doing this. More specifically, we show that by relying too much on previously sampled points, the adaptation can worsen the performance of the initial sampler.

We also propose a new algorithm (IR-MCMC) based on multiple chains where the approximation of the target disitrbution come from an auxiliary chain. The approach is similar to the Equi-Energy sampler. We observe that IR-MCMC performs slightly less well than the Equi-Energy but has the advantage of being simpler to implement. We believe that the multiple chain framework is a promising approach to adaptive Monte Carlo. But as currently designed, the Equi-Energy sampler and IR-MCMC hold an intrinsic limitation. At each time n and conditional on the past, the transition kernel $S_n^{(l)}$ do not maintain detailed balance with respect to π. This actually introduces an additional bias (converging to zero but at the same rate as the Monte Carlo error) in the sampler that eventually degrades its overall performance. We analysis this phenomenon in much greater detail in [4]. We expect that the next generation of samplers in this class will fix that problem.

5. Proof of Theorem 2.1

Let $\mathcal{F}_n$ be the σ-algebra generated by $(X_0, \dots, X_n)$. Let $\tau_0 = B$ and $B < \tau_1 < \tau_2 \dots$ the resampling times. We recall that $\tau_k = \tau_0 + km$. For $n > B$, define $N(n) := \max\{k \geq 0 :\ \tau_k \leq n\}$ the number of resampling times before n. Let $f \in L_V^\infty$ be given. Without any loss of generality, we assume

that $\pi(f) = 0$. Define the transition kernel

$$M = \frac{1}{m} \sum_{k=0}^{m-1} P^k T.$$

By conditioning on $\mathcal{F}_{\tau_{k-1}}$, it can be seen that for $k \geq 1$:

$$\mathbb{E}\left[f(X_{\tau_k})\right] = \mathbb{E}\left[Q_{k-1} f(X_{\tau_{k-1}})\right], \tag{10}$$

where

$$Q_k = \begin{cases} M & \text{if } k = 0 \\ \left(1 - \frac{1}{k}\right) I + \frac{1}{k} M & \text{if } k > 0. \end{cases} \tag{11}$$

A successive application of (10) yields:

$$\mathbb{E}\left[f(X_n)\right] = \mathbb{E}\left[\prod_{k=0}^{N(n)-1} Q_k P^{n-\tau_{N(n)}} f(X_B)\right], \tag{12}$$

with the convention that $\prod_{k=l}^{m} Q_k = I$ if $l > m$.

The transition kernels Q_k will play a role in the analysis to follow. Clearly, Q_k has invariant distribution π. The next lemma says that $Q_k \cdots Q_n$ converges to π as $n \to \infty$.

Lemma 5.1. *Assume (A). Take $\alpha \in (0, 1]$. For all $1 \leq k \leq n < \infty$*

$$\left\| Q_k \cdots Q_n - \pi \right\|_{V^\alpha} \leq C_\alpha \left(\frac{k}{n}\right)^{1-\rho_\alpha}, \tag{13}$$

for some finite constants $C_\alpha \in [0, \infty)$ and $\rho_\alpha \in (0, 1)$. Actually, $\rho_\alpha \leq \left\|\bar{T}\right\|_{V^\alpha} \frac{1}{m} \sum_{k=0}^{m-1} r_\alpha^k$, where r_α is as in (3).

Proof. Call ρ_α the rate of convergence to π of M^n in $L^\infty_{V^\alpha}$ as $n \to \infty$. Recall that $M = \frac{1}{m} \sum_{k=0}^{m-1} P^k T$. From (A), we have $\rho_\alpha \leq \left\|\bar{T}\right\|_{V^\alpha} \frac{1}{m} \sum_{k=0}^{m-1} r_\alpha^k < 1$. Let $c_0 \ldots, c_{n-k+1}$ be the constants such that:

$$\prod_{i=k}^{n} \left[\left(1 - \frac{1}{i}\right) + \frac{\rho_\alpha}{i}\right] = \sum_{i=0}^{n-k+1} c_i \rho_\alpha^i. \tag{14}$$

For $f \in L^\infty_{V^\alpha}$ such that $\pi(f) = 0$, we have precisely $Q_k \cdots Q_n(f) = \sum_{i=0}^{n-k+1} c_i M^i(f)$. Thus:

$$\begin{aligned} |Q_k \cdots Q_n(f)|_{V^\alpha} &\leq C_\alpha |f|_{V^\alpha} \sum_{i=0}^{n-k+1} c_i \rho_\alpha^i \\ &= C_\alpha |f|_{V^\alpha} \prod_{i=k}^{n} \left[\left(1 - \frac{1}{i}\right) + \frac{\rho_\alpha}{i} \right] \\ &\leq C_\alpha |f|_{V^\alpha} \exp\left(-(1-\rho_\alpha) \sum_{i=k}^{n} \frac{1}{i} \right) \\ &\leq C_\alpha |f|_{V^\alpha} \left(\frac{k}{n}\right)^{1-\rho_\alpha}, \end{aligned}$$

using the fact that $\sum_{i=k}^{n} 1/i \geq \log(n/k)$ for $n \geq k > 0$. This implies (13). □

Theorem 2.1 now follows.

Proof of Theorem 2.1 1. Without any loss of generality, we assume that $n > B + m$ so that $N(n) \geq 2$. For $f \in L^\infty_V$ such that $\pi(f) = 0$:

$$\begin{aligned} \left|\mathcal{L}^{(n)}(f)\right| &= |\mathbb{E}\left[f(X_n)\right]| \\ &= \left| \mathbb{E}\left[\prod_{k=0}^{N(n)-1} Q_k P^{n-\tau_{N(n)}} f(X_B) \right] \right| \quad (\text{ by (12)}) \\ &= \left| \mathbb{E}\left[P^B \prod_{k=0}^{N(n)-1} Q_k P^{n-\tau_{N(n)}} f(X_0) \right] \right| \\ &\leq C r_1^B \left(\frac{1}{N(n)-1}\right)^{1-\rho_1} r_1^{n-\tau_{N(n)}} |f|_V \, \mathbb{E}(V(X_0)) \quad (\text{ By (A) and Lemma 5.1}). \end{aligned}$$

The theorem follows by noting that $m(N(n)+1) > n - B$. □

Proof of Theorem 2.1 2. Here we assume that $m = 1$ and $T = P$. Therefore $M = P$. Let $f \in L^\infty_V$ such that $\pi(f) = 0$. Recall the notation $\pi_n(f) = \frac{1}{n}\sum_{k=1}^{n} f(X_k)$. We have:

$$\begin{aligned} \mathbb{E}\left[\pi_{n+1}(f)|\mathcal{F}_n\right] &= \frac{n}{n+1}\pi_n(f) + \frac{1}{n+1}\mathbb{E}\left[f(X_{n+1})|\mathcal{F}_n\right] \\ &= \frac{n}{n+1}\pi_n(f) + \frac{1}{n+1}\pi_n(Pf) \\ &= \pi_n\left(Q_{n+1} f\right), \end{aligned} \tag{15}$$

where $Q_n = (1 - 1/n)I + (1/n)P$, I the identity operator. A successive application of (15) yields:

$$\mathbb{E}\left[f(X_{n+k})|\mathcal{F}_n\right] = \pi_n\left[Q_{n+1} \cdots Q_{n+k-1} P f\right], \tag{16}$$

with the convention that $\prod_{i=l}^{m} Q_i = I$ if $l > m$.

The proof is based on the following decomposition of $\pi_n(f) = \frac{1}{n}\sum_{k=1}^{n} f(X_k)$:

$$\pi_n(f) = Q_1 \cdots Q_n f(X_0) + \sum_{k=1}^{n} \frac{1}{k} r_{n,k}, \tag{17}$$

where:

$$r_{n,k} = \begin{cases} f(X_n) - \mathbb{E}_x\left[f(X_n)|\mathcal{F}_{n-1}\right] & \text{if } k = n \\ Q_{k+1} \cdots Q_n f(X_k) - \mathbb{E}_x\left[Q_{k+1} \cdots Q_n f(X_k)|\mathcal{F}_{k-1}\right] & \text{if } k < n \end{cases} \tag{18}$$

To see why this hold, write $\pi_n(f) = \frac{n-1}{n}\pi_{n-1}(f) + \frac{1}{n}\mathbb{E}_x\left[f(X_n)|\mathcal{F}_{n-1}\right] + \frac{1}{n}\left(f(X_n) - \mathbb{E}_x\left[f(X_n)|\mathcal{F}_{n-1}\right]\right)$. Then observe that $\frac{n-1}{n}\pi_{n-1}(f) + \frac{1}{n}\mathbb{E}_x\left[f(X_n)|\mathcal{F}_{n-1}\right]$ can also be rewritten as $\frac{1}{n-1}\sum_{k=1}^{n-1} Q_n f(X_k)$ leading to:

$$\pi_n(f) = \frac{1}{n-1}\sum_{k=1}^{n-1} Q_n f(X_k) + \frac{1}{n}\left(f(X_n) - \mathbb{E}_x\left[f(X_n)|\mathcal{F}_{n-1}\right]\right).$$

A recursive application of this identity gives the announced representation. Now, note that $\{\sum_{k=1}^{m} \frac{1}{k} r_{n,k},\ \mathcal{F}_m,\ 1 \leq m \leq n\}$ is a martingale array. Therefore:

$$\begin{aligned} \mathbb{E}\left[\left(\pi_n(f)\right)^2\right] &= \mathbb{E}\left[\left(Q_1 \cdots Q_n f(X_0)\right)^2\right] + \mathbb{E}_x\left(\sum_{k=1}^{n} \frac{1}{k} r_{n,k}\right)^2 \\ &= \mathbb{E}\left[\left(Q_1 \cdots Q_n f(X_0)\right)^2\right] + \sum_{k=1}^{n} \frac{1}{k^2}\mathbb{E}_x\left(r_{n,k}^2\right). \end{aligned}$$

By Lemma 5.1, $\mathbb{E}\left[\left|Q_1 \cdots Q_n f(X_0)\right|^2\right] \leq C\left|f\right|_{V^{1/2}}^2 \mathbb{E}\left[V(X_0)\right]/n^{2(1-\rho_{1/2})}$ and

$$\begin{aligned} \mathbb{E}\left[\left|r_{n,k}\right|^2\right] &\leq \mathbb{E}\left[\left(Q_{k+1} \cdots Q_n f(X_k)\right)^2\right] \\ &\leq C\left|f\right|_{V^{1/2}}^2 \mathbb{E}\left[V(X_n)\right](k/n)^{2(1-\rho_{1/2})}, \end{aligned}$$

for some finite constant C, where $\rho_{1/2}$ is as in Lemma 5.1. We can then conclude that:

$$\mathbb{E}\left[\left(\pi_n(f)\right)^2\right] \leq C'\left|f\right|_{V^{1/2}}^2 \mathbb{E}\left[V(X_n)\right]\left(\frac{1}{n^{2(1-\rho_{1/2})}} + \frac{1}{n}\right),$$

for some finite constant C'. Since $\sup_n \mathbb{E}\left[V(X_n)\right] < \infty$, the result follows. □

References

[1] Andrieu, C. and Atchade, Y. F. (2007). On the efficiency of adaptive mcmc algorithms. *Electronic Communications in Probability* **12** 336–349.

[2] Andrieu, C., Jasra, A., Doucet, A. and Del Moral, P. (2007). On non-linear markov chain monte carlo via self-interacting approximations. *Technical report* .

[3] Andrieu, C. and Moulines, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.* **16** 1462–1505.

[4] Atchade, Y. (2007). A cautionary tale on the efficiency of some adaptive monte carlo schemes. *Technical Report, Dept of Statistics, University of Michigan* .

[5] Atchade, Y. F. (2006). An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. *Methodol Comput Appl Probab* **8** 235–254.

[6] Chauveau, D. and Vandekerkhove, P. (2001). Improving convergence of the hastings-metropolis algorithm with an adaptive proposal. *Scandinavian Journal of Statistics* **29** 13–29.

[7] Geyer, C. J. (1991). "markov chain monte carlo maximum likelihood" in computing in science and statistics. *Proceedings of the 23rd Symposium on the Interface* 156–163.

[8] Geyer, C. J. and Thompson, E. (1995). Annealing markov chain monte carlo with applications to pedigree analysis. *Journal of the American Statistical Association* **90** 909–920.

[9] Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli* **7** 223–242.

[10] Kirkpatrick, S., Gelatt, J. C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220** 671–680.

[11] Kou, S., Zhou, Q. and Wong, W. (2006). Equi-energy sampler with applications in statistical inference and statistical mechanics. *Annals of Statistics* **34** 1581–1619.

[12] Liang, F. and Wong, W. H. (2001). Real-parameter evolutionary monte carlo with applications to bayesian mixture models. *Journal of the American Statistical Association* **96** 653–666.

[13] Marinari, E. and Parisi, G. (1992). Simulated tempering: A new monte carlo schemes. *Europhysics letters* **19** 451–458.

[14] Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Springer-Verlag London Ltd., London.

[15] Neal, R. (2001). Annealed importance sampling. *Statistics and Computing* **11**.