ESAIM: PROCEEDINGS, September 2007, Vol.19, 79-84
Christophe Andrieu & Dan Crisan, Editors
DOI: 10.1051/proc:071911

NON-LINEAR MARKOV CHAIN MONTE CARLO

CHRISTOPHE ANDRIEU[1], AJAY JASRA[2], ARNAUD DOUCET[3] AND PIERRE DEL MORAL[4]

Abstract. In this paper we introduce a class of non-linear Markov Chain Monte Carlo (MCMC) methods for simulating from a probability measure π. Non-linear Markov kernels (e.g. Del Moral (2004)) can be constructed to admit π as an invariant distribution and have typically superior mixing properties to ordinary (linear) MCMC kernels. However, such non-linear kernels often cannot be simulated exactly, so, in the spirit of particle approximations of Feynman-Kac formulae (Del Moral 2004), we construct approximations of the non-linear kernels via *Self-Interacting Markov Chains* (Del Moral & Miclo 2004) (SIMC). We present several non-linear kernels and investigate the performance of our approximations with some simulations.

1. Introduction

Let $(E, \mathcal{E})$ be a measurable space and $\mathscr{P}(E)$ be the class of probability measures on E. In this paper we are interested in the problem of simulating from a probability measure $\pi \in \mathscr{P}(E)$, which is (potentially) known up to an unknown normalizing constant, and estimating expectations of $\pi-$integrable functions $V : E \to \mathbb{R}$ via Monte Carlo estimates; that is approximating

$$\pi(V) \quad := \quad \int_E V(x)\pi(dx) \tag{1}$$

by

$$S_n^X(V) \quad = \quad \frac{1}{n+1}\sum_{i=0}^{n} V(X_i)$$

where $S_n^X(du) := \frac{1}{n+1}\sum_{i=0}^{n}\delta_{X_i}(du)$ is the empirical measure based upon random variables $\{X_k\}_{0\leq k\leq n}$ drawn from π. This problem is important in many areas of science, including statistics, engineering and physics.

[1] Department of Mathematics, University of Bristol, UK
[2] Department of Mathematics, Imperial College London, UK
[3] Departments of Statistics & Computer Science, University of British Columbia, CA
[4] Department of Mathematics, University of Nice Sophia Antipolis, FR

© EDP Sciences, SMAI 2007

Article published by EDP Sciences and available at http://www.edpsciences.org/proc or http://dx.doi.org/10.1051/proc:071911

2. Current Solutions

The most common method used, certainly within statistics, for estimating (1) are MCMC methods (e.g. Robert & Casella (2004)). This technique proceeds by simulating an ergodic Markov chain $\{X_n\}_{n\geq 0}$ of invariant distribution π, and using the estimate (2). It is well known by Monte Carlo specialists that standard MCMC algorithms, such as the Metropolis-Hastings method, often have difficulties in simulating from complicated distributions: for example densities which possess multiple modes.

As a result, there have been a large number of alternative, generic, methods proposed in the literature, a brief list includes; adaptive MCMC (Andrieu & Moulines, 2006), sequential Monte Carlo samplers (Del Moral et al. 2006) and equi-energy methods (Kou et al. 2006).

In this paper, we consider another alternative: non-linear MCMC via self-interacting approximations. We note that related self-interacting ideas have appeared, directly in Brockwell & Doucet (2006) and indirectly in Kou et al. (2006). An algorithm closely related to the work presented here is the resampling from the past algorithm of Atchadé (2006). However, the framework presented here is far more general both methodologically and theoretically. Methodologically, non-linear MCMC allow us to create a large class of new stochastic simulation algorithms. Theoretically, the proofs presented in (Atchadé, 2006) are technically correct but do not correspond to the algorithm implemented; an alternative convergence study, including a theoretical justification of the algorithms presented below, can be found in Andrieu et al. (2006).

3. Non-Linear Markov Kernels via Self Interacting Approximations

Standard MCMC algorithms rely on Markov kernels of the form $K : E \to \mathscr{P}(E)$. These Markov kernels are *linear* operators on $\mathscr{P}(E)$; that is $\mu(dy) = \int_E \xi(dx)K(x, dy)$ where $\mu, \xi \in \mathscr{P}(E)$. A *non-linear* Markov kernel $K : \mathscr{P}(E) \times E \to \mathscr{P}(E)$ is defined as a non-linear operator on the space of probability measures. Non-linear Markov kernels, K_μ, can often be constructed to exhibit superior mixing properties to ordinary MCMC versions. For example, let

$$K_\mu(x, dy) \quad = \quad (1 - \epsilon)K(x, dy) + \epsilon\Phi(\mu)(dy) \; , \tag{2}$$

where K is a Markov kernel of invariant distribution π, $\epsilon \in (0, 1)$ and $\Phi : \mathscr{P}(E) \to \mathscr{P}(E)$ is a selection/mutation operator (Del Moral 2004), with $\Phi(\mu)(dy) := \mu(gK)/\mu(g)(dy)$. The potential function g is a bounded and measurable such that $\Phi(\pi) = \pi$ (in this simple case, $g \equiv 1$ to ensure that π is an invariant distribution.). In most cases, we will be unable to simulate from K_π and instead we propose a self-interacting approximation.

A self-interacting Markov chain generates a stochastic process $\{X_n\}_{n\geq 0}$ which is allowed to interact with values realized in the past. That is, we might approximate (at time $n+1$ of the process) the selection/mutation operator by:

$$\Phi(S_n^X)(dy) \quad = \quad \frac{\sum_{i=0}^n g(X_i)K(X_i, dy)}{\sum_{i=0}^n g(X_i)} \; .$$

This process corresponds to a backward in time selection step and then a mutation step, via the kernel K. Selection allows previous values with high potential to return.

In the context of stochastic simulation, SIMCs can be thought of as storing modes and then allowing the algorithm to return to them in a relatively simple way. It is thus the attractive idea of being able to fully exploit the information provided by the previous samples which has motivated us to investigate this stochastic simulation approach.

In summary, non-linear MCMC can be characterised by the following procedure:

- Identify a non-linear kernel, K_μ, that admits π as an invariant distribution and can be expected to mix faster than an ordinary MCMC kernel e.g. (2).
- Construct a stochastic process that approximates the kernel, which can be simulated in practice.

4. Some Non-Linear MCMC Algorithms

We will study the following non-linear kernels in this paper; the motivation for such kernels is discussed in Andrieu et al. (2006).

(NL1): Self Interacting Approximation. Let K be a Markov kernel of invariant distribution π, and $\Phi : \mathscr{P}(E) \to \mathscr{P}(E)$ a selection/mutation operator. We consider a self interacting approximation of (2); in this case, at time $n+1$, the approximated kernel becomes

$$K_{S_n^x}(x_n, dx_{n+1}) := (1-\epsilon)K(x_n, dx_{n+1}) + \epsilon\Phi(S_n^x)(dx_{n+1}). \tag{3}$$

Example 1. Selection/Mutation, Identity Potential. *Let $g(x) \equiv 1$ and K be an MCMC kernel of invariant distribution π. Then we may simulate (3) to estimate (1).*

(NL2): Auxiliary Self-Interaction. We introduce the following family of kernels $\{(K \times P)_\mu, \mu \in \mathscr{P}(E)\}$ defined on the extended state-space $(F := E \times E, \mathcal{F} := \mathcal{E} \times \mathcal{E})$ which can help the exploration ability of the simulated kernel and the selection mechanism,

$$\begin{aligned}
(K \times P)_{\mu\times\delta_y}((x,y), d(x',y')) &:= (1-\epsilon)K(x, dx') \times P(y, dy') + \epsilon\Psi(\mu \times \delta_y)((x,y), d(x',y')) \,, \\
\Psi(\mu \times \delta_y)((x,y), d(x',y')) &:= \frac{(\mu \times \delta_y)\big((g \times 1)(K \times P)\big)}{(\mu \times \delta_y)(g \times 1)}(d(x',y')) \\
&= \Phi(\mu)(dx) \times P(y, dy') \,.
\end{aligned}$$

with $P : E \to \mathscr{P}(E)$ a Markov kernel g and Φ as considered above.

Example 2. Selection/Mutation with Potential. *Let P be an MCMC kernel of invariant distribution η, and assume $\pi \ll \eta$. Let $g(x) = \frac{d\pi}{d\eta}(x)$ and set K to be an MCMC kernel of invariant distribution π. If we were able to sample exactly from η then one could sample exactly from $(K \times P)_\eta$ which has invariant distribution $\pi \times \eta$. However, for efficient algorithms, this will not be the case and instead we suggest using the following approximation, here given at time $n+1$:*

$$(K \times P)_{S_n^y \times \delta_{y_n}}((x_n, y_n), d(x_{n+1}, y_{n+1})) = \big[(1-\epsilon)K(x_n, dx_{n+1}) + \epsilon\Phi(S_n^y)(dx_{n+1})\big]P(y_n, dy_{n+1})$$

that is, we are 'feeding' the chain $\{X_n\}$ the empirical measure S_n^y.

(NL3): Auxiliary Self-Interaction with Genetic Moves. For any $\mu \in \mathscr{P}(E)$ we define a non-linear Markov kernel $Q_\mu : \mathscr{P}(E) \times E \to \mathscr{P}(E)$ with potential $g : E \times E \to (0, \infty)$ $(|g|_\infty < \infty)$ as

$$Q_\mu(x, dx') := \frac{\int_{E\times E} g(x,v)\widetilde{K}((x,v), dx')\mu(dv)}{\int_E g(x,v)\mu(dv)} \,,$$

where

$$\begin{aligned}
\widetilde{K}((u,v), dx) &:= \int_{E\times E} K^S((u,v), d(u',v'))K(u', dx) \\
K^S((u,v), d(x',y')) &:= \alpha(u,v)\delta_v(dx')\delta_u(dy') + \big[1 - \alpha(u,v)\big]\delta_u(dx')\delta_v(dy') \\
\alpha(u,v) &:= 1 \wedge \frac{\pi(v)\eta(u)}{\pi(u)\eta(v)}
\end{aligned}$$

with K^S an exchange kernel. We now define the following non-linear kernel

$$\begin{aligned}(K\times P)_{\mu\times\delta_y}((x,y),d(x',y')) &= (1-\epsilon)K(x,dx')P(y,dy')+\epsilon\Psi\big(\mu\times\delta_y\big)((x,y),d(x',y'))\\ \Psi\big(\mu\times\delta_y\big)((x,y),d(x',y')) &:= \frac{\int_{E\times E} g(x,v)\widetilde{K}((x,v),dx')P(w,dy')\mu(dv)\delta_y(dw)}{\int_{E\times E} g(x,v)\mu(dv)\delta_y(dw)}\\ &= Q_\mu(x,dx')\times P(y,dy')\ .\end{aligned}$$

Example 3. Simplified Equi-Energy Sampling (Kou et al. 2006) with Identity Potential. *Let $g(x,y)\equiv 1$, $\eta\sim\pi$, K (resp. P) be an MCMC kernel of invariant distribution π (resp. η). Then we have $(\pi\times\eta)(K\times P)_{\eta\times\delta_y}=\pi\times\eta$; that is, via Fubini:*

$$\begin{aligned}\pi Q_\eta(dx') &= \int_{E\times E}\left[\int_{E\times E}\pi(dx)\eta(dy)K^S((x,y),d(u,v))\right]K(u,dx')\\ &= \pi K(dx').\end{aligned}$$

We can then simulate the self-interacting Markov chain $(K\times P)_{S_n^Y\times\delta_y}$ at time n, where S_n^Y is the empirical measure that has been built by the chain with invariant distribution η.

4.1. The Algorithms

To summarize our algorithm for (NL1) is:

0. (Initialization): Set $n=0$ and $X_0=x$, $S_0^x=\delta_x$.
1. (Iteration): Set $n=n+1$, simulate $X_n\sim K_{S_{n-1}^x}(x_{n-1},\cdot)$.
2. (Update). $S_n^x=S_{n-1}^x+\frac{1}{n+1}[\delta_{x_n}-S_{n-1}^x]$ and return to 1.

For (NL2) and (NL3), our algorithm is:

0. (Initialization): Set $n=0$ and $X_0=x$, $Y_0=y$, $S_0^y=\delta_y$.
1. (Iteration): Set $n=n+1$, simulate $Y_n\sim P(y_{n-1},\cdot)$ and $X_n\sim K_{S_{n-1}^y}(x_{n-1},\cdot)$.
2. (Update). $S_n^y=S_{n-1}^y+\frac{1}{n+1}[\delta_{y_n}-S_{n-1}^y]$ and return to 1.

5. A Simulation Example

We now present some simulations; we begin by describing the target, then the kernels, simulation parameters and then the results. A similar example can be found in Andrieu et al. (2006) and Kou et al. (2006).

5.1. Simulation Details

We consider the target probability measure:

$$\begin{aligned}\pi(dx) &= \frac{1}{Z}\exp\Big\{-h(x)\Big\}dx\\ h(x) &= -\log(f(x))\\ f(x) &= \sum_{l=1}^{20} w_l\phi_2(x;\mu_l,\Sigma_l)\end{aligned}$$

where Z is the normalizing constant, $w_l=1/20\ \forall l$, $\phi_2(\cdot;\mu,\Sigma)$ is the bivariate Gaussian density of mean μ and covariance Σ, which is assumed diagonal. We adopt the $\{\mu_l\}$ and $\{\Sigma_l\}$ used by Kou et al. (2006).

(NL1), example 1	True	$\epsilon = 0.05$	$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$	$\epsilon = 1.0$
$\mathbb{E}[X_1]$	4.478	4.703	4.900	4.890	4.949	4.763
$\mathbb{E}[X_2]$	4.905	5.113	5.503	5.324	5.606	5.470
CPU (sec)		107	107	109	107	107
(NL2), example 2		4.318	4.391	4.332	4.327	4.399
		4.754	4.734	4.801	4.734	4.691
		108	110	109	114	118
(NL3), example 3		4.423	4.731	4.515	4.418	4.277
		4.628	5.124	4.933	4.936	4.554
		108	107	108	108	108

TABLE 1. Estimates from mixture comparison for Non-Linear MCMC. We ran each algorithm 5 times for 2 million iterations after a 50000 iteration burn-in and allowed the possibility of self-interaction every 200^{th} iteration.

We will use a population MCMC algorithm (e.g. Jasra et al. (2005)) to approximate η in (NL2-3). This will be an MCMC algorithm that samples from a target density on an augmented space and will admit η as a marginal; see Andrieu et al. (2006) for details. The MCMC step (i.e. K) in the non-linear kernels was taken as 200 iterates of the random walk/population kernel.

5.2. **Results**

We ran the algorithms five times for 2 million iterations after burn-in (50000 iterations, that is, we did not use the selection step until this time), storing only 10000 samples (all results averaged over the chains). The chains were run for a similar CPU time, as reported in Table 1.

We estimated $\mathbb{E}[X]$ using the approximation S_{9999} for each non-linear MCMC method (using 5 different settings of ϵ); the results are in Table 1.

In Table 1 we can observe that the fully self-interacting approximation (NL1) has performed quite poorly for all ϵ; the estimates of the means become even more inaccurate as ϵ goes to 1. We can intuitively explain this poor performance as follows. Despite the fact that an approximation of π seems optimal, no property of π is used in the selection step and hence the algorithm suffers from very slow convergence properties.

Conversely, (NL2-3) both perform reasonably well, with quite similar parameter estimates for all values of ϵ. The algorithms (NL2-3) are able to avoid most of the difficulties of (NL1) because they rely on more sophisticated selection schemes (NL2-3) and exchange step (NL3). This allows them to exploit the information of the empirical measure more efficiently.

A final point is that when we ran the population MCMC kernel to sample from the target for a similar CPU (110sec), the results were significantly poorer than for the self-interacting algorithms with estimates of 4.00 for $\mathbb{E}[X_1]$ and 3.73 for $\mathbb{E}[X_2]$. This emphasizes the importance of self-interacting mechanisms.

6. SUMMARY

In this paper we have presented several non-linear MCMC kernels, as well as a practical means to simulating these kernels. A full methodological and convergence study can be found in Andrieu et al. (2006).

REFERENCES

ANDRIEU, C. & MOULINES E. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Prob.*, **16**, 1462–1505.

ANDRIEU, C., JASRA, A., DOUCET, A., & DEL MORAL, P. (2006). Non-Linear Markov chain Monte Carlo via self-interacting approximations. Technical Report, University of Bristol.

Atchadé, Y. F. (2006). Resampling from the past to improve MCMC algorithms. Technical Report, Department of Mathematics & Statistics, University of Ottawa.

Brockwell, A. & Doucet, A. (2006). Sequentially interacting Markov chain Monte Carlo. Technical Report, Canergie Mellon University.

Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer: New York.

Del Moral, P. & Miclo, L. (2004). On convergence of chains with occupational self-interactions. *Proc. R. Soc. Lond. A*, **460**, 325–46.

Del Moral, P., Doucet, A. & Jasra, A. (2006). Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, **68**, 411–436.

Jasra, A., Stephens, D. A. & Holmes, C. C. (2005). Population-based reversible jump Markov chain Monte Carlo. Technical Report, Imperial College London.

Kou, S. C, Zhou, Q., & Wong, W. H. (2006). Equi-energy sampler with applications to statistical inference and statistical mechanics (with discussion). *Ann. Statist.* (in press).

Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer: New York.