British Journal of Mathematics & Computer Science
4(24): 3387-3402, 2014
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Model Selection of Stochastic Simulation Algorithm Based on Generalized Divergence Measures

Papa Ngom[*1] **and B. Don Bosco Diatta**[1]

[1] *LMA-Laboratoire de Mathématiques Appliquées, Université Cheikh Anta Diop BP 5005 Dakar-Fann Sénégal, Sénégal.*

Article Information
DOI: 10.9734/BJMCS/2014/12020
Editor(s):
(1) Xiaodi Li, School of Mathematical Sciences, Shandong Normal University Ji'nan, 250014, Shandong, P. R. China.
(2) Paul Bracken, Department of Mathematics, The University of Texas-Pan American Edinburg, TX 78539, USA.
Reviewers:
(1) Anonymous, G. Pulla Reddy Engineering College, India.
(2) Anonymous, Harbin Engineering University, China.
(3) Anonymous, Institute of Science and Technology Information, Republic of Korea.
(4) Anonymous, Liaoning Medical University, Jinzhou 121000, China.
Complete Peer review History:
http://www.sciencedomain.org/review-history.php?iid=699id=6aid=6302

Original Research Article

Received: 13 June 2014
Accepted: 08 September 2014
Published: 01 October 2014

Abstract

We consider the generalized divergence measure approach to compare different simulation strategies such as the Independent Sampler (IS), the Random Walk of Metropolis Hastings (RWMH), the Gibbs Sampler(GS), the Adaptive Metropolis (AM), and Metropolis within Gibbs (MWG). From a selected set of simulation algorithm candidates, the statistical analysis allows us to choose the best strategy in the sense of rate of convergence. We use the informational criteria such as the Rényi divergence measure $R_\alpha(p,q)$, the Tsallis divergence $T_\alpha(p,q)$, and the α-divergence $D_\alpha(p,q)$, where p and q are probability density functions, to show in some examples of synthetic models with target distributions in one dimensional, and two dimensional cases, the consistency and applicability of these α-divergence measures for stochastic simulation selection.

Keywords: MCMC methods, Metropolis-Hastings algorithm, Gibbs Sampler, Adaptive Metropolis,

Corresponding author: E-mail: papa.ngom@ucad.edu.sn

Metropolis Within Gibbs, simulation strategy, target density, proposal density, α-divergence measure.

2010 Mathematics Subject Classification: 60J60, 62F03, 62F05, 94A17

1 Introduction

The stochastic simulation methods (MCMC, recent hybrid, and adaptive methods) have many simulation algorithms with different convergence speeds. These convergence rates are unknown in practice. Recall that the MCMC methods are used to simulate a probability distribution Π having a density function f. These methods, consist to generate a Markov chain whose Π is the invariant measure. For generating a Markov chain one needs a proposal distribution Φ having a density function q. Consider, for example, the RWMH strategy of the Metropolis - Hastings sampler for seeing how are used the densities f, q and the probability distribution Φ in this method. As any iterative algorithm, this algorithm (RWMH) has an initial distribution density function denoted p^0, which generates X_0 the first element of the Markov chain. This Markov chain $(X_n)_{n\geq 0}$ is then then built by iteration. Suppose that for $n \geq 0$, $X_n = x_n$ is already simulated, then we want to get X_{n+1}. We generate random independent variables Y_n and U_n such as

- $Y_n \sim \Phi(x_n, .)$, distribution having as parameter the current state $X_n = x_n$. If $\Phi(x_n, .)$ is the normal distribution we will have $\Phi(x_n, .) = \mathcal{N}(x_n, \sigma^2)$ ie x_n is the mean of normal distribution.
- $U_n \sim \mathcal{U}([0, 1])$ ie U_n has the uniform law on $[0, 1]$.
- Note by $\alpha(x, y) = \min(1, \frac{f(y)q(y,x)}{f(x)q(x,y)})$ with the convention $\alpha(x, y) = 1$ if $f(x)q(x, y) = 0$.
- If $U_n \leq \alpha(X_n, Y_n)$ then, $X_{n+1} = Y_n$ i.e. we accept the transition.
- If $U_n > \alpha(X_n, Y_n)$ then, $X_{n+1} = X_n$ i.e. we reject the transition.

The study concerning the comparison of stochastic simulation strategies is introduced by Chauveau[1], Chauveau and Vandekerkhove [2]. These authors have used the Kullback-Leibler divergence measure to compare simulation strategies. The study made by these authors is based on two strategies of Metropolis - Hastings sampler the Independent Sampler (IS) and the Random Walk of Metropolis Hastings (RWMH). Then the convergence Theorem of the Kullback divergence estimator required the Lipschitz condition for densities p^n of X_n , $n = 1, \ldots$ generated by the IS, or RWMH strategy. To obtain this Lipschitz property on densities p^n , $n = 1, \ldots$, many assumptions, not always verifiable in practice, were made on the densities f, q, and p^0.

In order to do a unified study of their statistical properties, the contribution of this paper is to propose some generalized divergences, called α-divergence measures [3], which include as particular case the above mentioned divergence measure. These divergence measures are the Rényi divergence $R_\alpha(p, q)$, the Tsallis divergence $T_\alpha(p, q)$ and the α-divergence $D_\alpha(p, q)$ where p and q are probability density functions. Each divergence measure is characterized by a certain value of the parameter α.

The methodology presented in our paper has the advantage to show, by a graphical study, the rate of convergence of the simulation strategies IS, RWMH, Gibbs Sampler (GS), and as well as hybrid and adaptive simulation methods (AM and MWG).

Thus our study provides the ability to compare different methods of stochastic simulation qualified for a given problem. The use of these α-divergence measures gives a variety of divergence measures for different values of the parameter α. Using these α-divergences is particularly advantageous since there is no major requirements on densities p^n from simulation strategies and the target density f. Here the assumptions on the target density f and proposal density q are very often verifiable in practice. The divergence measures that we use in our study have a common part that is the integral

p^n is not the nth power of p but it is the density function of X_n

which appears in the definition of the α-divergence measures. For our illustrative examples we will use this integral estimator given by Póczos and Schneider [4].

Now we give more detail about our general study framework. Suppose we have a probability distribution with density function f from which we want to obtain samples. Suppose also that the methods of direct simulations are out of reach. Then we use the MCMC methods and derived methods (adaptive or hybrid simulation methods). The stochastic simulation methods that we compare in this paper are *Idependent Sampler (IS)*, *Random Walk of Metropolis-Hastings (RWMH)* and *Gibbs Sampler (GS)* which are MCMC methods but also with *Adaptive Metropolis (AM)* that is an adaptive method and *Metropolis Within Gibbs (MWG)* which is hybrid method. These two latter methods (*AM* and *MWG*) are explained more fully in Section 4.

For implementing these different simulation algorithms, one needs in each case a proposal distribution that generates samples. Note also that each algorithm, depending on the choice of the proposal law, converges more or less promptly.

This is highlighted when one uses a simulation method with two different proposal distributions and compares the convergence time of the two resulting algorithms. We recall that for a given simulation strategy each proposal distribution that one proposes for simulation corresponds to one algorithm. The convergence time of a stochastic simulation algorithm, here, is actually the time that the densities p^n, $n = 1, \ldots$ will put to converge to the target density f. The density function p^n is precisely the density of the nth element of the stochastic process $(X_n)_{n\geq 1}$ generated by one simulation strategy ie, at each iteration n, p^n is the density function of the random variable X_n.

This is what explains the interest, if we consider for example the Rényi divergence, to calculate for each iteration n the value $R_\alpha(p^n, f)$. For different iterations n quantities $R_\alpha(p^n, f)$ will be represented by a curve; then the interest is to study the evolution of this curve with respect to the value 0.

In this respect, consider two simulation strategies S_1 and S_2 for a given target distribution having density function f. We assume that these strategies S_1 and S_2 generate respectively the samples $X_{1,1}, X_{1,2}, \ldots, X_{1,n}, \ldots$ and $X_{2,1}, X_{2,2}, \ldots, X_{2,n}, \ldots$. The densities p_1^n and p_2^n are respectively density functions of $X_{1,n}$ and $X_{2,n}$. If we choose for example the Rényi divergence we can compare the curves of values $R_\alpha(p_1^n, f)$ and $R_\alpha(p_2^n, f)$, $n = 1, \ldots$ The curve which converges more rapidly to 0 indicates that the corresponding simulation strategy is more efficient.

The paper is organised as follows : in Section 2 we give the definition of estimators of the α-divergence measures and the methodology for comparing two simulation strategies, and how to choose an efficient simulation strategy. We show in Section 3 that our divergence measures have a gaussian asymptotic distribution. Then we have some application examples in order to illustrate our methodology in Section 4. Finally, in Section 5, we explain the results of our various illustrative examples.

2 Description and Estimators of Divergence Measures and Methodology for Finding an Optimal Simulation Method

2.1 Description of divergence measures

Let $(\mathcal{X}, \mathcal{A}, \lambda)$ be an arbitrary measure space with λ being a finite or $\sigma-$finite measure. Let also μ_1, μ_2 probability measures on $\mathcal{X}$ such that $\mu_1, \mu_2 \ll \lambda$ (absolutely continuous).
Denote the Randon-Nikodym derivatives (densities) of μ_i with respect to λ by $p_i(x)$:

$$p_i(x) = \frac{\mu_i(dx)}{\lambda(dx)}, \quad i = 1, 2.$$

Definition 2.1. The Kullback-Leibler relative divergence (also called relative entropy) between two probability measures μ_1 ,and μ_2 is defined by

$$K(\mu_1,\mu_2)=\int_{\mathcal{X}} p_1(x)\log\big(\frac{p_1(x)}{p_2(x)}\big)\lambda(dx)=\mathbb{E}_{\mu_1}\left[\log\frac{p_1(X)}{p_2(X)}\right] \tag{2.1}$$

We can also write $K(p_1,p_2)$.

On this paper we emphasize on a particular family of *Csiszár ϕ-divergence*, that is the family of α-*divergence* measures. The D_α divergence Cichocki[5], Ngom[6], the Rényi α-divergence and the Tsallis α-divergence Cichocki[3] are also part of this family .

Definition 2.2 (α-divergence)**.** The α-divergence is defined by

$$D_\alpha(\mu_1,\mu_2)=\frac{1}{\alpha(1-\alpha)}\Big(1-\int_{\mathcal{X}} p_1^\alpha(x)p_2^{1-\alpha}(x)\lambda(\mathrm{d}x)\Big),\ \alpha>0 \text{ and } \alpha\neq 1. \tag{2.2}$$

Definition 2.3 (Rényi α-divergence)**.** Rényi (1961) for the first time gave one generalization of the relative entropy given in (2.1). It is defined by

$$R_\alpha(\mu_1,\mu_2)=\frac{1}{\alpha-1}\log\int_{\mathcal{X}} p_1^\alpha(x)p_2^{1-\alpha}(x)\lambda(dx),\ \alpha>0 \text{ and } \alpha\neq 1. \tag{2.3}$$

Definition 2.4 (Tsallis α-divergence)**.** Tsallis α-divergence is defined by

$$T_\alpha(\mu_1,\mu_2)=\frac{1}{\alpha-1}\Big(\int_{\mathcal{X}} p_1^\alpha(x)p_2^{1-\alpha}(x)\lambda(\mathrm{d}x)-1\Big),\ \alpha>0 \text{ and } \alpha\neq 1 \tag{2.4}$$

In these definitions of divergence measures it appears one integral. This integral $M_\alpha(p_1,p_2)=\int_{\mathcal{X}} p_1^\alpha(x)p_2^{1-\alpha}(x)\lambda(\mathrm{d}x)$, is common to all of three measures, and it is difficult to be determined. That is the resason in practice, we need estimators of these α-divergence measures to assess the distance (similarity) between two density functions.

2.2 Estimators of divergence measures

We first make an estimator of the divergence measure $D_\alpha(p^n,f)$, afterward we can have in the same manner an estimator of Tsallis and Rényi divergences. We work now with λ the Lebesgue measure defined on $\mathcal{X}=\mathbb{R}^d$. Then we have

$$D_\alpha(p^n,f)=\frac{1}{\alpha(1-\alpha)}\Big(1-\int_{\mathbb{R}^d}\Big(\frac{f(x)}{p^n(x)}\Big)^{1-\alpha}p^n(x)\mathrm{d}x\Big) \tag{2.5}$$

We can initially think to write the integral like a mathematical expectation and apply the method of Monte Carlo integration. We would have then

$$D_\alpha(p^n,f)=\frac{1}{\alpha(1-\alpha)}\Big(1-\mathbb{E}\Big(\big(\frac{f(X)}{p^n(X)}\big)^{1-\alpha}\Big)\Big) \tag{2.6}$$

If f and p^n are such $(f(x)/p^n(x))^{1-\alpha}$ is measurable we can use the *Strong Law of Large Numbers* we will have the following estimator

$$\hat{D}_\alpha(p^n,f)=\frac{1}{\alpha(1-\alpha)}\Big(1-\frac{1}{N}\sum_{i=1}^{N}\big(\frac{f(X_i)}{p^n(X_i)}\big)^{1-\alpha}\Big), \text{with } X_i\sim p^n \tag{2.7}$$

which converges almost surely to $D_\alpha(p^n,f)$.

However, in practice, it is more complex because densities p^n are often untractables ie it is difficult to establish their analytical and simple expressions. In addition, in the case of row data, target densities f are often known up to a multiplicative constant ie $f(x)=c_0\varphi(x)$ where c_0 is

unknown real number. That is why the expression in (2.7) can not be used here . Even if the density f is entirely known, as in some of examples in Section 4, we prefer using its estimator in order to be able to use the divergence estimator proposed in Poczos[4]. Then we use the methods of nonparametric estimation of probability density functions. In fact, Poczos[4] have proposed in their paper an estimator of probability density based on *k-NN method* (k Nearest Neighbor). This probability density estimation method has been introduced by Loftsgaarden[7]. The authors Póczos and Shneider have used it in Poczos[4], [8]. If we apply their results to our densities f and p^n, we obtain the following density estimators

$$\hat{p}_{k,n,N}(X_i) = \frac{k/(N-1)}{V(H(X_i,\rho_k(X_i)))} = \frac{k}{(N-1)c\,\rho_k^d(X_i)} \tag{2.8}$$

$$\hat{f}_{k,M}(X_i) = \frac{k/M}{V(H(X_i,\gamma_k(X_i)))} = \frac{k}{M\,c\,\gamma_k^d(X_i)} \tag{2.9}$$

Denote by $V(H(z,r)) = \pi^{d/2}r^d/\Gamma((d/2)+1)$ is the volume of d-dimensional sphere around $z \in \mathbb{R}^d$ with radius $r > 0$, $\Gamma(.)$ is the Gamma function and c stands for the volume of a d-dimensional unit ball.

Recall how these density estimators are constructed in Poczos[4]. Let $X_{1:N} = (X_1, ..., X_N)$ be a sample simulated from distribution having density function p^n where X_i , $i = 1, \ldots, N$ are i.i.d (independent identically distributed). One can choose one realization X_i on the sample $X_{1:N}$ and calculate the value $\rho_k(X_i)$ that is the euclidean distance between X_i and its *kth* nearest neighbor on the sample. If the sample $Y_{1:M} = (Y_1, ..., Y_M)$, where the Y_i , $i = 1, \ldots, M$ are i.i.d, is generated from the density f, then $\gamma_k(X_i)$ is the euclidean distance between X_i and its *kth* nearest neighbor on the sample $Y_{1:M}$. Let us note that in practice as in some of our illustrative examples the density function f can not produce i.i.d samples, then we use an efficient MCMC method for generating a Markov chain $(X_n)_{n\geq 1}$. We know that the elements of the generated Markov chain are not independent. For that we will overcome this little problem by using the following method. Thus the samples that we use to estimate the density f are from this Markov chain and are built as follow: when n is sufficiently large, we choose $X_{n+r}, X_{n+2r}, \ldots, X_{n+ir}, \ldots, X_{n+Mr}$ where r is an integer for example such as $r \geq 10$. The fact to make this leap of r elements between samples that we choose allows us to obtain samples more or less independent.

We can notice in their respective definitions that divergence measures used here have a common part which is the integral $M_\alpha(p_1,p_2) = \int_{\mathbb{R}}^d p_1^\alpha(x)p_2^{1-\alpha}(x)\mathrm{d}x$ where p_1 and p_2 are density functions. As we have already mentioned in the introduction, Poczos[4] have given an estimator of $M_\alpha(p_1,p_2)$ that we use in this paper. This estimator is

$$\hat{M}_{\alpha,N,M,k}(p_1,p_2) = \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{\hat{p}_{2,k,M}(X_i)}{\hat{p}_{1,k,N}(X_i)}\Big)^{1-\alpha} \times B_{k,\alpha} \tag{2.10}$$

where $\hat{p}_{1,k,N}$ and $\hat{p}_{2,k,M}$ are respective density estimators of p_1 and p_2; the constant $B_{k,\alpha}$ is defined as follow

$$B_{k,\alpha} = \frac{(\Gamma(k))^2}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)} \quad \text{with } \Gamma(.) \text{ the Gamma function} \tag{2.11}$$

If we replace p_1 by p^n and p_2 by f we can apply the method proposed in Poczos[4] to obtain our divergence measures estimators

$$\hat{D}_{\alpha,N,M,k}(p^n,f) = \frac{1}{\alpha(1-\alpha)}\Big(1 - \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{(N-1)\rho_k^d(X_i)}{M\gamma_k^d(X_i)}\Big)^{1-\alpha} \times B_{k,\alpha}\Big) \tag{2.12}$$

$$\hat{T}_{\alpha,N,M,k}(p^n,f) = \frac{1}{\alpha-1}\Big(\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{(N-1)\rho_k^d(X_i)}{M\gamma_k^d(X_i)}\Big)^{1-\alpha} \times B_{k,\alpha} - 1\Big) \tag{2.13}$$

$$\hat{R}_{\alpha,N,M,k}(p^n,f)=\frac{1}{\alpha-1}\log\Big(\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{(N-1)\rho_k^d(X_i)}{M\gamma_k^d(X_i)}\Big)^{1-\alpha}\times B_{k,\alpha}\Big) \tag{2.14}$$

that are respectively the divergence estimators of $D_\alpha(p^n,f)$, $T_\alpha(p^n,f)$ and $R_\alpha(p^n,f)$.

The following theorem contains results which prove the convergence of these divergence estimators. This theorem was developped by Poczos[4] which consist in replacing the densities by the density functions p^n of sampled observations X_n, $n=1,\dots$ and the target density function f.

Theorem 2.1 (L_2 consistency)**.** *We have the following assumptions:* $k\geq 2$, $0<\gamma=1-\alpha<(k-1)/2$, p^n *is bounded away from 0,* p^n *is uniformly Lebesgue approximable,* $\exists\,\delta_0$ *such that* $\forall\,\delta\in(0,\delta_0)$ $\int F(x,p^n,\delta,1/2)p^n(x)\mathrm{d}x<\infty$,
$\int\|x-y\|^\gamma p^n(y)\mathrm{d}y<\infty$ *for almost all* $x\in\mathbb{R}^d$, $\int\int\|x-y\|^\gamma p^n(y)p^n(x)\mathrm{d}y\mathrm{d}x<\infty$, *and that* f *is bounded above. Then*

$$\lim_{N,M\to\infty}\mathbb{E}\Big(\big(\hat{M}_{\alpha,N,M,k}(p^n,f)-M_\alpha(p^n,f)\big)^2\Big)=0 \tag{2.15}$$

that is, the estimator is L_2 *consistent.*

The function $F(x,q,\delta,1/2)$, with q a probability density, defined in Poczos[4].

Once the divergence measures and their estimators described, one needs to know how to use these α-divergence measures. In the following section we show the methodology of comparison of two simulation strategies using the curve described by these divergence measures.

2.3 Methodology

In this present paper we propose to study an optimal stochastic simulation algorithm which may come from Metropolis-Hastings methods (*IS* and *RWMH*), *Gibbs Sampler*, or recent adaptive *(AM)* and hybrid *(MWG)* methods. For that, we will use our α-divergence measures. The interest of these divergences is that its subtracts the target density f, the proposal density q_i and initial density p_i^0 (corresponding to one simulation strategy S_i) of many assumptions like in Chauveau[2].

We develop here our methodology that we have slightly discussed in the introduction. The novelty of our approach lies in the fact it is easy to be implemented, because the divergence estimators that we use are built independently of any simulation strategy unlike in Chauveau[2]. This is why we can compare, in addition to *IS* and *RWMH* methods compared in Chauveau[2] various simulation methods. Among these methods we choose the *GS*, *AM*, and *MWG* strategies presented in the Introduction. However, one can not compare simulation strategies that are not empowered to solve the same problem. For example, the Gibbs Sampler method, is used when the target density f is a function defined on $E\subseteq\mathbb{R}^d$ with $d\geq 2$ and when the conditional densities are available. Thus, to compare two or more stochastic simulation strategies, one must ensure to have a common target distribution. Among simulation strategies some are more effective than others, and are intended to generate samples $X_1,X_2,\dots,X_n,\dots$ such that when n is large (tends to ∞) then X_n tends to f as his density function ($X_n\sim f$). It is then appropriate to evaluate at each iteration n the measure of similarity between the density p^n of X_n and the target density f. Here we still use the Rényi divergence measure for illustration, enven though it is not the only divergence measure used in the paper. Our methodology allows to have multiple measures of divergence that we can use as needed. Among them there are well known divergence measures: it is the Hellinger divergence measure $D_{\frac{1}{2}}(p,q)$ and the Chi-square divergence measure $D_2(p,q)$ where p and q are density functions. Assume that we have two simulation strategies S_1 and S_2 which have respective probability densities p_1^n and p_2^n at time n and a target density f. We can use these divergence measures to compare S_1 and S_2. Indeed, we can compare $R_\alpha(p_1^n,f)$ and $R_\alpha(p_2^n,f)$, for each iteration n, to see which of S_1 and S_2 is more efficient. The comparison is made by using curves.

- If the curve of $R_\alpha(p_1^n, f)$ values is closer to 0 than the curve of $R_\alpha(p_2^n, f)$ values (ie $0 < R_\alpha(p_1^n, f) < R_\alpha(p_2^n, f)$), then we can say that densities p_1^n converge faster than densities p_2^n to the stationary and target density f. Consequently, the simulation strategy S_1 associated to p_1^n is more effective than S_2.
- If the reverse happens, ie the curve of $R_\alpha(p_2^n, f)$ values is closer to 0 than the curve of $R_\alpha(p_1^n, f)$, $(0 < R_\alpha(p_2^n, f) < R_\alpha(p_1^n, f))$ we will say that the strategy S_2 is more efficient than S_1.
- If the curve of $R_\alpha(p_1^n, f)$ is more or less similar to the curve of $R_\alpha(p_2^n, f)$ and all are very close to 0 ($R_\alpha(p_1^n, f) \backsimeq R_\alpha(p_2^n, f)$), we can say that the two simulation strategies S_1 and S_2 are equivalent and both efficient.
- If the curve of $R_\alpha(p_1^n, f)$ and the curve of $R_\alpha(p_2^n, f)$ are far from 0 and slow to approach 0 or not approaching 0, then both strategies are ineffective.

After presenting the methodology, the knowledge of the distribution of divergence measures estimators can be useful, but we will not use these distributions, we only present it.

3 Asymptotic Distribution of Divergence Estimators

We seek to know the asymptotic distribution of our estimators. The asymptotic distribution of these estimators is studied under the assumption of measurability of densities p^n and f. If densities p^n , f, and their respective estimators are measurable, we choose $N = M$ and we get the following results.

3.1 Asymptotic distribution of D_α and T_α divergence estimators

Theorem 3.1. *If estimators of densities f and p^n are measurable, X_i are i.i.d and $\sigma^2 = \lim_{N\to\infty} Var(\bar{Y}_N) < \infty$, then*

$$\hat{D}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\Big(D_\alpha(p^n, f), \frac{\sigma^2 B_{k,\alpha}^2}{\alpha^2(1-\alpha)^2}\Big), \text{ when } N \to \infty. \tag{3.1}$$

$$\hat{T}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\Big(T_\alpha(p^n, f), \frac{\sigma^2 B_{k,\alpha}^2}{(\alpha-1)^2}\Big), \text{ when } N \to \infty. \tag{3.2}$$

Proof. We know that X_i are i.i.d, now consider $\bar{Y}_N = \frac{1}{N}\sum_{i=1}^N h_N(X_i)$ with $h_N(X_i) = \Big(\frac{(N-1)\rho_k^d(X_i)}{N\gamma_k^d(X_i)}\Big)^{1-\alpha}$. Function h_N which is the ratio of two measurable functions is also measurable. We have therefore the independence of $h_N(X_i)$, $i = 1, \ldots, N$. We can now apply the *Central Limit Theorem* to have asymptotical normal distribution

$$\frac{\bar{Y}_N - \mathbb{E}(\bar{Y}_N)}{\sqrt{Var(\bar{Y}_N)}} \xrightarrow{d} \mathcal{N}(0,1) \Longrightarrow \bar{Y}_N \times B_{k,\alpha} \xrightarrow{d} \mathcal{N}\Big(M_\alpha(p^n, f), \sigma^2 B_{k,\alpha}^2\Big)$$

We have now

$$\hat{D}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\Big(D_\alpha(p^n, f), \frac{\sigma^2 B_{k,\alpha}^2}{\alpha^2(1-\alpha)^2}\Big), \text{ when } N \to \infty.$$

$$\hat{T}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\Big(T_\alpha(p^n, f), \frac{\sigma^2 B_{k,\alpha}^2}{(\alpha-1)^2}\Big), \text{ when } N \to \infty.$$

□

3.2 Asymptotic distribution for Rényi divergence estimator

Theorem 3.2. *If estimators of densities f and p^n are measurable, X_i are i.i.d. $\sim> p^n$, $\sigma^2 = \lim_{N\to\infty} Var(\bar{Y}_N) < \infty$ and $\delta^2 = \lim_{N\to\infty} Var(h_N(X) < \infty$ with $X \sim> p^n$, then*

$$\hat{R}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\Big(\frac{1}{\alpha-1}\log[M_\alpha(p^n, f)]\,, \frac{\delta^2 B_{k,\alpha}^2}{(\alpha-1)^2[M_\alpha(p^n, f)]^2}\Big) \tag{3.3}$$

Proof. For determining the asymptotic distribution of the estimator of the Rényi divergence, we use the *Delta method*. Next, onsider this mean $\bar{Y}_N = \frac{1}{N}\sum_{i=1}^N h_N(X_i)$. We know that X_i are i.i.d. and h_N is measurable, then $h_N(X_i)$ are also i.i.d. The *Central Limit Theorem* gives us the following result
$\frac{\bar{Y}_N - \mathbb{E}(\bar{Y}_N)}{\sqrt{Var(\bar{Y}_N)}} \xrightarrow{d} \mathcal{N}(0,1)$ implying,

$$\sqrt{N}\big(\bar{Y}_N B_{\alpha,k} - M_\alpha(p^n, f)\big) \xrightarrow{d} \mathcal{N}\big(0, \delta^2 B_{k,\alpha}^2\big), \tag{3.4}$$

Relation (3.4) allows us to apply the *Delta method* to obtain

$$\sqrt{N}\big(\log(\bar{Y}_N B_{k,\alpha}) - \log(M_\alpha(p^n, f)\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{\delta^2 B_{k,\alpha}^2}{[M_\alpha(p^n, f)]^2}\Big).$$

Hence we obtain the result

$$\hat{R}_{\alpha,N,k}(p^n, f) \xrightarrow{d} \mathcal{N}\Big(\frac{1}{\alpha-1}\log[M_\alpha(p^n, f)]\,, \frac{\sigma^2 B_{k,\alpha}^2}{(\alpha-1)^2[M_\alpha(p^n, f)]^2}\Big)$$

□

4 Examples

We will now illustrate our methodology with simple examples. That's why we only give some examples in 1-dimensional and 2-dimensional cases. In the following examples, we will use α-divergence measures to compare firstly proposal densities corresponding to one given simulation strategy. Recall that for drawing samples from density function f by a given simulation method one needs a proposal distribution with density function q, an initial distribution with density p^0. The interest here is to compare, for a given target law, the different proposal distributions that are candidates. The comparison is more relevant if the different generated chains have the same starting point ie the same initial distribution. One can now try to find among these proposal densities $q_i\,, i = 1, \ldots, N$ those that allow to obtain better results (the speedy convergence of the algorithm). After this first example (Fig. 1) we will mainly compare different simulation strategies in the four others examples. As we have already said, two simulation strategies can not be compared if they do not have the same target density. In all our examples the proposal distributions that we use are just examples to illustrate our methodology. Someone might choose to take others; for example in Example 1 he could choose to change the two proposal laws. The sizes of sample drawn from distributions with density functions p^n and f are equal for all examples, ie $M = N$. The value of k must be known, now empirical experiments have suggested to take k equals to the integer part of $\sqrt{N}$ Loftsgaarden[7]. In this section we only present the results, the explanations will be made in the Section Discussion.

4.1 One-dimensional case

4.1.1 Target density f fully known

For the two first examples we have chosen target densities that are fully known. It comes the normal distribution in the first case and a Gaussian mixture in the second case. We will in the

first instance use the IS strategy to compare two proposal densities; then in the second example we compare the IS and RWMH strategies.

a) Independence Sampler (IS): comparison of proposal densities

For a given simulation strategy, the choice of an efficient proposal density (or proposal distribution) is important. Indeed, for achieving satisfactory simulation results it is important to choice an efficient proposal density. For simplicity, we consider densities that differ by the value of their parameters. Assume that the standard normal law $\mathcal{N}(0,1)$ is the target distribution here, consider that its density function is f. To implement the IS simulation method we propose, for example here, two candidate proposal distributions that are $\mathcal{N}(-3,2)$ and $\mathcal{N}(0,3)$. We want to know if these two proposal distributions are all efficient or which of the two is more effective. If densities p_1^n and p_2^n are respectively generated by $\mathcal{N}(-3,2)$ and $\mathcal{N}(0,3)$ we compare now for all n the curves $D_2(p_1^n, f)$ and $D_2(p_2^n, f)$. The Figure 1 shows that the curve associated to the law $\mathcal{N}(0,3)$ converges faster to 0 than the other curve (more details in Section Discussion).

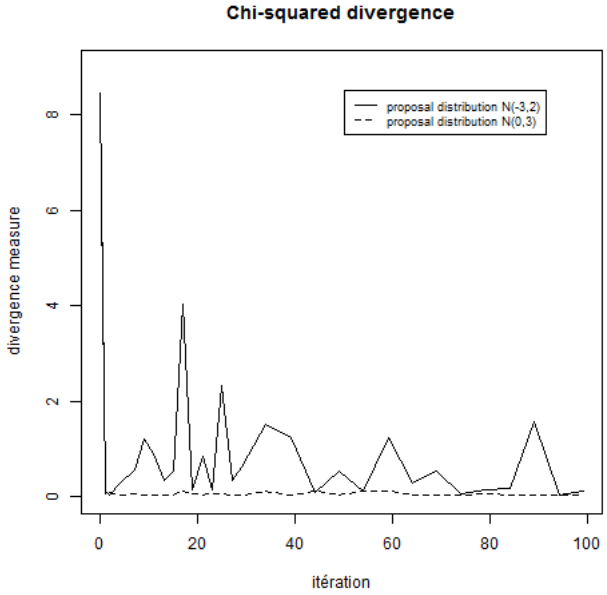

Figure 1: Comparison of two proposal densities, using the Independence Sampler and the Chi-squared divergence (D_α with $\alpha = 2$)

b) IS - RWMH : comparison of simulation strategies

After finding a good proposal distribution for each strategy, we compare here the two main strategies of Metropolis Hastings algorithm that are *IS* and *RWMH*. It may happen in an experiment that the *IS* strategy trumps *RWMH* strategy and in another experiment the opposite occurs. Everything depends on the instrumental distribution but also the target distribution to some extent, even if the initial law is the same for both strategies.

Here the chosen target distribution is a gaussian mixture $0.4\mathcal{N}(-8,2) + 0.6\mathcal{N}(0,6)$. For the *IS* we use the proposal distribution $\mathcal{N}(-2.5, 15)$ whereas for *RWMH* method we propose to take $\mathcal{N}(x, 15)$. This distribution has a mean equal to the current element $X_n = x$. We show the comparison of *IS* and *RWMH* strategies (Fig. 2). Note that the curve associated with *IS* strategy is below curve associated with *RWMH*. The explanations are given in section Discussion.

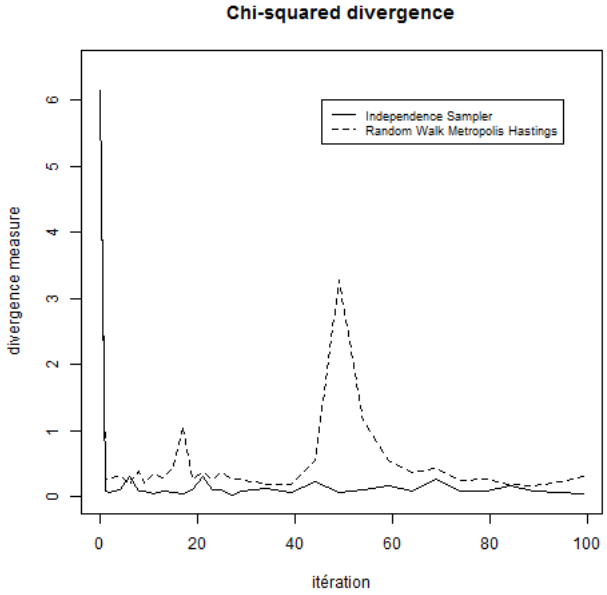

Figure 2: Comparison of two simulation strategies: *IS* vs *RWMH*, using the Chi-squared divergence (D_α with $\alpha = 2$)

4.1.2 Target density f is known up to a multiplicative constant

Adaptive Metropolis - RWMH: comparison of simulation strategies

In most real situations, the target density f is not analytically known. This is the case, for example Bayesian context where f is the density of the posterior. Then f is written as $f = c\varphi$ where c is unknown constant. The target density f, used here, is known up to a multiplicative constant. The *AM* simulation strategy less known than Metropolis - Hastings samplers (*IS* and *RWMH*) is presented here in more details.

We present some *Adaptive Metropolis (AM)* strategy proposed by Haario[9]. First recall that the stochastic process generated by this simulation method is not a Markov chain. However it has well ergodicity properties. The assumption on the target density is that it is bounded from above and has a bounded support.

The target density has a support $E \subset \mathbb{R}^d$, $d \geq 1$. Suppose, that at time t we have sampled the states $X_0, X_1, \ldots, X_{t-1}$, where X_0 is the initial state. Then a candidate point Y is sampled from the proposal distribution $Q_t(\,.\,|X_0, ..., X_{t-1})$, which now may depend on the whole history $(X_0, X_1, ..., X_{t-1})$. The candidate point Y is accepted with probability

$$\alpha(X_{t-1}, Y) = \min\left(1, \frac{\pi(Y)}{\pi(X_{t-1})}\right)$$

in which case we set $X_t = Y$, and otherwise $X_t = X_{t-1}$. Observe that the chosen probability for the acceptance resembles the acceptance probability of the Metropolis - Hastings algorithm in symmetric case. The proposal distribution $Q_t(\,.\,|X_0, \ldots, X_{t-1})$ employed in the *AM* algorithm is a Gaussian distribution with mean equal to the current point X_{t-1} and variance

$$C_t = \begin{cases} C_0\,, & t \leq t_0 \\ S_d Cov(X_0, \ldots, X_{t-1}) + S_d \epsilon I_d\,, & t > t_0\,. \end{cases}$$

where S_d is a parameter that depends only on dimension d and $\epsilon > 0$ is a constant that we may choose very small compared to the dimension of the space E. Here I_d denotes the d-dimensional identity matrix.

For these two simulation methods *(AM and RWMH)* we see that the respective divergence measures are very closely and very quickly all tend to 0 (Fig. 3). Here the divergence used is the *Hellinger divergence* measure $D_{1/2}$. The target density is $f(x) \propto \exp(-x^2)(2 + \sin(5x) + \sin(2x))$.

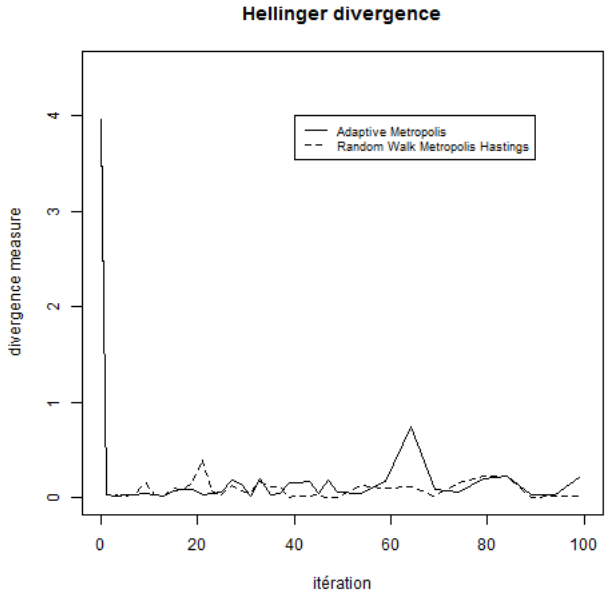

Figure 3: Comparison of two simulation strategies: *AM* vs *RWMH*, using the Hellinger divergence (D_α with $\alpha = 1/2$).

4.2 Two-dimensional case

We consider again here a target density function known up to a multiplicative constant. We choose now samples $X_1, \ldots, X_n$ which are i.i.d such that $X_i \sim \mathcal{N}(m, \sigma^2)$. So we have the following likelihood

$$L(x|m, \sigma^2) \propto (\sigma^2)^{(-n/2)} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2 \Big) \tag{4.1}$$

the prior distributions are

$$m \sim \mathcal{N}(m_0, \sigma_0^2)$$
$$\sigma^2 \sim \mathcal{IG}(\alpha, \beta),$$

the full posterior density is known up to a constant

$$\Pi(m, \sigma^2|x) \propto (\sigma^2)^{-\frac{n}{2}-(\alpha+1)} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2 - \frac{(m - m_0)^2}{2\sigma_0^2} - \frac{\beta}{\sigma^2} \Big),$$

the conditional distributions of parameters are

$$m|\sigma^2, x \sim \mathcal{N}(M, \Sigma^2)$$

where

$$M = \frac{\sigma_0^2 \sum_{i=1}^n x_i + \sigma^2 m_0}{\sigma^2 + n\sigma_0^2} \quad \text{and} \quad \Sigma^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$\sigma^2 | m, x \sim \mathcal{IG}\Big(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^n (x_i - m)^2 + \beta\Big)$$

So let's compare firstly *RWMH* and *Gibbs Sampler (GS)* and secondly we will compare *RWMH* and *Metropolis Within Gibbs (MWG)*. The likelihood in (4.1) will be our target density function.

a) RWMH - GS: comparison of simulation strategies

If *RWMH* is applied in dimension $d \geq 1$, the *GS* is only applied in dimension $d > 1$ ie for a multivariate probability distribution. However, we study *GS* here only in dimension 2. Note that this method *(Gibbs Sampler)* has been used by Geman [10] to generate observations from a Gibbs distribution (Boltzmann distribution) (Latuszyński [11]). It is an efficient MCMC method and is widely used in Bayesian analysis. For drawing observations from a probability density $f(\theta)$ with $\theta = (\theta_1, \ldots, \theta_p)$ we can use the following algorithm

Algorithm 4.1. *1. Initialisation: generating a vector* $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$ *from one initial law* Π_0.
2. Repeat for $j = 0, 1, 2, \ldots, M$ *simulation from the conditional distributions* $f_i(\theta_i | \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$ $i = 1, 2, \ldots, p$,

- *generate* $\theta_1^{(j+1)} \sim f_1(\theta_1 | \theta_2^{(j)}, \ldots, \theta_p^{(j)})$
- *generate* $\theta_2^{(j+1)} \sim f_2(\theta_2 | \theta_1^{(j+1)}, \theta_3^{(j)}, \ldots, \theta_p^{(j)})$

  $\vdots$
- *generate* $\theta_p^{(j+1)} \sim f_p(\theta_p | \theta_1^{(j+1)}, \theta_2^{(j+1)}, \ldots, \theta_{p-1}^{(j+1)})$

3. Return the values $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}\}$

We see that the curve corresponding to *RWMH* strategy is well above curve representing the *GS* method (Fig. 4). As mentioned in the legend, the dashed curve is associated to *RWMH* while the solid curve is associated to the *GS* algorithm.

b) RWMH - Metropolis Within Gibbs: comparison of simulation strategies

Metropolis Within Gibbs (MWG) is a hybrid simulation method that combines stages of the *Gibbs Sampler* and *Metropolis Hastings* method. It is used in some cases where we use GS and have conditional distributions for which we can't directly sample. There are several versions of this sampler, so we present the following.

Assume that $\pi(.)$ is the target density, $\pi(.|z_{-i})$ denote now the conditional distribution of $Z|Z_{-i} = z_{-i}$ where $Z \sim \pi$. $X_n := (X_{n,1}, \ldots, X_{n,d})$; $X_{n,-i} := (X_{n,1}, \ldots, X_{n,i-1}, X_{n,i+1}, X_{n,d})$; $\alpha := (\alpha_1, \ldots, \alpha_d)$. Now we have the following algorithm.

Algorithm 4.2. *1. Choose coordinate* $i \in \{1, \ldots, d\}$ *according to selection probabilities* α*, that is, with* $\mathbb{P}(i = j) = \alpha_j$.
2. Draw $Y \sim q(X_{n-1,i}, .)$.
3. Accepte the candidate Y *with probability*

$$\min\Big(1, \frac{\pi(Y | X_{n-1,-i}) q(Y, X_{n-1,i})}{\pi(X_{n-1,i} | X_{n-1,-i}) q(X_{n-1,i}, Y)}\Big)$$

and set $X_n := (X_{n-1,1}, \ldots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \ldots, X_{n-1,d})$
otherwise reject Y *and set* $X_n = X_{n-1}$.

Starting from a common point, the curves stay away from the value 0 (Fig. 5). The solid curve is associated with *MWG* strategy and the dashed curve associated with *RWMH* strategy. Here we use the Rényi divergence measure for doing comparison between strategies.

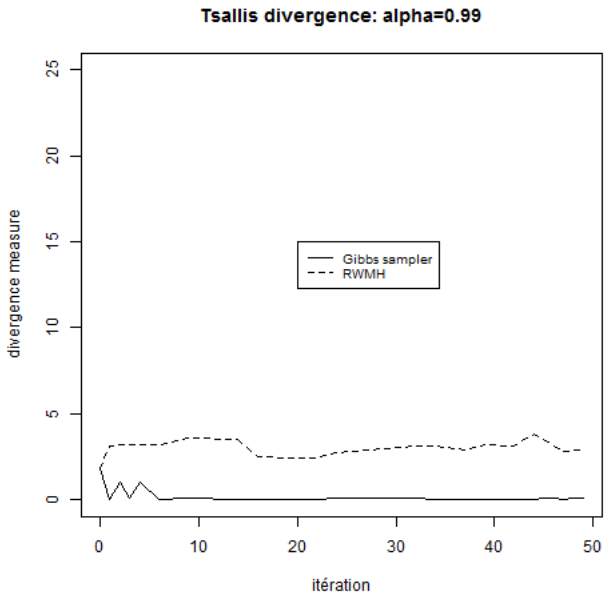

Figure 4: Comparison of two simulation strategies: *GS* vs *RWMH*, using the Tsallis divergence with $\alpha = 0.99$.

5 Discussion

5.1 Independence Sampler : comparison of proposal densities

The divergence measures $D_2(p_1^n, f)$ and $D_2(p_2^n, f)$ are functions of the number of iterations n and are represented by the curves in Figure 1. As already stated, the fact that these curves have the same starting point can be explained by relevance to begin with a same state x_0 or same initial law π_0 to compare two simulation strategies. However, we have chosen one same starting point drawn from the initial distribution. The dashed curve (IS with $\mathcal{N}(0,3)$) is below the solid curve (*IS* with $\mathcal{N}(-3,2)$). It follows that, the *IS* strategy with the proposal law $\mathcal{N}(0,3)$ is more efficient than *IS* with $\mathcal{N}(-3,2)$ because its mean $m = 0$ is equal to the mean of the target distribution $\mathcal{N}(0,1)$. Its variance $\sigma^2 = 3$ is greater than the variance of the target distribution which is $\sigma^2 = 1$. It follows that its support covers the support of the target density. Thus the proposal law $\mathcal{N}(0,3)$ is closer to $\mathcal{N}(0,1)$ than proposal distribution $\mathcal{N}(-3,2)$. This is why the samples $X_0, X_1, \ldots, X_n, \ldots$ generated by the strategy *IS* with $\mathcal{N}(0,3)$ converge in distribution more quickly to a random variable X having as probability distribution the target law $\mathcal{N}(0,1)$, than the random observations $Y_0, Y_1, \ldots, Y_n, \ldots$ drawn from the *IS* with $\mathcal{N}(-3,2)$.

5.2 Independence Sampler - RWMH

In the Figure 2 the solid curve is below the dashed curve. This shows that *IS* strategy is more efficient than the *RWMH* strategy. This is explained by the right choice of the proposal distribution $\mathcal{N}(-2.5, 15)$ for the IS strategy. With a mean $m = -2.5$ and a variance $\sigma^2 = 15$, the density function of this law is centered relatively to the target density $f(x)$. Thus the support of the target density is well covered by this proposal density.

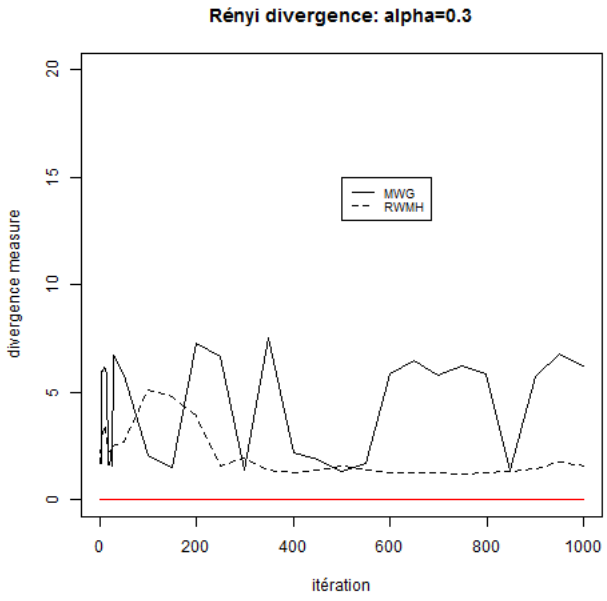

Figure 5: Comparison of two simulation strategies: *MWG* vs *RWMH*, using the Rényi α-divergence with $\alpha = 0.3$.

Regarding the *RWMH* strategy, after a first jump before the 20th iteration the dashed curve makes a big jump to reach the value 3.28 at the 50th iteration (Fig. 2). This implies that this *RWMH* algorithm is not still stable. Therefore the convergence to the target distribution will be slowly. This is explained by the fact that the proposal distribution $\mathcal{N}(x_n, 15)$ depend, at each iteration n, on the current state $X_n = x_n$. Unlike the density function of the distribution $\mathcal{N}(-2.5, 15)$ which is all the time centered on a well chosen value -2.5, the density function of the distribution $\mathcal{N}(x_n, 15)$ is centered, at each iteration n, on the variable value x_n. Consequently the convergence of the process $(X_n,\, 0 \leq n \leq N)$ generated by the *RWMH* algorithm to the target distribution $0.4\mathcal{N}(-8, 2)+0.6\mathcal{N}(0, 6)$ depends greatly on the initial state $X_0 = x_0$.

5.3 Adaptive Metropolis - RWMH

The curves which describe the effectiveness of simulation strategies *AM* and *RWMH* are shown in Figure 3. We note that these two curves are almost similar. Besides this, these curves evolve while remaining close to 0 when n is large even if we are limited, here, to 100 iterations. Thus, the two corresponding simulation strategies are all very efficient. The efficiency of *AM* strategy is explained, by the fact that it adapts its proposal density to the target density $f(x) = c\exp(-x^2)(2 + \sin(5x) + \sin(2x))$, where c is unknown constant . The adjustment mechanism is performed on the variance of proposal distribution. In this respect, if $X_1, X_2, \ldots, X_{n-1}$ have already been simulated and one wants to obtain a new point X_n, he draws it from a Gaussian proposal distribution $\mathcal{N}(x_{n-1}, \sigma^2)$ having mean equal to the current state $X_{n-1} = x_{n-1}$ and variance $\sigma^2 = C_t$. C_t is the covariance matrix described in section 4.1.2. But here the variance $\sigma^2 = C_t$ is a real value (one-dimensional case). Recall that the *AM* strategy acts on the variance σ^2 of the proposal law. The updating of this variance starts from the 16th iteration (arbitrary choice in this example). We have chosen the value of the variance

$\sigma^2 = 5$ between the first and the 15-th iteration. The behavior of the curves in Figure 3 shows that this value of the variance $\sigma^2 = 5$ is acceptable for the both simulation strategies *AM* and *RWMH* since the proposal law for *RWMH* method is $\mathcal{N}(x_{n-1}, 5)$. Also, the adjustments on the variance σ^2, made by th *AM* method, enable to have good drawn points $X_0, X_1, \ldots, X_n, \ldots$. Therefore the two simulation methods *AM* and *RWMH* are all efficient here.

5.4 Gibbs Sampler - RWMH

We study here two simulation methods in 2-dimensional case using the Tsallis α-divergence measure, with the parameter $\alpha = 0.99$ (this value is a simple choice). The dashed curve *(RWMH)* is above the solid curve *(GS)* and far from the value 0 after 50 iterations. It shows, here, that our *RWMH* strategy is ineffective. This inefficiency is due to the bad choice of covariance matrix

$$M = \begin{bmatrix} 12 & 8 \\ 8 & 15 \end{bmatrix}$$

of the proposal distribution. This matrix, which is an implementation parameter, is not optimal. The choice of this matrix is merely illustrative. Thus, with a bad chosen parameter (matrix), as it is the case here, the resulting algorithm will be inefficient. The solid curve *(GS)* tends rapidly to 0 (after the 7th iteration). Recall that the Gibbs Sampler is generally very effective. Its implementation is made possible when the conditional distributions having density functions $f_i(\theta_i|\theta_1, \ldots \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$ (ref. Algorithm 4.1) are available. The strength of this sampler is mainly due to the fact that the components of the vector $X^{(n)} = (m^{(n)}, \sigma^{2(n)})$ are generated directly from the conditional distributions of the target distribution $L(x|m, \sigma^2)$ defined in equation (4.1).

5.5 Metropolis Within Gibbs - RWMH

Here we compare the previous *RWMH* strategy in section 5.4 and *MWG* algorithm. We now study the efficiency of a classical Metropolis - Hastings algorithm *(RWMH)* and a hybrid algorithm which combines Metropolis - Hastings algorithm and Gibbs Sampler *(MWG)*. This simulation method (MWG) is used in specific situations, ie in case where one uses the Gibbs Sampler and, for a few conditional distributions $f_i(\theta_i|\theta_1, \ldots \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$ used in Algorithm 4.1, he can't directly simulate the observations. Then he can introduce in the Gibbs Sampler algorithm a few steps of RWMH sampler (ref. Algorithm 4.2).

In this example (MWG - RWMH) we use a version of the *MWG* algorithm which systematically gives proposal probability distribution to each conditional distribution in Gibbs sampler (Metropolis sampler steps), so that its convergence time is very elongated (Fig. 5). Our *RWMH* strategy also used in section 5.4 is ineffective. This is the reason that the two curves are all far from 0 even after 1000 iterations (Fig 5), reflecting, here, the ineffectiveness of these two simulation algorithms.

5.6 Conclusion

We have shown that with various divergence measures we can compare two different simulation strategies solving one stochastic simulation problem for determining the optimal algorithm. This comparison lead us to determine the effectiveness or no of a stochastic simulation algorithm. For doing this study, our paper has been structured as follows. In Section 2 we have described the α-divergence measures, estimators of these divergence measures and we have given a methodology for choosing an optimal simulation strategy. In Section 3 we have given the asymptotic distributions of D_α, T_α and R_α divergence estimators. Then in Section 4, we have given some examples for implementing our diverse simulation strategies, thus it allowed us to illustrate our methodology. Finally

we have given, in Section 5, the explanations about the behavior of curves in different graphics.

Acknowledgment

This research was supported, in part, by grants from Cheikh Anta Diop University and SIMONS-NLAGA Project. We are grateful to many seminar participants and to anonymous referees for comments.

Competing Interests

The authors declare that no competing interests exist.

References

[1] Chauveau D, Vandekerkhov P. Smoothness of Metropolis-Hastings algorithm and application to entropy estimation. ESAIM: Probability and Statistics. 2012;(17):419-431.

[2] Chauveau, Vandekerkhove. How to compare MCMC simulation strategies? 2007. hal.inria.fr/hal-00019174 (version 3), arXiv:math/0605263.

[3] Cichocki, Amari S. Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities.2010;1-41. Entropy.

[4] Póczos B, Schneider J. On the Estimation of α-Divergences. Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS). 2011;15:609-617.

[5] Cichocki A, Lee H, Kim YD, Choi S. Non-negative matrix factorization with α-divergence. Pattern Recognition Letters; 2008.

[6] Ngom P, Ntep B. Minimum Penalized Hellinger Distance for Model Selection in Small Samples. Open Journal of Statistic. 2012;2(4):369-382.

[7] Loftsgaarden DO. Quesenberry CP. A Nonparametric Estimate of a Multivariate Density Function. Annals of Mathematical Statistic. 1965;36(3):1049-1051.

[8] Póczos B, Xiong J, Schneider J. Nonparametric Divergence Estimation with Applications to Machine Learning on Distributions. arXiv:1202.3758v1 [cs.LG]. 2012;599-608.

[9] Haario H, Saksman E, Tamminen J. An adaptive Metropolis algorithm. Bernoulli. 2001;7(2):223-242.

[10] Geman S, Geman D. Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. IEE Transactions on Pattern Analysis and Machine Intelligence. 1984;6:721741.

[11] Latuszyński K, Roberts GO, Rosenthal JS. Gibbs sampler and related MCMC methods. The Annals of Applied Probability. 2013;23(1):66-98.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid=699&id=6&aid=6302