The Canadian Journal of Statistics
Vol. 34, No. 4, 2006, Pages ???-???
La revue canadienne de statistique

Optimal scaling of Metropolis algorithms: Heading toward general target distributions

Mylène BÉDARD and Jeffrey S. ROSENTHAL

Key words and phrases: acceptance rate; Langevin diffusion; marginal process; scaling term; weak convergence.
MSC 2000: Primary 60F05; secondary 65C40.

Abstract: The authors provide an overview of existing optimal scaling results for the Metropolis algorithm with Gaussian proposal distribution; they address in more depth the case of high-dimensional target distributions formed of independent, but not identically distributed components. The paper attempts to give an intuitive explanation as to when the previously-derived optimal acceptance rate of 0.234 is indeed optimal, and when it is unsuitable. In the latter case, it also explains how to find the correct asymptotically optimal acceptance rate, and why it is sometimes necessary to turn to inhomogeneous proposal variances in order to obtain an efficient algorithm. This is all illustrated with a simple example.

Échelonnage optimal de l'algorithme Metropolis: en route vers des distributions cibles générales

Résumé: Les auteurs survolent différents résultats sur l'échelonnage optimal de l'algorithme Metropolis avec distribution instrumentale gaussienne. Ils adressent particulièrement le cas des distributions cibles multidimensionnelles aux composantes indépendantes, mais non identiquement distribuées. Cet article tente d'expliquer intuitivement les conditions sous lesquelles le taux d'acceptation optimal 0.234, dérivé dans le passé, peut être appliqué. Dans le cas où ce dernier n'est pas applicable, les auteurs fournissent une méthode qui permet de déterminer le bon taux d'acceptation optimal, et expliquent également la raison pour laquelle il est parfois indispensable d'avoir recours à des distributions instrumentales non-homogènes. Ces résultats sont illustrés à l'aide d'un exemple.

1. INTRODUCTION

This paper surveys optimal scaling results for the Metropolis algorithm, in particular those introduced in Bédard (2006a). Metropolis-Hastings algorithms constitute the most popular class of MCMC algorithms; they generate values from a target distribution of interest by building a Markov chain $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \ldots$ having the target distribution as its stationary distribution. Optimal scaling refers to the need to tune the parameters of the algorithm to make the Markov chain converge as fast as possible to stationarity. Nowadays, more sophisticated algorithms are available in the literature for practitioners to apply to their specific case. Nonetheless, understanding the nature of the simpler algorithms may provide insight into the workings of other algorithms. We then consider a d-dimensional Metropolis algorithm with a Gaussian proposal distribution, and use it to sample from a target distribution having density π with respect to Lebesgue measure. Given the time-t position of the chain, $\mathbf{X}_t^{(d)}$, we propose a new position $\mathbf{Z}_{t+1}^{(d)} \sim N\left(0, \sigma^2 I_d\right)$ for the next time step; here I_d is the d-dimensional identity matrix. This proposed value is then accepted with probability $\alpha(\mathbf{X}_t^{(d)}, \mathbf{X}_t^{(d)} + \mathbf{Z}_{t+1}^{(d)}) = 1 \wedge \pi(\mathbf{X}_t^{(d)} + \mathbf{Z}_{t+1}^{(d)})/\pi(\mathbf{X}_t^{(d)})$. More formally,

$$\mathbf{X}_{t+1}^{(d)} = \mathbf{X}_t^{(d)} + \mathbf{Z}_{t+1}^{(d)} \, \mathbf{1}\left(U_{t+1} < \alpha\left(\mathbf{X}_t^{(d)}, \mathbf{X}_t^{(d)} + \mathbf{Z}_{t+1}^{(d)}\right)\right), \tag{1}$$

where $U_{t+1} \sim \mathrm{Uniform}[0,1]$.

The variance of the proposal distribution, σ^2, turns out to have a significant impact on the speed of convergence of the algorithm to its stationary distribution. Indeed, small values of σ^2 cause the algorithm to explore its state space very slowly, while large variances only rarely generate acceptable moves and this results in a chain remaining still for long periods of time. Since extremal variances lead to an algorithm exploring its state space in a lazy fashion and converging slowly to stationarity, we would expect the existence of an optimal σ^2 for which the mixing of states is maximised; that optimal value is the subject of this paper. Hereafter, the term "optimality" thus refers to the best possible mixing of states or, in other words, the fastest convergence to stationarity (see the Appendix for a brief review of convergence of Markov chains to stationarity). In this state of mind, a reasonable efficiency criterion would then consist in maximising the average squared jumping distance of the algorithm.

In the next section, we shall describe the first theoretical optimal scaling results to have appeared in the literature. We shall also mention generalisations addressed in later papers. In Sections 3, 4 and 5, we present the optimal scaling results in Bédard (2006a): the target model studied is first described in Section 3, and the optimal scaling results are discussed in Sections 4 and 5. The aim of Section 6 is to illustrate how the results from the previous sections can be applied through an example. To conclude, we discuss possible avenues for future research on this problem.

2. HISTORY OF OPTIMAL SCALING

The issue of optimal scaling of Metropolis algorithms was recognised in the original paper by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953), where proposals of the form $Z_{t+1} \sim \mathbf{Uniform}[X_t - \alpha,\ X_t + \alpha]$ were considered, and it was noted (p. 1089):

> It may be mentioned in this connection that the maximum displacement α must be chosen with some care; if too large, most moves will be forbidden, and if too small, the configuration will not change enough. In either case it will then take longer to come to equilibrium.

Historically, the tuning of Metropolis algorithms was typically performed by trial and error, or by users trusting their own intuition/judgement. In the 1990s, Besag and Green (1993) and Besag, Green, Higdon, and Mergensen (1995) provided rules of thumb to select reasonable values for σ^2; these methods promoted variances corresponding to acceptance rates between 30% and 70%. It then came as a surprise when, a couple of years later, it was proved in Roberts, Gelman, and Gilks (1997) (see also Gelman, Roberts, and Gilks 1996) that high-dimensional algorithms should be tuned to accept 23% of the proposed moves only in order to perform optimally. It should be made clear that the models considered in the practically oriented papers Besag and Green (1993) and Besag et al. (1995) are more general than those for which the 23% figure is valid. One difference is that they include low-dimensional algorithms, which are now known to feature greater optimal acceptance rate than higher-dimensional ones. This however illustrates how counter-intuitive an acceptance rate as small as 23% might appear; indeed, this says that high-dimensional algorithms should stay still 77% of the time in order to obtain the fastest mixing chain!

As mentioned previously, the result of Roberts, Gelman, and Gilks (1997) applies in the asymptotic limit of infinite-dimensional target distributions only, with proofs relying heavily on laws of large numbers as the dimension goes to infinity. This however does not restrict the applicability of this result to extremely large-dimensional problems; in general, distributions with as few as 15 dimensions behave according to their asymptotics. It is difficult to derive theoretical optimal scaling results for finite-dimensional target distributions; the few exceptions existing in the literature are discussed in Sherlock (2006); among them we find the nice case of finite-dimensional Gaussian target distributions.

The optimal scaling result of Roberts, Gelman, and Gilks (1997) also assumes a very simple *iid* form for the target distribution, i.e. $\pi(\mathbf{x}) = \prod_{i=1}^{d} f(x_i)$ for some smooth one-dimensional density f (specifically, we require that f be a C^2 density and that $(\log f)'$ be Lipschitz continuous). Of course, due to the independence among the components, such a multidimensional problem could easily be considered as many univariate problems by sampling from each target component individually. The optimal scaling issue is a complex one, and an *iid* target density was the natural starting point. To this day, there exist very few optimal scaling results for correlated targets, and those were derived for very specific models. In spite of the peculiarity of the *iid* model, the result derived might still be used to provide some intuition as to how algorithms applied to sample from slightly correlated target densities should be scaled.

Theoretical justification - Outline: Before proceeding, we roughly sketch the proof of this result. For the algorithm to converge to a nontrivial limiting process as the dimension of the target density increases to infinity, a space-time rescaling is required; this was achieved in Roberts, Gelman, and Gilks (1997) by letting the proposal variance be $\sigma^2(d) = \ell^2/d$ for some positive constant ℓ and by studying the sped up process $\{\mathbf{W}^{(d)}(t);t \geq 0\} = \{\mathbf{X}^{(d)}([td];t \geq 0)\}$. The process $\{\mathbf{W}^{(d)}(t)\}$ is thus an accelerated version of the algorithm, which proposes d moves instead of only one in each unit time interval. As $d \to \infty$, the proposed moves become both smaller and closer in time, eventually resulting in an asymptotically continuous process. By studying each component of the d-dimensional rescaled process separately (i.e. by studying every one-dimensional path followed by the process, given the past moves of the whole d-dimensional algorithm), it was proved in Roberts, Gelman, and Gilks (1997) that each component asymptotically behaves independently from the others according to a Langevin diffusion process $\{W(t);t \geq 0\}$.

The speed measure of this diffusion, $\upsilon(\ell)$, is the only part of the asymptotic process depending on the proposal variance. This means that the following relation holds: $dW(t) = dW(\upsilon(\ell)s)$, where $dW(s) = dB(s) + \frac{1}{2}(\log f(W(s)))'\ ds$ is the stochastic differential equation (SDE) of the limiting Langevin diffusion process with speed measure unity. In order to find the optimal value $\hat{\ell}$, it suffices to choose the diffusion which goes fastest; by optimising $\upsilon(\ell)$ with respect to ℓ, the optimal scaling value was then found to be $\hat{\ell} = 2.38/\sqrt{I}$, where $I = E_f[\left((\log f)'\right)^2]$. Here, I measures the "roughness" of the density. The smoother the density is (so the smaller I is), the more aggressive we can afford to be in the magnitude of the proposed steps (so the larger $\hat{\ell}$ is). Note however that I is not the Fisher information since the derivative of $(\log f)$ is with respect to the variable and not the parameter.

This result can also be expressed in terms of an asymptotically optimal acceptance rate (AOAR) rather than an asymptotically optimal scaling $\hat{\ell}$. To be precise, define the average acceptance rate of the algorithm as

$$\begin{aligned} a_d(\ell) &= E\left[\alpha\left(\mathbf{X}^{(d)}(t), \mathbf{X}^{(d)}(t) + \mathbf{Z}^{(d)}(t+1)\right)\right] \\ &= \int\int \alpha\left(\mathbf{x}^{(d)}, \mathbf{x}^{(d)} + \mathbf{z}^{(d)}\right)\phi\left(d, \mathbf{z}^{(d)}\right)\pi\left(\mathbf{x}^{(d)}\right) d\mathbf{z}^{(d)} d\mathbf{x}^{(d)}, \end{aligned} \tag{2}$$

where $\phi(d,\cdot)$ is the probability density function (*pdf*) of a $N\left(\mathbf{0}, \ell^2 I_d/d\right)$ random variable and π is the d-dimensional density with *iid* components. It turns out that

$$\lim_{d\to\infty} a_d(\ell) = 2\Phi\left(-\frac{\ell}{2}\sqrt{I}\right) \equiv a(\ell),$$

and therefore we find that $a(\hat{\ell}) \doteq 2\,\Phi(-1.19) \doteq 0.234$. This result is very simple to apply in practice: monitor the acceptance rate of the algorithm and adjust the proposal variance such that the algorithm accepts roughly 23% of the proposed moves. This shall yield the Markov chain converging as fast as possible to its invariant distribution. This is quite convenient as the AOAR is always the same, regardless of the form of f.

After these results were published, a number of researchers attempted to relax the *iid* assumption for the target distribution. In particular, Breyer and Roberts (2000) showed that an acceptance rate of 0.234 holds for suitably behaved sequences of target densities with partial correlations of finite range (i.e. when no phase transition occurs). Around the same time, Roberts and Rosenthal (2001) showed that for inhomogeneous target densities of the form $\pi(\mathbf{x}^{(d)}) = \prod_{i=1}^{d} C_i\, f(C_i x_i)$, where $C_i > 0$ are chosen from some fixed distribution, the magic number 0.234 still holds.

Christensen, Roberts, and Rosenthal (2003) studied high-dimensional Metropolis algorithms in their initial transient phase. They showed that when selecting the proposal variance as prescribed in Roberts, Gelman, and Gilks (1997), the convergence of this algorithm in the transient phase is extremely regular, and in fact resembles a deterministic trajectory. Recently, Neal and Roberts (2006) considered the case where updates of high-dimensional Metropolis algorithms are lower dimensional than the target density itself. They found that the optimal acceptance rate 0.234 holds for the Metropolis-within-Gibbs algorithm, as well as for lower dimensional updates of the Metropolis algorithm. This thus implies that lower-dimensional updates are to be preferred since high-dimensional updates are generally computationally more demanding.

The same method of proof has also been applied to derive optimal scaling results for other types of MCMC algorithms, such as the Metropolis-adjusted Langevin algorithm (MALA); see Roberts and

Rosenthal (1998, 2001), Breyer, Piccioni, and Scarlatti (2002), Christensen, Roberts, and Rosenthal (2003), Neal and Roberts (2006). In this paper, we consider Metropolis algorithms only; we do not give an exhaustive account of the literature about the MALA.

In the next section, we introduce a target model for which the limiting behaviour of the algorithm sometimes differs from that established in Roberts, Gelman, and Gilks (1997), producing different asymptotically optimal acceptance rates (AOARs).

3. A MORE GENERAL FRAMEWORK

We now introduce a class of target distributions that generalises the *iid* assumption of Roberts, Gelman, and Gilks (1997). Specifically, we study the natural extension which consists of multidimensional target distributions with independent, but not identically distributed components. The models considered are tractable enough to allow for an asymptotic analysis of the algorithm as the dimension goes to infinity, but still add in complexity as the various components of the algorithm might converge to their stationnary distribution at different speeds. This extension might seem artificial; nonetheless, studying these components jointly rather than considering them as several univariate problems is the first step towards understanding and developing optimal scaling results for target distributions with correlated components.

Indeed, there exist many correlated targets for which some components converge faster than others; from an optimal scaling viewpoint, the different speeds of convergence might even be more important than correlation itself. An example where this issue is of prime importance would be hierarchical models; the limiting behaviour and the optimal scaling issue of Metropolis algorithms with hierarchical target distributions are currently under study by one of the authors. It should be noted that the target distributions considered hereafter also include, as a special case, multivariate normal target distributions with a non-trivial correlation structure. We shall also see that the asymptotic optimal acceptance rate of target models with independent components constitutes an upper bound for the AOAR of *any* other high-dimensional target distribution.

The proposal distribution considered in this paper is the same as that in Roberts, Gelman, and Gilks (1997), i.e. a Gaussian distribution with independent components. There are two main reasons for using this setting. Firstly, since the target distributions considered are made of independent components, it would be suboptimal to use a correlated proposal. Secondly, even if we were studying correlated target distributions, using a correlated proposal would require studying/estimating the correlation structure of the target distribution, hence involving more computations. The choice of a correlated proposal with an inappropriate correlation structure would bring very few advantages over a proposal with independent components.

3.1 Target distribution

We suppose that we want to sample from the d-dimensional target density given by

$$\pi\left(d, \mathbf{x}^{(d)}\right) = \prod_{i=1}^{d} \frac{1}{\sqrt{\theta_i^2(d)}} f\left(\frac{x_i}{\sqrt{\theta_i^2(d)}}\right). \tag{3}$$

That is, we consider d independent components each based upon the same smooth density f, but each possessing its own scaling term $\theta_i^2(d)$, $i = 1, \ldots, d$. The regularity conditions on f stated in Bédard (2006a), although weaker than those in Roberts, Gelman, and Gilks (1997), are still stronger than required; all that is needed is that f be a C^2 density, $(\log f)'$ be Lipschitz continuous, and that $E[|f''(X)/f(X)|^{1+\epsilon}] < \infty$ for some $\epsilon > 0$.

To obtain optimal scaling results for the Metropolis algorithm, we shall be interested in the infinite-dimensional version of the target density in (3), and hence in a family of scaling vectors $\mathbf{\Theta}^2(d)$ as $d \to \infty$. To make sense of this, we suppose that each scaling term takes the form $\theta^2(d) = K/d^\lambda$ for some constant $K > 0$ and power $-\infty < \lambda < \infty$. As $d \to \infty$, we might be interested in different patterns for the scaling terms in the infinite-dimensional scaling vector. For instance, we might want to preserve the proportion occupied by each scaling term in the vector $\mathbf{\Theta}^2(d)$, or else we might be interested in keeping a certain number of scaling terms fixed (e.g. $\theta_1(d) \equiv 1$ is unique for any $d \geq 1$).

To illustrate the former, suppose that we are faced in practice with a 20-dimensional target distribution with half of its scaling terms equal to 1/20, and the other half equal to 1. In this situation, the limiting behaviour of a d-dimensional target distribution for which the proportion of the scaling terms is preserved (i.e. half equal to $1/d$, half equal to 1) would be much more representative of the behaviour of our 20-dimensional target than the limiting behaviour of a d-dimensional target distribution with only the first 10 scaling terms equal to $1/20$ and the other $d-10$ equal to 1. Similarly, an example where scaling terms rather than proportions should be considered as fixed could be provided.

To allow for such generality, we let

$$\mathbf{\Theta}^2(d) = \left(\frac{K_1}{d^{\lambda_1}}, \ldots, \frac{K_n}{d^{\lambda_n}}, \underbrace{\frac{K_{n+1}}{d^{\gamma_1}}, \ldots, \frac{K_{n+1}}{d^{\gamma_1}}}_{c(1,d)}, \ldots, \underbrace{\frac{K_{n+m}}{d^{\gamma_m}}, \ldots, \frac{K_{n+m}}{d^{\gamma_m}}}_{c(m,d)} \right). \tag{4}$$

There are thus $0 \leq n < \infty$ scaling terms appearing only once each, and $1 \leq m < \infty$ other scaling terms which appear $c(1,d), \ldots, c(m,d)$ times respectively in the d-dimensional vector, where $\lim_{d\to\infty} c(i,d) = \infty$ for each i.

To ease notation, we assume that the $n+m$ distinct scaling terms appearing in (4) appear at the first $n+m$ positions in $\mathbf{\Theta}^2(d)$ as $\left(K_1/d^{\lambda_1}, \ldots, K_n/d^{\lambda_n}, K_{n+1}/d^{\gamma_1}, \ldots, K_{n+m}/d^{\gamma_m}\right)$. Without loss of generality, we also suppose that the first n and next m scaling terms are respectively arranged according to an increasing order, i.e. $\lambda_n \leq \ldots \leq \lambda_1$ and $\gamma_m \leq \ldots \leq \gamma_1$. We let $b = \max(j \in \{1, \ldots, n\}; \lambda_j = \lambda_1)$ represent the number among the first n scaling terms which are of the same order as $\theta_1^2(d)$.

The scaling terms in (4) are not the most general form under which the theorems of the following sections are valid, but hopefully they allow the reader to get more intuition about the algorithm's behaviour. Optimal scaling results have been derived for target distributions as in (3), but where the θ_i's can be basically any function of d. For more details, we refer the reader to Bédard (2006a, 2006b, 2006c, 2007).

3.2 Space-time rescaling

In order to study the limit of the algorithm as $d \to \infty$, it is necessary to apply a space-time rescaling factor to the process. This was briefly mentioned in Section 2 for the *iid* model, but we now justify this step and adjust it for the target distribution introduced in Section 3.1.

By definition, the Metropolis algorithm is a discrete-time process. This process makes sense in finite dimensions, but becomes degenerate at the starting value $\mathbf{X}(0)$ in an infinite-dimensional setting. Indeed, a large number of dimensions means a large number of moves that are proposed independently of each other at every step. If we propose an infinite number of increments each coming from some fixed $N\left(0, \sigma^2\right)$ distribution, this implies that at least one of the proposed increments will be unacceptable and thus cause the rejection of the proposed infinite-dimensional move. The resulting process will then be a constant one. To compensate for this, we must then let the variance of proposed increments converge to 0 as $d \to \infty$, in a similar fashion to the approach of Roberts, Gelman, and Gilks (1997) who set $\sigma^2(d) = \ell^2/d$ for the *iid* model.

The present case is somewhat more complicated: the proposal variance selected should converge to 0 faster than any of the $\theta_i^2(d)$. It turns out that the optimal form for the proposal variance as a function of d is $\sigma^2(d) = \ell^2/d^{\varphi}$, where ℓ is some positive constant and φ is the smallest number satisfying

$$\lim_{d\to\infty} \frac{d^{\lambda_1}}{d^{\varphi}} < \infty \qquad \text{and} \qquad \lim_{d\to\infty} \frac{d^{\gamma_i} c(i,d)}{d^{\varphi}} < \infty, \quad \text{for } i = 1, \ldots, m. \tag{5}$$

With this modification, the magnitude of the jumps becomes smaller and smaller as d grows; since we are still dealing with an algorithm proposing jumps at every time unit, this algorithm with $\sigma^2(d) = \ell/d^{\varphi}$ eventually converges to a constant function (again!). To understand the limiting behaviour of the process, we must then speed up the process by the same factor as for the variance, i.e. d^{φ}.

The theory used to prove the weak convergence results of Sections 4 and 5 involves continuous-time Markov processes. The only way to preserve the Markov property of the discrete-time process while making it a continuous-time process is to resort to the memoryless property of the exponential distribution. That is, we let the time between each step be exponentially distributed with mean $1/d^{\varphi}$ so that the process jumps according to a Poisson process with rate d^{φ}; on average, the sped up process $\{\mathbf{W}^{(d)}(t)\}$ thus moves about d^{φ} times in every time unit. As $d \to \infty$, the jumps become closer in time (since the expected time between each proposed move is $d^{-\varphi}$) and we are then dealing with an asymptotically continuous process. This space-time rescaling of the original algorithm $\{\mathbf{X}^{(d)}(t)\}$ thus yields nontrivial limiting processes, which we study in the next section.

3.3 Adjusting $\boldsymbol{\Theta}^2(d)$

In Sections 4 and 5, we shall study the weak convergence of the rescaled process $\{\mathbf{W}^{(d)}(t)\}$ introduced in Section 3.2, for target densities as in Section 3.1. In particular, we choose a component of interest of the generated (rescaled) Markov chain, say $W_{i^*}^{(d)}$ ($i^* \in \{1, \ldots, n+m\}$), and study its limit as $d \to \infty$. Eventually, we shall study the limiting behaviour of each of the first $n+m$ components, but those are studied separately given the path of the d-dimensional algorithm up to time t.

The analysis of the process depends heavily on the selected component of interest. Indeed, every target component depends on the dimension through its own function $\theta_i^2(d)$. Some components are likely to have their probability mass concentrated in much narrower intervals of the state space than others (those with a smaller scaling term). These components have the opportunity to explore their space more efficiently, resulting in a faster convergence towards their invariant distribution. Because of this characteristic of the model, it is necessary to apply a scalar transformation to the target in order for the scaling term of the component of interest, $\theta_{i^*}^2(d)$, to be of order 1. For instance, if we have the scaling vector $\boldsymbol{\Theta}^2(d) = (4/d, 1, \ldots, 1)$ then we can easily study $\{W_2^{(d)}(t)\}$, the second component of the rescaled algorithm (or equivalently any of the last $d-1$ components), without performing any transformation to the target density. Studying $\{W_1^{(d)}(t)\}$ would however not make sence since the density of the first target component converges to 0 in probability as $d \to \infty$. A solution to this problem would thus be to transform the density and instead deal with $\boldsymbol{\Theta}^2(d) = (4, d, \ldots, d)$. Such transformations allow us to obtain nontrivial limiting processes for each of the $n+m$ different components included in the algorithm.

The target density in (3) is an extension of the *iid* model considered in Roberts, Gelman, and Gilks (1997); in particular, the densities they consider are special cases of (3) where the $\theta_i^2(d)$'s are all equal. We thus know that 0.234 shall be the AOAR for at least some scaling vectors $\boldsymbol{\Theta}^2(d)$, and we expect this conclusion to hold as long as the scaling terms in $\boldsymbol{\Theta}^2(d)$ are not too different from each other. A question of interest is then: How big a discrepancy between the $\theta_i^2(d)$'s must there be for the limiting behaviour of the algorithm to be affected? It turns out that the scaling terms of smaller order play an important part in answering this question: if none of the first n scaling terms in (4) is significantly smaller than any of the other $\theta_i^2(d)$'s, then the limiting behaviour of the algorithm is unaffected (with respect to that for the corresponding *iid* target density), and the AOAR is still 0.234. We shall see in Theorem 1 below that the key condition is that $\sum_{i=1}^{b} \theta_i^{-2}(d)$ is of strictly smaller order than $\sum_{i=1}^{d} \theta_i^{-2}(d)$, where b is as defined in Section 3.1.

4. OPTIMAL SCALING RESULTS

We now present the optimal scaling results, which can be separated into three distinct cases: in the first one, the limiting behaviour of the algorithm is the same as for the *iid* case; in the second and third cases, the algorithm is affected by the b components having significantly small scaling terms.

We denote weak convergence (in the Skorokhod topology) by $\Rightarrow$, standard Brownian motion at time t by $B(t)$, and the standard normal cumulative distribution function (*cdf*) by $\Phi(\cdot)$. In order to ease notation, we adopt the following convention for defining vectors: $\mathbf{X}^{(b-a)} = (X_{a+1}, \ldots, X_b)$; furthermore, $\mathbf{X}^{(b-a)-}$ means that the component of interest, X_{i^*}, is excluded from the vector.

4.1 An AOAR equal to 0.234

Theorem 1. *Consider a Metropolis algorithm whose moves are updated according to (1) with* $\mathbf{Z}^{(d)}(t+1) \sim N(0, \ell^2 I_d / d^{\varphi})$*, where* φ *satisfies (5). Suppose the algorithm is applied to the target density introduced in (3), with* $\boldsymbol{\Theta}^2(d)$ *as in (4) and* $\theta_{i^*} \equiv \sqrt{\theta_{i^*}^2(d)} = K_{i^*}^{1/2}$.

Consider the i^**-th component of the sped up process* $\{\mathbf{W}^{(d)}(t)\}$*, that is* $\{W_{i^*}^{(d)}(t)\}$*, and let the algorithm start in stationarity (i.e.,* $\mathbf{X}^{(d)}(0)$ *is distributed according to the target density* π *in (3)).*

Then $\{W_{i^*}^{(d)}(t)\} \Rightarrow \{W(t)\}$*, where* $W(0)$ *is distributed according to the density* $f(x/\theta_{i^*})/\theta_{i^*}$ *and* $\{W(t)\}$ *satisfies the Langevin stochastic differential equation (SDE)*

$$dW(t) = \upsilon_1(\ell)^{1/2}\, dB(t) + \frac{1}{2}\upsilon_1(\ell)\left(\log f\left(\frac{W(t)}{\theta_{i^*}}\right)\right)' dt,$$

if and only if

$$\lim_{d\to\infty}\frac{\sum_{i=1}^{b}\theta_i^{-2}(d)}{\sum_{i=1}^{d}\theta_i^{-2}(d)} = \lim_{d\to\infty}\frac{\sum_{j=1}^{b} d^{\lambda_j}}{\sum_{j=1}^{n} d^{\lambda_j} + \sum_{i=1}^{m} c(i,d)\, d^{\gamma_i}} = 0. \tag{6}$$

Here, $\upsilon_1(\ell) = 2\ell^2\Phi\left(-\ell\sqrt{E_R}/2\right)$ *and*

$$E_R \;=\; \lim_{d\to\infty}\sum_{i=1}^{m}\frac{c(i,d)}{d^{\varphi}}\frac{d^{\gamma_i}}{K_{n+i}}E\left[\left(\frac{f'(X)}{f(X)}\right)^2\right]. \tag{7}$$

Proof of Theorem 1. The main lines of the proof are similar to the theoretical justification provided in Section 2. To prove the weak convergence of the Metropolis algorithm to a Langevin diffusion process, we use the convergence theory of stochastic processes expounded in Ethier and Kurtz (1986). Specifically, we prove that the rescaled Metropolis algorithm is relatively compact, and that its generator converges to the generator of a Langevin diffusion. The proof relies heavily on Taylor expansions and the law of large numbers. For more details, see Bédard (2006a).

Remark. From the ordering of the scaling terms specified in Section 3.1, it is obvious that the smallest scaling term is either $\theta_1^2(d)$ or $\theta_{n+1}^2(d)$. However, since there are infinitely many components with scaling $\theta_{n+1}^2(d)$ in the limit, it is clear that $\theta_{n+1}^2(d)$ cannot be declared significantly smaller than the other scaling terms. This explains why the numerator in (6) depends on $\theta_1^2(d),\ldots,\theta_b^2(d)$ only.

The essence of Theorem 1 is that each component of the rescaled Metropolis algorithm asymptotically behaves according to a Langevin diffusion process. Even though the different components of $\mathbf{X}^{(d)}(t)$ depend upon each other in finite dimensions, they become independent in the limit as $d \to \infty$. As explained in Section 2, it suffices to choose the fastest diffusion in order to find the optimal value $\hat{\ell}$.

We compute that the speed measure of the Langevin diffusion $\upsilon_1(\ell)$ is maximised when $\ell = \hat{\ell} \doteq 2.38/\sqrt{E_R}$, implying that $\hat{\ell}$ varies inversely proportionally with $\sqrt{E_R}$. The latter is influenced both by the density f and the scaling vector $\mathbf{\Theta}^2(d)$. The smoother f is, the smaller is E_R and thus the larger is $\hat{\ell}$; this makes sense, since a smoother density indicates that we might afford to propose larger, less conservative moves. As far as the scaling vector is concerned, we conclude that $\theta_1^2(d),\ldots,\theta_n^2(d)$ do not affect the limiting process whatsoever in the present case. In fact, the only scaling terms impacting on $\hat{\ell}$ are those belonging to the groups for which $O(c(i,d)\,d^{\gamma_i}) = O(d^{\varphi})$, where $i \in \{1,\ldots,m\}$. For those $\theta_i^2(d)$'s affecting the value of E_R, the general rule is the following: the larger is the corresponding constant K_{n+i}, and thus the variance of the component, the larger is $\hat{\ell}$.

To illustrate this intuition, consider a d-dimensional normal target distribution centered around 0. For the standard normal density f, we first find that $E[(f'(X)/f(X))^2] = 1$. Now, suppose that the scaling vector is equal to $\mathbf{\Theta}^2(d) = (1,4,1,4,\ldots,1,4)$; that is, half of the components ($d/2$) have a variance equal to 1, and the other half have a variance equal to 4. From (5) we find that $\sigma^2(d) = \ell^2/d$, and from (7) we obtain $E_R = 1/2 + 1/2\cdot 1/4 = 5/8$, implying that $\hat{\ell} = 2.38/\sqrt{5/8} = 3.01$. If we had instead $\mathbf{\Theta}^2(d) = (1,4,4,1,4,4\ldots,1,4,4)$, i.e. two components with a variance of 4 for each component with a variance equal to 1, we would find $E_R = 1/3 + 2/3\cdot 1/4 = 1/2$, resulting in $\hat{\ell} = 2.38/\sqrt{1/2} = 3.37$. Since there is a larger proportion of smoother components in the second case, we can afford to be more aggressive in terms of proposed moves and that is why $\hat{\ell}$ is higher. In the case of a gamma target distribution with parameters $\alpha = 4$ and $\lambda = 1$, then $f(x) = x^{\alpha-1}\exp(-x)/\Gamma(\alpha)$ and we find that $E[(f'(X)/f(X))^2] = 1/2$; based on the measure of "roughness" $I = E[(f'(X)/f(X))^2]$ discussed in Section 2, we then deduce that the gamma(4,1) density is smoother than the standard normal density. Therefore, for the previously considered scaling vectors, we would find values of E_R smaller than those that were found for the normal targets respectively, which would result in higher values of $\hat{\ell}$.

As explained in Section 2, these results can also be expressed in terms of asymptotically optimal acceptance rates (AOARs) rather than asymptotically optimal variances of the proposal distribution. By using (2), where $\phi(d,\cdot)$ is now the *pdf* of a $N\left(\mathbf{0},\ell^2 I_d/d^{\varphi}\right)$ random variable and π is as in (3), we can show that

$$\lim_{d\to\infty} a_d(\ell) = 2\Phi\left(-\frac{\ell}{2}\sqrt{E_R}\right) \equiv a_*(\ell)\,;$$

hence, $a_*(\hat{\ell}) \doteq 0.234$ once again. As for the *iid* case, Theorem 1 is generally applicable to target distributions with as few as 15 dimensions.

4.2 A dilemma

When there exist b scaling terms that are significantly smaller than the others (that is, when (6) is violated and instead $\lim_{d\to\infty}\sum_{i=1}^{b}\theta_i^{-2}(d)/\sum_{i=1}^{d}\theta_i^{-2}(d) > 0$), we unfortunately cannot reach the same conclusion as in the previous section. Indeed, the components linked to these b scaling terms converge substantially faster than the others to their invariant distribution. Due to the size of their $\theta_i^2(d)$ they directly affect the proposal variance, i.e. the magnitude of the proposed steps; it is easy to see that when (6) is violated then $\varphi = \lambda_1$ in (5). In fact the components $\mathbf{W}^{(b)} = (W_1^{(d)},\ldots,W_b^{(d)})$ do not average out in the limit as was the case in Section 4.1 and their impact remains significant.

The case where (6) is violated can be divided in two further cases: the first one where $\theta_1^2(d),\ldots,\theta_b^2(d)$ are reasonably small compared to the other scaling terms (so the limit in (6) takes values in (0,1)), and the case where they are not (where the ratio in (6) converges to 1). We first consider the former; the latter shall be discussed in Section 5.

The limiting behaviour of the algorithm depends upon the component of interest selected. In particular, the limiting behaviour of the path followed by one of the components linked to $\theta_1^2(d),\ldots,\theta_b^2(d)$ differs from the limiting behaviour of any other component in the algorithm. Since there are two different asymptotic processes involved, and that components are collectively accepted or rejected at every step, we have to choose one of these processes only in order to derive the AOAR for the algorithm. To justify our choice, it is necessary to say more about each of these asymptotic processes.

If we select $\mathbf{W}^{(b)} = (W_1^{(d)},\ldots,W_b^{(d)})$ as the components of interest and study the limiting behaviour of these components simultaneously as $d\to\infty$ (through the convergence of generators, as usual), we find that $\mathbf{W}^{(b)}$ converges to the continuous-time version of a discrete-time Metropolis algorithm with acceptance rule

$$\begin{aligned}\alpha^*\left(\ell^2,\mathbf{X}^{(d)},\mathbf{Z}^{(d)}\right) \quad &= \quad \Phi\left(\frac{\sum_{j=1}^{b}\varepsilon\left(X_j,X_j+Z_j\right)-\ell^2E_R/2}{\sqrt{\ell^2E_R}}\right)\\ &\quad +\prod_{j=1}^{b}\frac{f\left(\left(X_j+Z_j\right)/\theta_j\right)}{f\left(X_j/\theta_j\right)}\Phi\left(\frac{-\sum_{j=1}^{b}\varepsilon\left(X_j,X_j+Z_j\right)-\ell^2E_R/2}{\sqrt{\ell^2E_R}}\right),\end{aligned}$$

where $\varepsilon\left(X_j,X_j+Z_j\right) = \log(f\left(\left(X_j+Z_j\right)/\theta_j\right)/f\left(X_j/\theta_j\right))$. Here, $\theta_j = K_j^{1/2}$ is just the square root of the scaling term $\theta_j^2(d)$, but we wish to emphasise the fact that this term is presently independent of d for the first b components (as explained in Section 3.3).

The components in $\mathbf{W}^{(b)}$ are thus asymptotically independent of the remaining $d-b$ components, but they are not independent of each other. In addition, they converge rapidly to their invariant distribution (in $O(1)$ iterations). This convergence to a discrete-time limiting process is related to the adjustment of $\mathbf{\Theta}^2(d)$ explained in Section 3.3. We remind the reader that in order to be able to study the first b components of the algorithm, we must set $\theta_1^2(d) = K_1,\ldots,\theta_b^2(d) = K_b$ (i.e. no dependence on d). Hence, $\varphi = \lambda_1 = 0$ and a space-time rescaling factor is not required for studying the limiting behaviour of the algorithm; the process is thus not accelerated and remains discrete.

It turns out that the components $W_{b+1}^{(d)},\ldots,W_d^{(d)}$ are asymptotically conditionally independent of each other given $W_1^{(d)},\ldots,W_b^{(d)}$. Once again this might be assessed by checking the convergence of the generator of the algorithm, for which $W_{b+1}^{(d)}$ is now the component of interest, to the generator of the corresponding diffusion process. In particular, given $\mathbf{W}^{(b)}$, each of these components independently behaves according to a diffusion process with drift

$$\ell^2 E\left[\exp\left(\sum_{j=1}^{b}\varepsilon\left(X_j,X_j+Z_j\right)\right)\Phi\left(\frac{-\sum_{j=1}^{b}\varepsilon\left(X_j,X_j+Z_j\right)-\ell^2E_R/2}{\sqrt{\ell^2E_R}}\right)\middle|\mathbf{X}^{(b)}\right]$$

and volatility

$$\begin{aligned}\ell^2\ E\Bigg[&\Phi\left(\frac{\sum_{j=1}^{b}\varepsilon\left(X_j,X_j+Z_j\right)-\ell^2E_R/2}{\sqrt{\ell^2E_R}}\right)\\&+\exp\left(\sum_{j=1}^{b}\varepsilon\left(X_j,X_j+Z_j\right)\right)\Phi\left(\frac{-\sum_{j=1}^{b}\varepsilon\left(X_j,X_j+Z_j\right)-\ell^2E_R/2}{\sqrt{\ell^2E_R}}\right)\Bigg|\mathbf{X}^{(b)}\Bigg].\end{aligned}$$

Furthermore, they each converge to their invariant distribution at the rate $O\left(d^{\lambda_1}\right)$. The convergence to their stationary distribution is slower than for $\mathbf{W}^{(b)}$, since their density is significantly more spread out over the state space (this is ensured mathematically by the fact that $\lambda_1 > \lambda_{b+1}$ and $\lambda_1 > \gamma_{n+1}$).

The impact of $\mathbf{W}^{(b)}$ on the limiting distribution of the algorithm keeps the components from being independent of each other; moreover, the differences between the scales of the target components imply that some components converge substantially faster than others to their invariant distribution. Determining the optimal scaling value $\hat{\ell}$ and the AOAR is thus a delicate task. Since there are two different types of limiting processes involved, which one should we choose to determine $\hat{\ell}$? For the algorithm to have converged to its invariant distribution, we should be confident that every distinct component has done so; considering that the speed of convergence is not the same for all components, we should then base our analysis on components that take longer to reach their invariant distribution.

We also face a second problem: the diffusive limit found does not possess a speed measure as was the case before, but instead nontrivial drift and volatility terms which both depend on $\mathbf{W}^{(b)}$. It is unclear how this process can be optimised to determine $\hat{\ell}$ (should we maximise the drift, minimise the volatility, etc.) In addition, even if we knew how to find $\hat{\ell}$, we would obtain an optimal scaling $\hat{\ell}(\mathbf{W}^{(b)})$ which is conditional on $\mathbf{W}^{(b)}$. To find a global optimal value $\hat{\ell}$, one solution would be to take the expectation of the diffusive process obtained with respect to $\mathbf{W}^{(b)}$. It turns out that this yields a Langevin diffusion process with speed measure $\upsilon_2\left(\ell\right)$, which can be optimised to find a global $\hat{\ell}$ and AOAR.

4.3 An AOAR Smaller than 0.234

We note that the averaging over $\mathbf{W}^{(b)}$ discussed in the previous section is equivalent to studying the marginal processes of the algorithm, i.e. studying each one-dimensional path of the algorithm *given its own past path only*. Focusing on the marginal processes makes sense since the last $d-b$ components, on which we base our analysis to determine $\hat{\ell}$, all possess the same limit. We can then infer a common value $\hat{\ell}$ for the algorithm. We have the following result, expressed in terms of marginal processes, again from Bédard (2006a).

Theorem 2. *Consider a Metropolis algorithm as in Theorem 1, applied to the same type of target distribution.*

We have that the marginal process $\{W_{i^*,M}^{(d)}\left(t\right)\}\Rightarrow\{W_M\left(t\right)\}$, *where* $W_M\left(0\right)$ *is distributed according to the density* $f(x/\theta_{i^*})/\theta_{i^*}$ *and* $\{W_M\left(t\right)\}$ *is as below, if and only if*

$$0<\lim_{d\to\infty}\frac{\sum_{i=1}^{b}\theta_i^{-2}\left(d\right)}{\sum_{i=1}^{d}\theta_i^{-2}\left(d\right)}<1. \qquad (8)$$

For $i^*=1,\ldots,b$ *with* $b=\max\left(j\in\{1,\ldots,n\}\,;\lambda_j=\lambda_1\right)$, *the limiting process* $\{W_M\left(t\right)\}$ *is the continuous-time version of a Metropolis algorithm with acceptance rule*

$$\alpha^*\left(\ell^2,X_{i^*},X_{i^*}+Z_{i^*}\right)\ =\ E\left[\Phi\left(\frac{\sum_{j=1}^{b}\varepsilon\left(X_j,X_j+Z_j\right)-\ell^2E_R/2}{\sqrt{\ell^2E_R}}\right)\right.$$

$$+\prod_{j=1}^{b}\frac{f\left(\left(X_j+Z_j\right)/\theta_j\right)}{f\left(X_j/\theta_j\right)}\Phi\left(\frac{-\sum_{j=1}^{b}\varepsilon\left(X_j,X_j+Z_j\right)-\ell^2E_R/2}{\sqrt{\ell^2E_R}}\right)\Bigg|\,X_{i^*},Z_{i^*}\Bigg]. \tag{9}$$

For $i^*=b+1,\ldots,d$, $\{W_M(t)\}$ satisfies the Langevin stochastic differential equation (SDE)

$$dW_M(t) \;=\; \upsilon_2(\ell)^{1/2}\,dB(t)+\frac{1}{2}\upsilon_2(\ell)\left(\log f\left(\frac{W_M(t)}{\theta_{i^*}}\right)\right)'dt,$$

where

$$\upsilon_2(\ell) \;=\; 2\,\ell^2\,E\left[\Phi\left(\frac{\sum_{j=1}^{b}\varepsilon\left(X_j,X_j+Z_j\right)-\ell^2E_R/2}{\sqrt{\ell^2E_R}}\right)\right]. \tag{10}$$

In both cases, $\varepsilon\left(X_j,X_j+Z_j\right)=\log\left(f\left(\left(X_j+Z_j\right)/\theta_j\right)/f\left(X_j/\theta_j\right)\right)$ and

$$E_R=\lim_{d\to\infty}\sum_{i=1}^{m}\frac{c\left(\mathcal{J}(i,d)\right)}{d^{\lambda_1}}\frac{d^{\gamma_i}}{K_{n+i}}E\left[\left(\frac{f'(X)}{f(X)}\right)^2\right]. \tag{11}$$

Furthermore, $\lim_{d\to\infty}a_d(\ell)=a_{**}(\ell)\equiv\upsilon_2(\ell)/\ell^2$.

Proof of Theorem 2. The main lines of the proof are similar to the proof of Theorem 1. For more details, see Bédard (2006a).

Besides determining whether or not there exist components converging to their stationary distribution significantly faster than others, condition (6) also is a statement about the dependence relationship among the chain components in the limit. Whenever (6) is satisfied, the limiting independence among the chain components is ascertained; when (6) is violated, then we can be positive that a dependence relationship among the components exists in the limit.

The acceptance rule (9) of the rescaled Metropolis algorithm is considerably more complex than the usual one (see Section 1). It still belongs to the Metropolis class (i.e., it yields an algorithm which is reversible with respect to π), but proposed moves are more difficult to accept under this rule than under the usual one. Initially, the acceptance probability of the d-dimensional Metropolis algorithm is taken to be $1\wedge\prod_{i=1}^{d}\{f\left(\left(X_j+Z_j\right)/\theta_j(d)\right)/f\left(X_j/\theta_j(d)\right)\}$. The ratio of densities involving the last $d-b$ components, $\prod_{i=b+1}^{d}\{f\left(\left(X_j+Z_j\right)/\theta_j(d)\right)/f\left(X_j/\theta_j(d)\right)\}$, behaves according to an $\exp\left(N\left(-\mu/2,\mu\right)\right)$ random variable, where μ is some positive constant. Therefore, the marginal acceptance rate in (9) is the solution of the following expectation,

$$E\left[1\wedge\prod_{i=1}^{b}\frac{f\left(X_j+Z_j\right)}{f\left(X_j\right)}e^{N(-\mu/2,\mu)}\right],$$

where the expectation is taken with respect to every random variable but the components of interest (X_1 and Z_1, say). Even though the resulting asymptotic process does not depend explicitly on the last $d-b$ components, these components still restrict this algorithm from mixing as freely as in the case of the regular Metropolis algorithm. We then realise that, although the last $d-b$ components do not have a big enough impact to keep the first b components from collectively behaving according to a Metropolis algorithm, they are numerous enough to cause the rejection of certain proposed moves and thus still play a role in the acceptance probability of the algorithm.

As mentioned earlier, the second marginal limiting process obtained is a (continuous-time) Langevin diffusion process as before, but is now governed by a different speed measure $\upsilon_2(\ell)$ given by (10). This function can be optimised to find $\hat{\ell}$, and applied to the limiting average acceptance rate to find the AOAR $a_{**}(\hat{\ell})$. Unfortunately, we cannot obtain closed-form solutions for these quantities; however, $\upsilon_2(\ell)$ can easily be optimised numerically. The AOARs resulting from this equation are not constant anymore, but vary with the target distribution.

As a general rule, both $\hat{\ell}$ and the AOAR decrease as the number of components affecting the speed measure increases (i.e. b and/or the number of components playing a role in the value of E_R). Conversely, $\hat{\ell}$ and the AOAR are larger when the target density is smoother (either through f or $\mathbf{\Theta}^2(d)$). It is worth noticing however that the AOAR cannot exceed 0.234; indeed, since the proposal distribution is formed of *iid* components, then the proposed algorithm can hardly do better than for the case of *iid* target distributions.

5. INHOMOGENEOUS PROPOSALS

5.1 The problem - An inefficient algorithm

We finally consider the case where the smaller scaling terms, $\theta_1^2(d), \ldots, \theta_b^2(d)$, are unreasonably small compared to the other ones (i.e., the ratio in (6) converges to 1). In such a situation, the probability mass of the components $\mathbf{W}^{(b)}$ is concentrated in such a narrow interval of the state space that the algorithm has to propose tiny steps in order for the proposed moves to be accepted. Unfortunately, the size of the proposed steps is not sufficiently large to ensure a reasonable convergence rate for the other components; consequently, in large dimensions, $\mathbf{W}^{(b)}$ are the only components to affect the acceptance rate of the algorithm.

We have the following result, again from Bédard (2006a). As discussed in Sections 4.2 and 4.3, we present the weak convergence result in terms of marginal distributions for each component of the process.

Theorem 3. *In the setting of Theorem 2 except with condition (8) replaced by*

$$\lim_{d\to\infty} \frac{\sum_{i=1}^{b} \theta_i^{-2}(d)}{\sum_{i=1}^{d} \theta_i^{-2}(d)} = 1, \tag{12}$$

the conclusions of Theorem 2 are preserved, except that the acceptance rule (9) for the modified Metropolis algorithm is now replaced by

$$\alpha^*\left(X_{i^*}, X_{i^*}+Z_{i^*}\right) = E\left[1 \wedge \prod_{j=1}^{b} \frac{f\left(\left(X_j+Z_j\right)/\theta_j\right)}{f\left(X_j/\theta_j\right)} \middle| X_{i^*}, Z_{i^*}\right],$$

and the speed measure $\upsilon_2(\ell)$ in (10) for the limiting Langevin diffusion is replaced by

$$\upsilon_3(\ell) \;=\; 2\,\ell^2\, P\left(\sum_{j=1}^{b} \varepsilon\left(X_j, X_j+Z_j\right) > 0\right).$$

Furthermore, $\lim_{d\to\infty} a_d(\ell) = a_{***}(\ell) \equiv \upsilon_3(\ell)/\ell^2$.

Proof of Theorem 3. The main lines of the proof are similar to the proof of Theorem 2. For more details, see Bédard (2006a).

From the previous result, we see that $\mathbf{W}^{(b)}$ asymptotically behaves according to a b-dimensional Metropolis algorithm with the usual acceptance rule. In large dimensions, the vector of components $\mathbf{W}^{(d-b)}$ has thus no effect whatsoever on the algorithm; indeed, the proposed increments are so small compared to the region where $\mathbf{W}^{(d-b)}$ has the majority of its probability mass that the components forming this vector always accept the proposed moves.

This phenomenon is also observed through the continuous-time limiting Langevin diffusion process obtained for the components in $\mathbf{W}^{(d-b)}$. Again, given the vector $\mathbf{W}^{(b)}$ of components ruling the chain, every component in $\mathbf{W}^{(d-b)}$ is conditionally independent of the others. The difference with the previous case lies in the fact that the term E_R, which used to affect the value of the speed measure of the diffusion in Theorems 1 and 2, is now equal to 0. This results in a speed measure $\upsilon_3(\ell)$ that is unbounded as a function of ℓ, and thus in an optimal value $\hat{\ell}$ approaching ∞ and an AOAR converging towards 0 as d grows. Since the increments proposed are of order $O\left(d^{-\lambda_1}\right)$, the optimal value $\hat{\ell}$ is pushed towards higher values to compensate for the fact that the order of the proposed moves is very small. Contrarily to the preceding case in Section 4.3, it is however not possible to reach an equilibrium between $\hat{\ell}$ and the order of $\sigma^2(d)$, hence the divergence of $\hat{\ell}$. In large dimensions, the homogeneous Metropolis algorithm considered in this paper is not only inefficient, but completely useless.

5.2 The solution - Modifying the proposals

The obvious solution to this problem is the use of inhomogeneous Metropolis algorithms; that is, we want to allow the proposal variance to vary for different components of the algorithm. Instead of dealing with $\mathbf{Z}^{(d)} \sim N\left(0, \ell^2 I_d / d^{\varphi}\right)$ we work with $\mathbf{Z}^{(d)} \sim N\left(0, \mathbf{v}^T(d)\, I_d\right)$, where $\mathbf{v}\,(d)$ is a d-dimensional vector of variances. The principle is simple: the problematic of the previous section lies in the fact that the algorithm is totally ruled by a finite number of components only as $d \to \infty$; the solution is then to adjust the proposal scalings for the components in $\mathbf{W}^{(d-b)}$ (i.e., make $E_R > 0$) so that these components can mix faster and have a significant impact on the overall behaviour of the algorithm.

There are many ways of choosing an inhomogeneous scheme that will yield a finite optimal scaling value $\hat{\ell}$. The simplest extension to the homogeneous approach is to use a vector of proposal variances exhibiting only two different values. In particular, the variance of a given component would either be ℓ^2/d^{φ} or ℓ^2/d^{φ_*}, which use the same constant ℓ, but different powers of d. In particular, we would have $v_i\,(d) = \ell^2/d^{\varphi_i}$ where $\varphi_i = \varphi$ as in (5) for $i = 1, \ldots, n$; that is, the proposal variance for the first n terms remains the one that was initially determined through the homogeneous method. For $i = n+1, \ldots, d$, we would then let $\varphi_i = \varphi_*$, where φ_* is the power of d that would be selected in (5) if we were ignoring the first n scaling terms $\theta_1^2\,(d), \ldots, \theta_n^2\,(d)$. In other words, φ_* is the smallest number such that

$$\lim_{d\to\infty} \frac{d^{\gamma_i} c\,(i,d)}{d^{\varphi_*}} < \infty, \quad \text{for } i = 1, \ldots, m. \tag{13}$$

This method is the simplest one ensuring a finite value $\hat{\ell}$ since it makes sure that besides $\mathbf{W}^{(b)}$, the target components associated with at least one of the m groups of scaling terms appearing an infinite number of times in the limit also asymptotically affect the accept/reject ratio of the algorithm. Under this scheme, the algorithm adopts the behaviour described in Theorem 2 but with

$$E_R \quad = \quad \lim_{d\to\infty} \sum_{i=1}^{m} \frac{c\,(i,d)}{d^{\varphi_{n+i}}} \frac{d^{\gamma_i}}{K_{n+i}} E\left[\left(\frac{f'\,(X)}{f\,(X)}\right)^2\right], \tag{14}$$

and therefore it suffices to optimise (10) in order to find $\hat{\ell}$ (which shall be finite, since E_R is now strictly positive).

It is worth mentioning that basically any reasonable inhomogeneous scheme with $v_i\,(d) = \ell^2/d^{\varphi_i}$ will yield similar results. At the other extreme the most complex scheme, where the orders of the proposal variances are exactly suited to the corresponding group of components, also behaves as in Theorem 2. Under this scheme, we have $\varphi_i = \lambda_i$ for $i = 1, \ldots, n$; for the last $d-n$ components, we then allow each of the m groups of scaling terms growing with d to have a different proposal variance. Each group being formed of *iid* components, it thus suffices to choose the m different powers of d to be the smallest m values such that the m limits in (13) be all finite. The only difference among these various inhomogeneous methods lies in the value of E_R, which gets bigger as we personalise the proposal variances to better suit the corresponding target components. As a result of this better fitted proposal distribution, larger values of $\hat{\ell}$, and correspondingly of the AOAR, are tolerated as optimal. The inhomogeneous methods discussed here all contribute to improve the rate of convergence of various groups of target components to their invariant distribution.

For further details about these results under more general assumptions for $\mathbf{\Theta}^2\,(d)$ and for the proofs of the theorems, we refer the reader to Bédard (2006a).

6. EXAMPLE

We now present a simple example illustrating how the previous results can be applied to optimise the speed of convergence of a Metropolis algorithm with a Gaussian proposal distribution.

Let f be the standard normal density, and consider a d-dimensional target density as in (3); that is, π is a multivariate normal distribution with a diagonal covariance matrix. This toy example is used extensively in the literature, and thus allows us to compare the present theory with previously published results. Furthermore in the case of a normal target, the scaling terms are the variances of each of the d individual components, which gives an intuitive feeling for what is happening.

We shall consider three different scaling vectors, $\mathbf{\Theta}_1^2\,(d)$, $\mathbf{\Theta}_2^2\,(d)$, and $\mathbf{\Theta}_3^2\,(d)$, each one exemplifying one of the three different cases introduced in Sections 4 and 5. In each situation, we shall perform simulation studies to illustrate the validity of the theorems.

6.1 First case

Firstly, let the component variances be $\mathbf{\Theta}_1^2(d) = (d, 9/d, 1, 9/d, 1, \ldots)$; that is, there is only one non-replicated variance ($n = b = 1$), and two groups of variances that are replicated $O(d/2)$ times each. Specifically, $m = 2$ with cardinality functions $c(1, d) = c(2, d) = (d-1)/2$. From (5) we realise that $\varphi = 2$, hence the proposed increments for the algorithm shall be generated from a normal distribution with mean 0 and variance $\sigma^2(d) = \ell^2/d^2$.

The first target component possesses a density function which is much more spread out over the real line than the other components. As the dimension of the target density gets larger, this discrepancy is amplified. Nonetheless, this situation does not prevent condition (6) from being satisfied; indeed, $W_1^{(d)}$ can hardly cause the rejection of a move since pretty much any move generated from the proposal distribution is suitable for this component:

$$\frac{\theta_1^{-2}(d)}{\sum_{i=1}^{d} \theta_i^{-2}(d)} = \frac{1/d}{1/d + d(d-1)/18 + (d-1)/2} \to 0 \quad \text{as} \quad d \to \infty.$$

From Theorem 1, it then follows that tuning the algorithm to accept 23.4% of the proposed increments is asymptotically optimal. What is the corresponding optimal value $\hat{\ell}$? Given that f is the standard normal density, $E[((\log f(X))')^2] = 1$ is easily computed. It directly follows that E_R in (7) is equal to 1/18, and thus $\hat{\ell} \approx 2.38\sqrt{18} = 10.1$.

Note here that the scaling vector $(9/d, 1, 9/d, 1 \ldots)$ would yield identical results; therefore $W_1^{(d)}$, the component which is scaled according to d, does not affect the asymptotic behavior of the algorithm in any way. In fact, the only terms affecting the value of E_R are those having a variance which is $O(d^{-1})$. In the present setting, components with a variance of 1 have a density which is too spread out over $\mathbf{R}$ to affect the accept/reject ratio of the algorithm in the limit. The only role they play here is to limit the cardinality function $c(1, d)$ to be $O(d/2)$ (rather than $O(d)$ in the *iid* case), which results in a larger value for $\hat{\ell}$ ($\hat{\ell} = 10.1$ versus $\hat{\ell} = 7.14$ for the *iid* case).

6.2 Second case

Secondly, we define

$$\mathbf{\Theta}_2^2(d) = (\underbrace{1/d, \ldots, 1/d}_{5}, 25, 25, \ldots);$$

this time there are many variances which are not replicated as d grows ($n = 5$), and only one group ($m = 1$) of replicated variances with cardinality function $c(1, d) = d - 5$.

With this new setting, we realise that the proposal variance is of the form $\sigma^2(d) = \ell^2/d$. There is also a slight noticeable change in the application of (5) when determining φ; the first five scaling terms now yield a finite limit in (5) (contrarily to the previous case with $\mathbf{\Theta}_1^2(d)$). This means that they play an active role in the determination of φ, and that they actually affect the accept/reject ratio of the method. This shall, as we may guess, be embodied when verifying

$$\lim_{d\to\infty} \frac{d}{5d + (d-5)/25} \to 0.1984 > 0 \quad \text{as} \quad d \to \infty.$$

In other words, the first five scaling terms in $\mathbf{\Theta}_2^2(d)$ are small enough compared to the other ones so that (8) be satisfied, and this difference in size is not large enough so that (12) be satisfied. Consequently, Theorem 2 applies and we must now solve for the exact $\hat{\ell}$ and AOAR using (10). It is found that $\hat{\ell} = 1.95$ and AOAR $= 0.177$.

Note that the AOAR found is fairly close to 0.234, a phenomenon due to the excessive regularity of the target distribution. In order to find an AOAR somewhat different in the case of a normal target, it was necessary to have a large group of scaling terms substantially smaller than the others ($b = 5$); that is, a large number of target components whose density is concentrated in a small portion of the state space. For other types of target density, it is worth mentioning that the AOAR is not necessarily close to 0.234. In fact, the further the target is from the normal distribution, the more likely it is to find an AOAR notably distinct from that value.

To illustrate this, let us see what happens when dealing with a non-normal density f. Specifically, consider the case of a log-gamma density with parameters α, μ, and σ

$$f(x \,|\alpha, \mu, \sigma) = \frac{e^{\alpha\left(\frac{x-\mu}{\sigma}\right) - e^{\frac{x-\mu}{\sigma}}}}{\sigma\Gamma(\alpha)}, \quad \alpha > 0, \sigma > 0.$$

In the case of a standard log-gamma, with $\mu = 0$ and $\sigma = 1$, we then have that $X = \log Y$, where $Y \sim \text{gamma}\,(\alpha, 1)$. We decide to focus on a d-dimensional target distribution with $\alpha = 4$, $\mu = 0$, and scaling vector $\sigma^2 = \mathbf{\Theta}^2\,(d) = (1/d, 1/d, 25, \ldots, 25)$. This scaling vector is similar to the previous one, except that $n = 2$, and $c\,(1, d) = d - 2$. The form of the proposal variance still is $\sigma^2\,(d) = \ell^2/d$; (8) is verified (with a limit now converging to 0.4902), and thus Theorem 2 is applicable once again. Before optimising (10) to find $\hat{\ell}$, we have to compute E_R in (11). For our standard f, i.e. $\alpha = 4$, $\mu = 0$, and $\sigma = 1$, it is found that $E[((\log f\,(X))')^2] = \alpha = 4$; this yields $E_R = \lim_{d \to \infty} 4(d-2)/25d = 0.16$. Optimising (10), it is thus found that $\hat{\ell} = 2.08$ and AOAR $= 0.091$, a value significantly smaller to 0.234. Indeed from the graph in Figure 1, we realise that settling for an acceptance rate of 0.234 in this case would result in an algorithm that performs suboptimally. From this example, we also observe that the theorems presented previously seem quite robust to the violation of some of the regularity assumptions on the density f. Although the log-gamma density does not satisfy the Lispchitz continuity assumption of $(\log f)'$, it is interesting to remark that the theoretical conclusion obtained with Theorem 2 closely agrees with the simulations. This suggests that some of the assumptions stated in Section 3.1 could be even further relaxed.

6.3 Third case

Had we chosen to define a fixed number of target components with scaling terms smaller than $O\,(1/d)$ in the previous cases, we would have faced a situation where the first b scaling terms rule the algorithm and the AOAR converges to 0. This is the case we now discuss.

Let $\mathbf{\Theta}_3^2\,(d) = \left(1/d^2, 1/d, 1, d, 1, d, \ldots\right)$; the variance of the first component is much smaller than the variances of the other components. In fact in large dimensions, $W_1^{(d)}$ totally governs the algorithm. Thus, (12) is satisfied and Theorem 3 applies. As illustrated in Figure 1, this results in the inefficiency of the algorithm (i.e. $\hat{\ell} \to \infty$ and AOAR $\to 0$ as $d \to \infty$).

To improve the situation, we therefore turn to inhomogeneous proposal distributions. Under the simplest scheme described in Section 5.2, we would use the vector of proposal variances $\mathbf{v}^T(d) = \ell^2 \left(1/d^2, 1/d^2, 1/d, 1/d, \ldots\right)$; increasing in complexity, we could also use the vector $\mathbf{v}^T(d) = \ell^2 \left(1/d^2, 1/d, 1/d, 1, 1/d, 1, \ldots\right)$. In both cases it suffices to optimise (10), the speed measure of Theorem 2, to find the optimal scaling values and AOARs (according to (14), $E_R = 1/2$ and $E_R = 1$ for the simpler and more complex schemes respectively). Of course, the second scheme produces a more efficient method as each target component converges as fast as possible. The values obtained are as follows: under the simpler scheme, $\hat{\ell} = 7.2$ and the AOAR is 0.1827; under the other scheme, $\hat{\ell} = 3$ and the AOAR is 0.1866. It is interesting to remark that these two AOARs are quite close to each other, but the optimal scaling value of the simpler scheme is noticeably higher. Since the proposal variances suit the target components better in the more complex case, the value $\hat{\ell}$ needs not compensate for the lack of precision present in the simpler case.

Before concluding this section, we present a simulation study depicting the previous conclusions. The graphs in Figure 1 illustrate the efficiency of the Metropolis algorithm versus the acceptance rate for each of the three cases considered. In each graph the solid line represents the theoretical curve of $\upsilon\,(\ell)$ versus $a\,(\ell) = \upsilon\,(\ell)\,/\ell^2$. The dotted curves have been obtained by running the algorithm in R, and each curve in a given graph is the result of a simulation study for a particular value of d. The curves themselves are obtained by running many Metropolis algorithms with different proposal variances; each point is thus the result of a 100,000-iteration run for 50 different values of ℓ^2 (but for a constant d).

For each of the runs, we estimated the efficiency of the method using the average squared jumping distance of the algorithm, $\sum_{j=1}^{d} \sum_{i=1}^{N} (X_j^{(d)}\,(i) - X_j^{(d)}\,(i-1))^2/Nd$, where N is the number of iterations performed (see Pasarica and Gelman 2003; Roberts and Rosenthal 1998); we also estimated the acceptance rate by the proportion of accepted moves in the algorithm, $\sum_{i=1}^{N} \mathbf{1}(\mathbf{X}^{(d)}\,(i) \neq \mathbf{X}^{(d)}\,(i-1))/N$.

In each of the graphs, we see that even in low dimensions, the behaviour of the algorithm is rather close to its asymptotic counterpart. Furthermore, the peak of the curves are very close to the theoretical values specified earlier. To obtain the curves in the first and second cases, we respectively used the variances $\ell^2 = 35 + (1 : 50) * 3$ and $\ell^2 = (1 : 50) * .15$. For the log-gamma, we used $\ell^2 = (1 : 20) * .15, (21 : 25) * .18, (26 : 50) * .25$. For the last case, $\ell^2 = (1 : 20) * 2 + (1 : 50) * 40$ was used for the homogeneous case, while $\ell^2 = (1 : 50) * .6$ and $\ell^2 = (1 : 50) * .25$ were used for the simplest and more complex inhomogeneous schemes respectively.

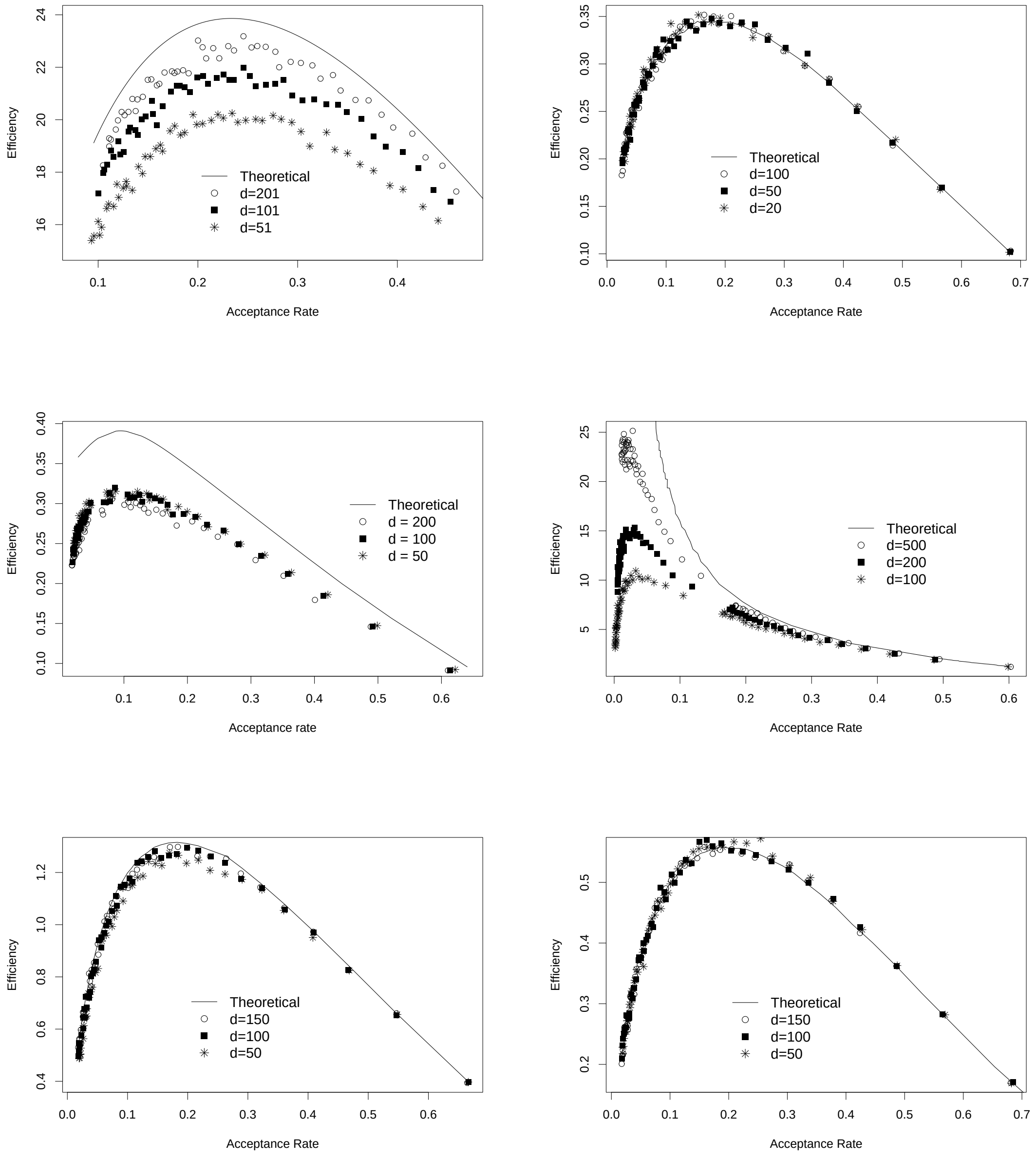

Figure 1: Efficiency versus acceptance rate for the Metropolis algorithm. The plotted symbols are the results of simulation studies in different dimensions, while the solid lines represent the theoretical curves [i.e., $\upsilon(\ell)$ versus $\upsilon(\ell)/\ell^2$]. From left to right and top to bottom, we have: 1st case, 2nd case - normal, 2nd case - log-gamma, 3rd case with homogeneous proposal variances, 3rd case with the simpler inhomogeneous scheme, and 3rd case with the complex inhomogeneous scheme.

7. DISCUSSION

This paper has presented optimal scaling results for the Metropolis algorithm with a Gaussian proposal distribution, and has focused on the case where the target density is formed of independent but not identically distributed components (see Bédard 2006a, 2006b, 2006c, 2007). The results obtained in Bédard (2006a) are the first instance of AOARs that differ from the usual constant 0.234.

From the theoretical results presented along with the example, we concluded that AOARs for target distributions that are similar to the Gaussian proposal are fairly robust to moderate and large discrepancies in the scale of f in (3). By opposition, target distributions whose behaviour greatly differs from that of a Gaussian density are not very robust to moderate discrepancies in the scale of f, and yield AOARs that are much smaller than 0.234.

We might then wonder if, in the same way 0.234 is robust to small discrepancies in the scale of f in (3), our results are also robust to weak correlation structures among the target components. It turns out that they are for very weak dependency, but as soon as the dependency gets stronger, the AOAR gets further from (in fact, smaller than) those prescribed by Theorems 1, 2, and 3. Generally speaking, algorithms that propose smarter moves enjoy higher AOARs (e.g. the MALA). This makes sense since such algorithms use properties from the target distribution in order to propose moves that are suitable for the target (at the cost of extra computations). In the present case, the chosen proposal distribution with independent components is obviously best suited for target distributions with independent components. Therefore the algorithm considered cannot, overall, propose moves that suit a correlated target better than they suit a target with no correlation. Consequently, AOARs for the algorithm considered are lower in the case of correlated target distributions than they are for target distributions with independent components. Nonetheless, when it comes to derive general optimal scaling results, proposal distributions with independent components constitute the natural choice since they lie at the middle ground between positive and negative correlation structures.

Besides providing optimal scaling results for the target distribution in (3) as well as a better understanding of the limiting behaviour of multidimensional Metropolis algorithms, the results introduced in this paper might then serve as a guideline to scale Metropolis algorithms for more complex target distributions. Particularly, we remind the reader that the AOARs found in Theorems 1, 2, and 3 act as an upper bound for target distributions with correlated components. For the log-gamma density of Section 6 for instance, introducing correlation between the components would result in decreasing the AOAR of 0.091; we might then want to use an acceptance rate between say 0.05 and 0.09 for sampling from such a target distribution, depending on the strength of the correlation. As far as other types of algorithms are concerned, then a very crude rule of thumb would be to favour higher acceptance rates for algorithms with smarter proposals, and lower ones for algorithms that would propose moves that are in general not as good as those considered here.

In the special case where the joint target density can be expressed as a multivariate normal density it is worth mentioning that the results discussed in this paper are valid, regardless of the covariance matrix of the distribution. This is due to the invariance of multivariate normal distributions under orthogonal transformations; we can thus transform the target density into a d-dimensional normal density with independent components, and then apply the appropriate theorem to obtain the exact $\hat{\ell}$ and AOAR. For more details on this, we refer the reader to Bédard (2006a, 2006c).

Contrarily to results previously published in the literature, the theorems presented herein require that we study not only each one-dimensional path of the joint process, but also each of these one-dimensional processes marginally.

This has thus raised the question as to whether or not we could use the same marginal approach to derive results for target distributions with dependent components. Unfortunately, it does not seem sensible nor possible to ignore the dependence among the various target components by focusing on the marginal processes generated by the algorithm to derive asymptotic optimal scaling results. In fact, ongoing investigation is telling us that we should persist with the usual method and deal with target components that shall often be asymptotically dependent on each other. In some cases, for example hierarchical target distributions, the asymptotics of the algorithm even seem to point towards more efficient alternatives to sample from the target distribution, such as the Gibbs sampler or Metropolis-within-Gibbs. Optimal scaling results for these methods are difficult to obtain (though see Neal and Roberts 2006), due to the fact that they involve updating a finite number of components at every step. According to simulation studies, the Metropolis algorithm discussed in this paper seems nonetheless a reasonably efficient sampling method for hierarchical target distributions, and optimal scaling results can be developed for this case.

As the target distribution becomes more complex, the proofs of associated optimal scaling results become considerably more tedious. Proving weak convergence of processes using Dirichlet forms, as discussed in Bédard and Kendall (2008), might be an interesting approach as it involves first-order asymptotics only and thus simplify the proofs considerably. The optimal scaling results published in the literature have one main point in common: they rely heavily on the assumption that f, the target components density, is continuous over $\mathbf{R}$. Recently, discontinuous target distributions have received some attention; in particular, Neal, Roberts, and Yuen (2007) showed that AOARs exist for such targets and depend on the points of discontinuity of the target density.

Finally, it seems possible to derive optimal scaling results for fancier types of algorithms; Bédard, Fort, and Moulines (2008) features a comparison of the asymptotic behaviour of the Multiple Try Metropolis algorithm and the Delayed Rejection Metropolis algorithm (for more information on these sampling methods, see Liu, Liang, and Wong (2000) and Mira (2001), respectively). Both these algorithms rely on the idea that many moves should be generated at every step, and one of them chosen as the proposed value. It turns out that the second of these methods behaves erratically in the limit, while the other admits a nice asymptotic process.

APPENDIX - CONVERGENCE OF MARKOV CHAINS TO STATIONARY DISTRIBUTIONS

We briefly review the notion of convergence of Markov chains to their stationary distribution.

Consider a Markov chain $\{X_n; n \geq 0\}$ on a continuous state space $\mathcal{X}$, with k-step transition law $P^k(x, A) = P(X_k \in A | X_0 = x)$ for all $x \in \mathcal{X}$, $k \geq 0$, and measurable $A \subseteq \mathcal{X}$.

Definition 1. A probability measure $\pi(\cdot)$ is a stationary distribution if

$$\pi(A) = \int_{x \in \mathcal{X}} \pi(dx)\, P^k(x, A)$$

for all measurable $A \subseteq \mathcal{X}$ and for all $k \geq 0$.

Although $\pi(\cdot)$ is stationary for the Markov chain, this does not necessarily ensure that $\pi(\cdot)$ and $P^k(x, \cdot)$ are close for large k. We introduce below two properties of Markov chains, irreducibility and aperiodicity: the first one ensures that the chain can eventually visit any measurable region of the state space, while the other one keeps the chain from cycling through some sets of states.

Definition 2. A Markov chain is ϕ-irreducible if there exists a non-zero measure ϕ on $\mathcal{X}$ such that for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$, and for all $x \in \mathcal{X}$, there exists a postive integer k such that $P^k(x, A) > 0$.

Definition 3. A Markov chain is aperiodic if there do not exist $d \geq 2$ disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d \subseteq \mathcal{X}$ with $\pi(\mathcal{X}_i) > 0$, such that $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i$ $(1 \leq i \leq d-1)$, and $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$.

We then have the following result about the convergence in total variation distance of $P^k(x, A)$ to $\pi(\cdot)$.

Theorem 4. *If a Markov chain is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, then for π-a.e. $x = X_0 \in \mathcal{X}$,*

$$\lim_{k \to \infty} ||P^k(x, \cdot) - \pi(\cdot)|| = \lim_{k \to \infty} \sup_A |P^k(x, A) - \pi(A)| = 0.$$

In particular,

$$\lim_{k \to \infty} P^k(x, A) = \pi(A), \;\; A \subseteq \mathcal{X}.$$

Proof of Theorem 4. See, for instance, Nummelin (1984).

It is interesting to note that Theorem 4 applies to virtually any Metropolis-Hastings algorithm. In MCMC applications, we always start with the target distribution $\pi(\cdot)$ being stationary. Furthermore, it is usually easy to verify that the chain is irreducible by selecting ϕ to be the Lebesgue measure on the appropriate region. Finally, aperiodicity generally holds since we sometimes have a non-zero probability of staying at the same state during two consecutive steps. The various algorithms, however, do not all converge to their stationary distribution with the same speed.

ACKNOWLEDGEMENTS

Both authors received the support of the Natural Sciences and Engineering Research Council of Canada for pursuing this research project.

REFERENCES

M. Bédard (2006a). *On the Robustness of Optimal Scaling for Random Walk Metropolis Algorithms.* Ph.D. dissertation, Department of Statistics, University of Toronto.

M. Bédard (2006b). Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234. *To appear in Stochastic Process. Appl.*

M. Bédard (2006c). Efficient sampling using Metropolis algorithms: Applications of optimal scaling results. *To appear in J. Comput. Graph. Statist.*

M. Bédard (2007). Weak convergence of Metropolis algorithms for non-*iid* target distributions. *Ann. Appl. Probab.*, 17, 1222-44.

M. Bédard, G. Fort & E. Moulines (2008). Optimal scaling for the multiple-try Metropolis algorithm. *In preparation.*

M. Bédard & W. S. Kendall (2008). Weak convergence of RWM algorithms using Dirichlet forms. *In preparation.*

J. Besag & P. J. Green (1993). Spatial statistics and Bayesian computation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 55, 25-38.

J. Besag, P. J. Green, D. Higdon & K. Mergensen (1995). Bayesian computation and stochastic systems. *Statist. Sci.* 10, 3-66.

L. A. Breyer, M. Piccioni & S. Scarlatti (2002). Optimal scaling of MALA for nonlinear regression. *Ann. Appl. Probab.* 14, 1479-1505.

L. A. Breyer & G. O. Roberts (2000). From Metropolis to diffusions: Gibbs states and optimal scaling. *Stochastic Process. Appl.* 90, 181-206.

O. F. Christensen, G. O. Roberts & J. S. Rosenthal (2003). Scaling limits for the transient phase of local Metropolis-Hastings algorithms. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67, 253-69.

S. N. Ethier & T. G. Kurtz (1986). *Markov Processes: Characterization and Convergence.* Wiley.

A. Gelman, G. O. Roberts & W. R. Gilks (1996). Efficient Metropolis jumping rules. *Bayesian Statistics V*, 599-608, Clarendon Press, Oxford.

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

J. S. Liu, F. Liang & W. H. Wong (2000). The multiple-try method and local optimization in Metropolis sampling. *J. Amer. Statist. Assoc.* 95,121-34.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller & E. Teller (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087-92.

A. Mira (2001). On Metropolis-Hastings algorithms with delayed rejection. *Metron.* 59, 231-41.

P. Neal & G. O. Roberts (2006). Optimal scaling for partially updating MCMC algorithms. *Ann. Appl. Probab.* 16, 475-515.

P. Neal, G. O. Roberts & J. Yuen (2007). Optimal scaling of random walk Metropolis algorithms with discontinuous target densities. Technical Report.

E. Nummelin (1984). *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge Univ. Press.

C. Pasarica & A. Gelman (2003). Adaptively scaling the Metropolis algorithm using expected squared jumped distance. Technical report, Department of Statistics, Columbia University.

G. O. Roberts, A. Gelman & W. R. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* 7, 110-20.

G. O. Roberts & J. S. Rosenthal (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 60, 255-68.

G. O. Roberts & J. S. Rosenthal (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.* 16, 351-67.

C. Sherlock (2006). *Methodology for Inference on the Markov Modulated Poisson Process and Theory for Optimal Scaling of the Random Walk Metropolis.* Ph.D. dissertation, Department of Mathematics and Statistics, Lancaster University.

Received July 2007
Accepted June 2008

Mylène BÉDARD: bedard@dms.umontreal.ca
Département de mathématiques et de statistique
Université de Montréal
Montréal, Québec
Canada, H3C 3J7

Jeffrey S. ROSENTHAL: jeff@math.toronto.edu
Department of Statistics
University of Toronto
Toronto, Ontario
Canada, M5S 3G3