



**HAL**  
open science

## Selection of a Metropolis-Hastings simulation strategy via an entropy convergence criterion

Didier Chauveau, Pierre Vandekerkhove

► **To cite this version:**

Didier Chauveau, Pierre Vandekerkhove. Selection of a Metropolis-Hastings simulation strategy via an entropy convergence criterion. 2004. hal-00019174v1

**HAL Id: hal-00019174**

**<https://hal.science/hal-00019174v1>**

Preprint submitted on 17 Feb 2006 (v1), last revised 10 Apr 2007 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Selection of a Metropolis-Hastings simulation strategy via an entropy convergence criterion

Didier CHAUVEAU      Pierre VANDEKERKHOVE

Université de Marne-la-Vallée

Laboratoire d'Analyse et de Mathématiques Appliquées  
CNRS UMR 8050  
5 Bd Descartes, Champs-sur-Marne  
77454 Marne-la-Vallée Cedex 2, France.  
chauveau@math.univ-mlv.fr, vandek@math.univ-mlv.fr

June 13rd, 2004

**Abstract.** Various convergence properties of the Metropolis-Hastings (MH) algorithm have been proved under conditions relative to the instrumental (proposal) distribution. To implement this algorithm, many different choices for the instrumental distribution can be made, associated in practice to unknown rates of convergence. In this paper we propose a simulation-based methodology to compare these rates of convergence, grounded on an entropy criterion computed from parallel (i.i.d.) simulated Markov chains coming from each candidate strategy. Theoretically, we give general conditions under which the successive densities of the MH algorithm satisfy adequate smoothness and tail properties, so that this entropy criterion can be estimated consistently using kernel density estimate and Monte Carlo integration. Examples are provided to illustrate this convergence criterion.

**Keywords.** Entropy, Kullback divergence, Markov Chain Monte Carlo, Metropolis-Hastings algorithm.

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

## 1 Introduction

The Metropolis-Hastings (MH) algorithm (Hastings [22]) is one of the most popular algorithm used in Markov Chain Monte Carlo (MCMC) methods. A 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 situations

where direct simulation from  $f$  is not tractable, or where integrals like  $\mathbb{E}_f[h] = \int h(x)f(x) dx$  are not available in closed form, a MCMC method is appropriate since, for  $T$  large enough,  $x^{(T)}$  is approximately  $f$  distributed, and  $\mathbb{E}_f[h]$  can be approximated by ergodic averages from the chain. Another commonly used MCMC methods is the Gibbs sampler (first introduced by Geman and Geman [15]; see also Gelfand and Smith [13]). An account of definitions and convergence properties of Gibbs and MH algorithms can be found, e.g., in Gilks *et al.* [17].

In this paper, we focus on the MH algorithm. For this algorithm, the “target” pdf  $f$  needs to be known only up to a (normalizing) multiplicative constant. Each step is based on the generation of the proposed next move  $y$  from a general conditional density  $q(y|x)$ , called the *instrumental distribution* or *proposal density* (hence a practical requirement is that simulations from  $q$  should be done easily). For a starting value  $x^{(0)} \sim p^0$ , the  $n$ -th step  $x^{(n)} \rightarrow x^{(n+1)}$  of the algorithm is as follows:

1. **generate**  $y \sim q(\cdot|x^{(n)})$
2. **compute**  $\alpha(x^{(n)}, y) = \min \left\{ 1, \frac{f(y)q(x^{(n)}|y)}{f(x^{(n)})q(y|x^{(n)})} \right\}$
3. **take**  $x^{(n+1)} = \begin{cases} y & \text{with probability } \alpha(x^{(n)}, y), \\ x^{(n)} & \text{with probability } 1 - \alpha(x^{(n)}, y). \end{cases}$

The two dominating MH strategies are (i) the (so-called) *Independence Sampler* (IS), i.e. the MH algorithm 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 usual choice for the latter is a gaussian perturbation with a fixed variance matrix (e.g., in the one-dimensional case,  $q(y|x)$  is the pdf of  $\mathcal{N}(x, \sigma^2)$  where  $\sigma^2$  is the scaling parameter of the perturbation, that has to be tuned).

Ergodicity and convergence properties of the MH algorithm have been intensively studied in recent literature, and conditions have been given for its geometric convergence (see, e.g., Mengersen and Tweedie [25], or Roberts and Tweedie [29]). In particular, Mengersen and Tweedie proved geometric convergence in total variation norm of the IS, under the condition  $q(y) \geq af(y)$  for some  $a > 0$ . The associated geometric rate is  $(1 - a)^n$ , not surprisingly pointing out the link between the convergence rate and the proximity of  $q$  to  $f$ .

To actually implement the MH algorithm, a virtually unlimited number of choices for the instrumental distribution can be made, with the goal of improving mixing and convergence properties of the resulting Markov chain. If one wants to use the IS strategy, selection of a reasonably “good” proposal density can be done using several procedures, among which: numerical analysis or a priori knowledge about the target to approximate the shape of  $f$  (modes, . . .); preliminary MCMC experiment, or adaptive methods to dynamically build a proposal density on the basis of the

chain(s) history (Gelfand and Sahu [14], Gilks *et al.* [16], [18], Chauveau and Vandekerkhove [5], Haario *et al.* [21]). If one wants to use the RWMH strategy, “good” scaling constants must be found, since the mixing depends dramatically on the variance matrix of the perturbation (see, e.g., Roberts and Rosenthal [28]). However, these various choices are associated in general to unknown rates of convergence, because of the complexity of the kernel, and of the associated theoretical computations of bounds.

The motivation of this paper is thus to propose a method to compare the rates of convergence of several candidate MH algorithms (designed for the same target  $f$ ), solely on the basis of the simulated output from each Markov chain. Note that the question of selecting the best MH strategy among a family of proposal densities is the subject of recent developments (see, e.g., Gåsemyr [12] for a heuristical solution using adaptation).

We suggest to use an entropy criterion between the pdf of each algorithm at time  $n$  and the target density  $f$ . The computation of an estimate of this criterion requires the simulation, for a short duration  $n_0$ , of  $N$  parallel (i.i.d.) chains coming from each strategy. This can be seen as a pre-run, to determine the best algorithm before running it for the long duration required by the MCMC methodology.

More precisely, let for instance  $q_1$  and  $q_2$  be two proposal distributions associated to two simulation strategies, generating two MH algorithms with densities denoted by  $p_1^n$  and  $p_2^n$  at time (iteration)  $n$ . For the comparison, both algorithms are started with the same initial distribution, i.e.  $p_1^0 = p_2^0$ . Define the relative entropy of a probability density  $p$  by

$$\mathcal{H}(p) = \int p(x) \log p(x) dx. \quad (1)$$

A natural measure of the algorithm’s quality is the evolution in time ( $n$ ) of the Kullback-Leibler “divergence” between  $p_i^n$ ,  $i = 1, 2$ , and  $f$ , given by

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

The behavior of the application  $n \rightarrow \mathcal{K}(p_i^n, f)$  will be precised in section 2.

When  $f$  is analytically known, an a.s. consistent estimation of  $\mathbb{E}_{p_i^n}[\log f]$  is obtained easily by Monte-Carlo integration using  $N$  i.i.d. realisations from the MH algorithm at time  $n$ . Unfortunately in the MCMC setup,  $f$  is usually the posterior density of a Bayesian model, so that  $f(\cdot) = C\varphi(\cdot)$  where the normalisation constant  $C$  cannot be computed, hence this direct estimation cannot be done. However, if we want to compare two strategies, knowing  $C$  is not needed to estimate the *difference* of the divergences with respect to  $f$ . Define

$$\begin{aligned} D(p_1^n, p_2^n, f) &= \mathcal{K}(p_1^n, f) - \mathcal{K}(p_2^n, f) \\ &= \mathcal{H}(p_1^n) - \mathcal{H}(p_2^n) + \mathbb{E}_{p_2^n}[\log \varphi] - \mathbb{E}_{p_1^n}[\log \varphi]. \end{aligned} \quad (2)$$

The Kullback criterion is the only known divergence insuring this property, hence it motivates our choice for applying it.

We propose to use the  $N$  i.i.d. simulations from each of the two strategies at time  $n$ ,

$$(X_1^{i,n}, \dots, X_N^{i,n}) \text{ i.i.d. } \sim p_i^n, i = 1, 2.$$

These simulations are first used to estimate  $\mathbb{E}_{p_i^n}[\log \varphi]$  via Monte-Carlo integration. Denote these estimates by

$$p_i^n(\overline{\log \varphi})_N = \frac{1}{N} \sum_{j=1}^N \log \varphi(X_j^{i,n}) \xrightarrow{a.s.} \mathbb{E}_{p_i^n}[\log \varphi], \quad (3)$$

where the convergence comes from the strong law of large numbers.

The remaining problem is then the estimation of the entropies  $\mathcal{H}(p_i^n)$ ,  $i = 1, 2$ . One classical approach is to build a nonparametric kernel density estimate of  $p_i^n$ , and to compute the Monte-Carlo integration of this estimate. Techniques based on this approach has been suggested by Ahmad and Lin [1], and studied by several authors under different assumptions (see, e.g., Ahmad and Lin [2], Eggermont and LaRiccia [11], Mokkadem [27]). An interesting point is that in our setup, we can “recycle” the simulations already used to compute the  $p_i^n(\overline{\log \varphi})_N$ ’s. We denote in the sequel our entropy estimate of  $\mathcal{H}(p_i^n)$  by  $\mathcal{H}_N(p_i^n)$ , which will be defined and studied in Section 4. We define accordingly

$$D_N(p_1^n, p_2^n, f) = \mathcal{H}_N(p_1^n) - \mathcal{H}_N(p_2^n) + p_2^n(\overline{\log \varphi})_N - p_1^n(\overline{\log \varphi})_N.$$

Our methodology can be applied in actual situations in the following way: assume we have  $k$  possible MH simulation strategies to sample from  $f$ , related to  $k$  proposal densities  $q_1, \dots, q_k$ , and resulting in  $k$  successive densities  $p_i^n$ ,  $i = 1, \dots, k$ ,  $n \geq 0$ . Let  $p_i^0 = p^0$  be the common initial distribution for the  $k$  algorithms. The determination of the best MH algorithm among the  $k$  candidates can be done using the steps below:

1. select the best strategy  $q_b$  between  $q_1$  and  $q_2$ , on the basis of the sign of (the plot of)  $n \mapsto D_N(p_1^n, p_2^n, f)$ , for  $n = 1, \dots, n_0$ , where  $n_0$  is the simulation duration.  $n_0$  can be chosen small, since the best strategy is in general determined during the very first iterations (see the examples in Section 5, where  $n_0 = 30$  is often sufficient);
2. store the sequence of estimates  $\{\mathcal{H}_N(p_b^n) - p_b^n(\overline{\log \varphi})_N\}_{1 \leq n \leq n_0}$ ;
3. for  $i = 3, \dots, k$ ,
  - (a) select the best strategy between  $q_b$  and  $q_i$ , as in step 1. Notice that the computation of  $D_N(p_b^n, p_i^n, f)$  just requires now that of  $\{\mathcal{H}_N(p_i^n) - p_i^n(\overline{\log \varphi})_N\}_{1 \leq n \leq n_0}$ ;

(b) update  $b$  and the sequence  $\{\mathcal{H}_N(p_b^n) - p_b^n(\overline{\log \varphi})_N\}_{1 \leq n \leq n_0}$ .

Storing at each step the sequence of estimates of the best strategy  $\{\mathcal{H}_N(p_b^n) - p_b^n(\overline{\log \varphi})_N\}_{1 \leq n \leq n_0}$  clearly saves computing time; the total number of simulations required is thus  $N \times k \times n_0$ . Concerning the computer investment, a C program for doing the parallel (i.i.d.) simulations together with entropy estimation for a generic MH algorithm is available (from the first author) as a starting point.

Section 2 outlines some links between ergodicity and convergence to zero of  $\mathcal{K}(p^n, f)$  as  $n \rightarrow \infty$ . In Section 3, we establish assumptions on  $q, f$  and the initial density  $p^0$  to insure that, at each time  $n$ , adequate smoothness conditions hold for the successive densities  $p^n, n \geq 0$ . These conditions are stated for the general (multi-dimensional) case, and detailed precisely in the Appendix for the one-dimensional situation, for which practical examples are given. In Section 4, these conditions are used to define an estimate of  $\mathcal{H}(p^n)$  on the basis of the i.i.d. simulation output. Finally, Section 5 illustrates the behavior of our methodology for synthetic one and two-dimensional examples.

## 2 Kullback divergence to stationarity

In this section we show a property of the evolution in time of the Kullback-Leibler divergence between the distributions  $p^n$  of the MH algorithm and the target distribution  $f$ . It has been proved (see, e.g., Miclo [26]) that for countable discrete Markov chains, the Kullback-Leibler divergence between the measure at time  $n$  and the stationary measure  $\mu$  decreases with time, i.e. that  $\mathcal{K}(mP, \mu) \leq \mathcal{K}(m, \mu)$ , where  $mP$  is the transportation of a measure  $m$  with the Markov kernel  $P$ , defined by  $mP(\cdot) = \int m(dx)P(x, \cdot)$ .

We denote in the sequel the supremum norm of a real-valued function  $\varphi$  by

$$\|\varphi\|_\infty := \sup_{x \in \Omega} |\varphi(x)|. \quad (4)$$

We first recall a result due to Holden [23] 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:

$$\forall y \in \Omega, \quad \left| \frac{p^n(y)}{f(y)} - 1 \right| \leq (1 - a)^n \left\| \frac{p^0}{f} - 1 \right\|_\infty. \quad (5)$$

We use this result to show that the Kullback-Leibler divergence between  $p^n$  and  $f$  decreases geometrically fast in this case:

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

$$\mathcal{K}(p^n, f) \leq \kappa \rho^n (1 + \kappa \rho^n), \quad (6)$$

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

*Proof.* Using equation 5, we have:

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

□

More generally, the question of the convergence in entropy of a Markov process is an active field of current research; see, e.g., Ball *et al.* [3], Del Moral *et al.* [7], and Chauveau and Vandekerkhove [6].

### 3 Smoothness of the MH algorithm densities

For estimating the entropy of the MH algorithm successive densities  $p^n$ ,  $n \geq 0$ , we have to check that appropriate smoothness and tails technical conditions on these successive densities hold.

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:

$$p(x, y) = \begin{cases} q(y|x)\alpha(x, y) & \text{if } x \neq y, \\ 0 & \text{if } x = y, \end{cases} \quad (7)$$

and set the probability of staying at  $x$ ,

$$r(x) = 1 - \int p(x, y) dy.$$

The MH kernel can be written as:

$$P(x, dy) = p(x, y)dy + r(x)\delta_x(dy), \quad (8)$$

where  $\delta_x$  denotes the point mass at  $x$ .

#### 3.1 The Independence Sampler case

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, which will be assumed to be sufficiently smooth. We will assume also that the proposal density  $q$  and the target p.d.f.  $f$  are also sufficiently smooth.

From (8), the successive densities of the IS are given by the recursive formula

$$p^{n+1}(y) = q(y) \int p^n(x) \alpha(x, y) dx + p^n(y) \int q(x) (1 - \alpha(y, x)) dx \quad (9)$$

$$= q(y) I_n(y) + p^n(y) (1 - I(y)), \quad (10)$$

where

$$I_n(y) = \int p^n(x) \alpha(x, y) dx, \quad n \geq 0 \quad (11)$$

$$I(y) = \int q(x) \alpha(y, x) dx. \quad (12)$$

For convenience, we introduce the notations

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

We consider the first iteration of the algorithm. From (10), the regularity properties of the density  $p^1$  are related to the regularity properties of the two parameter-dependent integrals  $I_1$  and  $I$ . Regularity properties of such integrals are classically handled by the theorem of continuity under the integral sign (see, e.g., Billingsley [4] Theorem 16.8 p. 212). Continuity is straightforward here:

**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  $\Omega$  for  $n \geq 1$ .*

*Proof.* It suffices to prove continuity for  $p^1(y)$  at any  $y_0 \in \Omega$ . The integrand of  $I_0$ ,  $p^0(x) \alpha(x, y)$ , is continuous in  $y$  at  $y_0$  for any  $x \in \Omega$ , and

$$|p^0(x) \alpha(x, y)| \leq p^0(x), \quad \forall x, y \in \Omega.$$

Then  $I_0$  is continuous at  $y$  by the Lebesgue's dominated convergence theorem (since  $\Omega$  is a metric space, so that continuity can be stated in term of limit of sequence). The same reasoning applies to  $I(y)$  by using  $q$  for the dominating function.  $\square$

In our setup, it appears tractable to apply results on entropy estimation based on a *Lipschitz condition*. Remember that a function  $\varphi : \Omega \rightarrow \mathbb{R}$  is called  $c$ -Lipschitz if there exists a constant  $c > 0$  such that, for any  $x, y \in \Omega$ ,  $|\varphi(y) - \varphi(x)| \leq c \|y - x\|$ .

From equation (10), we have directly that

$$\begin{aligned} |p^{n+1}(y) - p^{n+1}(z)| &\leq \|q\|_\infty |I_n(y) - I_n(z)| + \|I_n\|_\infty |q(y) - q(z)| \\ &+ \|1 - I\|_\infty |p^n(y) - p^n(z)| \\ &+ \|p^n\|_\infty |I(y) - I(z)|, \end{aligned} \quad (13)$$

so that, to prove recursively that  $p^{n+1}$  is Lipschitz, we have first to prove that  $I_n$  and  $I$  are both Lipschitz.



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

- (i)  $\int p^n h < \infty$ ;
- (ii)  $I_n$  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 (9) (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. Indeed, one has to choose a proposal density  $q$  with adequate tails for the MH to be efficient, i.e. to converge quickly. As recalled in the introduction, it has been proved that the IS is uniformly geometrically ergodic if  $q(y) \geq af(y)$  for some  $a > 0$  (Mengersen and Tweedie [25]). 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 (e.g., mixtures of multidimensional Student distributions with small degrees of freedom parameters), so that  $f/q$  is typically a continuous and positive function which goes to zero when  $\|x\| \rightarrow \infty$ . It can then reasonably 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  $\Omega$ , and  $q$  has heavier tails than  $f$ , so that  $\lim_{\|y\| \rightarrow \infty} h(y) = +\infty$ .*

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

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

and since we have made the “good” choice for the proposal density (assumption A),  $h = q/f$  is obviously *not* Lipschitz.

A direct study of  $\alpha(\cdot, x) = [h(\cdot)/h(x)] \wedge 1$ , as it appears in  $I(y)$  (equations (10) and (12)) is needed here. Consider a fixed  $x \in \Omega$  in the sequel. Clearly, 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

$$\forall y \notin K(x), \quad \alpha(y, x) = 1.$$

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*

$$\forall y, z \in K(x), \quad |\alpha(y, x) - \alpha(z, x)| \leq c(x)\|y - z\|, \quad (14)$$

where  $c(x)$  satisfies

$$\int q(x)c(x) dx < \infty, \quad (15)$$

then  $I$  satisfies the Lipschitz condition:

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

Examples where lemma 3 holds will be given in the Appendix, for the one-dimensional situation.

**Proposition 2** *If 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  $n \geq 0$ , there exists  $k(n) < \infty$  such that

$$\forall (y, z) \in \Omega^2, \quad |p^n(y) - p^n(z)| \leq k(n)\|y - z\|. \quad (16)$$

*Proof.* Using equation (13), 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, (16) is satisfied, with

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

□

### 3.2 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

$$\begin{aligned} p^{n+1}(y) &= \int p^n(x)q(y|x)\alpha(x, y) dx + p^n(y) \int q(x|y)(1 - \alpha(y, x)) dx \\ &= J_n(y) + p^n(y) (1 - J(y)), \end{aligned} \tag{17}$$

where

$$J_n(y) = \int p^n(x)q(y|x)\alpha(x, y) dx, \tag{18}$$

$$J(y) = \int q(x|y)\alpha(y, x) dx. \tag{19}$$

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

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

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

then  $p^n$  is continuous on  $\Omega$  for  $n \geq 1$ .

*Proof.* As for Lemma 1, it is enough to check the dominating conditions of, e.g., Billingsley [4], p.212. However, for  $J$ , we need the local condition (ii) to prove the continuity of  $J(y)$  at any  $y_0 \in \Omega$ .  $\square$

Note that the additional condition (ii) is reasonable. For instance, we refer to the most-used case of the RWMH with gaussian perturbation of scale parameter  $\sigma^2 > 0$ . In the one-dimensional case,  $q(x|y)$  is the pdf of  $\mathcal{N}(y, \sigma^2)$  evaluated at  $x$ , and one can simply take for condition (ii)

$$\begin{aligned} \varphi_{y_0, \delta}(x) = & q(x|y_0 - \delta)\mathbb{I}_{x < y_0 - \delta} + q(y_0 - \delta|y_0 - \delta)\mathbb{I}_{[y_0 - \delta, y_0 + \delta]}(x) \\ & + q(x|y_0 + \delta)\mathbb{I}_{x > y_0 + \delta}, \end{aligned} \quad (20)$$

(i.e. the tail of the leftmost gaussian pdf on the left side, the tail of the rightmost gaussian pdf on the right side, and the value of the gaussian at the mode inside  $[y_0 - \delta, y_0 + \delta]$ ).

To prove that the successive densities  $p^n$  of the general MH algorithm are Lipschitz, we proceed using conditions at a higher level than for the IS case, because the successive densities are more complicated to handle

**Proposition 3** *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

$$\forall (y, z) \in \Omega^2, \quad |p^n(y) - p^n(z)| \leq \ell(n) \|y - z\|. \quad (21)$$

*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

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

equation (21) 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 2.  $\square$

Proposition 3 may look artificial since the conditions are clearly “what is needed” to insure the Lipschitz property for  $p^n$ . However, we show in the Appendix (section 7.2) that these conditions are reasonable, in the sense that they are satisfied, e.g., in the one-dimensional case for usual RWMH algorithms with gaussian proposal densities.

## 4 Relative entropy estimation

Let  $\mathbf{X}_N = (X_1, \dots, X_N)$  be an i.i.d.  $N$ -sample of random vectors taking values in  $\mathbb{R}^s$ ,  $s \geq 1$ , with common probability density function  $p$ . Suppose we want to estimate the relative entropy of  $p$ ,  $\mathcal{H}(p)$  given by (1), assuming that it is well defined and finite. Various estimators for  $\mathcal{H}(p)$  based on  $\mathbf{X}_N$  have been proposed and studied in the literature, mostly for the case  $s = 1$ . One method to estimate  $\mathcal{H}(p)$  consists in obtaining a suitable density estimate  $\hat{p}_N$  for  $p$ , and then substituting  $p$  by  $\hat{p}_N$  in an entropy-like functional of  $p$ . This approach has been adopted by Dmitriev and Tarasenko [8][9], Ahmad and Lin [1][2], Györfi and Van Der Meulen [19][20], and Mokkadem [27] who prove strong consistency of their estimators in various frameworks. More recently Eggermont and LaRiccia [11] prove, that they get the 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 [30], and Dudewicz and Van Der Meulen [10].

In our case, and due to the rather poor smoothness properties that can be proved for the densities  $p^n$  we have to consider, we use the entropy estimate proposed by Györfi and Van Der Meulen [20], but with smoothness conditions of Ivanov and Rozhkova [24]: a Lipschitz condition which appeared tractable in our setup, as shown in Section 3.

Following Györfi and Van Der Meulen [20], we decompose the sample  $\mathbf{X}_N$  into two subsamples  $\mathbf{Y}_N = \{Y_i\}$  and  $\mathbf{Z}_N = \{Z_i\}$ , defined by

$$Y_i = X_{2i} \quad \text{for } i = 1, \dots, [N/2] \quad (22)$$

$$Z_i = X_{2i-1} \quad \text{for } i = 1, \dots, [(N+1)/2], \quad (23)$$

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

$$\hat{p}_N(x) = \frac{1}{h_N^s (N+1)/2} \sum_{i=1}^{[(N+1)/2]} K_{h_N} \left( \frac{x - Z_i}{h_N} \right), \quad x \in \mathbb{R}^s, \quad (24)$$

where the kernel  $K$  is a density and  $h_N > 0$  with  $\lim_{N \rightarrow \infty} h_N = 0$ , and  $\lim_{N \rightarrow \infty} N h_N^s = \infty$ . The entropy estimate  $\mathcal{H}_N(p) = \mathcal{H}_{N, \mathbf{Y}, \mathbf{Z}}(p)$  introduced by Györfi and Van Der Meulen [20], is then defined by:

$$\mathcal{H}_N(p) = \frac{1}{[N/2]} \sum_{i=1}^{[N/2]} \log \hat{p}_N(Y_i) \mathbb{I}_{\{p_N(Y_i) \geq a_N\}} \quad (25)$$

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

**Theorem 1** *Assume that  $\mathcal{H}(f) < \infty$ . For all  $n \geq 0$ , let  $\mathbf{X}_N^n$  be a  $N$ -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 (24), based on the subsample  $\mathbf{Z}_N^n$  defined in (23). 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 (25) with*

$$a_N = (\log N)^{-1}. \quad (26)$$

Assume that there are positive constants  $C, r_0, a, A$  and  $\epsilon$ , such that either:

- (i) *in the case of the Independance Sampler:  $f, q$  and  $p_0$  satisfy conditions of Proposition 2;  $q$  satisfies the minoration condition  $q(y) \geq a f(y)$ , and  $f$  satisfies the tail condition*

$$f(y) \leq \frac{C}{\|y\|^s (\log \|y\|)^{2+\epsilon}}, \quad \text{for } \|y\| > r_0; \quad (27)$$

- (ii) *in the general MH case:  $f, q$  and  $p_0$  satisfy conditions of Proposition 3;  $q$  is symmetric ( $q(x|y) = q(y|x)$ );  $\|p^0/f\|_\infty \leq A$ , and  $f$  satisfies the tail condition*

$$f(y) \leq \frac{C}{1 + \|y\|^{s+\epsilon}}. \quad (28)$$

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

*Proof.* This result uses directly Györfi and Van Der Meulen's Theorem in [20] p. 231. Conditions (27) or (28) 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 (1) and of the positivity of  $\mathcal{K}$ . For the tail condition (27), case (i), it suffices to notice that from (5) 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 (17) 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

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

which is stricter than Györfi and Van Der Meulen's tail condition. As to smoothness, the conditions of our Proposition 2 for case (i), and Proposition 3 for case (ii) give the Lipschitz condition of Ivanov and Rozhkova [24] 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 [20].  $\square$

## 5 Examples

We give in this section several examples for synthetic models, with target densities which are one and two-dimensional mixtures of gaussian distributions. The advantage of taking a mixture is that it is an easy way to build multimodal target densities with "almost disconnected" modes, i.e. separated modes with regions of low probability in between (see figures 1 and 6).

The difficulty for classical RWMH algorithm is then to properly calibrate the variance of the random walk to propose jumps under all the modes in a reasonable amount of time. The difficulty for the IS is to use a good proposal density  $q$ , hopefully allowing sufficient mass over the modal regions.

It is important to understand that in all the examples below, the target density is completely known so that, instead of estimating the difference  $\mathcal{K}(p_1^n, f) - \mathcal{K}(p_2^n, f)$  between any two given strategies, we are able to estimate directly  $\mathcal{K}(p_i^n, f)$  for each strategy  $i$  leading to the successive densities  $p_i^n$  separately. Actually, we can compute the strongly consistent estimate

$$\mathcal{K}_N(p_i^n, f) = \mathcal{H}_N(p_i^n) - \frac{1}{N} \sum_{j=1}^N \log f(X_j^{i,n}),$$

where the  $X_j^{i,n}$ 's,  $j = 1, \dots, N$  are i.i.d.  $\sim p_i^n$ .

We give the results in terms of these estimates, since they provide easier comparisons and illustrate more clearly the behaviour of our method. However, one should keep in mind that in real-size situations, only the plots of the differences are accessible to computation. This is not a flaw in the method since clearly, the better algorithm can be deduced from these plots. For complete illustration, however, we have also provided for the first example the plots of  $n \mapsto D(p_1^n, p_2^n, f)$  for comparing three strategies. The only information *not* provided by the plot of the difference is the “convergence time” of each chain (in the sense of the convergence assessment of MCMC, see, e.g., Gilks *et al.* [17]). Indeed, even if the difference goes about zero at time  $n$ , there is always a possibility that both MH algorithms fail to converge at that time, with  $\mathcal{K}(p_1^n, f) \approx \mathcal{K}(p_2^n, f)$ .

Since the models are quite simple here, we could run a large number of i.i.d. markov chains to obtain precise estimates. So we tried up to  $N = 1000$  chains for the one-dimensional model, and up to  $N = 200$  chains for the two-dimensional model. This number of parallel chains can be reduced without compromising the decision for real-size applications (we tried  $N = 100$  chains with satisfactory results). Note also that the computing time needed, even with large  $N$ , is not long since the duration  $n_0$  of the parallel simulation is itself short: the best algorithm is quickly detected, as shown in the figures below. Finally, for computing the estimate of the entropy (25), we use a threshold  $a_N = \mathcal{O}(\log(N)^{-1})$  instead of (26), to avoid rejection of too many observations for small values of  $N$ .

## 5.1 A one-dimensional example

To illustrate the relevance of our comparison of the convergence rates, we first choose a very simple but meaningful situation, consisting in MH algorithms for simulation from a mixture of 3 gaussian distributions, with density

$$f(x) = \sum_{i=1}^3 \alpha_i \varphi(x; \mu_i, \sigma_i^2), \quad (29)$$

where  $\varphi(\cdot; \mu, \sigma^2)$  is the pdf of  $\mathcal{N}(\mu, \sigma^2)$ . The chosen parameters  $\alpha_1 = 0.5$ ,  $\alpha_2 = 0.3$ ,  $\mu_1 = 0$ ,  $\mu_2 = 9$ ,  $\mu_3 = -6$ ,  $\sigma_1^2 = 2$ , and  $\sigma_2^2 = \sigma_3^2 = 1$ , result in the trimodal pdf depicted in figure 1.

**Independence Sampler** We first ran the independence sampler with a gaussian proposal density  $q = \mathcal{N}(0, \sigma^2)$ , for several settings of the variance parameter. None of these MH are optimal, since  $q$  does not have modes at  $-6$  and  $9$ , whereas  $f$  has. If the variance is too small (e.g.,  $\sigma \leq 1$ ), the algorithm can “almost never” propose jumps outside, say,  $[-4; +4]$ , so that it can (almost) never visit the right or left modes. The algorithm requires then a dramatically long time to converge. Our Kullback divergence estimate reflect this fact (figure 2, left). For more adapted settings like, e.g.,  $\sigma = 3$ , the algorithm converges faster, and the convergence again deteriorates



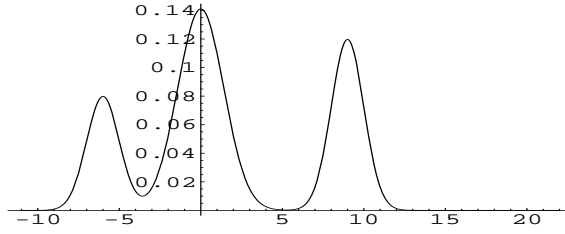
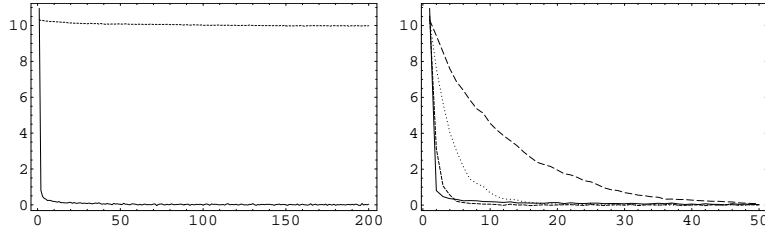


Figure 1: True mixture density (29).


 Figure 2: Plots of  $n \mapsto \mathcal{K}_N(p^n, f)$  for the IS with a gaussian proposal  $\mathcal{N}(0, \sigma^2)$ . Left:  $\sigma = 3$  (solid) vs.  $\sigma = 1$  (dashed); Right:  $\sigma = 3$  (solid) vs.  $\sigma = 10$  (dashed),  $\sigma = 30$  (dotted),  $\sigma = 100$  (long dashed).

as  $\sigma$  increases above, say, 10, since the proposal density is then overdispersed (see figure 2, right).

For this example, we also provide in figure 3 two examples of the plots available in actual situations, i.e. that of  $n \mapsto D(p_1^n, p_2^n, f)$ . The sign of the plots in both cases, and for the first iterations, clearly indicate that, in both cases, the first strategy ( $q_1 = \mathcal{N}(0, 3^2)$ ) is preferable. Moreover, the comparison of the two plots indicate that  $\sigma = 100$  is even worse than  $\sigma = 30$ .

To check our estimates with heavy-tailed proposal densities, we also ran the independence sampler with a Student proposal density  $t(d)$ , for  $d = 1$  (the Cauchy distribution), up to  $d = 100$  (for which the Student is almost the normal distribution). As expected, the algorithms converge faster when they use Student distributions with heavier tails, since in that case they can still propose jumps in the left or right mode. When  $d \rightarrow \infty$ , the proposal converges to the  $\mathcal{N}(0, 1)$ , and the IS shows the same behavior as the previous one, with  $\sigma = 1$  (compare figure 2, left with figure 4, right).

**RWMH** We also ran on the same example a random walk MH algorithm with a gaussian proposal  $q(\cdot|x) \equiv \mathcal{N}(x, \sigma^2)$ , and several settings for  $\sigma^2$ . As expected, a good choice is about  $\sigma = 10$ . For too small settings (e.g.,  $\sigma = 0.1$ ), the jumps of the random walk (of order  $\pm 3\sigma$ ) are too small, so that the chain needs a dramatically long time to reach the rightmost mode. This is clearly indicated by our estimate in figure 5, right, where up to  $n = 600$  iterations are needed for this unefficient

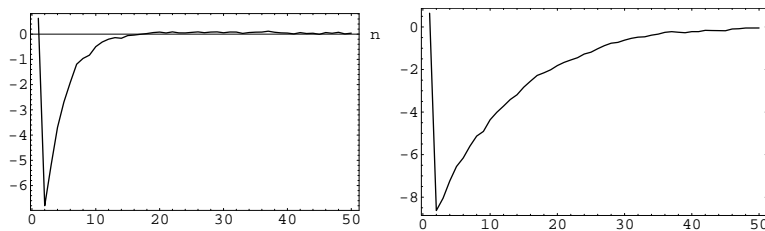


Figure 3: Plots of the difference  $n \mapsto D(p_1^n, p_2^n, f)$  for the IS with a gaussian proposal  $q_i = \mathcal{N}(0, \sigma_i^2)$ . Left:  $\sigma_1 = 3$  vs.  $\sigma_2 = 30$ ; Right:  $\sigma_1 = 3$  vs.  $\sigma_2 = 100$ .

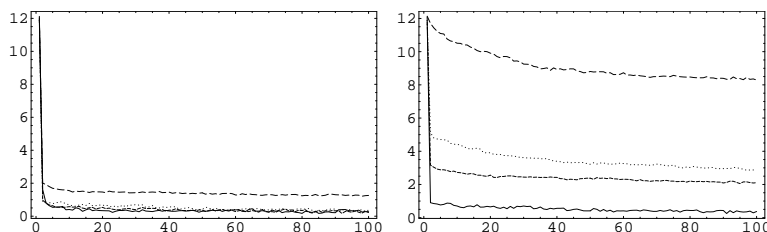


Figure 4: Plots of  $n \mapsto \mathcal{K}_N(p^n, f)$  for the IS with a Student proposal  $t(d)$ : Left:  $d = 1$  (solid),  $d = 2$  (short dashed),  $d = 3$  (dotted),  $d = 10$  (long dashed). Right:  $d = 3$  (solid),  $d = 20$  (short dashed),  $d = 50$  (dotted),  $d = 100$  (long dashed).

algorithm to converge.

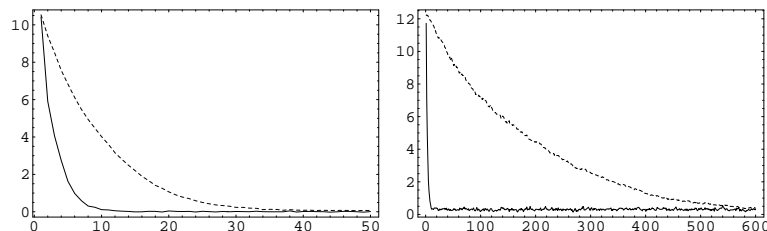


Figure 5: Plots of  $n \mapsto \mathcal{K}_N(p^n, f)$  for the RWMH with gaussian proposal  $q(\cdot|x) \equiv \mathcal{N}(x, \sigma^2)$ : left:  $\sigma = 1$  (dashed) vs.  $\sigma = 10$  (solid); right:  $\sigma = 0.1$  (dashed) vs.  $\sigma = 10$  (solid).

## 5.2 A two-dimensional example

We also tried for the target density a two-dimensional gaussian mixture, depicted in figure 6, left (the true parameters are not given for brevity). For this example, we compare three “good” strategies of different types: (i) An IS with a uniform proposal density over the compact  $[-20; 20]^2$ ; this algorithm is “almost geometric” since the mass of the tails of  $f$  outside the compact are negligible (the minoration condition  $q \geq af$  is fulfilled on the compact). (ii) A RWMH with a bivariate gaussian proposal  $q(\cdot|x) = \mathcal{N}(x, \sigma^2 I)$ , with a “good” setting  $\sigma = 17$ , founded using our

Kullback divergence method. (iii) An adaptive MH algorithm following the ideas in Chauveau and Vandekerckhove [5]. In short, parallel chains started with the IS (i) are ran and kernel density estimates are built at some specified times using the past of these i.i.d. chains. For example, the proposal density built at time  $n = 48$  is depicted in figure 6, right.

The estimates of  $\mathcal{K}(p^n, f)$  for this setup (and  $N = 200$  chains) are given in figure 7. As expected, the IS with the uniform proposal density performs better than the calibrated RWMH. But the adaptive proposal is even better than the two others. The times at which the adaptive proposal density is updated ( $n = 5$  and  $9$ ), are even visible in the plot of  $\mathcal{K}(p^n, f)$  for this strategy, which means that it has an immediate effect on this divergence.

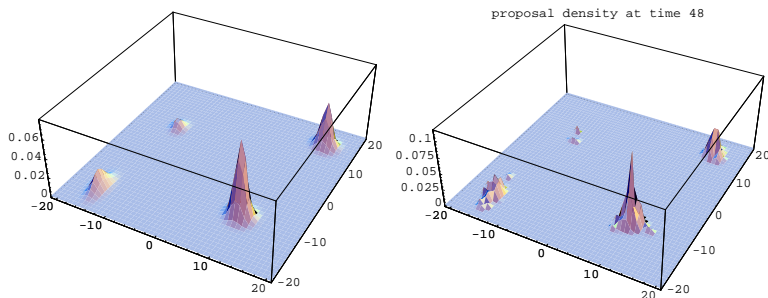


Figure 6: left: true target pdf; right: adaptive proposal density.

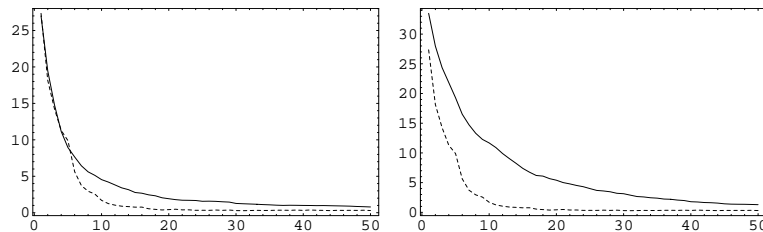


Figure 7: Plots of  $n \mapsto \mathcal{K}_N(p^n, f)$  based on  $N = 200$  chains for the IS with adaptive proposal (dashed) vs. left: IS with  $q = \text{Uniform distribution}$ ; right: RWMH with  $q(\cdot|x) = \mathcal{N}(x, \sigma^2 I)$ ,  $\sigma = 17$ .

## 6 Conclusion

We have proposed a methodology to precisely quantify and compare several MH simulation strategies only on the basis of the simulations output. A procedure for applying our method in practice has been given in Section 1.

A novelty is that this methodology is based upon the use of the relative entropy of the successive densities of the MH algorithm. Hence a consistent estimate of this entropy in the MH setup has been proposed, and general conditions insuring its convergence have been detailed. Indeed, the conditions of propositions 2 and 3 are difficult to verify in practice. However, the theoretical study of Section 3 has been developed to support the idea that, if the “ingredients” of the MH algorithm ( $q$ ,  $p^0$  and  $f$ ) have sufficient tails and smoothness conditions, then one can reasonably expect that the successive densities  $p^n$ ,  $n \geq 1$ , of the MH algorithm will also satisfy these conditions, so that our usage of the estimates of the  $\mathcal{H}(p_i^n)$ 's will indicate the most efficient MH algorithm to use.

Our methodology is an alternative to the adaptive MH strategies, which represent an active field of the current literature on MCMC. The advantage of our approach is that it avoids the difficulties associated to adaptation, namely the preservation of the convergence to the desired limiting distribution, and the theoretical guarantee that the adaptive algorithm will perform better than any other heuristical approach. Finally, if several adaptive strategies are available, our method can also be used to find the best method (see example 5.2).

## 7 Appendix

The purpose of this appendix is to show that some of the conditions required in Proposition 2 and Proposition 3, which look difficult to check in actual situations, are satisfied at least in simple situations, for classically used MH algorithms in the one-dimensional case i.e. when  $x \in \mathbb{R}$ .

### 7.1 The one-dimensional independence sampler case

In the IS case, the difficult conditions are conditions (14) and (15) of Lemma 3. These conditions are simpler to handle in the one-dimensional case. First, note that we can prove under additional conditions the derivability of  $p^n$  for all  $n \geq 1$  (that proof is not given since we are not using it here). In the one-dimensional case, when  $q$  and  $f$  are in addition derivable, and have non-oscillating tails, assumption (A) leads to

$$\exists m_1 < m_2 : \forall x < m_1, h'(x) < 0, \text{ and } \forall x > m_2, h'(x) > 0. \quad (30)$$

For a fixed  $x \in \mathbb{R}$ , there exists by (30) 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

$$c(x) = \sup_{y \in K(x)} \left| \frac{\partial \phi(y, x)}{\partial y} \right| < \infty. \quad (31)$$

and Lemma 3 holds if the integrability condition (15) is satisfied. Note that  $|a(x)|$  and  $b(x)$  both go to  $+\infty$  as  $|x| \rightarrow \infty$ ; in particular,  $b(x) = x$  for  $x > m_2$ . Hence  $c(x) \rightarrow \infty$  as  $|x| \rightarrow \infty$ , and condition (15) 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 (15) 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 (15) 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 (15) 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.

## 7.2 The one-dimensional general MH case

In the general MH case, the difficult conditions are conditions (ii) and (iii) of Proposition 3. 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 the target distribution  $\mathcal{N}(0, 1)$ .

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  $x$ ,  $\lim_{|y| \rightarrow \infty} \alpha(x, y) = 0$ . In other words, both  $\alpha(x, y)$  and  $q(y|x)\alpha(x, y)$  have tails behavior for large  $y$ . The non-truncated 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 3 holds.

We turn now to condition (iii) of Proposition 3, i.e. we have to show that  $J(y)$  given by (19) 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,

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

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

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

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

$$\begin{aligned} \int_{K(y)} |q(x|y) - q(x|z)| dx &\leq \int |q(x|y) - q(x|z)| dx \\ &\leq |y - z| \int \frac{|x - y^*|}{2\pi} \exp(-(x - y^*)^2/2) dx \\ &\leq \sqrt{\frac{2}{\pi}} |y - z|, \end{aligned} \tag{32}$$

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

$$\int_y^z \left| q(x|y) \frac{f(x)}{f(y)} - q(x|z) \right| dx \leq \delta |y - z|. \quad (33)$$

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

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

where, as in Proposition 3,  $Q = \|q\|_\infty$ , and since  $\sup_{x \in K(z)^c} |f(x)/f(z)| = 1$ . Using the mean value theorem as for the first term,

$$\int_{K(z)^c} |q(x|y) - q(x|z)| dx \leq \sqrt{\frac{2}{\pi}} |y - z|. \quad (34)$$

Finally,

$$\begin{aligned} \int_{K(z)^c} \left| \frac{f(x)}{f(y)} - \frac{f(x)}{f(z)} \right| dx &= 2 \int_z^\infty \left| \frac{f(x)}{f(y)} - \frac{f(x)}{f(z)} \right| dx \\ &\leq 2 \left| \frac{1}{f(y)} - \frac{1}{f(z)} \right| \int_z^\infty f(x) dx \\ &\leq 2\sqrt{2\pi} z e^{z^2/2} \frac{e^{-z^2}}{z + \sqrt{z^2 + 4/\pi}} |y - z|, \quad (35) \end{aligned}$$

$$\leq D |y - z|, \quad (36)$$

where the left term in (35) comes from the mean value theorem applied to the function  $1/f(\cdot)$ , the rightmost term in (35) 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 (32), (33), (34) and (36) 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

- [1] Ahmad, I. A. and Lin, P. E. (1976), A nonparametric estimation of the entropy for absolutely continuous distributions, *IEEE Trans. Inform. Theory*, vol. 22, 372–375..
- [2] Ahmad, I. A. and Lin, P. E. (1989), A nonparametric estimation of the entropy for absolutely continuous distributions,” *IEEE Trans. Inform. Theory*, vol. 36, 688–692.
- [3] Ball, K., Barthe, F. and Naor, A., (2003), Entropy jumps in the presence of a spectral gap, *Duke Mathematical Journal*, **119**, 1, 41–63.
- [4] Billingsley (1995), *Probability and Measure*, 3rd Edition, Wiley, New York.
- [5] Chauveau, D. and Vandekerkhove, P. (2002), Improving convergence of the Hastings-Metropolis algorithm with an adaptive proposal, *Scandinavian Journal of Statistics*, **29**, 1, 13–29.
- [6] Chauveau, D. and Vandekerkhove, P. (2004), A Monte Carlo estimation of the entropy for Markov chains, *preprint*.
- [7] Del Moral P., Ledoux M., Miclo, L., (2003), Contraction properties of Markov kernels. *Probab. Theory and Related Fields*, **126**, pp. 395–420.
- [8] Dmitriev, Y. G., and Tarasenko, F. P. (1973), On the estimation of functionals of the probability density and its derivatives, *Theory Probab, Appl.* **18**, 628–633.
- [9] Dmitriev, Y. G., and Tarasenko, F. P. (1973), On a class of non-parametric estimates of non-linear functionals of density, *Theory Probab, Appl.* **19**, 390–394.
- [10] Dudevicz, E. J. and Van Der Meulen, E. C. (1981), Entropy-based tests of uniformity, *J. Amer. Statist. Assoc.*, **76** 967–974.
- [11] 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*, vol. 45, no. 4, 1321–1326.
- [12] Gåsemyr, J. (2003), On an adaptive version of the Metropolis-Hastings algorithm with independent proposal distribution, *Scand. J. Statist.*, **30**, no. 1, 159–173.
- [13] Gelfand, A.E. and Smith, A.F.M. (1990), Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.
- [14] Gelfand, A.E. and Sahu, S.K. (1994), On Markov chain Monte Carlo acceleration, *Journal of Computational and Graphical Statistics* **3**, 261–276.



- [15] Geman, S. and Geman, D. (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741.
- [16] Gilks, W.R., Roberts, G.O. and George, E.I. (1994), Adaptive direction sampling, *The statistician*, **43**, 179–189.
- [17] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996), *Markov Chain Monte Carlo in practice*. Chapman & Hall, London.
- [18] Gilks, W.R., Roberts, G.O. and Sahu, S.K. (1998), Adaptive Markov chain Monte carlo through regeneration, *Journal of the American Statistical Association* **93**, 1045–1054.
- [19] 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.
- [20] 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.
- [21] Haario, H., Saksman, E and Tamminen, J. (2001), An adaptive Metropolis Algorithm, *Bernoulli* **7**, 2, 223–242.
- [22] Hastings, W.K. (1970), Monte Carlo sampling methods using Markov Chains and their applications, *Biometrika* **57**, 97–109.
- [23] Holden, L. (1998), Geometric Convergence of the Metropolis-Hastings Simulation Algorithm, *Statistics and Probability Letters*, **39**, 1998.
- [24] 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.
- [25] Mengersen, K.L. and Tweedie, R.L. (1996), Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24**, 101–121.
- [26] Miclo, L. (1997) Remarques sur l’hypercontractivité et l’évolution de l’entropie des chaînes de Markov finies. *Séminaire de Probabilités XXXI, Lecture Notes in Mathematics*, Springer, 136–168.
- [27] Mokkaedem, A. (1989), Estimation of the entropy and information of absolutely continuous random variables, *IEEE Trans. Inform. Theory* **23** 95–101.
- [28] Roberts, G.O. and Rosenthal, J.S. (2001), Optimal scaling for various Metropolis-Hastings algorithms, *Statistical Science*, **16**, 351–367.

- [29] Roberts, G.O. and Tweedie, R.L. (1996), Geometric convergence and Central Limit Theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83**, 95–110.
- [30] 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.