ESAIM: PROCEEDINGS, September 2007, Vol.19, 1-5
Christophe Andrieu & Dan Crisan, Editors
DOI: 10.1051/proc:071901

CONVERGENCE OF THE EQUI-ENERGY SAMPLER

CHRISTOPHE ANDRIEU, AJAY JASRA, ARNAUD DOUCET & PIERRE DEL MORAL[1]

Abstract. In a recent paper 'The equi-energy sampler with applications statistical inference and statistical mechanics' Kou, Zhou & Wong (2006) have presented a new stochastic simulation method called the equi-energy (EE) sampler. This technique is designed to simulate from a probability measure π, perhaps only known up to a normalizing constant. The authors demonstrate that the sampler performs well in quite challenging problems but their convergence results (Theorem 2) appear incomplete. This was pointed out, in the discussion of the paper, by Atchadé & Liu (2006) who proposed an alternative convergence proof. However, this alternative proof, whilst theoretically correct, does not correspond to the algorithm that is implemented. In this note we consider the difficulties of the proofs as well as pointing out an alternative convergence result established by the authors (Andrieu et al. 2007b).

1. Introduction

In this note we consider the convergence properties of a new stochastic simulation technique, the equi-energy sampler introduced in (Kou, et al. 2006). This is a method designed to draw samples from a probability measure $\pi \in \mathcal{P}(E)$ (where $\mathcal{P}(E)$ denotes the class of probability measures) on measurable space $(E, \mathcal{E})$, where E may be a high dimensional space and the density, is known pointwise up to a potentially unknown constant. In particular, the algorithm generates a non-Markovian stochastic process $\{X_n\}_{n\geq 0}$ whose stationary distribution is ultimately π; this algorithm is described fully in Section 2.

In the paper of Kou et al. (2006), an attempt to analyze the algorithm is made (in Theorem 2). However, it was noticed in the discussion by Atchadé & Liu (2006) that this result is incomplete. We note the points that were stated by Atchadé & Liu and further expand upon their point; see Section 3. The work of Atchadé & Liu provides an alternative convergence proof: Although this proof is correct, the authors study a stochastic process which does not correspond to the algorithm.

The objective of this note is to point out the difficulties of the proofs as well to point towards a result that has been established by the authors.

2. Notation and Algorithm

We now outline the notation that is adopted throughout the paper as well as the algorithm that is analyzed.

[1] University of Bristol, Imperial College London,
University of British Columbia & University of Nice

2.1. **Notation**

Define a measurable space $(E, \mathcal{E})$, with $\pi \in \mathcal{P}(E)$ a target probability measure of interest.

For a stochastic process $\{X_n\}_{n\geq 0}$ on $(E^{\mathbb{N}}, \mathcal{E}^{\otimes\mathbb{N}})$, $\mathcal{G}_n = \sigma(X_0, \ldots, X_n)$ is the natural filtration. $\mathbb{P}_x$ is taken as a probability law of a stochastic process with initial distribution δ_x and $\mathbb{E}_x$ the associated expectation. We use $X_n \xrightarrow{a.s}_{\mathbb{P}} X$ to denote almost sure convergence of X_n to X. The equi-energy sampler generates a stochastic process on $(\Omega, \mathcal{F})$, which is defined in the next Section.

Throughout, $K : E \to \mathcal{P}(E)$ is taken as a generic Markov kernel; the standard notations, for measurable $f : E \to \mathbb{R}$, $K(f)(x) := \int_E f(y)K(x, dy)$ and for $\mu \in \mathcal{P}(E)$ $\mu K(f) := \int_E K(f)(x)\mu(dx)$ are used. $\mathcal{B}_b(E)$ is used to represent the bounded measurable functions and for $f \in \mathcal{B}_b(E)$, $\|f\|_\infty := \sup_{x\in E} |f(x)|$ is used to denote the supremum norm.

We will denote by $K_\mu : \mathcal{P}(E) \times E \to \mathcal{P}(E)$ a generic *non-linear* Markov kernel and its invariant measure (given its existence) as $\omega(\mu)$ ($\omega : \mathcal{P}(E) \to \mathcal{P}(E)$). For a sequence of probability measures $\{\mu_n\}_{n\geq 0}$ we denote the composition $\int_{E^{n-1}} K_{\mu_1}(x, dy_1) \ldots K_{\mu_n}(y_{n-1}, A)$ as $K_{\mu_1:\mu_n}(x, A)$. The empirical measure of an arbitrary stochastic process $\{X_n\}_{n\geq 0}$ is defined, at time n, as: $S_n(du) := \frac{1}{n+1}\sum_{i=0}^n \delta_{x_i}(du)$.

In addition, $a \vee b := \max\{a, b\}$ (resp. $a \wedge b := \min\{a, b\}$). The indicator function of $A \in \mathcal{E}$ is written $\mathbb{I}_A(x)$. Note also that $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $\mathbb{T}_m := \{1, \ldots, m\}$.

2.2. **Algorithm**

We introduce a sequence of probability measures, for $r \geq 2$, $\{\pi_n\}_{n\in\mathbb{T}_r}$, $\pi_n \in \mathcal{P}(E)$, $n \in \mathbb{T}_r$ and $\pi_r \equiv \pi$ which are assumed to be absolutley continuous, wrt some reference measure λ^*, and, in an abuse of notation, write the Radon-Nikodym derivatives as $d\pi_n/d\lambda^*(x) = \pi_n(x)$ also. The EE sampler will generate a stochastic process $\{Y_n^r\}_{n\geq 0}$, with $Y_n^r = (X_n^1, \ldots, X_n^r)$, with $X_n^i : E \to \mathbb{R}^k$, $i \in \mathbb{T}_r$, $k \geq 1$ (that is $\{Y_n^r\}_{n\geq 0}$ is a stochastic process on $(\Omega, \mathcal{F}) = ((E^r)^{\mathbb{N}}, (\mathcal{E}^{\otimes r})^{\otimes\mathbb{N}})$). Central to the construction of the EE sampler is the concept of the energy rings; this will correspond to the partition $E = \bigcup_{i=1}^d E_i$.

For each X_n^i we associate a non-linear Markov kernel $\{K_{\mu,n}\}_{n\in\mathbb{T}_r}$ with $K_{\mu,1} \equiv K_1$ (i.e. K_1 is an ordinary Markov kernel) and $\mu \in \mathcal{P}(E)$. Additionally, assume that for $i = 2, \ldots, r-1$:

$$\omega_i(\pi_{i-1})K_{\pi_{i-1},i}(dy) \quad = \quad \omega_i(\pi_{i-1})(dy) = \pi_i(dy) \tag{1}$$

and that $\pi_1 K_1 = \pi_1$. Here, it is assumed that, given that we input the invariant probability measure for $K_{\pi_{i-2},i-1}$ into the non-linear kernel $K_{\mu,i}$, the target probability measure π_i is obtained. Define:

$$K_{\mu,i}(x, dy) \quad := \quad (1-\epsilon)K_i(x, dy) + \epsilon Q_{\mu_x,i}(x, dy) \tag{2}$$

$i = 2, \ldots, r$, $\epsilon \in [0, 1]$, with K_i a Markov kernel of invariant distribution π_i and also:

$$\begin{aligned} Q_{\mu_x,i}(x, dy) &:= \int_E \mu_x(dz)K_i^S(K_i(dy))(x, z) \\ \mu_x(A) &:= \sum_{i=1}^d \mathbb{I}_{E_i}(x)\frac{\mu(E_i \cap A)}{\mu(E_i)} \end{aligned}$$

where it is assumed $\mu(E_i) > 0$; let $\mathcal{P}_d(E) = \{\mu \in \mathcal{P}(E) : \mu(E_i) > 0\ \forall i \in \mathbb{T}_d\}$. Finally define:

$$\begin{aligned} K_i^S((x, y), d(x', y')) &:= \delta_x(dy')\delta_y(dx')\alpha_i(x, y) + \delta_x(dx')\delta_y(dy')[1 - \alpha_i(x, y)] \\ \alpha_i(x, y) &= 1 \wedge \frac{\pi_i(y)\pi_{i-1}(x)}{\pi_i(x)\pi_{i-1}(y)} \end{aligned}$$

which is the swapping kernel. It is easily seen that the kernels (2) satisfy the equation (1). However, it is often the case that such a system cannot be simulated exactly. The idea is to approximate the correct probability measures π_n via the empirical measures generated by the previous chain.

The algorithm which corresponds to the equi-energy sampler is as follows. Define predetermined integers $N_1, \ldots, N_r$ and assume that for all $i \in \mathbb{T}_r$, $j = \mathbb{T}_d$ (recall d corresponds to the number of energy levels) we have $S^i_{N_{1:i}}(E_j) > 0$ with S^i the empirical measure of the i^{th} process and $N_{1:i} = \sum_{j=1}^{i} N_j$. The algorithm is in Figure 1.

0.: Set $n = 0$ and $X_0^{1:r} = x_0^{1:r}$, $S_0^l = \delta_{x_0^l}$, $l = 1, \ldots, r$. Set $i = 1$.
1.: Perform the following for $i = 1$ until $i = r$. Set $j = 1$.
2.: Perform the following for $j = 1$ until $j = N_i$, then set $i = i + 1$ and go to 1.
3.: Set $n = n + 1$, $k = 1$.
4.: Perform the following for $k = 1$ until $k = i$, then set $k = i + 1$ and go to 5.
$X_n^k \sim K_{S_n^{k-1},k}(x_{n-1}^k, \cdot)$, $S_n^k = S_{n-1}^k + \frac{1}{n+1}[\delta_{x_n^k} - S_{n-1}^k]$, set $k = k + 1$ and go to 4.
5.: Perform the following for $k = i + 1$ until $k \geq r$, then set $j = j + 1$ and go to 2.
6.: $X_n^k \sim \delta_{x_{n-1}^k}(\cdot)$ then set $k = k + 1$ and go to 5.

FIGURE 1. An equi-energy sampler.

Remark 1. *We point out here that our algorithm is slightly different from that of Kou et al. However, it should be noted that, from a technical point of view, changing the algorithm back to the EE sampler presents no difficulties, in terms of the arguments in Andrieu et al. (2007b).*

3. Discussion of the Previous Proofs

The difficulties of the convergence proofs of Kou et al (2006) and Atchadé & Liu (2006) are now discussed.

3.1. Theorem 2 of Kou et al. (2006)

We begin with the proof of Theorem 2 of Kou et al. Recall that the Theorem states, under some assumptions, that the steady state distribution of $\{X_n^i\}_{n \geq 0}$ is π_i. The authors use induction and start by using the ergodicity of the M-H chain which verifies the case $r = 1$ and continue from there.

The main difficulty of the proof is as follows, quoting Kou et al (2006), pp-1590:

> *Therefore, under the induction assumption, $X^{(i)}$ is asymptotically equivalent to a Markovian sequence governed by $S^{(i)}(x, \cdot)$.*

Here the kernel $S^{(i)}(x, \cdot)$ is the theoretical kernel corresponding to $K_{\pi_{i-1},i}$. The authors then state that $S^{(i)}(x, \cdot)$ is an ergodic Markov kernel which then yields the convergence of $X^{(i)}$. This is the difficulty of the proof: the authors verify that the transitions of the stochastic process are asymptotically equivalent to that of an ergodic Markov kernel, however, this is not enough to provide the required convergence of the process. That is, Kou et al. (2006) prove that (suppressing the notation $N_{1:i-1}$)

$$\lim_{n \to \infty} |K_{S_n^{i-1},i}(x, A) - K_{\pi_{i-1},i}(x, A)| \xrightarrow{a.s}_{\mathbb{P}^{(i-1)}} 0$$

where $\mathbb{P}^{(i-1)}$ is the probability law of the process with $i-1$ chains. However, this convergence property essentially means that when the input probability measure S_n^{i-1} is converging to the 'correct' probability measure π_{i-1} then a set-wise convergence of the non-linear kernel $K_{\cdot,i}$ is induced. This is far from sufficient as the law of the process at iteration n is, for $A \in \mathcal{E}$

$$K_{S_1^{i-1},i}[K_{S_2^{i-1},i}[\cdots K_{S_n^{i-1},i}(A)],$$

where $S_1^{i-1}, S_2^{i-1}, \ldots, S_n^{i-1}$ are empirical distributions constructed from the same realisation of the process at level $i-1$. It is clear that if the algorithm is to converge, then the joint distributions of $X_{n-\tau}^{(i)}, \ldots, X_n^{(i)}$ for any (in fact increasing with n) lag τ should converge to

$$K_{\pi_{i-1},i} \times K_{\pi_{i-1},i} \times \cdots \times K_{\pi_{i-1},i},$$

which as we shall see is far from trivial. This remark indicates an appropriate approach to a proof; via standard Markov chain convergence theorems. As a result, using the arguments of Kou et al. (2006), we cannot even say that

$$\lim_{n\to\infty} |K_{S_1:S_{n+N_{1:i-1}},i}(x,A) - \pi_i(A)| \xrightarrow[\mathbb{P}^{(i-1)}]{a.s} 0$$

via the ergodicity of $K_{\pi_{i-1},i}(x,A)$; i.e. a set-wise convergence of the kernel that is *simulated*.

3.2. **Theorem 3.1 of Atchadé & Liu (2006)**

Atchadé & Liu state (pp-1625, in the proof of Theorem 3.1):

> *Note that the i^{th} chain is actually a non-homogeneous Markov chain with transition kernels $K_0^{(i)}, K_1^{(i)}, \ldots,$ where $K_n^{(i)}(x,A) = \mathbb{P}(X_{n+1}^{(i)} \in A | X_n^{(i)} = x)$.*

This statement is not quite accurate. The i^{th} chain is a non-homogeneous Markov chain only conditional upon a realization of the previous chain; unconditionally, it is not a Markov chain. As a result, Atchadé & Liu analyze the process of kernel:

$$K_n^{(i)}(x,dy) \quad = \quad (1-\epsilon)K_i(x,dy) + \epsilon \mathbb{E}\left[R_n^{(i)}(x,dy)\right]$$

where $R_n^{(i)}$ is defined in Atchadé & Liu. This is not the kernel corresponding to the algorithm; the algorithm simulates:

$$Q_{S_x^{i-1},i}(x,dy) \quad = \quad \int_E S_x^{i-1}(dy) K_i^S(K_i(dy))(x,y)$$

that is, we do not integrate over the process $\{X_n^{i-1}\}$, we condition upon it. The algorithm that they study corresponds to an implementation of the EE algorithm where the empirical measures $S_1^{i-1}, S_2^{i-1}, \ldots, S_n^{i-1}$ are constructed from n independent realisations of the process at level i. Therefore, the proofs of Atchadé & Liu do not provide a theoretical validation of the equi-energy sampler.

4. Ergodicity Results

The SLLN is now presented; see Andrieu et al. (2007a) for further details and greater generality. Note that our proof is non-trivial and relies on a SLLN for $U-$statistics of stationary ergodic stochastic processes (Aaronson et al. 1996).

4.1. **Assumptions**

We make the following assumptions (it is assumed that for any $i \in \mathbb{T}_r$, $j \in \mathbb{T}_d$, $\pi_i(E_j) > 0$ throughout).

(**A1**) • (*Stability of Algorithm*): There is a universal constant $\theta > 0$, such that for any $n \geq 0$, $j \in \mathbb{T}_d$, $i \in \mathbb{T}_{r-1}$ we have, recalling that $N_{1:i} = \sum_{j=1}^{i} N_j$:

$$S_{N_{1:i}+n}^i(E_j) \quad \geq \quad \theta \qquad \mathbb{P}_{x_0^{1:r}} - a.s.$$

(**A2**) • (*Uniform Ergodicity*): The $\{K_n\}_{n\in\mathbb{T}_r}$ are uniformly ergodic Markov kernels with a one step minorization condition. That is: $\forall n \in \mathbb{T}_r$, $\exists(\phi_n, \nu_n) \in \mathbb{R}^+ \times \mathcal{P}(E)$ such that $\forall(x, A) \in E \times \mathcal{E}$:

$$K_n(x, A) \quad \geq \quad \phi_n \nu_n(A).$$

(**A3**) • (*State-Space Constraint*): E is polish (separable complete metrisable topological space).

4.2. **SLLN**

Theorem 4.2. *Assume (A1-2). Then for any $p \geq 1$, $\exists B_p < \infty$ such that for any $n \geq N_{1:r-1}$ and $f \in \mathcal{B}_b(E)$ we have that:*

$$\mathbb{E}_{x_0^{1:r}}\left[|[S_n^r - S_{n,r}^{\omega}](f)|^p\right]^{1/p} \quad \leq \quad \frac{B_p \|f\|_\infty}{(n - N_{1:r-1} + 1)^{\frac{1}{2}}}.$$

if, in addition, (A3) holds then for any $f \in \mathcal{B}_b(E)$:

$$S_n^2(f) \xrightarrow{a.s}_{\mathbb{P}_{x_0^{1:2}}} \pi_2(f).$$

REFERENCES

Aaronson, J., Burton, R., Dehling, H., Gilhat, D., Hill, T. & Weiss B. (1996). Strong laws for $L-$ and $U-$ statistics. *Trans. Amer. Math. Soc.*, **348**, 2845–2866.

Andrieu, C., Jasra, A., Doucet, A. & Del Moral P. (2007a). Non-Linear Markov chain Monte Carlo via self interacting approximations. Technical Report, University of Bristol.

Andrieu, C., Jasra, A., Doucet, A. & Del Moral P. (2007b). A note on the convergence of the equi-energy sampler. Technical Report, University of Bristol.

Atchadé, Y. & Liu, J. S. (2006). Discussion of Kou, Zhou & Wong. *Ann. Statist.*, **34**, 1620–1628.

Kou, S. C, Zhou, Q., & Wong, W. H. (2006). Equi-energy sampler with applications to statistical inference and statistical mechanics (with discussion). *Ann. Statist.*, **34**, 1581–1619.