Smoothness of Metropolis-Hastings algorithm and application to entropy estimation

Didier Chauveau[1] Pierre Vandekerkhove[2]

[1] *MAPMO - CNRS UMR 6628*
Fédération Denis Poisson – Université d'Orléans
didier.chauveau@univ-orleans.fr
[2] *LAMA - CNRS UMR 8050, Université de Marne-la-Vallée*
Pierre.Vandekerkhove@univ-mlv.fr

April 13, 2011

Abstract

The transition kernel of the well-known Metropolis-Hastings (MH) algorithm has a point mass at the chain's current position, which prevent direct smoothness properties to be derived for the successive densities of marginals issued from this algorithm. We show here that under mild smoothness assumption on the MH algorithm "input" densities (the initial, proposal and target distributions), propagation of a Lipschitz condition for the iterative densities can be proved. This allows us to build a consistent nonparametric estimate of the entropy for these iterative densities. This theoretical study can be viewed as a building block for a more general MCMC evaluation tool grounded on such estimates.

Keywords. Entropy, Kullback divergence, Metropolis-Hastings algorithm, nonparametric statistic.

AMS 2000 subject Classification 60J22, 62M05, 62G07.

1 Introduction

A Markov Chain Monte Carlo (MCMC) method generates an ergodic Markov chain $x^{(t)}$ for which the stationary distribution is a given probability density function (pdf) f over a state space $\Omega \subseteq \mathbb{R}^s$. In, e.g., Bayesian inference, f is a posterior distribution typically known only up to a multiplicative normalization constant, hence simulation or integration w.r.t. f can be approximated by ergodic averages from the chain. The Metropolis-Hastings (MH) algorithm (Hastings (1970) and Metropolis *et al.* (1953)) is one of the most popular MCMC algorithm. An account

of definitions and convergence properties of MCMC algorithms can be found, e.g., in Gilks *et al.* (1996).

In this paper we are concerned with smoothness properties of the MH transition kernel. Each MH step is based on the generation of the proposed next move y from a general conditional *proposal density* $q(y|x)$. For a starting value $x^{(0)} \sim p^0$, the n-th step $x^{(n)} \to x^{(n+1)}$ of the algorithm is as follows:

1. `generate` $y \sim q(\cdot|x^{(n)})$

2. `compute` $\alpha(x^{(n)}, y) = \min\left\{1, \dfrac{f(y)q(x^{(n)}|y)}{f(x^{(n)})q(y|x^{(n)})}\right\}$

3. `take` $x^{(n+1)} = \begin{cases} y & \texttt{with probability } \alpha(x^{(n)}, y), \\ x^{(n)} & \texttt{with probability } 1 - \alpha(x^{(n)}, y). \end{cases}$

It is clear from the definition that the "target" pdf f needs to be known only up to a (normalizing) multiplicative constant. Two well-known MH strategies are (i) the *Independence Sampler* (IS), with proposal distribution $q(y|x) = q(y)$ independent of the current position, and (ii) the Random Walk MH algorithm (RWMH), for which the proposal is a random perturbation u of the current position, $y = x^{(n)} + u$. The common choice for the latter is a gaussian perturbation with a fixed variance matrix which acts as the scaling parameter of the perturbation and has to be tuned. Ergodicity and convergence properties of the MH algorithm have been intensively studied in the literature, and conditions have been given for its geometric convergence (see, e.g., Mengersen and Tweedie (1996), Roberts and Tweedie (1996), and Jarner and Hansen (2000)).

In practice, many choices for the proposal density can be made, with the goal of improving mixing and convergence properties of the HM algorithm. In the IS case, Mengersen and Tweedie (1996) proved geometric convergence with rate $(1 - a)^n$ under the minoration condition $q(y) \geq af(y)$ for some $a > 0$, pointing out the link between the convergence rate and how q resembles f. For the RWMH strategy, "good" scaling constants must be found, since the mixing depends dramatically on the variance matrix of the perturbation (Roberts and Rosenthal, 2001). Hence selection of "good" proposal densities in these senses can be done using several procedures, either *a priori* (numerical approximation of the shape of f, e.g., modes), or more recently *adaptive* methods to dynamically build a proposal density on the basis of the chain(s) history (see, e.g., Gilks *et al.* (1998), Haario *et al.* (2001), Chauveau and Vandekerkhove (2002), Atchadé and Rosenthal (2005), or Andrieu and Thoms (2008) for a recent survey on adaptive MCMC). One practical difficulty is that these various strategies are often associated to unknown rates of convergence because of the complexity of the MCMC kernels.

The objective of this paper is twofold: (i) provide a detailed study of the smoothness of the successive marginal densities induced by the MH kernel, assuming only

mild smoothness conditions of the "input ingredients" of the MH algorithm that are the initial, proposal and target densities, and (ii) taking advantage of this smoothness property to propose a simulation-based estimate of the entropy of the marginal density p^n of a MH algorithm at time n, $\mathcal{H}(p^n)$, where

$$\mathcal{H}(p) = \int p(x) \log p(x)\, dx \tag{1}$$

is the relative entropy of a probability density p. A motivation for (ii) is related to the evaluation of the often unknown rates of convergence of MH algorithms as discussed above: indeed an estimate of $\mathcal{H}(p^n)$ — more precisely a monitoring of $n \mapsto \mathcal{H}(p^n)$ — can be used to evaluate the rate of convergence of a (eventually adaptive) MH algorithm. The algorithm efficiency can also be monitored through the evolution in time (n) of the Kullback-Leibler divergence

$$\mathcal{K}(p^n, f) = \int \log\left(\frac{p^n(x)}{f(x)}\right) p^n(x)\, dx = \mathcal{H}(p^n) - \mathbb{E}_{p^n}[\log f],$$

for which the estimation of $\mathcal{H}(p^n)$ is a building block. The Kullback divergence is currently used as a criterion in other simulation approaches (see Douc *et al.* (2007)), and Holden (1998)'s uniform condition implies that $\mathcal{K}(p^n, f)$ decreases geometrically (see Proposition 3 in the Appendix).

For point (ii), we propose to simulate N i.i.d. copies of Markov chains from the MH algorithm and to use the N chains locations at time n, $(X_1^n, \ldots, X_N^n)$ i.i.d. $\sim p^n$. Note that when f is analytically known, an *a.s.* consistent estimate of $\mathbb{E}_{p^n}[\log f]$ is obtained easily by Monte Carlo integration

$$\frac{1}{N}\sum_{j=1}^{N} \log f(X_j^n) \xrightarrow{a.s.} \mathbb{E}_{p^n}[\log f], \tag{2}$$

where the convergence comes from the strong law of large numbers. When f is known only up to a multiplicative constant $f(\cdot) = C\varphi(\cdot)$, $\mathbb{E}_{p^n}[\log \varphi]$ can be estimated similarly, hence evaluation of $\mathcal{K}(p^n, f)$ can be made up to the constant $\log(C)$. Hence we focus here only on the estimation of $\mathcal{H}(p^n)$ for MH algorithms.

Various estimators for $\mathcal{H}(p)$ based on an i.i.d. sample from p have been proposed and studied in the literature, mostly for the univariate case $s = 1$. One approach consists in obtaining a suitable density estimate $\hat{p}_N$ for p, and then susbtituting p by $\hat{p}_N$ in an entropy-like functional of p. This approach have been adopted by Dmitriev and Tarasenko (1973a) (1973b), Ahmad and Lin (1976, 1989), Györfi and Van Der Meulen (1987) (1989), and Mokkadem (1989) who prove strong consistency of their estimators in various framework. More recently Eggermont and LaRiccia (1999) proved best asymptotic normality for the Ahmad and Lin's estimator for $s = 1$, this property being lost in higher dimension. Another method used to estimate $\mathcal{H}(p)$ is based on considering the sum of logarithms of spacings of order statistics.

This approach was considered by Tarasenko (1968), and Dudewicz and Van Der Meulen (1981).

For the MH case, a major difficulty comes from the fact that the MH kernel has a point mass at the current position, which prevent strong smoothness properties to be provable. Hence we choose to use the entropy estimate proposed by Györfi and Van Der Meulen (1989), but with compatible smoothness conditions of Ivanov and Rozhkova (1981): a Lipschitz condition which appeared tractable in our setup. Note that another appproach to estimate $\mathcal{H}(p^n)$ in Markov chains setup has been proposed by Chauveau and Vandekerkhove (2007). It is based on a "double" Monte Carlo integration, and does not require the regularity assumptions needed for the kernel density methods. This double Monte Carlo consistent estimator (also based on the simulation of i.i.d. chains) applies when p^n is the pdf at time n of a Gibbs sampler, whenever the full conditional distributions are known (which is the usual case). Unfortunately, this estimate cannot be applied in the MH case since its kernel has a point mass in the current position (see Section 2).

In Section 2, we establish assumptions on the proposal density q, f and the initial density p^0 to insure that, at each time n, adequate smoothness conditions hold for the successive densities p^n, $n \geq 1$. We give in Section 3 theoretical conditions under which our simulation-based estimate of $\mathcal{H}(p^n)$ is proved to converge. Finally, Section 4 illustrates the behavior of our estimator for a synthetic example in moderate dimension.

2 Smoothness of MCMC algorithms densities

For estimating the entropy of a MH algorithm successive densities, we start by showing that a mild smoothness assumption, a Lipschitz condition, can propagate to the sequence of marginals (p^n), $n \geq 1$.

2.1 The MH Independence Sampler case

From the description of the MH algorithm in Section 1, we define the off-diagonal transition density of the MH kernel at step n by:

$$p(x,y) = \begin{cases} q(y|x)\alpha(x,y) & \text{if } x \neq y, \\ 0 & \text{if } x = y, \end{cases} \tag{3}$$

and set the probability of staying at x, $r(x) = 1 - \int p(x,y)dy$. The MH kernel can be written as:

$$P(x,dy) = p(x,y)dy + r(x)\delta_x(dy), \tag{4}$$

where δ_x denotes the point mass at x.

We focus first on the IS case ($q(y|x) \equiv q(y)$) since it allows for simpler conditions. Let p^0 be the density of the initial distribution of the MH algorithm with proposal

density q and target f. We will assume that these densities are sufficiently smooth in a sense that will be precised. From (4), the successive densities of the IS are given by the recursive formula

$$\begin{aligned} p^{n+1}(y) &= q(y)\int p^n(x)\alpha(x,y)\,dx + p^n(y)\int q(x)(1-\alpha(y,x))\,dx \quad (5)\\ &= q(y)I_n(y) + p^n(y)\,(1-I(y))\,, \quad (6)\end{aligned}$$

where

$$\begin{aligned} I_n(y) &= \int p^n(x)\alpha(x,y)\,dx, \quad n\geq 0 \quad (7)\\ I(y) &= \int q(x)\alpha(y,x)\,dx. \quad (8)\end{aligned}$$

We consider the first iteration of the algorithm. From (6), the regularity properties of the density p^1 are related to the regularity properties of the two parameter-dependent integrals I_1 and I, that are classically handled by standard results (see, e.g., Billingsley (1995) Theorem 16.8 p. 212). Continuity is straightforward here, and the proof is omitted since it is simply an application of Lebesgue's dominated convergence theorem:

Lemma 1 *If q and f are strictly positive and continuous on $\Omega\subseteq\mathbb{R}^s$, and p^0 is continuous, then p^n is continuous on Ω for $n\geq 1$.*

From equation (6), we have directly that

$$\begin{aligned} |p^{n+1}(y)-p^{n+1}(z)| &\leq ||q||_\infty\,|I_n(y)-I_n(z)| + ||I_n||_\infty\,|q(y)-q(z)|\\ &+ ||1-I||_\infty\,|p^n(y)-p^n(z)| + ||p^n||_\infty\,|I(y)-I(z)|. \quad (9)\end{aligned}$$

To prove recursively that p^{n+1} is Lipschitz, we have first to prove that I_n and I are both Lipschitz. For convenience, we denote (where $a\wedge b=\min\{a,b\}$)

$$\alpha(x,y)=\phi(x,y)\wedge 1, \quad \phi(x,y)=\frac{h(x)}{h(y)}, \quad h(x)=\frac{q(x)}{f(x)}.$$

Lemma 2 *If f/q is c_1-Lipschitz, and $\int p^0 h<\infty$, then for all $n\geq 1$, $\int p^n h<\infty$, and I_n is $(c_1\int p^n h)$-Lipschitz.*

Proof. First we have to check that $\int p^0 h<\infty$ can be iterated. This comes directly from the recursive definition (5) (since $0\leq r(x)\leq 1$):

$$\begin{aligned} \int p^1(y)h(y)\,dy &= \int\left[\int p^0(x)p(x,y)\,dx + p^0(y)r(y)\right]h(y)\,dy\\ &\leq \int\frac{q(y)^2}{f(y)}\left[\int p^0(x)\phi(x,y)\,dx\right]dy + \int p^0(y)\frac{q(y)}{f(y)}\,dy\\ &= 2\int p^0(y)h(y)\,dy<\infty.\end{aligned}$$

Hence $\int p^0 h < \infty \Rightarrow \int p^n h < \infty$ for $n \geq 1$. Then, we have

$$\begin{aligned} |I_n(y) - I_n(z)| &\leq \int p^n(x)|\alpha(x,y) - \alpha(x,z)|\,dx \\ &\leq \int p^n(x)|\phi(x,y) - \phi(x,z)|\,dx \\ &\leq \int p^n(x)h(x)\left|\frac{f(y)}{q(y)} - \frac{f(z)}{q(z)}\right| dx \leq \left(c_1 \int p^n h\right) ||y-z||. \end{aligned}$$

□

Note that the hypothesis that f/q is Lipschitz is reasonable in the IS context. As recalled in the introduction, the IS is uniformly geometrically ergodic if $q(y) \geq af(y)$ for some $a > 0$ (Mengersen and Tweedie 1996). Actually, these authors also proved that the IS is not even geometrically ergodic if this condition is not satisfied. But satisfying this minoration condition requires q to have tails heavier than the tails of the target f. Hence, common choices for implementing the IS make use of heavy-tailed proposal densities, so that f/q is typically a continuous and positive function which goes to zero when $||x|| \to \infty$. It can then be assumed to be Lipschitz. This condition in Lemma 2 may thus be viewed as a consequence of the following assumption, which will be used below:

Assumption (A): *q and f are strictly positive and continuous densities on Ω, and q has heavier tails than f, so that* $\lim_{||y||\to\infty} h(y) = +\infty$.

We turn now to the second integral $I(y) = \int q(x)\alpha(y,x)\,dx$. The difficulty here comes from the fact that the integration variable is now the *second* argument of $\alpha(\cdot,\cdot)$. Hence, applying the majoration used previously gives

$$|I(y) - I(z)| \leq \int q(x)|\phi(y,x) - \phi(z,x)|\,dx = \int f(x)|h(y) - h(z)|\,dx,$$

but with assumption (A), $h = q/f$ is obviously *not* Lipschitz. A direct study of $\alpha(\cdot,x) = [h(\cdot)/h(x)] \wedge 1$ is needed here. Clearly, for each fixed $x \in \Omega$ there exists by (A) a compact set $K(x)$ such that for any $y \notin K(x)$, $h(y) \geq h(x)$. This entails that $\forall y \notin K(x)$, $\alpha(y,x) = 1$. Now, for any $y \in K(x)$, $\alpha(y,x)$ is a continuous function truncated at one, so that it is uniformly continuous. If we assume slightly more, i.e. that $\alpha(\cdot,x)$ is $c(x)$-Lipschitz, we have proved the following Lemma:

Lemma 3 *If assumption (A) holds, and if for each x there exists $c(x) < \infty$ such that*

$$\forall (y,z) \in K^2(x), \quad |\alpha(y,x) - \alpha(z,x)| \leq c(x)||y-z||, \tag{10}$$

where $c(x)$ satisfies

$$\int q(x)c(x)\,dx < \infty, \tag{11}$$

then I satisfies the Lipschitz condition:

$$\forall (y,z) \in \Omega^2, \quad |I(y) - I(z)| \leq \left(\int q(x)c(x)\,dx \right) ||y - z||.$$

We have checked that Lemma 3 holds in some simple (one-dimensional) MH situations. An example is provided in Appendix 6.1.

Proposition 1 *If the conditions of Lemmas 1, 2 and 3 hold, and if*

(i) $||q||_\infty = Q < \infty$ *and* q *is* c_q*-Lipschitz;*

(ii) $||p^0||_\infty = M < \infty$ *and* p^0 *is* c_0*-Lipschitz;*

then the successive densities of the Independance Sampler satisfy a Lipschitz condition, i.e. for any $n \geq 1$*, there exists* $k(n) < \infty$ *such that*

$$\forall (y,z) \in \Omega^2, \quad |p^n(y) - p^n(z)| \leq k(n)\,||y - z||. \tag{12}$$

Proof. Using equation (9), and the fact that

$$||I_n||_\infty \leq \int p^n(x)\,dx = 1, \quad ||I||_\infty \leq \int q(x)\,dx = 1,$$

and

$$\begin{aligned} ||p^n||_\infty &\leq Q||I_{n-1}||_\infty + ||p^{n-1}||_\infty\,||1 - I(y)||_\infty \\ &\leq nQ + M, \end{aligned}$$

we obtain

$$\begin{aligned} |p^{n+1}(y) - p^{n+1}(z)| &\leq Q|I_n(y) - I_n(z)| + |q(y) - q(z)| \\ &\quad + |p^n(y) - p^n(z)| + (nQ + M)|I(y) - I(z)|. \end{aligned}$$

Thus, applying this recursively, (12) is satisfied, with

$$\begin{aligned} k(n) &= Qc_1 \int p^n(x)h(x)\,dx + c_q \\ &\quad + ((n-1)Q + M) \int q(x)c(x)\,dx + k(n-1), \quad n \geq 2 \\ k(1) &= Qc_1 \int p^0(x)h(x)\,dx + c_q + M \int q(x)c(x)\,dx + c_0. \end{aligned}$$

□

2.2 The general Metropolis-Hastings case

When the proposal density is of the general form $q(y|x)$ depending on the current position of the chain, the successive densities of the MH algorithm are given by

$$\begin{aligned} p^{n+1}(y) &= \int p^n(x)q(y|x)\alpha(x,y)\,dx + p^n(y)\int q(x|y)(1-\alpha(y,x))\,dx \\ &= J_n(y) + p^n(y)\,(1-J(y))\,, \end{aligned} \tag{13}$$

where

$$J_n(y) = \int p^n(x)q(y|x)\alpha(x,y)\,dx, \tag{14}$$

$$J(y) = \int q(x|y)\alpha(y,x)\,dx. \tag{15}$$

In comparison with the IS case, the continuity already requires some additional local conditions. Let $B(y_0,\delta)$ denotes the closed ball centered at $y_0 \in \Omega$, with radius δ.

Lemma 4 *If $q(x|y)$ and f are strictly positive and continuous everywhere on both variables, and p^0 is continuous, and if:*

(i) $\sup_{x,y} q(x|y) \le Q < \infty$ *;*

(ii) for any $y_0 \in \Omega$ and some $\delta > 0$, $\sup_{y\in B(y_0,\delta)} q(x|y) \le \varphi_{y_0,\delta}(x)$, where $\varphi_{y_0,\delta}$ is integrable;

then p^n is continuous on Ω for $n \ge 1$.

Proof. As for Lemma 1, it is enough to check the dominating conditions of, e.g., Billingsley (1995), p.212. However, for J, we need the local condition (ii) to prove the continuity of $J(y)$ at any $y_0 \in \Omega$. □

Note that condition (ii) is reasonable. For instance, in the one-dimensional RWMH with gaussian perturbation $q(x|y) =$ pdf of $\mathcal{N}(y,\sigma^2)$ evaluated at x, one can simply take

$$\begin{aligned} \varphi_{y_0,\delta}(x) &= q(x|y_0-\delta)\mathbb{I}_{x<y_0-\delta} + q(y_0-\delta|y_0-\delta)\mathbb{I}_{[y_0-\delta,y_0+\delta]}(x) \\ &\quad + q(x|y_0+\delta)\mathbb{I}_{x>y_0+\delta}. \end{aligned} \tag{16}$$

Proposition 2 *If conditions of Lemma 4 hold, and if*

(i) $||p^0||_\infty = M < \infty$ *and p^0 is c_0-Lipschitz;*

(ii) $q(\cdot|x)\alpha(x,\cdot)$ is $c_1(x)$-Lipschitz, with $\int p^n(x)c_1(x)\,dx < \infty$,

(iii) $J(\cdot)$ *is* c_2*-Lipschitz,*

then the successive densities of the general MH satisfy a Lipschitz condition, i.e. for any $n \geq 0$*, there exists* $\ell(n) < \infty$ *such that*

$$\forall (y,z) \in \Omega^2, \quad |p^n(y) - p^n(z)| \leq \ell(n)\, ||y - z||. \tag{17}$$

Proof. First, it is easy to check that, similarly to the IS case, $||J_n||_\infty \leq Q$, $||J||_\infty \leq 1$, and $||p^n||_\infty \leq nQ + M$. Then, using the decomposition

$$\begin{aligned} |p^{n+1}(y) - p^{n+1}(z)| \quad \leq \quad & |J_n(y) - J_n(z)| + 2|p^n(y) - p^n(z)| \\ & + ||p^n||_\infty |J(y) - J(z)|, \end{aligned}$$

equation (17) is clearly a direct consequence of conditions (ii) and (iii), and the $\ell(n)$'s can be determined recursively as in the proof of Proposition 1. □

We have checked that these conditions are satisfied, e.g., in the one-dimensional case for usual RWMH algorithms with gaussian proposal densities (see Appendix 6.2).

3 Relative entropy estimation

Let $\mathbf{X}_N = (X_1, \ldots, X_N)$ be an i.i.d. N-sample of random vectors taking values in $\mathbb{R}^s$, $s \geq 1$, with common probability density function p.

Following Györfi and Van Der Meulen (1989), we decompose the sample $\mathbf{X}_N$ into two subsamples $\mathbf{Y}_N = \{Y_i\}$ and $\mathbf{Z}_N = \{Z_i\}$, defined by

$$Y_i = X_{2i} \quad \text{for } i = 1, \ldots, [N/2], \tag{18}$$

$$Z_i = X_{2i-1} \quad \text{for } i = 1, \ldots, [(N+1)/2], \tag{19}$$

where $[N]$ denotes the largest integer inferior to N. Let $\hat{p}_N(x) = \hat{p}_N(x, \mathbf{Z}_N)$ be the Parzen-Rosenblatt kernel density estimate given by

$$\hat{p}_N(x) = \frac{1}{h_N^s (N+1)/2} \sum_{i=1}^{[(N+1)/2]} K_{h_N}\left(\frac{x - Z_i}{h_N}\right), \quad x \in \mathbb{R}^s, \tag{20}$$

where the kernel K is a density and $h_N > 0$ with $\lim_{N\to\infty} h_N = 0$, and $\lim_{N\to\infty} N h_N^s = \infty$. The entropy estimate $\mathcal{H}_N(p) = \mathcal{H}_{N,\mathbf{Y},\mathbf{Z}}(p)$ introduced by Györfi and Van Der Meulen (1989), is then defined by:

$$\mathcal{H}_N(p) = \frac{1}{[N/2]} \sum_{i=1}^{[N/2]} \log \hat{p}_N(Y_i) \mathbb{I}_{\{p_N(Y_i) \geq a_N\}} \tag{21}$$

where $0 < a_N < 1$ and $\lim_{N\to\infty} a_N = 0$.

Theorem 1 *Assume that $\mathcal{H}(f) < \infty$. For all $n \geq 0$, let $\mathbf{X}_N^n$ be an i.i.d. N-sample from p^n, the p.d.f. of the MH algorithm at time n, and consider the kernel density estimate $\widehat{p^n}_N$ given in (20), based on the subsample $\mathbf{Z}_N^n$ defined in (19). Let the kernel K be a bounded density, vanishing outside a sphere S_r of radius $r > 0$, and set $h_N = N^{-\alpha}$, $0 < \alpha < 1/s$. Consider the entropy estimate $\mathcal{H}_N$ defined in (21) with $a_N = (\log N)^{-1}$. Assume that there are positive constants C, r_0, a, A and ϵ, such that either:*

(i) in the case of the Independance Sampler: f, q and p_0 satisfy conditions of Proposition 1; q satisfies the minoration condition $q(y) \geq af(y)$, and f satisfies the tail condition

$$f(y) \leq \frac{C}{||y||^s (\log ||y||)^{2+\epsilon}}, \quad \text{for } ||y|| > r_0; \tag{22}$$

(ii) in the general MH case: f, q and p_0 satisfy conditions of Proposition 2; q is symmetric ($q(x|y) = q(y|x)$); $\left|\left|p^0/f\right|\right|_\infty \leq A$, and f satisfies the tail condition

$$f(y) \leq \frac{C}{1 + ||y||^{s+\epsilon}}. \tag{23}$$

Then, for all $n \geq 0$, $\mathcal{H}_N(p^n) \xrightarrow{a.s.} \mathcal{H}(p^n)$, as $N \to \infty$.

Proof. This result uses directly the Theorem given in Györfi and Van Der Meulen (1989) p. 231. Conditions (22) or (23) and the fact that $\mathcal{H}(f) < \infty$ implies, for all $n \geq 0$, the same conditions on the densities p^n in either cases (i) or (ii). Actually, $\mathcal{H}(f) < \infty$ is a direct consequence of Proposition 3 and of the positivity of $\mathcal{K}$. For the tail condition (22), case (i), it suffices to notice that from (25) we have for all $x \in \Omega$:

$$\begin{aligned} 0 \leq p^n(x) &\leq f(x) + \kappa\rho^n f(x) \\ &\leq \frac{C(1 + \kappa\rho^n)}{||x||^s (\log ||x||)^{2+\epsilon}}. \end{aligned}$$

The tail condition for the general case (ii) comes directly from the recursive formula (13) since

$$\begin{aligned} p^1(y) &= J_0(y) + p^0(y)(1 - J(y)) \leq \int p^0(x) q(y|x) \alpha(x,y)\, dx + p^0(y) \\ &\leq \int p^0(x) q(y|x) \frac{f(y)}{f(x)}\, dx + p^0(y) \\ &\leq Af(y) \int q(x|y)\, dx + p^0(y) \leq 2Af(y). \end{aligned}$$

Applying this recursively gives

$$p^n(y) \leq 2^n A f(y) \leq \frac{2^n AC}{1 + ||y||^{s+\epsilon}},$$

which is stricter than Györfi and Van Der Meulen's tail condition. As to smoothness, the conditions of our Proposition 1 for case (i), and Proposition 2 for case (ii) give the Lipschitz condition of Ivanov and Rozhkova (1981) for p^n, which in turn is stricter than Györfi and Van Der Meulen's smoothness condition, as stated in Györfi and Van Der Meulen (1989). □

4 An illustrative example

This section shows on a simple but multivariate example the good behavior of our entropy estimate of $n \mapsto \mathcal{H}(p^n) = \mathbb{E}_{p^n}[\log(p^n)]$. We choose as a target density a $s = 5$-dimensional Gaussian density with zero mean and diagonal covariance matrix with diagonal elements $(1, 2, 3, 4, 5)$. One advantage of this simple example is that the true relative entropy is known for the gaussian: for this particular case $\mathcal{H}(f) = \mathbb{E}_f(\log f) = -9.488$.

We ran $N = 200$ and $N = 500$ i.i.d. copies of Markov chains issued from two different MH samplers: (i) A RWMH strategy with gaussian proposal of variance matrix $\sigma^2 I_5$, where $\sigma^2 = 1$ and I_5 is the 5-dimensional identity matrix; (ii) an Independence Sampler (IS) strategy with gaussian proposal of variance $\sigma^2 I_5$ with $\sigma^2 = 25$. In addition to the purpose of illustrating the performance of our estimator, this example also shows that this entropy estimate can be used to evaluate the convergence rate of a MCMC sampler, since $\mathcal{H}(p^n)$ converges to $\mathbb{E}_f(\log f)$ as $n \to \infty$. Indeed, Fig. 1 shows that the RWMH is more efficient than the IS: the HMRW entropy estimates stabilizes after about 50 iterations whereas for the IS it requires about 150 iterations. The reason is that the calibration of the HMRW variance is appropriate for this target f, whereas the IS is using a gaussian proposal density with a too large variance.

These simple examples have been run using a R (R Development Core Team, 2010) software package for MCMC comparisons, under development (see Discussion section). CPU time is about few minutes on a today laptop. The (diagonal) bandwidth matrix we used in practice in the multivariate kernel density estimate (20) is optimal for multivariate Gaussian distributions (Scott, 1992). Note that we did not use in Equation (21) the treshold $a_N = \log(N)^{-1}$ suggested in the original theorem from Györfi and Van Der Meulen (1989), but too large in practice for N about hundreds, the important point being that $\lim_{N\to\infty} a_N = 0$. Preliminary experiments we did suggest to use instead a numerically reasonable scheme consisting in removing from the subsample $\mathbf{Y}_N$ the lowest α_N percent of the $\hat{p}_N(Y_i)$'s. For these simulations we set $\alpha_N = 2\%$. This technical point deserves a more comprehensive

numerical study to determine on this basis an appropriate a_N, which is part of the software tool under development (see Discussion).

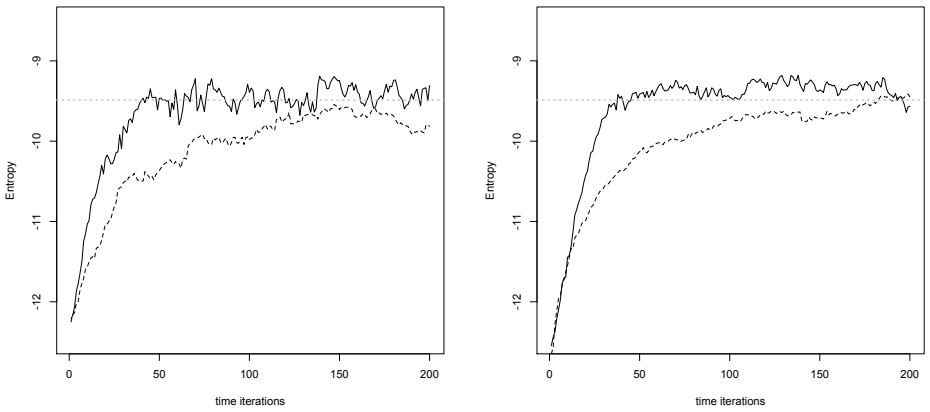

Figure 1: *Plots of* $n \mapsto \mathcal{H}_N(p^n)$ *from Left:* $N = 200$, *Right:* $N = 500$ *i.i.d. chains up to* $n = 200$ *iteration for (i) a RWMH strategy with gaussian proposal of diagonal variance matrix* $\sigma^2 I_5$ *with* $\sigma^2 = 1$ *(solid line); and (ii) Independence Sampler strategy with gaussian proposal of diagonal variance matrix* $\sigma^2 I_5$ *with* $\sigma^2 = 25$ *(dashed line). The horizontal dotted line is the true limiting (in* n*) entropy* $\mathcal{H}(f) = \mathbb{E}_f(\log f)$.

5 Discussion

In this paper we have shown that, despite the fact that the Metropolis-Hastings kernel has a point mass at the chain's current position (corresponding to the probability of rejection of each proposed next move) a Lipschitz smoothness property can "reasonably" be assumed for the iterative marginal densities. This allows us to build a simulation-based, consistent nonparametric estimate of the entropy for these iterative densities. "Reasonably" here means that the technical conditions of Theorem 1 are not meant to be verified by tedious calculations, as we did for two simple examples (Appendix 6.1 and 6.2). These examples support the idea that, if the input (initial, proposal and target) densities of the MH algorithm have usual tail and smoothness properties, then one can expect our estimate to behave well.

This theoretical study can be viewed as a building block for a methodological software tool for general (including adaptive) MCMC convergence evaluation. Within this in-progress R software package, the simulations output (e.g., stabilization of estimates of $\mathcal{H}(p^n)$ or Kullback distances) is intended to provide numerical evidence of convergence, without actual checking of the technical conditions. Note also that the MH homogeneous Markov property of the simulated chains does not play any role in the convergence of the entropy estimates, since these estimates are

based on i.i.d. copies at time n only. Hence adaptive MCMC efficiency can also be evaluated by this criterion.

On the negative side, our estimate is obviously sensitive to the "curse of dimensionality", since the involved kernel density estimates deteriorate as the dimension of the sample space increases. We are currently considering alternative entropy estimation methods, and reduction of dimension techniques, but these approaches are beyond the scope of this article.

6 Appendix

Proposition 3 *If the proposal density of the Metropolis-Hastings algorithm satifies $q(y|x) \geq af(y)$, for all $x, y \in \Omega$, and $a \in (0,1)$, then*

$$\mathcal{K}(p^n, f) \leq \kappa\rho^n(1 + \kappa\rho^n), \tag{24}$$

where $\kappa = ||p^0/f - 1||_\infty > 0$, f^0 is the initial pdf, and $\rho = (1-a)$.

Proof. We use a result due to Holden (1998) assessing the geometric convergence of the MH algorithm under a uniform minoration condition: If there exists $a \in (0,1)$ such that $q(y|x) \geq af(y)$ for all $x, y \in \Omega$, then

$$\forall y \in \Omega, \quad \left|\frac{p^n(y)}{f(y)} - 1\right| \leq (1-a)^n \left|\left|\frac{p^0}{f} - 1\right|\right|_\infty . \tag{25}$$

Using equation 25, we have:

$$\begin{aligned}
\mathcal{K}(p^n, f) &= \int \log\left(\frac{p^n(y)}{f(y)}\right) p^n(y)\, dy \\
&\leq \int \log\left(\left|\frac{p^n(y)}{f(y)} - 1\right| + 1\right)\left(\left|\frac{p^n(y)}{f(y)} - 1\right| + 1\right) f(y)\, dy \\
&\leq \log(\kappa\rho^n + 1)(\kappa\rho^n + 1) \leq \kappa\rho^n(\kappa\rho^n + 1).
\end{aligned}$$

□

The two last sections of this appendix illustrates that some of the conditions required in Proposition 1 and Proposition 2, which look difficult to check in actual situations, are satisfied for standard MH algorithms in the one-dimensional case.

6.1 The one-dimensional independence sampler case

In the IS case, the technical assumptions are conditions (10) and (11) of Lemma 3. These conditions are simpler to handle in the one-dimensional case. In this case,

when q and f are in addition derivable, and have non-oscillating tails, assumption (A) leads to

$$\exists m_1 < m_2 : \forall x < m_1, h'(x) < 0, \text{ and } \forall x > m_2, h'(x) > 0. \tag{26}$$

For a fixed $x \in \mathbb{R}$, there exists by (26) a compact set $K(x) = [a(x), b(x)]$ such that: (i) $[m_1, m_2] \subseteq K(x)$; (ii) $h(a(x)) = h(b(x)) = h(x)$; (iii) for any $y \notin K(x)$, $h(y) \geq h(x)$. As in the general case, this entails that $\forall y \notin K(x)$, $\alpha(y, x) = 1$. If we have the Lipschitz condition on $K(x)$:

$$\forall y, z, \quad |\alpha(y, x) - \alpha(z, x)| \leq c(x)|y - z|,$$

the expression of $c(x)$ can be precised

$$c(x) = \sup_{y \in K(x)} \left| \frac{\partial \phi(y, x)}{\partial y} \right| < \infty. \tag{27}$$

and Lemma 3 holds if the integrability condition (11) is satisfied. Note that $|a(x)|$ and $b(x)$ both go to $+\infty$ as $|x| \to \infty$; in particular, $b(x) = x$ for $x > m_2$. Hence $c(x) \to \infty$ as $|x| \to \infty$, and condition (11) is not always true, but merely depends on the relative decreasing rate of the tails of q and f.

For an illustrative example, assume that the tails of f are of order $x^{-\beta}$, and the tails of q are of order $x^{-\alpha}$. Satisfying assumption A requires that $\beta > \alpha$. Now, one can always use the fact that

$$c(x) \leq \sup_{y \in \mathbb{R}} \left| \frac{\partial \phi(y, x)}{\partial y} \right|,$$

so that if $\beta - 1 < \alpha < \beta$, then $c(x)$ is of order $x^{\alpha - \beta}$ for large x and (11) is satisfied. The condition $\alpha \in [\beta - 1, \beta]$ states that the tails of q should be "not too heavy", compared with the tails of f. This requirement is obviously stronger than what is needed, but more precise conditions require some analytical expression of $c(x)$ for $x \notin [m_1, m_2]$, and this expression depends on $a(x)$ and h'.

Fortunately, condition (11) is satisfied in much more general settings. For instance, consider situations where f and q are both symmetric w.r.t. 0, so that $K(x) = [-|x|, |x|]$ for x outside $[m_1, m_2]$, and $c(x)$ can be expressed in closed form. Then it is easy to verify that (11) holds for, e.g., $f \equiv \mathcal{N}(0, 1)$ and $q \equiv t(d)$, the Student distribution with d degrees of freedom, for $d \geq 2$ (even if, for $d = 2$ the tails of q are of order x^{-3}). In this example, the proposal density has tails much more heavier than f, but Lemma 3 holds i.e., I is still Lipschitz.

6.2 The one-dimensional general MH case

In the general MH case, the difficult conditions are conditions (ii) and (iii) of Proposition 2. Our aim is to show that these conditions hold in the simple RWMH case

with gaussian proposal density. In order to obtain a tractable case, let $q(y|x)$ be the p.d.f. of the gaussian $\mathcal{N}(x,1)$, and f be the density of $\mathcal{N}(0,1)$.

For condition (ii) we have to prove that $q(\cdot|x)\alpha(x,\cdot)$ is $c(x)$-Lipschitz, with $\int p^n(x)c(x)\,dx < \infty$. Here $q(y|x) = q(x|y)$, so that

$$\alpha(x,y) = 1 \wedge \frac{f(y)}{f(x)} \leq \frac{f(y)}{f(x)},$$

which is a truncated function such that, for any x, $\lim_{|y|\to\infty} \alpha(x,y) = 0$. In other words, both $\alpha(x,y)$ and $q(y|x)\alpha(x,y)$ have tails behavior for large y. The non-troncated function $\varphi_x(y) = q(y|x)f(y)/f(x)$ is then Lipschitz, with $c(x) = \sup_{y\in\mathbb{R}} |\varphi'_x(y)|$. A direct calculation (feasible in this simple case) gives $c(x) \propto \exp(x^2-2)/4$. Since to ensure the tails conditions of the successive densities p^n we have to assume that the initial distribution itself has tails lighter or equal to that of f (i.e. that $||p^0/f||_\infty < A$, see Theorem 1) then by the recursive definition of p^n we have, as in the proof of Theorem 1, $p^n(y) \leq 2^n A f(y)$, so that $\int p^n(x)c(x)\,dx < \infty$, i.e. condition (ii) of Proposition 2 holds.

We turn now to condition (iii) of Proposition 2, i.e. we have to show that $J(y)$ given by (15) is Lipschitz. For fixed $y, z \in \mathbb{R}$,

$$|J(y) - J(z)| \leq \int |q(x|y)\alpha(y,x) - q(x|z)\alpha(z,x)|\,dx.$$

As for the IS case, we need a precise study of the truncated function here. We assume first that $z > y > 0$. Since q is symmetric, $\alpha(y,x) = (f(x)/f(y)) \wedge 1$, and we can define two compact sets $K(y)$ and $K(z)$ by

$$K(t) = \{x \in \mathbb{R} : \alpha(t,x) = 1\} = \{x \in \mathbb{R} : f(x) \geq f(t)\}$$

which, in the present situation, are just $K(y) = [-y,y]$, $K(z) = [-z,z]$, and satisfy $K(y) \subset K(z)$. Hence

$$\begin{aligned} |J(y) - J(z)| \;\leq\; & \int_{K(y)} |q(x|y) - q(x|z)|\,dx \\ & + \int_{K(z)\setminus K(y)} \left| q(x|y)\frac{f(x)}{f(y)} - q(x|z) \right| dx \\ & + \int_{K(z)^c} \left| q(x|y)\frac{f(x)}{f(y)} - q(x|z)\frac{f(x)}{f(z)} \right| dx, \end{aligned}$$

where $K(z)^c = \mathbb{R} \setminus K(z)$. Using the mean value theorem, the first term can be written

$$\begin{aligned} \int_{K(y)} |q(x|y) - q(x|z)|\,dx \;&\leq\; \int |q(x|y) - q(x|z)|\,dx \\ &\leq\; |y-z| \int \frac{|x-y^*|}{2\pi} \exp\left(-(x-y^*)^2/2\right) dx \\ &\leq\; \sqrt{\frac{2}{\pi}}\,|y-z|, \end{aligned} \tag{28}$$

where the last inequality comes from the absolute first moment of the normal density.

For the second term, consider first the integral on the right side of $K(z) \setminus K(y)$, that is $\int_y^z |\varphi_{y,z}(x)|\, dx$, where

$$\varphi_{y,z}(x) = q(x|y)\frac{f(x)}{f(y)} - q(x|z).$$

In this simple setting, it is easy to check that $\varphi_{y,z}(\cdot)$ is a bounded function, monotonically decreasing from $\varphi_{y,z}(y) = \delta - q(y|z) > 0$ to $\varphi_{y,z}(z) = q(y|z) - \delta < 0$, where $\delta = q(y|y)$ is the value of the gaussian density at its mode. Hence

$$\int_y^z \left| q(x|y)\frac{f(x)}{f(y)} - q(x|z) \right| dx \leq \delta |y - z|. \tag{29}$$

The symmetric term $\int_{-z}^{-y} |\varphi_{y,z}(x)|\, dx$ is handled in a similar way.

The third term can in turn be decomposed into

$$\begin{aligned} \int_{K(z)^c} \left| q(x|y)\frac{f(x)}{f(y)} - q(x|z)\frac{f(x)}{f(z)} \right| dx &\leq Q \int_{K(z)^c} \left| \frac{f(x)}{f(y)} - \frac{f(x)}{f(z)} \right| dx \\ &+ \int_{K(z)^c} |q(x|y) - q(x|z)|\, dx, \end{aligned}$$

where, as in Proposition 2, $Q = ||q||_\infty$, and since $\sup_{x \in K(z)^c} |f(x)/f(z)| = 1$. Using the mean value theorem as for the first term,

$$\int_{K(z)^c} |q(x|y) - q(x|z)|\, dx \leq \sqrt{\frac{2}{\pi}}\, |y - z|. \tag{30}$$

Finally,

$$\begin{aligned} \int_{K(z)^c} \left| \frac{f(x)}{f(y)} - \frac{f(x)}{f(z)} \right| dx &= 2 \int_z^\infty \left| \frac{f(x)}{f(y)} - \frac{f(x)}{f(z)} \right| dx \\ &\leq 2 \left| \frac{1}{f(y)} - \frac{1}{f(z)} \right| \int_z^\infty f(x)\, dx \\ &\leq 2\sqrt{2\pi} z e^{z^2/2} \frac{e^{-z^2}}{z + \sqrt{z^2 + 4/\pi}} |y - z|, \qquad (31) \\ &\leq D|y - z|, \qquad (32) \end{aligned}$$

where the left term in (31) comes from the mean value theorem applied to the function $1/f(\cdot)$, the rightmost term in (31) is a well-known bound of the tail of the normal distribution, and

$$D = \sup_{z \in \mathbb{R}} \left| 2\sqrt{2\pi} z e^{z^2/2} \frac{e^{-z^2}}{z + \sqrt{z^2 + 4/\pi}} \right| < \infty.$$

Collecting (28), (29), (30) and (32) together shows that

$$|J(y) - J(z)| \leq k|y - z| \quad \text{for } z > y > 0 \text{ and } 0 < k < \infty.$$

The other cases are done similarly, so that $J(\cdot)$ is Lipschitz.

References

Ahmad, I. A. and Lin, P. E. (1976), A nonparametric estimation of the entropy for absolutely continuous distributions, *IEEE Trans. Inform. Theory*, **22**, 372–375.

Ahmad, I. A. and Lin, P. E. (1989), A nonparametric estimation of the entropy for absolutely continuous distributions," *IEEE Trans. Inform. Theory*, **36**, 688–692.

Andrieu C. and Thoms J. (2008), A tutorial on adaptive MCMC, *Stat. Comput.*, **18**, 343–373.

Atchadé, Y.F., and Rosenthal, J. (2005), On adaptive Markov chain Monte Carlo algorithms, *Bernoulli*, **11**, 815–828.

Billingsley (1995), *Probability and Measure*, 3rd Edition, Wiley, New York.

Chauveau, D. and Vandekerkhove, P. (2002), Improving convergence of the Hastings-Metropolis algorithm with an adaptive proposal, *Scand. J. Statist.*, **29**, 13–29.

Chauveau, D. and Vandekerkhove, P. (2007), A Monte Carlo estimation of the entropy for Markov chains, *Meth. Comput. Appl. Probab.*, **9**, 133–149.

Dmitriev, Y. G., and Tarasenko, F. P. (1973a), On the estimation of functionals of the probability density and its derivatives, *Theory Probab. Appl.*, **18**, 628–633.

Dmitriev, Y. G., and Tarasenko, F. P. (1973b), On a class of non-parametric estimates of non-linear functionals of density, *Theory Probab. Appl.*, **19**, 390–394.

Douc, R., Guillin, A., Marin, J.M. and Robert, C.P. (2007) Convergence of adaptive mixtures of importance sampling schemes, *Ann. Statist.*, **35**, 420–448.

Dudevicz, E. J. and Van Der Meulen, E. C. (1981), Entropy-based tests of uniformity, *J. Amer. Statist. Assoc.*, **76**, 967–974.

Eggermont, P. P. B. and LaRiccia, V. N. (1999), Best asymptotic Normality of the Kernel Density Entropy Estimator for Smooth Densities, *IEEE trans. Inform. Theory,* **45**, 1321–1326.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996), *Markov Chain Monte Carlo in practice*. Chapman & Hall, London.

Gilks, W.R., Roberts, G.O. and Sahu, S.K. (1998), Adaptive Markov chain Monte carlo through regeneration, *J. Amer. Statist. Assoc.* **93**, 1045–1054.

Györfi, L. and Van Der Meulen, E. C. (1987), Density-free convergence properties of various estimators of the entropy, *Comput. Statist. Data Anal.*, **5**, 425–436.

Györfi, L. and Van Der Meulen, E. C. (1989), An entropy estimate based on a kernel density estimation, *Colloquia Mathematica societatis János Bolyai 57. Limit Theorems in Probability and Statistics Pécs (Hungary)*, 229–240.

Haario, H., Saksman, E and Tamminen, J. (2001), An adaptive Metropolis Algorithm, *Bernoulli* **7**, 223–242.

Hastings, W.K. (1970), Monte Carlo sampling methods using Markov Chains and their applications, *Biometrika*, **57**, 97–109.

Holden, L. (1998), Geometric Convergence of the Metropolis-Hastings Simulation Algorithm, *Statist. Probab. Letters* , **39**, 371–377.

Ivanov, A. V. and Rozhkova, M.N. (1981), Properties of the statistical estimate of the entropy of a random vector with a probability density (in Russian), *Probl. Peredachi Inform*, **17**, 33-43. Translated into English in *Problems Inform. Transmission*, **17**, 171–178.

Jarner, S.F. and Hansen, E. (2000), Geometric ergodicity of Metropolis algorithms. *Stoch. Proc., and Their Appl.*, **85**, 341–361.

Mengersen, K.L. and Tweedie, R.L. (1996), Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, **24**, 101–121.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953), Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

Mokkadem, A. (1989), Estimation of the entropy and information of absolutely continuous random variables, *IEEE Trans. Inform. Theory*, **23**, 95–101.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Roberts, G.O. and Rosenthal, J.S. (2001), Optimal scaling for various Metropolis-Hastings algorithms, *Statistical Science*, **16**, 351–367.

Roberts, G.O. and Tweedie, R.L. (1996), Geometric convergence and Central Limit Theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83**, 95–110.

Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley, New York.

Tarasenko, F. P. (1968), On the evaluation of an unknown probability density function, the direct estimation of the entropy from independent observations of a continuous random variable, and the distribution-free entropy test of goodness-of-fit, *Proc. IEEE.*, **56**, 2052–2053.