Controlled MCMC for Optimal Sampling

Christophe Andrieu

Department of Mathematics, University of Bristol, Bristol, U.K.

Christian P. Robert

Ceremade - Université Paris-Dauphine, Paris , France

Summary. In this paper we develop an original and general framework for automatically optimizing the statistical properties of Markov chain Monte Carlo (MCMC) samples, which are typically used to evaluate complex integrals. The Metropolis-Hastings algorithm is the basic building block of classical MCMC methods and requires the choice of a proposal distribution, which usually belongs to a parametric family. The correlation properties together with the exploratory ability of the Markov chain heavily depend on the choice of the proposal distribution. By monitoring the simulated path, our approach allows us to learn "on the fly" the optimal parameters of the proposal distribution for several statistical criteria.
Keywords: Monte Carlo, adaptive MCMC, calibration, stochastic approximation, gradient method, optimal scaling, random walk, Langevin, Gibbs, controlled Markov chain, learning algorithm, reversible jump MCMC.

1. Motivation

1.1. Introduction

Markov chain Monte Carlo (MCMC) is a general strategy for generating samples x_i ($i = 0, 1, \ldots$) from complex high-dimensional distributions, say π defined on the space $\mathcal{X} \subset \mathbb{R}^{n_x}$, from which integrals of the type

$$I(f) = \int_{\mathcal{X}} f(x)\,\pi(x)\,dx,$$

can be calculated using the estimator

$$\widehat{I}_N(f) = \frac{1}{N+1}\sum_{i=0}^{N} f(x_i),$$

provided that the Markov chain produced is ergodic. The main building block of this class of algorithms is the Metropolis-Hastings (MH) algorithm. It requires the definition of a proposal distribution q whose role is to generate possible transitions for the Markov chain, say from x to y, which are then accepted or rejected according to the probability†

$$\alpha(x,y) = \min\left\{1, \frac{\pi(y)\,q(y,x)}{\pi(x)\,q(x,y)}\right\}.$$

The simplicity and universality of this algorithm are both its strength and weakness. The choice of the proposal distribution is crucial: the statistical properties of the Markov chain

†Therefore if the chain is in state x it remains in this state with probability $r(x) = 1 - \int_{\mathcal{X}} \alpha(x,y)\,q(x,y)\,dy$.

will heavily depend upon this choice. A careless choice will often result in poor performance of the Monte Carlo estimators. In practice, MCMC algorithms combine several mechanisms such as the one described above, for different proposal distributions, or *strategies*, q_i. Typically the transition kernel is a mixture of different strategies, *i.e.*

$$K(x, dy) = \sum_{i=1}^{n} \omega_i K_i(x, dy), \tag{1}$$

where $\sum_{i=1}^{n} \omega_i = 1$, $\omega_i \geq 0$ and the K_i's are "simple" MH kernels.

In many cases the proposal and mixture distributions belong to parametric families $\{\omega_{i,\theta_i}, q_{i,\theta_i}; \theta_i \in \Theta_i \ \subset \mathbb{R}^{n_{\theta}{}_i}\}$, and it is of interest to choose $\theta = \{\theta_i \in \Theta_i\} \in \mathbb{R}^{n_\theta}$ such that the statistical properties of the Markov chain meet some criterion. When a criterion is chosen, it is then of interest to define a procedure that will determine, either automatically or "by hand", the set of best parameters. As we shall see, we propose here to address the automatic adaptation problem. Before presenting our results, we briefly review both existing criteria (Subsections 1.2 and 1.3) and algorithms that have been proposed either to implement these criteria or on a heuristic basis (Subsection 1.4). The reader familiar with such techniques can skip to Subsection 1.5 where we first introduce our approach.

1.2. Criteria for Global Adaptation

The aim is to learn a global strategy for the whole target distribution π. Three families of criteria have been proposed. The first one, particularly relevant for independence samplers, consists of matching some of the moments (*e.g.* mean, covariance) of the proposal distribution with those of the target distribution (Haario *et al.* 2001), (Laskey and Myers 2001). The motivation is typically to determine a parametric approximation of the target distribution, from which it is routine to sample from, *e.g.* mixture of normal distributions as in (Kim *et al.* 1998).

The second category aims at reducing the variance of $\widehat{I}_N(f)$, or in other words at maximizing the efficiency (Besag and Green 1992) of the Markov chain, which takes the following form in the scalar case

$$\text{eff} = \frac{var_\pi(f(x))}{var_K(f(x))}.$$

Here $var_K(f(x))$ is approximately N times the variance of $\widehat{I}_N(f)$ (the estimator of $I(f)$ built from samples generated with the transition kernel K) for N large. Motivated by this idea of efficiency maximization, significant theoretical developments have resulted in approximate optimal scaling of random walk Metropolis algorithms, when the target distribution is either a univariate Gaussian distribution, or a specific spherical Gaussian distribution, with a proposal distribution that needs to be scaled. These asymptotic results, supported by numerical simulations, show that optimal scaling in this case is achieved for a scaling of $2.38/\sqrt{n_x}$ (we recall that n_x is the dimension of x) and leads to an expected acceptance probability of 0.234, when n_x is large (Gelman *et al.* 1995), (Roberts *et al.* 1997). This suggests a practical way of tuning the scaling parameters in order to achieve this probability. This is certainly the most routinely used calibration method nowadays. Similar results exist for the Langevin based MH algorithm (Roberts *et al.*1998).

The third category was introduced in (Sahu and Zhigljavsky 1998). It involves building a proposal distribution, in the example developed by the authors a possibly infinite discrete mixture of distributions, which ensures that the acceptance probability of the sampler is greater than a given threshold. Although interesting, this approach relies on increasing both the computational burden and memory requirements. Indeed, the number of components of the mixture from which the samples are drawn increases and will typically go to infinity.

1.3. Criteria for Local Adaptation

In this case the aim is to learn local characteristics of the target distribution π. Such techniques include local analytical approximations of the target distribution. This, typically, involves the search for a local mode and the evaluation of the Hessian of the distribution at this point. It can, therefore, be interpreted as a local version of the moment matching criterion presented above (Bennet *et al.* 1996) (Chib *et al.* 1998). The Monte Carlo analogue of this approach is presented in (Haario *et al.* 1999), where a sliding window is used in order to estimate such quantities. However it is shown in (Haario *et al.* 1999) that the averages produced from the samples are biased. Another category consists of multiple tries that typically balance local and global information about the target distribution. This approach is presented in (Liu *et al.* 2000), where several attempts are performed in parallel, whereas the approach illustrated in (Geyer and Thompson 1995) consists of delaying the rejection of samples. More precisely, the idea is to develop strategies that allow for several successive tries in order to improve mixing of the chain. However this later approach is rather heuristic, and there is no explicit criterion being optimized (Green and Mira 2000), (Tierney and Mira 1999).

1.4. Learning Techniques

Calibrating the proposal distribution in view of the past history of the chain, although tempting, can lead to a loss of its ergodic properties. This is why one can distinguish between two main categories of algorithms.

- *Markovian algorithms.* This category includes the delayed rejection and multiple try approaches which preserve the correct target distribution (Geyer and Thompson 1995), and for which general Markov chain ergodicity theorems apply. "Population Monte Carlo" or parallel Monte Carlo techniques (Chauveau and Vandekerkhove 2001), (Gilks *et al.* 1994), (Laskey and Myers 2001) also belong to this class of algorithms. The idea consists of running a Markov chain with target distribution

$$\pi^{\otimes M}\left(x^1,\ldots,x^M\right)=\prod_{i=1}^{M}\pi\left(x^i\right),$$

  with M sufficiently large. In (Chauveau and Vandekerkhove 2001), M increases with the number of iterations, and the distance between an estimate of the current distribution of the Markov chain and the target distribution is calculated, and drives the adaptation of θ. In (Gilks *et al.* 1994) and (Laskey and Myers 2001), M is fixed and the mean, the covariance or any integral of the type $I\left(f\right)$ can be estimated, and used to update the parameter θ of the transition kernel K_θ. Note that these estimates are biased as long as equilibrium is yet to be reached. Regeneration techniques (Mykland *et al.* 1994) have also been proposed in order to design mechanisms that ensure that ergodicity is preserved by updating the proposal distribution only when visiting a regeneration set (Gilks *et al.* 1998). However, as pointed out by the authors, regeneration techniques are difficult to apply, especially for high dimensional problems.

- *Non-Markovian algorithms.*‡ These procedures use a single chain - or at least as we shall see later a reduced number of parallel chains - and rely on the idea that at iteration i, the history of the chain, *i.e.* the path $x_0, x_1, \ldots, x_{i-1}$, brings some useful information about the target distribution. A natural idea would then consist of using this path in order to estimate quantities of the type $I(f)$ which can then themselves be used in order to adjust the parameters of the proposal distribution. However it is clear that the chain might in this case lose both its markovianity and ergodicity (Gelfand and Sahu 1994). It is shown in (Haario *et al.* 2001) that ergodicity can be preserved for a restricted class of algorithms.

1.5. A Controlled Markov Chain Approach

In this paper we present a general framework which allows for the development of new MCMC algorithms that are able to automatically learn the best strategy among a set of proposed strategies q_θ, in order to explore the target distribution π. According to the taxonomy given in Section 1, the approach is global and uses the path of the chain, but can be combined with population Monte Carlo strategies in a straightforward manner. It is very close in spirit to what may be found in the automatic control literature, where it is used to optimize systems that can be modelled as Markov chains for example. We now explain the main features of the approach:

Firstly the definition of a cost function $h(\theta)$ is required. This cost function expresses some measure of the statistical performance of the Markov chain in its stationary regime, *e.g.* favor negative correlation between x_i and x_{i+l} for some lag l and $i = 0, 1, \ldots$. Two precise and useful examples are given in Subsection 2.1. The cost function is defined in such a way that any - unknown - optimum value θ_* is a root of $h(\theta) = 0$. For our MCMC control problem the optimal exploration of the target distribution can be formalized in most cases as that of finding the solutions of an equation of the type,

$$h(\theta) \triangleq \int_{\mathcal{X}\times\mathcal{Y}} H(\theta, (x, y))\, \pi_\theta(y|\, x)\, \pi(x)\, dxdy = 0. \tag{2}$$

The function H can be explicitly chosen against one of several statistical criteria for the chain. The distribution π_θ typically depends upon the proposal distribution. Examples of $H(\theta, (x, y)), y$ and π_θ are given in Subsection 2.1. For simplicity, we will use the notation $w = (x, y) \in \mathcal{W} \triangleq \mathcal{X} \times \mathcal{Y}$ and $\mu_\theta(w) = \pi_\theta(y|\, x)\, \pi(x)$.

Secondly an algorithm is needed in order to find the roots of this equation, which involves both integration and optimization. Stochastic approximation (SA) techniques have been especially tailored for this purpose (Robbins and Monro 1951) and are now well documented. SA algorithms are by nature iterative, and their basic recursion can be interpreted as a noisy gradient iteration,

$$\begin{aligned} \theta_{i+1} &= \theta_i + \gamma_{i+1} H(\theta_i, w_{i+1}) \\ &= \theta_i + \gamma_{i+1} h(\theta_i) + \gamma_{i+1} \underbrace{[H(\theta_i, w_{i+1}) - h(\theta_i)]}_{e_{i+1}}. \end{aligned} \tag{3}$$

‡We mean here that the chain x_i is not markovian, whereas (θ_i, x_i) might be.

The γ_i's are slowly decreasing steps and x_i, y_i are updated according to the transition probabilities

$$\begin{aligned}\mathbb{P}\left(dx_i \middle| w_0, \ldots, w_{i-1}, \theta_0, \ldots, \theta_{i-1}\right) &= K_{\theta_{i-1}}\left(x_{i-1}; dx_i\right)\\ \mathbb{P}\left(dy_i \middle| w_0, \ldots, w_{i-1}, x_i, \theta_0, \ldots, \theta_{i-1}\right) &= K'_{\theta_{i-1}}\left(x_i, y_{i-1}; dy_i\right),\end{aligned}$$

such that for any $(\theta, x) \in \Theta \times \mathcal{X}$,

$$\begin{aligned}\pi K_\theta &= \pi\\ \pi_\theta\left(\cdot \middle| x\right) K'_\theta &= \pi_\theta\left(\cdot \middle| x\right),\end{aligned}$$

that is π (resp. $\pi_\theta\left(\cdot| x\right)$) is the invariant distribution of the transition kernel K_θ (resp. K'_θ). Intuitively, if we rearrange terms in Eq. (3)

$$\frac{\theta_{i+1} - \theta_i}{\gamma_{i+1}} = h\left(\theta_i\right) + e_{i+1},$$

we understand that if the noise e_i "cancels out on the average", then the trajectories $\theta_0, \theta_1, \ldots$ of the recursion should behave more or less like the solutions $\theta\left(t\right)$ of the ordinary differential equation

$$\dot{\theta}\left(t\right) = h\left(\theta\left(t\right)\right),$$

whose stationary points are precisely such that $h\left(\theta\right) = 0$.

Therefore in the algorithm that we propose, both the Markov chain of interest $x_0, x_1, \ldots$ and the proposal parameter θ are updated along the iterations in order to produce samples from π while meeting some statistical performance defined by $h\left(\theta\right) = 0$. Intuitively, the validity of the algorithm is ensured if the sequence θ_i converges to θ_*. Indeed the chain x_i then becomes "more and more" homogeneous, and will produce samples from π in an $h-$optimal manner. Conditions ensuring the convergence of this scheme are discussed in Section 2 and in further details in (Andrieu *et al.* 2001).

We note that, to the best of our knowledge, the only references where such techniques are touched upon in order to optimize MCMC samplers are (Geyer and Thompson 1995) and (Haario *et al.* 2001), either to estimate moments of the distribution or find the "best" normalizing constants for the tempering algorithm. However, in neither of these papers were the general framework and the potential of the approach realized.
We sum up here the advantages of the approach

- It is extremely general and flexible.
- Our criterion can be adapted in order to match the statistical properties of the Markov chain to the type of function f that we wish to integrate.
- The modification of standard MCMC code is trivial.
- The computational overhead is negligeable and the operations involved extremely simple.
- The computational and memory costs are constant and do not increase with the iterations.
- The input from the user is very limited, resulting in a truly self-learning algorithm.

- General convergence results for this type of algorithms abound in the literature and can be refined in our specific case.

The paper is organized as follows: we first describe in detail some of the criteria that have been proposed in the literature (Section 2), and we show how it is possible to define a general framework that encompasses these criteria as special cases. We briefly review practical implementation and theoretical aspects of stochastic approximation in Section 3. Then we move on to developing the algorithm for the case $w_1 = 1$ in Eq. (1). We firstly control the expected acceptance probability of the algorithm (Section 4) and then turn to the optimization of the efficiency of the sampler (Section 5). The general mixture case is addressed in Section 6. Computer simulations are given in Section 7 that demonstrate the interest of the approach. Finally some conclusions are drawn and further research directions suggested in Section 8.

2. Controlled MCMC for Adaptation

In this section we first consider two statistical criteria proposed in the literature to optimize MCMC samplers and show that they can be interpreted as particular cases of Eq. (2) (Subsection 2.1). Then we briefly describe the recursion associated to each of these criteria in Subsection 2.2.

2.1. Illustrative Criteria

We start this section with two specific criteria $h(\theta)$ that have been - explicitly or implicitly - suggested in the literature. Both examples focus on single MH algorithms, with target distribution π, proposal distribution q_θ, and aim at selecting θ in order to solve the equation $h(\theta) = 0$.

EXAMPLE 1. *(*Coerced acceptance probability*) In the cases of the random walk and Langevin algorithms, as recalled in Subsection 1.3, theoretical results have been derived that set approximate optimal expected acceptance rates $\overline{\alpha}_{th}$. Here optimal is to be understood as minimizing the variance of the estimators of integrals (Gelman et al. 1995), (Roberts et al.1998). More precisely, let*

$$\overline{\alpha}_\theta \triangleq \int_{\mathcal{X}^2} \min\left\{1, \frac{\pi(y)\, q_\theta(y,x)}{\pi(x)\, q_\theta(x,y)}\right\} \pi(x)\, q_\theta(x,y)\, dx dy$$

be the expected acceptance probability in the stationary regime for θ. There typically exists an optimal unknown value θ_* such that

$$\overline{\alpha}_{\theta_*} = \overline{\alpha}_{th},$$

which ensures the minimization of the asymptotic variance

$$\Sigma(\theta) = var_\pi(x_0) + 2\sum_{i=1}^{+\infty} cov_\theta(x_0, x_i),$$

of the random variable

$$\sqrt{N} \cdot \frac{1}{N}\sum_{i=1}^{N} x_i,$$

where $cov(x_0, x_i)$ *is the covariance between* x_0 *and* x_i *with* $x_0 \sim \pi$. *Schemes based on the idea of regeneration have been proposed in order to adaptively learn the value of this unknown parameter* θ_* *(Gilks et al. 1998). The main problem of these approaches is that they require the identification of regeneration times which leads to difficult practical problems. In order to embed this example in our framework we can define*

$$\eta(\theta) = \overline{\alpha}_\theta - \overline{\alpha}_{th},$$

and search the parameter space Θ *for solutions of the equation* $\eta(\theta) = 0$. *However the algorithm that we are to use is of the gradient type, and it is clearly difficult to relate the variations of* θ *with those of* $\eta(\theta)$ a priori. *A simple way of solving this problem consists of replacing* $\eta(\theta)$ *with* $\psi(\eta(\theta))$, *where* ψ *is such that* ψ *reaches its minimum at* 0 *and is convex. Our aim is then to find the minima of* $\psi(\eta(\theta))$, i.e. *localize zeros of* $\nabla_\theta \psi(\eta(\theta))$ *(provided that the derivative exists). Here we can choose* $\psi(x) = x^2$, *which leads to*

$$\begin{aligned} h(\theta) &= -\nabla_\theta \psi(\eta(\theta)) \\ &= -\nabla_\theta \eta^2(\theta) = -2\eta(\theta)\nabla_\theta \eta(\theta), \end{aligned}$$

and find the set of solutions of the equation

$$h(\theta) = 0,$$

which contains that of $\eta(\theta) = 0$.

EXAMPLE 2. *(*Moment matching*) Here we consider*

$$h(\theta) = \int_{\mathcal{X}} (\phi(x) - \theta)\pi(x)\,dx = \int_{\mathcal{X}} \phi(x)\pi(x)\,dx - \theta,$$

where ϕ *is an arbitrary function. This criterion means that the proposal distribution is parametrized in terms of a generalized moment of the target distribution. As we shall see, the implicit choice made in (Haario et al. 2001) corresponds for instance to*

$$\phi(x) = \left(xx^{\mathrm{T}}, x\right),$$

and the parameter effectively used in the proposal distribution of the MH step is $\theta_{xx^{\mathrm{T}}} - \theta_x\theta_x^{\mathrm{T}}$, *where* $\theta_{xx^{\mathrm{T}}}$ *(resp.* θ_x*) consists of the components of* θ *corresponding to* xx^{T} *(resp.* θ_x*) in* $\phi(x)$. *It is then clear that* $\theta_{xx^{\mathrm{T}}} - \theta_x\theta_x^{\mathrm{T}}$ *is an estimate of the covariance matrix of the target distribution, once the stationary regime has been reached.*

2.2. The Robbins-Monro Algorithm

Determining the roots of $h(\theta) = 0$ can be a difficult task in practice, as it simultaneously involves an optimization and an integration problem. The Robbins-Monro procedure was proposed in the 50's in order to solve this kind of problem, originally in a more restrictive framework than the one needed here. However the basic recursion used is the same and consists of the update§

$$\theta_{i+1} = \theta_i + \gamma_{i+1} H(\theta_i, w_{i+1}),$$

§In fact one can be slightly more general, see below.

where in our case w_{i+1} is distributed according to a distribution

$$\mathcal{K}_{\theta_i}\left(dw_{i+1}\right|w_i) = K_{\theta_i}\left(x_i; dx_{i+1}\right) K'_{\theta_i}\left(x_{i+1}, y_i; dy_{i+1}\right),$$

conditional upon the past of the chain.

EXAMPLE 3. (Example 1 continued) *Here the following recursion can be used*

$$\theta_{i+1} = \theta_i - \gamma_{i+1} 2\widehat{\eta\left(\theta\right)\nabla_\theta\eta}\left(\theta\right).$$

where $\widehat{\eta\left(\theta\right)\nabla_\theta\eta}\left(\theta\right)$ *is an estimate of* $\eta\left(\theta\right)\nabla_\theta\eta\left(\theta\right)$. *The calculation of the gradient* $\nabla_\theta\eta\left(\theta\right)$ *and the computation of its estimate are detailed in Section 4. Numerical simulations are presented in Section 7.*

EXAMPLE 4. *(*Example 2 continued*) The recursion used in (Haario et al. 2001) to adapt the proposal distribution of the algorithm is*

$$\begin{aligned} \theta_{i+1} &= \left(1-\gamma_{i+1}\right)\theta_i + \gamma_{i+1}\phi\left(x_{i+1}\right) \\ &= \theta_i + \gamma_{i+1}\left(\phi\left(x_{i+1}\right) - \theta_i\right), \end{aligned}$$

which corresponds exactly to the generic iteration described above. However neither the cost function formulation nor Robbins-Monro are mentioned in that paper.

3. Practical and Theoretical Aspects of Stochastic Approximation Algorithms

In this section, we first recall some classical results related to the existence of some of the gradients which might be required for Robbins-Monro type algorithms, and alternatives when such gradients do not exist or are hard to compute (Subsection 3.1). Then in Subsection 3.2 we recall some acceleration techniques, that allow the algorithm to achieve optimum asymptotic performance. We conclude the section with a simple and natural extension of the algorithm, Chen *et al.*'s projection method, for which theoretical results of convergence can be established under fairly general conditions (Subsection 3.3). The reader not interested in these rather technical - but practically important - aspects might at first directly skip to Section 4 for some applications.

3.1. Existence of the Gradient and Gradient-free Algorithms

As pointed out in the example above, in many situations the primary problem that we want to solve is that of the minimization/maximization of a given loss function. However one generally tries to reduce the initial problem to that of finding the roots of a gradient. This is the so-called Lagrangian case. Typically one needs to assess the existence and equality of the following quantities

$$\nabla_\theta \int_{\mathcal{W}} f\left(\theta, w\right)\lambda\left(dw\right) = \int_{\mathcal{W}} \nabla_\theta f\left(\theta, w\right)\lambda\left(dw\right),$$

for some measure λ and measurable function $f\left(\theta, w\right)$. The Lebesgue dominated convergence theorem gives us sufficient conditions:

THEOREM 1. *(*Lebesgue Dominated Convergence*) Assume that $f(\theta, w)$ is a λ-almost everywhere differentiable function (in θ) and that there exists a positive real valued summable function g such that in a neighborhood of θ_0,*

$$\|\nabla_\theta f(\theta, w)\| \leq g(w),$$

then

$$\nabla_{\theta_0} \int_{\mathcal{W}} f(\theta_0, w)\, \lambda(dw) = \int_{\mathcal{W}} \nabla_{\theta_0} f(\theta_0, w)\, \lambda(dw).$$

In Section 5 and Section 6 we will detail the conditions of applications of this theorem to our problem.

When the gradient is not a well-defined quantity, or difficult to compute, it is possible to use the Kiefer-Wolfowitz algorithm or the more efficient "simultaneous perturbation" (SP) algorithm (see (Spall 2000) and references therein), which both introduce finite differences rather than derivatives.

3.2. Acceleration Techniques

In order to improve the convergence properties of the algorithm, two techniques are available. The first one is the stochastic approximation equivalent of the classical deterministic Newton-Raphson algorithm which makes use of the curvature of $h(\theta)$. The recursions are then of the form

$$\begin{aligned} \theta_{i+1} &= \theta_i + \gamma_{i+1} \widetilde{\delta}_i^{-1} H(\theta_i, w_{i+1}) \\ \delta_{i+1} &= \left(1 - \gamma'_{i+1}\right) \delta_i + \gamma'_{i+1} \widehat{\delta}(\theta_{i+1}) \\ \widetilde{\delta}_{i+1} &= \xi(\delta_{i+1}), \end{aligned}$$

where $\widehat{\delta}(\theta)$ is a local estimate of the Hessian around θ and ξ is a projection on positive definite matrices, see (Spall 2000) for details and a review. This algorithm achieves an optimum convergence rate. It should be noticed that it can be extended to Kiefer-Wolfowitz and SP algorithms (Spall 2000).

The second approach consists of averaging the values of θ_i in order to build an estimate $\widehat{\theta}_i$ of θ_*. The algorithm proceeds as follows

$$\begin{aligned} \theta_{i+1} &= \theta_i + \gamma_{i+1} H(\theta_i, w_{i+1}) \\ \widehat{\theta}_{i+1} &= \left(1 - \frac{1}{i+1}\right) \widehat{\theta}_i + \frac{1}{i+1} \theta_{i+1}. \end{aligned}$$

It can be shown that this procedure achieves the asymptotic optimum rate of convergence of the second order algorithm presented above. However, as discussed in (Spall 2000), one should be careful when using this averaging procedure as it might initially slow down convergence of the algorithm towards the main attraction basin.

3.3. Stability and Convergence results

General convergence results exist for this type of algorithm, which corresponds to the so called Markovian dynamic case of stochastic approximation (Benveniste *et al.* 1990), (Delyon 1996). Theorems especially relevant to our case can be found in (Andrieu *et al.* 2001), together with convergence rates of estimators of the type $\widehat{I}_N(f)$. However, we would like to point out the projection technique due to Chen *et al.* (Chen *et al.* 1988), which stabilizes the algorithm and prevents it from diverging. Classical stabilizing procedures rely on reprojections into a compact set $\mathcal{K} \subset \Theta$ of θ_i, when $\theta_i \notin \mathcal{K}$. There is no methodology to chose $\mathcal{K}$ *a priori*, such that it contains θ_* **and the trajectories that lead to these optimizers**. Chen *et al.*'s procedure allows for such a $\mathcal{K}$ to be automatically determined by the algorithm. In order to precisely describe this technique we first need two concepts. Noting $\overset{\circ}{\mathcal{A}}$ the interior of $\mathcal{A}$ we can define:

Definition 1. *An increasing sequence* $(\mathcal{K}_k)_{k=1,\ldots,}$ *of compact sets is an* increasing compact covering *of* Θ *if*

$$\mathcal{K}_{k-1} \subset \overset{\circ}{\mathcal{K}_k} \text{ and } \Theta = \bigcup_{k=1}^{+\infty} \mathcal{K}_k.$$

A sequence $(\rho(k))_{k=1,\cdots}$ *is a* recurrence time sequence *if* $\theta_{\rho(k)} \in \mathcal{K}_0$ *and* $\rho(k-1) < \rho(k)$.

Now it is possible to describe the projection procedure. We set $s(1) = 1$ and then

$$(\theta_{i+1}, w_{i+1}) = \begin{cases} \left(\theta'_{i+1}, w_{i+1}\right) & \text{if } \theta'_{i+1} \in \mathcal{K}_{s(i)} \\ \left(\theta''_{i+1}, w''_{i+1}\right) \in \mathcal{K}_0 \times Q \text{ and } s(i+1) = s(i) + 1 & \text{if } \theta'_{i+1} \notin \mathcal{K}_{s(i)} \end{cases},$$

with

$$\begin{aligned} \theta'_{i+1} &= \theta_i + \gamma_{i+1} \min\left\{1, \frac{C\gamma_{i+1}^{-\eta}}{|H(\theta_i, w_{i+1})|}\right\} H(\theta_i, w_{i+1}) \\ \left(\theta''_{i+1}, w''_{i+1}\right) &\text{ is any function of } \{\theta_k, w_k\}_{k=0,\ldots,i} \text{ contained in } \mathcal{K}_0 \times Q, \end{aligned}$$

where Q is a fixed compact set of $\mathcal{W}$, $C > 0$ and $\eta < 1$. In the remainder of the paper we will use the short notation

$$\theta_{i+1} \equiv \theta_i + \gamma_{i+1} H(\theta_i, w_{i+1})$$

for the algorithm described above. The exact conditions that ensure the convergence of this algorithm, which are satisfied by most of our examples, can be found in (Andrieu *et al.* 2001). However we recall here - almost - classical conditions on the γ_i's that ensure its almost sure convergence:

- $H(\theta, w_{i+1})$ must be an unbiased or asymptotically unbiased observation of $h(\theta)$.
- The sequence of γ_i's is required to go to zero neither too quickly nor too slowly, more precisely

$$\sum_{i=1}^{+\infty} \gamma_i = +\infty \text{ and } \sum_{i=1}^{+\infty} \gamma_i^{2\eta} < +\infty.$$

In the next sections we explore several possible criteria, and specific examples are presented in Section 7.

4. Coerced Acceptance Probability

In this section we want to find the zeros of the criterion

$$\eta(\theta) = \int_{\mathcal{X}^2} \pi(x) \min\left\{1, \frac{\pi(y) q_\theta(y,x)}{\pi(x) q_\theta(x,y)}\right\} q_\theta(x,y)\,dxdy - \overline{\alpha}_{th}.$$

Note that the first term is the expected acceptance probability of a MH algorithm in the stationary regime. It is difficult to design an SA algorithm in order to find the roots of $\eta(\theta) = 0$ (the criterion is real, θ is multidimensional, and it is not clear how $\eta(\theta)$ evolves as a function of θ). We therefore search the space Θ for zeros of the expression

$$\begin{aligned} h(\theta) &= -\nabla_\theta \eta^2(\theta) \\ &= -2\eta(\theta) \int_{\mathcal{X}^2} \pi(x) \left[\mathbb{I}_{\{\rho;\rho<1\}}(\rho_\theta(x,y)) \nabla_\theta \rho_\theta(x,y) q_\theta(x,y) \right. \\ &\qquad \left. + \min\{1, \rho_\theta(x,y)\} \nabla_\theta q_\theta(x,y)\right] dxdy, \end{aligned}$$

provided that derivation under the integral sign is allowed (see Appendix B). The estimation of this gradient requires two Markov chains $x_0^1, x_1^1, \ldots$ and $x_0^2, x_1^2, \ldots$, one to estimate $\eta(\theta)$ and the other one to compute $\nabla_\theta \eta(\theta)$ (see note in Subsection 5.3 for an alternative). Then the algorithm can be described as

---

Coerced acceptance probability

- Initialization of θ_0, x_0^1 and x_0^2.
- Iteration $i+1$

  (a) For $c = 1, 2$ sample $x_*^c \sim q_{\theta_i}(x_i^c; dx_*^c)$
  (b) Update θ_i

$$\begin{aligned} \theta_{i+1} &= \theta_i - \gamma_{i+1} 2\left(\min\left\{1, \rho_{\theta_i}\left(x_i^1, x_*^1\right)\right\} - \overline{\alpha}_{th}\right) \\ &\quad \times \left(\mathbb{I}_{\{\rho;\rho<1\}}\left(\rho_{\theta_i}\left(x_i^2, x_*^2\right)\right) \nabla_\theta\, \rho_\theta\left(x_i^2, x_*^c\right)\big|_{\theta=\theta_i} + \frac{\nabla_\theta\, q_\theta\left(x_i^2, x_*^2\right)\big|_{\theta=\theta_i}}{q_{\theta_i}\left(x_i^2, x_*^2\right)} \alpha_\theta\left(x_i^2, x_*^2\right)\right). \end{aligned}$$

  (c) For $c = 1, 2$ set $x_{i+1}^c = x_*^c$ with probability $\min\left\{1, \rho_{\theta_i}\left(x_i^c, x_*^c\right)\right\}$, otherwise $x_{i+1}^c = x_i^c$.

---

5. Efficiency Optimization of a Single MH Kernel

Here we propose the minimization of a criterion of the type $\|\Sigma_{f,\tau}(\theta)\|^2$ where

$$\Sigma_{f,\tau}(\theta) \triangleq var_\pi(f(x_0)) + 2\sum_{i=1}^{\tau} cov_\theta(f(x_0), f(x_i)),$$

and τ is a fixed integer, and $\|A\|^2 = Tr\left(AA^{\mathsf{T}}\right)$ for a square matrix A. Note that as $\tau \to +\infty$ this quantity approaches the true asymptotic variance of $\sqrt{N}\widehat{I}_N(f)$ and is therefore key to the performance of the estimator $\widehat{I}_N(f)$. More precisely, with this criterion we will optimize the value of θ in order to adapt the sampler to the estimation of quantities of the type

$$\int_{\mathcal{X}} f(x)\,\pi(x)\,dx.$$

As the minimum value of $\|\Sigma_{f,\tau}(\theta)\|^2$ is not known a priori we need to compute the gradient $\nabla_\theta \|\Sigma_{f,\tau}(\theta)\|^2$ and find its zeros. The algorithm will then consist of the following update

$$\theta_{i+1} = \theta_i - \gamma_{i+1} \widehat{\nabla_\theta \|\Sigma_{f,\tau}(\theta)\|^2}\Big|_{\theta=\theta_i},$$

($\widehat{\nabla_\theta \|\Sigma_{f,\tau}(\theta)\|^2}$ is an estimate of the true gradient) in order to minimize $\|\Sigma_{f,\tau}(\theta)\|^2$. The calculation of this gradient is rather technical and the remainder of this section is dedicated to its computation. We first start rewriting this loss function so as to remove the delta-Dirac mass of the MH algorithm, in the spirit of the derivation of the Rao-Blackwellized estimator of (Casella and Robert 1996) (Subsection 5.1). Then we derive an expression of $\nabla_\theta \|\Sigma_{f,\tau}(\theta)\|^2$ under appropriate conditions (Proposition 4 in Subsection 5.2) and finally provide the reader with some pseudo-code that describes the estimation of the gradient (Subsection 5.3) and the main structure of the algorithm (Subsection 5.4).

5.1. "Dirac free" Covariance

We first write

$$\begin{aligned}
cov_\theta\left(f(x_0), f(x_i)\right) &= \mathbb{E}_{\pi,K^i}\left[f(x_0) f^{\mathsf{T}}(x_i)\right] - \mathbb{E}_\pi\left[f(x_0)\right]\mathbb{E}_{\pi,K^i}\left[f^{\mathsf{T}}(x_i)\right] \\
&= \int_{\mathcal{X}^{i+1}} f(x_0) f^{\mathsf{T}}(x_i)\,\pi(dx_i)\prod_{j=1}^{i}\left[\alpha_\theta(x_{j-1},x_j)\,q_\theta(x_{j-1},dx_j)\right. \\
&\qquad\qquad \left. + \,\delta_{x_{j-1}}(dx_j)\,r_\theta(x_{j-1})\right] \\
&\quad - \int_{\mathcal{X}} f(x_0)\,\pi(dx_0)\int_{\mathcal{X}^{i+1}} f^{\mathsf{T}}(x_i)\,\pi(dx_0)\prod_{j=1}^{i} K_\theta(x_{j-1},dx_j),
\end{aligned}$$

where $\delta_x(dy)$ is the delta Dirac mass at x, in a simpler way. It can be noticed that the second term simplifies due to the invariance of π with respect to K_θ,

$$\int_{\mathcal{X}^{i+1}} f(x_i)\,\pi(dx_0)\prod_{j=1}^{i} K_\theta(x_{j-1},dx_j) = \int_{\mathcal{X}} f(x)\,\pi(dx),$$

which does not depend upon θ. We therefore concentrate on the first term, using a decomposition of the path according to all possible acceptance histories.

PROPOSITION 2. *The cross product $\mathbb{E}_{\pi,K^i}\left[f(x_0) f^{\mathsf{T}}(x_i)\right]$ can be written as follows*

$$\mathbb{E}_{\pi,K^i}\left[f(x_0) f^{\mathsf{T}}(x_i)\right] = \int_{\mathcal{X}^{i+1}} \pi(dx_0)\,f(x_0)\,f^{\mathsf{T}}(x_i)\sum_{k=0}^{i}\sum_{1\le j_1<\ldots<j_k\le i}\Lambda_k\left(dx_{j_1},\ldots,dx_{j_k}\right),$$

where

$$\begin{aligned}\Lambda_k\left(dx_{j_1},\ldots,dx_{j_k}\right) &\triangleq \prod_{n=1}^{j_1-1}\left(1-\alpha_\theta\left(x_0,x_n\right)\right)q_\theta\left(x_{j_l},dx_n\right)\left\{\prod_{l=1}^{k}\alpha_\theta\left(x_{j_{l-1}},x_{j_l}\right)q_\theta\left(x_{j_{l-1}},dx_{j_l}\right)\right.\\ &\quad\left.\times\prod_{m=j_l+1}^{j_{l+1}-1}\left(1-\alpha_\theta\left(x_{j_l},x_m\right)\right)q_\theta\left(x_{j_l},dx_m\right)\right\},\end{aligned}$$

with the convention $j_0 \triangleq 0$.

Proof. The expansion of the product

$$\prod_{j=1}^{i}\left[\alpha_\theta\left(x_{j-1},x_j\right)q_\theta\left(x_{j-1},dx_j\right)\delta_{x_{j-1}}\left(dx_j\right)r_\theta\left(x_{j-1}\right)\right], \tag{4}$$

leads to

$$\begin{aligned}&\sum_{k=0}^{i}\sum_{1\le j_1<\ldots<j_k\le i}\prod_{n=1}^{j_1-1}\delta_{x_0}\left(dx_n\right)r_\theta\left(x_0\right)\left\{\prod_{l=1}^{k}\alpha_\theta\left(x_{j_{l-1}},x_{j_l}\right)q_\theta\left(x_{j_{l-1}},dx_{j_l}\right)\right.\\ &\qquad\left.\times\prod_{m=j_l+1}^{j_{l+1}-1}\delta_{x_{j_l-1}}\left(dx_m\right)r_\theta\left(x_{j_l-1}\right)\right\}.\end{aligned}$$

Now we focus on the products of delta functions, *i.e.* terms of the form

$$\varphi\left(dx_{j_k}\right)\prod_{l=j_k+1}^{j_{k+1}-1}\delta_{x_{l-1}}\left(dx_l\right)r_\theta\left(x_{l-1}\right)\psi\left(x_{j_{k+1}-1},dx_{j_{k+1}}\right),$$

where φ and ψ are appropriate terms that can be identified from Eq. (4). One can finally apply the following manipulation to this "generic" term,

$$\begin{aligned}&\underbrace{\int_{\mathcal{X}}}_{j_k}\underbrace{\int_{\mathcal{X}}}_{j_k+1}\underbrace{\int_{\mathcal{X}}}_{j_k+2}\cdots\underbrace{\int_{\mathcal{X}}}_{j_{k+1}}\varphi\left(dx_{j_k}\right)\prod_{l=j_k+1}^{j_{k+1}-1}\delta_{x_{l-1}}\left(dx_l\right)r_\theta\left(x_{l-1}\right)\psi\left(x_{j_{k+1}-1},dx_{j_{k+1}}\right)\\ &= \underbrace{\int_{\mathcal{X}}}_{j_k}\underbrace{\int_{\mathcal{X}}}_{j_{k+1}}\varphi\left(dx_{j_k}\right)\prod_{l=j_k+1}^{j_{k+1}-1}r_\theta\left(x_{j_k}\right)\psi\left(x_{j_k},dx_{j_{k+1}}\right)\\ &= \underbrace{\int_{\mathcal{X}}}_{j_k}\underbrace{\int_{\mathcal{X}}}_{j_{k+1}}\varphi\left(dx_{j_k}\right)\prod_{l=j_k+1}^{j_{k+1}-1}\int_{\mathcal{X}}\left[1-\alpha\left(x_{j_k},u_l\right)\right]q_\theta\left(x_{j_k},du_l\right)\psi\left(x_{j_k},dx_{j_{k+1}}\right)\\ &= \underbrace{\int_{\mathcal{X}}}_{j_k}\underbrace{\int_{\mathcal{X}}}_{j_k+1}\underbrace{\int_{\mathcal{X}}}_{j_k+2}\cdots\underbrace{\int_{\mathcal{X}}}_{j_{k+1}}\varphi\left(x_{j_k}\right)\prod_{l=j_k+1}^{j_{k+1}-1}\left[1-\alpha\left(x_{j_k},u_l\right)\right]q_\theta\left(x_{j_k},du_l\right)\psi\left(x_{j_k},dx_{j_{k+1}}\right),\end{aligned}$$

and obtain the final result.

5.2. Expression of the Gradient

First we recall that the differential of a function $\Phi(\cdot)$ at point z is a linear functional $d_z\{\Phi(\cdot)\}(\cdot)$ such that

$$d_z\{\Phi(\cdot)\}(h)=\Phi(z+h)-\Phi(z)+o(\|h\|).$$

In our case we have

$$d_A\left\{\|\cdot\|^2\right\}(h_A)=2Tr\left(Ah_A^{\mathrm{T}}\right).$$

Then for a matrix $A(\theta)$ which is itself a function of a vector or matrix θ we obtain

$$d_{A(\theta)}\left\{\|\cdot\|^2\right\}(h_\theta)=2Tr\left[A(\theta)\,d_{A(\theta)}\{A(\theta)\}(h_\theta)\right].$$

The u,v-th coordinate of the gradient is obtained as $d_{A(\theta)}\left\{\|\cdot\|^2\right\}(E_{uv})$, where $[E_{uv}]_{k,l}=\mathbb{I}_{\{(u,v)\}}[(k,l)]$ is the u,v-th canonical matrix. Therefore, for our problem we need the expression of quantities of the type $\nabla_\theta cov_\theta(f(x_0),f(x_i))$, which can be thought of as a vector of matrices.

LEMMA 3. *Assuming that derivation under the sum sign is allowed, then the u,v-th coordinate of the gradient $\nabla_\theta cov_\theta(f(x_0),f(x_i))$ is*

$$\begin{aligned}\frac{\partial cov_\theta(f(x_0),f(x_i))}{\partial\theta_{u,v}} &= \int_{\mathcal{X}^{i+1}}\pi(dx_0)\,f(x_0)\,f^{\mathrm{T}}(x_i)\sum_{k=0}^{i}\sum_{1\leq j_1<\ldots<j_k\leq i}\frac{\partial\log\Lambda_k(x_{j_1},\ldots,x_{j_k})}{\partial\theta_{u,v}}\\ &\qquad\times\Lambda_k(dx_{j_1},\ldots,dx_{j_k}),\end{aligned}$$

where

$$\begin{aligned}\frac{\partial\log\Lambda_k(x_{j_1},\ldots,x_{j_k})}{\partial\theta_{u,v}} &= \sum_{n=1}^{j_1-1}\frac{-\frac{\partial\alpha_\theta(x_0,x_n)}{\partial\theta_{u,v}}}{1-\alpha_\theta(x_0,x_n)}+\frac{\frac{\partial q_\theta(x_0,x_n)}{\partial\theta_{u,v}}}{q_\theta(x_0,x_n)}\\ &+\sum_{l=1}^{k}\left\{\frac{\frac{\partial\alpha_\theta\left(x_{j_{l-1}},x_{j_l}\right)}{\partial\theta_{u,v}}}{\alpha_\theta\left(x_{j_{l-1}},x_{j_l}\right)}+\frac{\frac{\partial q_\theta\left(x_{j_{l-1}},x_{j_l}\right)}{\partial\theta_{u,v}}}{q_\theta\left(x_{j_{l-1}},x_{j_l}\right)}\right.\\ &\left.+\sum_{m=j_l+1}^{j_{l+1}-1}\frac{-\frac{\partial\alpha_\theta\left(x_{j_l},x_m\right)}{\partial\theta_{u,v}}}{1-\alpha_\theta\left(x_{j_l},x_m\right)}+\frac{\frac{\partial q_\theta\left(x_{j_l},x_m\right)}{\partial\theta_{u,v}}}{q_\theta\left(x_{j_l},x_m\right)}\right\}.\end{aligned}$$

In the lemma we have used the formal notation

$$\frac{\partial\alpha_\theta(x,y)}{\partial\theta_{u,v}}\triangleq\mathbb{I}_{\{\alpha;\alpha<1\}}(\alpha_\theta(x,y))\,\frac{\partial\rho_\theta(x,y)}{\partial\theta_{u,v}},$$

where

$$\rho_\theta(x,y)\triangleq\frac{\pi(y)\,q_\theta(y,x)}{\pi(x)\,q_\theta(x,y)}.$$

One should refer to Appendix A for more details on the validity of the assumption of the lemma.

PROOF. We can rewrite

$$
\begin{aligned}
\frac{\partial cov_\theta\left(x_0,x_i\right)}{\partial\theta_{u,v}} = & \int_{\mathcal{X}^{i+1}}\pi\left(dx_0\right)f\left(x_0\right)f^{\mathrm{T}}\left(x_i\right)\sum_{k=0}^{i}\sum_{1\leq j_1<\ldots<j_k\leq i}\left\{\prod_{l=1}^{k}\alpha_\theta\left(x_{j_{l-1}},x_{j_l}\right)q_\theta\left(x_{j_{l-1}},dx_{j_l}\right)\right. \\
& \left.\times\prod_{m=j_l+1}^{j_{l+1}-1}\left(1-\alpha_\theta\left(x_{j_l},x_m\right)\right)q_\theta\left(x_{j_l},dx_m\right)\right\} \\
& \times\left\{\sum_{l'=1}^{k}\frac{\frac{\partial\left[\alpha_\theta\left(x_{j_{l'-1}},x_{j_{l'}}\right)q_\theta\left(x_{j_{l'-1}},x_{j_{l'}}\right)\right]}{\partial\theta_{u,v}}}{\alpha_\theta\left(x_{j_{l'-1}},x_{j_{l'}}\right)q_\theta\left(x_{j_{l'-1}},x_{j_{l'}}\right)}\right. \\
& \left.+\sum_{m=j_{l'}+1}^{j_{l'+1}-1}\frac{\frac{\partial\left[\left(1-\alpha_\theta\left(x_{j_{l'}},x_m\right)\right)q_\theta\left(x_{j_{l'}},x_m\right)\right]}{\partial\theta_{u,v}}}{\left(1-\alpha_\theta\left(x_{j_{l'}},x_m\right)\right)q_\theta\left(x_{j_{l'}},x_m\right)}\right\}
\end{aligned}
$$

which leads to the expected result.

REMARK 1. *Notice that the result can be obtained using the classical formula,*

$$
\frac{\partial cov_\theta\left(x_0,x_i\right)}{\partial\theta_{u,v}}=\int_{\mathcal{X}^{i+1}}\pi\left(dx_0\right)f\left(x_0\right)f^{\mathrm{T}}\left(x_i\right)\sum_{l=1}^{i}\frac{\frac{\partial K_\theta\left(x_{l-1},dx_l\right)}{\partial\theta_{u,v}}}{K_\theta\left(x_{l-1},dx_l\right)}\prod_{l'=1}^{i}K_\theta\left(x_{l'-1},dx_{l'}\right) \qquad (5)
$$

Finally we have the following result

PROPOSITION 4. *The u,v-th coordinate of the gradient $\nabla_\theta\left\|\Sigma_{f,\tau}\left(\theta\right)\right\|^2$ is*

$$
\frac{\partial\left\|\Sigma_{f,\tau}\left(\theta\right)\right\|^2}{\partial\theta_{u,v}}=4Tr\left[\Sigma_{f,\tau}\left(\theta\right)\sum_{i=1}^{\tau}\frac{\partial cov_\theta^{\mathrm{T}}\left(f\left(x_0\right),f\left(x_i\right)\right)}{\partial\theta_{u,v}}\right],
$$

where $\frac{\partial cov_\theta^{\mathrm{T}}\left(f\left(x_0\right),f\left(x_i\right)\right)}{\partial\theta_{u,v}}$ is as in Lemma 3.

5.3. Estimation of the Gradient

An asymptotically unbiased estimator (*i.e.* in the sense that if $x_0\sim\pi$ then it is unbiased, which is guaranteed according to convergence results (Delyon 1996)) of the gradient $\nabla_\theta\left\|\Sigma_{f,\tau}\left(\theta\right)\right\|^2$, which is here an $m_\theta\times n_\theta$ matrix, is given by the following procedure. Note that in order to obtain unbiased estimates we need to run three separate Markov chains, x_l^1,x_l^2 and x_l^3, as the evaluation of products of two types of integrals is required. One of them is the covariance $\Sigma_{f,\tau}\left(\theta\right)$ which requires two chains.¶

¶Alternatively, in order to find the minimum of function of the type $\psi\left(h\left(\theta\right)\right)$ it could be also possible to consider recursions of the type

$$
h_{i+1}=\left(1-\gamma_i'\right)h_i+\gamma_i'H\left(\theta_{i+1},w_{i+1}\right)
$$

with $w_{i+1}\sim\mathcal{K}_{\theta_i}\left(w_i,dw\right)$ and

$$
\theta_{i+1}=\theta_i\pm\gamma_i\nabla_\theta\psi\left(h_{i+1}\right),
$$

where ψ is a non linear function. This approach does not require the simulation of three parallel chains.

Estimate of $\nabla_\theta \left\|\Sigma_{f,\tau}(\theta)\right\|^2$

- Initialization of $\theta, x_0^1, x_0^2, x_0^3$ and $G = 0$ and

$$\begin{aligned}\widehat{\Sigma}_{f,\tau}(\theta) \quad \leftarrow \quad & f\left(x_0^2\right) f^{\mathrm{T}}\left(x_0^2\right) + f\left(x_0^3\right) f^{\mathrm{T}}\left(x_0^3\right) \\ & -\frac{f\left(x_0^2\right)+f\left(x_0^3\right)}{\sqrt{2}}\left[\frac{f\left(x_0^2\right)+f\left(x_0^3\right)}{\sqrt{2}}\right]^{\mathrm{T}}.\end{aligned}$$

- For $l = 1, \ldots, \tau$

 (a) $x_l^c \sim K_\theta\left(x_{l-1}^c; dx_l^c\right)$ for $c = 1, 2, 3$.
 (b) If $x_l^1 \neq x_{l-1}^1$

$$G \leftarrow G + \frac{\nabla_\theta \alpha_\theta\left(x_{l-1}^1, x_l^1\right)}{\alpha_\theta\left(x_{l-1}^1, x_l^1\right)} + \frac{\nabla_\theta q_\theta\left(x_{l-1}^1, x_l^1\right)}{q_\theta\left(x_{l-1}^1, x_l^1\right)},$$

 else

$$G \leftarrow G - \frac{\nabla_\theta \alpha_\theta\left(x_{l-1}^1, x_l^1\right)}{1-\alpha_\theta\left(x_{l-1}^1, x_l^1\right)} + \frac{\nabla_\theta q_\theta\left(x_{l-1}^1, x_l^1\right)}{q_\theta\left(x_{l-1}^1, x_l^1\right)}.$$

 (c) For all $u = 1, \ldots, m_\theta$ and $v = 1, \ldots, n_\theta$

$$\frac{\partial \widehat{cov}_\theta^{\mathrm{T}}\left(f\left(x_0\right), f\left(x_l\right)\right)}{\partial \theta_{u,v}} \leftarrow f\left(x_0^1\right) f^{\mathrm{T}}\left(x_l^1\right) G_{u,v}.$$

 (d) Update the estimate of the covariance matrix,

$$\begin{aligned}\widehat{\Sigma}_{f,\tau}(\theta) \quad \leftarrow \quad & 2\left[f\left(x_0^2\right) f^{\mathrm{T}}\left(x_l^2\right) + f\left(x_0^3\right) f^{\mathrm{T}}\left(x_l^3\right)\right] \\ & -2\frac{f\left(x_0^2\right)+f\left(x_0^3\right)}{\sqrt{2}}\left[\frac{f\left(x_l^2\right)+f\left(x_l^3\right)}{\sqrt{2}}\right]^{\mathrm{T}}.\end{aligned}$$

- Finally for all $u = 1, \ldots, m_\theta$ and $v = 1, \ldots, n_\theta$

$$\frac{\partial \widehat{\left\|\Sigma_{f,\tau}(\theta)\right\|^2}}{\partial \theta_{u,v}} \leftarrow 4Tr\left[\widehat{\Sigma}_{f,\tau}(\theta) \sum_{l=1}^{\tau} \frac{\partial \widehat{cov}_\theta^{\mathrm{T}}\left(f\left(x_0\right), f\left(x_l\right)\right)}{\partial \theta_{u,v}}\right].$$

■

5.4. Main Iteration

Now we describe the main iteration of the algorithm, for which two versions are possible. The first one updates the parameter θ every τ iterations of the "MCMC" algorithm, whereas the second one updates the parameter θ every iteration of the "MCMC" algorithm. Here to simplify notation $x_i \triangleq \left(x_i^1, x_i^2, x_i^3\right)$.

Algorithm version 1

- Initialization of θ_0, x_0.
- Iteration i

  (a) For $c = 1, 2, 3$ and $l = 1, \ldots, \tau$
  $x^c_{(i-1)\tau+l} \sim K_{\theta_{i-1}}\left(x^c_{(i-1)\tau+l-1}; dx^c_{(i-1)\tau+l}\right)$
  (b) $\theta_i \equiv \theta_{i-1} + \gamma_i H\left(\theta_{i-1}, x_{(i-1)\tau:i\tau}\right)$

---

where $H\left(\theta, x_{0:\tau}\right)$ is the estimate $\nabla_\theta \left\|\widehat{\Sigma_{f,\tau}\left(\theta\right)}\right\|^2$, whose evaluation is detailed in previous section. The second version of the algorithm is as follows,

---

Algorithm version 2

- Initialization of θ_0, x_0.
- Iteration i

  (a) For $c = 1, 2, 3$
  (b) $x^c_i \sim K_{\theta_{i-1}}\left(x^c_{i-1}; dx^c_i\right)$
  (c) $\theta_i \equiv \theta_{i-1} + \gamma_i H\left(\theta_{i-1}, x_{i-\tau+1:i}\right)$.

---

6. Efficiency Optimization of a Mixture of MH Kernels

In this section, we consider a mixture of transition kernels, where the weights are to be optimized in order to maximize the efficiency of the algorithm. To simplify presentation, we assume that the proposal distributions do not depend upon θ. Applications of this type of strategies are numerous. The complete transition kernel could be a mixture of strategies for example and the aim would then be to select efficient strategies more often than others. In a MH one-variable-at-a-time type algorithm, this could be used in order to determine optimum blocking of the parameters. In the case of a random scan Gibbs sampler it is possible to determine optimal proportions. Note that an interesting extension not considered here would be to introduce a Markov chain on the choice of transition kernels, in order to favor interesting efficient sequences of transition kernels. In either case the transition kernel writes,

$$K_\theta\left(x, dy\right) = \frac{1}{1 + \sum_{j=2}^{p}\left(\theta_j^2 + \varepsilon\right)} K_1\left(x, dy\right) + \sum_{i=2}^{p} \frac{\theta_i^2 + \varepsilon}{1 + \sum_{j=2}^{p}\left(\theta_j^2 + \varepsilon\right)} K_i\left(x, dy\right),$$

where the K_i's for $i = 1, \ldots, p$ are transition kernels that admit π as invariant distributions. The parametrization of the mixture probability ensures that they are positive and sum to 1. Note that when $\varepsilon > 0$ the zero probability can only be asymptotically reached. We here again consider the criterion $\|\Sigma_{f,\tau}(\theta)\|^2$ which we aim at minimizing. We thus need the gradient of quantities like

$$\begin{aligned} cov_\theta(f(x_0), f(x_i)) &= \int_{\mathcal{X}^{i+1}} f(x_0) f^{\mathsf{T}}(x_i)\,\pi(dx_0) \prod_{j=1}^{i} K_\theta(x_{j-1}, dx_j) \\ &\quad - \int_{\mathcal{X}} f(x_0)\,\pi(dx_0) \int_{\mathcal{X}^{i+1}} f^{\mathsf{T}}(x_i)\,\pi(dx_0) \prod_{j=1}^{i} K_\theta(x_{j-1}, dx_j). \end{aligned}$$

Here we have that for $i \neq k$

$$\begin{aligned} \frac{\partial}{\partial\theta_i} \log\left(\frac{\theta_k^2+\varepsilon}{1+\sum_{j=2}^{p}(\theta_j^2+\varepsilon)}\right) &= \frac{-2\theta_i}{1+\sum_{j=2}^{p}(\theta_j^2+\varepsilon)}, \\ \frac{\partial}{\partial\theta_i} \log\left(\frac{\theta_i^2+\varepsilon}{1+\sum_{j=2}^{p}(\theta_j^2+\varepsilon)}\right) &= \frac{2\theta_i}{(\theta_i^2+\varepsilon)} - \frac{2\theta_i}{1+\sum_{j=2}^{p}(\theta_j^2+\varepsilon)}. \end{aligned}$$

The expression of the gradient follows directly from Eq. (5) and is estimated using the following procedure (here for $p=2$, with u_i^1 the label of the chosen kernel at iteration i for chain 1):

Estimate of $\nabla_\theta \|\Sigma_{f,\tau}(\theta)\|^2$

- Initialization of $\theta, x_0^1, x_0^2, x_0^3$ and $G=0$ and

$$\begin{aligned} \widehat{\Sigma}_{f,\tau}(\theta) &\leftarrow f(x_0^2) f^{\mathsf{T}}(x_0^2) + f(x_0^3) f^{\mathsf{T}}(x_0^3) \\ &\quad - \frac{f(x_0^2)+f(x_0^3)}{\sqrt{2}} \left[\frac{f(x_0^2)+f(x_0^3)}{\sqrt{2}}\right]^{\mathsf{T}}. \end{aligned}$$

- For $c=1,2,3$ and $l=1,\ldots,\tau$

(a) $x_l^c, u_l^c \sim K_\theta(x_{l-1}^c; dx_l^c)$.
(b) If $u_l^1 = 1$

$$G \leftarrow G - \frac{2\theta}{1+(\theta+\varepsilon)^2},$$

else

$$G \leftarrow G + \frac{2\theta}{(\theta+\varepsilon)^2} - \frac{2\theta}{1+(\theta+\varepsilon)^2}.$$

(c) Update the estimate of the gradient,

$$\frac{\partial \widehat{cov}_\theta^{\mathsf{T}}(f(x_0), f(x_l))}{\partial\theta} \leftarrow f(x_0^1) f^{\mathsf{T}}(x_l^1)\, G.$$

(d) Update the estimate of the covariance matrix,

$$\begin{aligned} \widehat{\Sigma}_{f,\tau}(\theta) &\leftarrow 2\left[f(x_0^2) f^{\mathsf{T}}(x_l^2) + f(x_0^3) f^{\mathsf{T}}(x_l^3)\right] \\ &\quad - \left[f(x_0^2)+f(x_0^3)\right]\left[f(x_l^2)+f(x_l^3)\right]^{\mathsf{T}}. \end{aligned}$$

- Finally,

$$\frac{\partial \widehat{\|\Sigma_{f,\tau}(\theta)\|^2}}{\partial\theta} \leftarrow 4Tr\left[\widehat{\Sigma}_{f,\tau}(\theta)\sum_{l=1}^{\tau}\frac{\partial \widehat{cov}_{\theta}^{\intercal}\left(f\left(x_0\right),f\left(x_l\right)\right)}{\partial\theta}\right].$$

■

7. Computer Simulations

In this section, we apply the general approaches developed above to specific cases: random walk Metropolis, Langevin algorithm and a mixture of Metropolis algorithms. For each problem, elements of the proof of existence of the required gradient are given in Appendix A. In all cases, we have chosen $\gamma_i \propto i^{-0.7}$ in order to favor a good exploration of Θ. We did not observe any divergence of the algorithm and no reprojection was therefore observed.

7.1. Coerced Acceptance Probability: Random Walk MH

Here we ran two experiments. First we imposed an expected acceptance probability of 0.4 for a random walk MH with Gaussian univariate target and proposal distributions. The parameter to be estimated is the variance of the proposal distribution. Results are presented on Fig. 1-3 for $40,000$ iterations. Both for the acceptance probabilities and the estimation of θ, convergence occurs quite quickly. We repeated the experiment for a bimodal target distribution which consists of a mixture of two normal distributions with parameters $\mu_1 = -5.0, \sigma_1^2 = 1.0, \mu_2 = 5.0, \sigma_2^2 = 2.0$ and weights $(\omega = 0.2, 1-\omega)$. Results are presented on Fig. 4-6 for $200,000$ iterations. Notice on Fig 4 that the initial proposal hardly covers the second mode of the target distribution as shown by the path of the Markov chain, whereas the final proposal has overcome this defect. The effect of the smoothing procedure suggested in Subsection 3.2 is particularly striking on Fig. 6. The variance of the proposal distribution is, as expected, larger for the bimodal distribution than for the simple normal distribution, for the same acceptance probability.

7.2. Coerced Acceptance Probability: Langevin MH Algorithm

An alternative to the random walk Metropolis proposal consists of using the gradient of the target distribution (Besag and Green 1992). Here we restrict ourselves to the univariate case, for which the proposal distribution and acceptance probabilities are

$$q_\theta(x,y) = \frac{1}{\sqrt{2\pi\theta^2}}\exp\left(-\frac{\left(y-x-\frac{\theta^2}{2}\nabla_x\log\pi(x)\right)^2}{2\theta^2}\right),$$

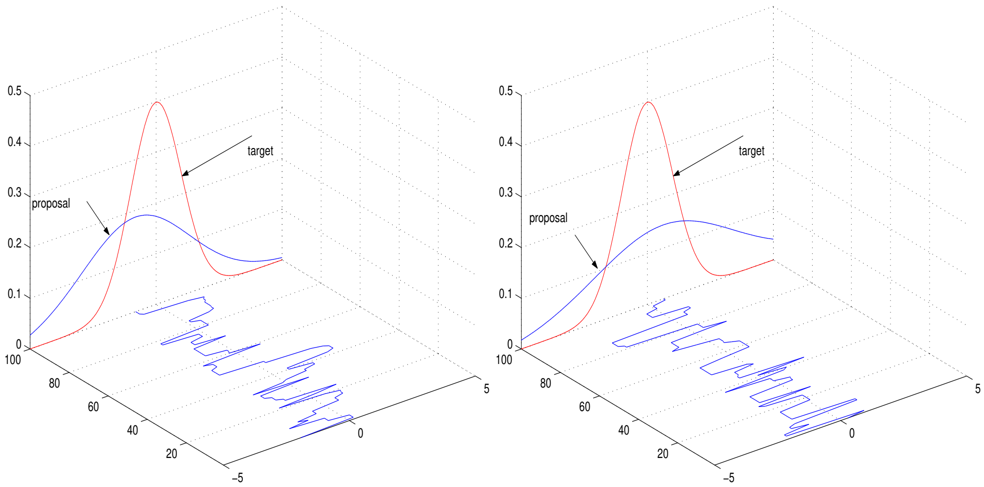

Fig. 1. 3D rendering of the Gaussian target distribution and the proposal distribution for the random walk example. The graph on the horizontal plane provides a snapshot of the Markov chain for 100 iterations (left: initial iterations/right: final iterations ending at $40,000$).

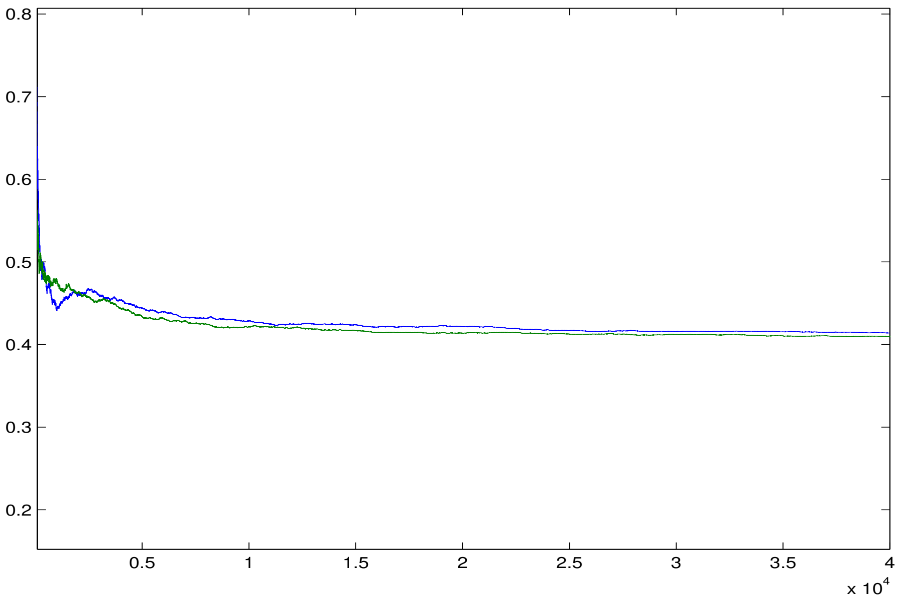

Fig. 2. Convergence of the empirical acceptance probabilities of chain 1 and 2 for the Gaussian target distribution and the random walk proposal.

$$\rho_\theta\left(x,y\right) = \frac{\pi\left(y\right)}{\pi\left(x\right)}\exp\left(-\frac{\left(x-y-\frac{\theta^2}{2}\nabla_y\log\pi\left(y\right)\right)^2-\left(y-x-\frac{\theta^2}{2}\nabla_x\log\pi\left(x\right)\right)^2}{2\theta^2}\right).$$

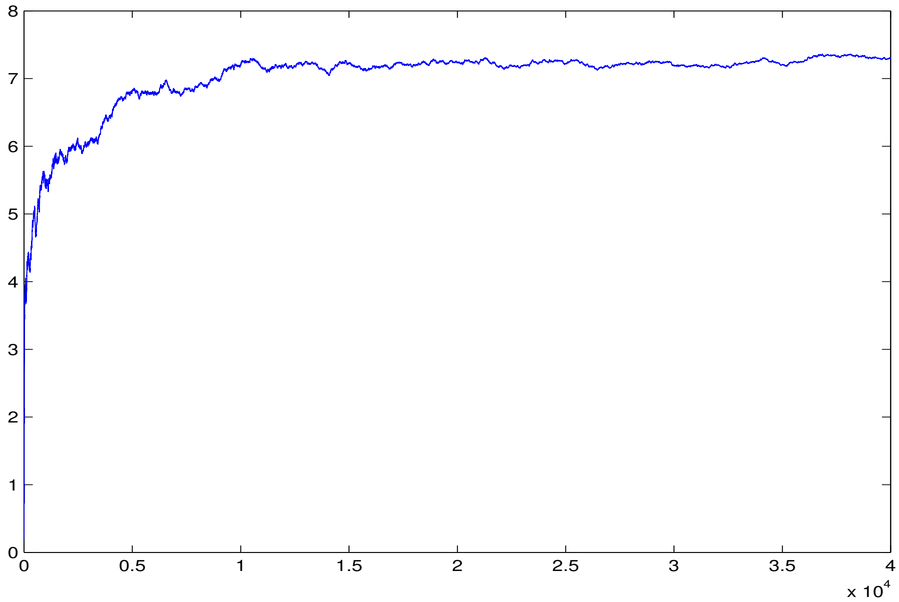

Fig. 3. Convergence of the variance of the proposal distribution for the Gaussian target distribution and the random walk proposal.

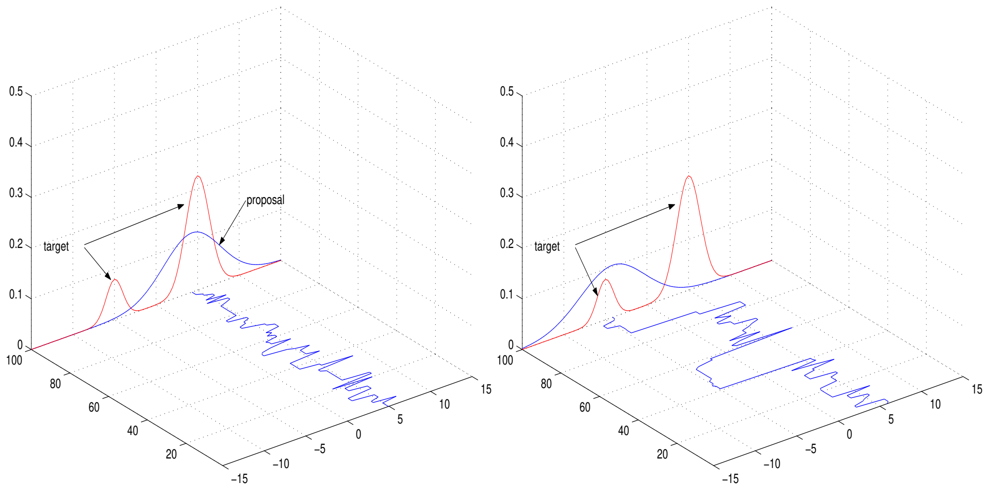

Fig. 4. 3D rendering of the mixture target distribution and the proposal distribution for the random walk example. The graph on the horizontal plane provides a snapshot of the Markov chain for 100 iterations (left: initial iterations/right: final iterations ending at $200,000$.

This case is referred to as the Metropolis adjusted Langevin algorithm in the literature. Therefore (for $\theta \neq 0$)

$$\frac{\partial \log q_\theta\left(x,y\right)}{\partial\theta} = \frac{-1}{\theta} + \frac{\left(y - x - \frac{\theta^2}{2}\nabla_x \log\pi\left(x\right)\right)^2}{\theta^3}$$

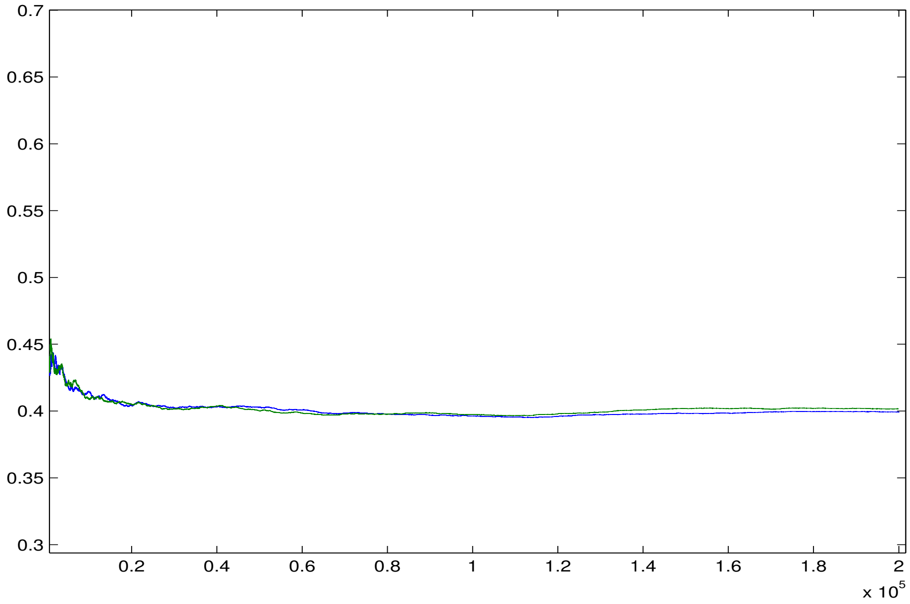

Fig. 5. Convergence of the empirical acceptance probabilities of chain 1 and 2 for the bimodal distribution and the random walk proposal.

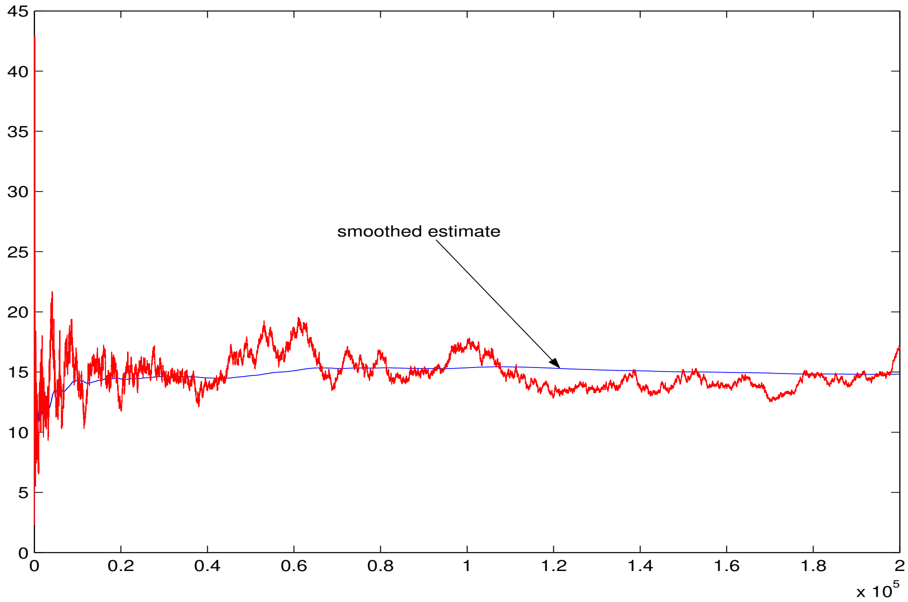

Fig. 6. Convergence of the variance of the proposal distribution for the bimodal target distribution (together with the smoothed estimate) and the random walk proposal.

$$+\frac{\nabla_x \log \pi(x)\left(y - x - \frac{\theta^2}{2}\nabla_x \log \pi(x)\right)}{\theta},$$

and

$$\frac{\partial \log \rho_\theta(x,y)}{\partial \theta} = \frac{\left(x - y - \frac{\theta^2}{2}\nabla_y \log \pi(y)\right)^2 - \left(y - x - \frac{\theta^2}{2}\nabla_x \log \pi(x)\right)^2}{\theta^3}$$
$$+\frac{\nabla_y \log \pi(y)\left(x - y - \frac{\theta^2}{2}\nabla_y \log \pi(y)\right) - \nabla_x \log \pi(x)\left(y - x - \frac{\theta^2}{2}\nabla_x \log \pi(x)\right)}{\theta}.$$

We ran two experiments, for two different target distributions which reproduced those used in the previous section. Results are presented on Fig. 7-8 for the Gaussian target distribution and Fig. 9-10 for the mixture of Gaussians. In both cases the number of iterations was $50,000$. Notice the difference between the convergencies of parameter θ of both the Gaussian and mixture cases: the bimodality of the target distribution explains the abrupt changes observed in the path of θ, since the corresponding Markov chain tends to remain stuck in one of the two modes for long periods of time. This is in contrast with the random walk Metropolis algorithm of previous section.

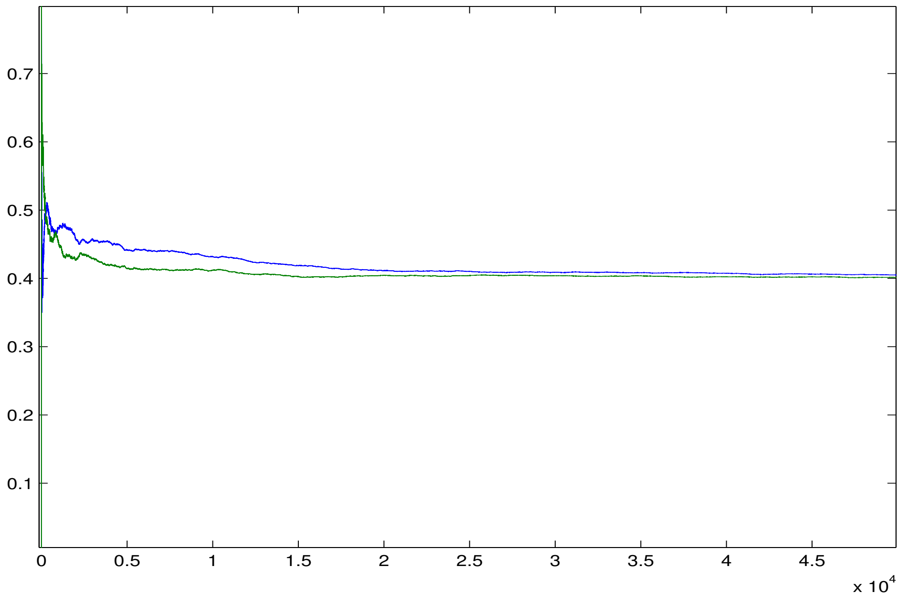

Fig. 7. Evolution of the expected acceptance probability for both chains, for the Langevin algorithm and the Gaussian target distribution.

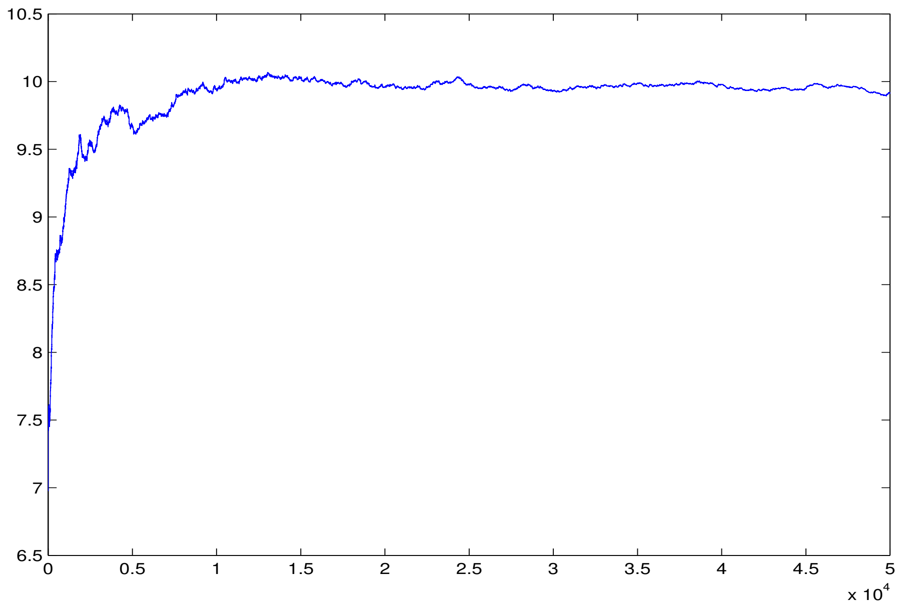

Fig. 8. Evolution of parameter θ for the Langevin algorithm and Gaussian target distribution.

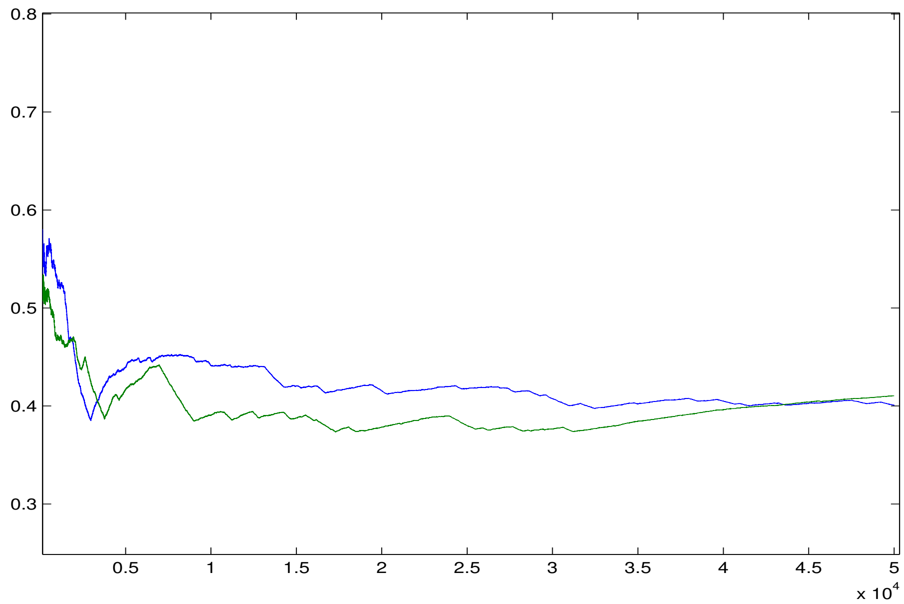

Fig. 9. Evolution of the estimate of the expected acceptance probability for the Langevin algorithm and the mixture target distribution.

7.3. Efficiency Maximization: Multivariate Gaussian Random Walk

Here we develop our strategy for the case of a Gaussian random walk and seek to optimize the covariance matrix Σ of this proposal distribution. In order to impose positiveness of

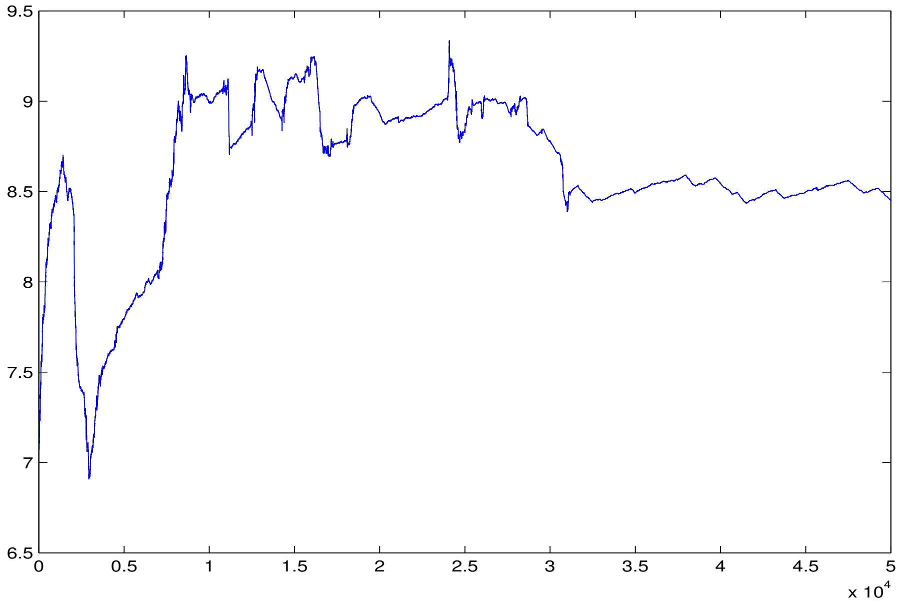

Fig. 10. Evolution of parameter θ for the Langevin algorithm and the mixture target distribution.

Σ, we consider the parameter $\theta = \Sigma^{-1/2}$ which is a lower triangular matrix such that $\Sigma^{-1} = \theta\theta^{\mathrm{T}}$, in other words the Cholesky decomposition of Σ^{-1}. Therefore the number of parameters in θ is $n_x (n_x + 1)/2$. The proposal distribution is in this case

$$q_\theta (x, y) = \left| \frac{\theta}{\sqrt{2\pi}} \right| \exp \left(\frac{-1}{2} (y - x)^{\mathrm{T}} \theta\theta^{\mathrm{T}} (y - x) \right)$$

and the acceptance probability does not depend upon θ,

$$\alpha (x, y) = \min \left\{ 1, \frac{\pi (y)}{\pi (x)} \right\}.$$

Now the differential of $q_\theta (x, y)$ with respect to θ can be calculated

$$\begin{aligned} d_\theta \{ q_\theta (x, y) \} (h) &= \left\{ d_\theta \{ |\theta| \} (h) - |\theta| \frac{(y - x)^{\mathrm{T}} d_\theta \{ \theta\theta^{\mathrm{T}} \} (h) (y - x)}{2} \right\} \\ &\quad \times \frac{1}{(2\pi)^{n_x/2}} \exp \left(\frac{-1}{2} (y - x)^{\mathrm{T}} \theta\theta^{\mathrm{T}} (y - x) \right) \\ &= \left\{ \frac{d_\theta \{ |\theta| \} (h)}{|\theta|} - \frac{(y - x)^{\mathrm{T}} d_\theta \{ \theta\theta^{\mathrm{T}} \} (h) (y - x)}{2} \right\} q_\theta (x, y), \end{aligned}$$

and we recall here the differential formulæ

$$\begin{aligned} d_\theta \{ |\theta| \} (h) &= |\theta| \, Tr \left(\theta^{-1} h \right) \\ d_\theta \{ \theta\theta^{\mathrm{T}} \} (h) &= \theta h^{\mathrm{T}} + h\theta^{\mathrm{T}}. \end{aligned}$$

from which the expressions of the gradient can be systematically obtained by using the canonical matrices $h = E_{ij}$. This yields

$$\frac{d_\theta \{|\theta|\} (E_{ij})}{|\theta|} = Tr\left(\theta^{-1} E_{ij}\right) = \left[\theta^{-1}\right]_{ji},$$

and

$$\begin{aligned}(y-x)^{\mathrm{T}} d_\theta \left\{\theta\theta^{\mathrm{T}}\right\}(E_{ij})(y-x) &= (y-x)^{\mathrm{T}}\left(\theta E_{ji} + E_{ij}\theta^{\mathrm{T}}\right)(y-x) \\ &= (y-x)^{\mathrm{T}} \theta E_{ji} (y-x) \\ &\quad + (y-x)^{\mathrm{T}} E_{ij}\theta^{\mathrm{T}} (y-x) \\ &= 2(y-x)^{\mathrm{T}} \theta_{1:n,j} (y_i - x_i).\end{aligned}$$

Therefore the $i,j-$th element of the required gradient is

$$\frac{d_\theta \{q_\theta (x,y)\}(E_{ij})}{q_\theta (x,y)} = \left[\theta^{-1}\right]_{ji} - (y-x)^{\mathrm{T}} \theta_{1:n,j} (y_i - x_i) \text{ for } i+j \geq 0.$$

Results are presented on Fig. 11-17 for two realizations of $200,000$ iterations on a Gaussian target distribution given in Fig. 11 and $\tau = 5$. The quantity θ was initialized at random from a Wishart distribution. Parameters a, b and α which are referred to in the captions correspond respectively to the eigenvalues of θ and the first principal direction of $\theta\theta^{\mathrm{T}}$ (in radians). As can be noticed from Fig. 14 and 17, and despite the truncation of the true covariance matrix, the results obtained agree with the general rules of thumb and theoretical rates reported in (Gelman *et al.* 1995) for low dimensions. The beneficial effect of the variance reduction technique suggested in Subsection 3.2 is illustrated on Fig. 15-16.

7.4. Efficiency Maximization: Optimal Mixture of Strategies

Here we investigate sampling from a bidimensional Gaussian distribution represented on Fig. 18 using a mixture of random walk Metropolis algorithms, whose proposal distributions are also normal. To ease the presentation, we introduce the following representation of the covariance matrices

$$\begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} a^2 & 0 \\ 0 & b^2 \end{pmatrix} \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix}^{\mathrm{T}}, \tag{6}$$

where a and b represent the length of an ellipse along the principal axes, which are a rotation of angle α of the x,y-axes. The main direction of the target distribution is $\pi/4$. The proposal distributions are normal distributions with orientations $\alpha = 0, \pm\pi/4, \pi/2$ and the same scale. Our aim is to find the optimum set of weights of the mixture of transition

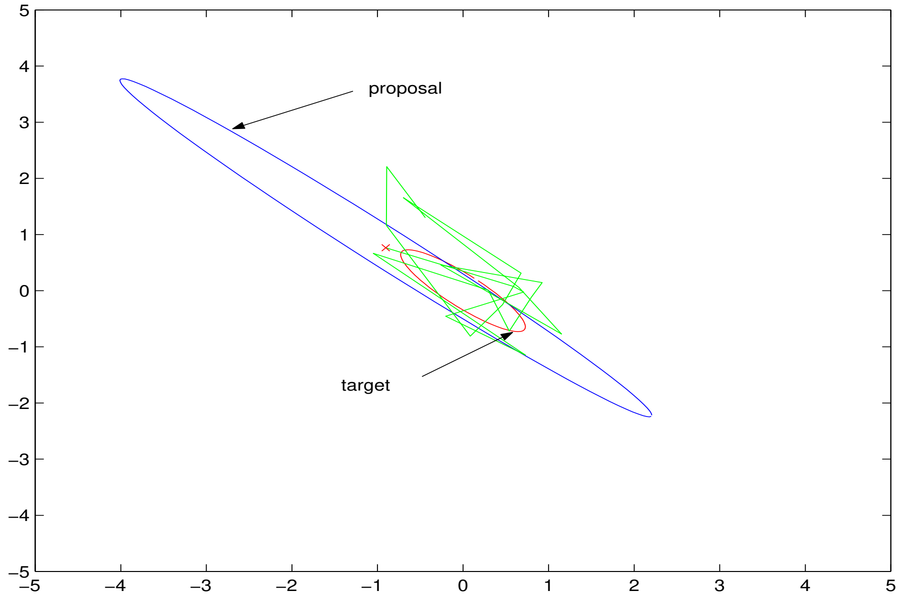

Fig. 11. The target Gaussian distribution (red ellipse with center $(0, 0)$). The Gaussian proposal distribution after $200,000$ iterations (blue) and the path of the Markov chain for 50 consecutive iterations (green).

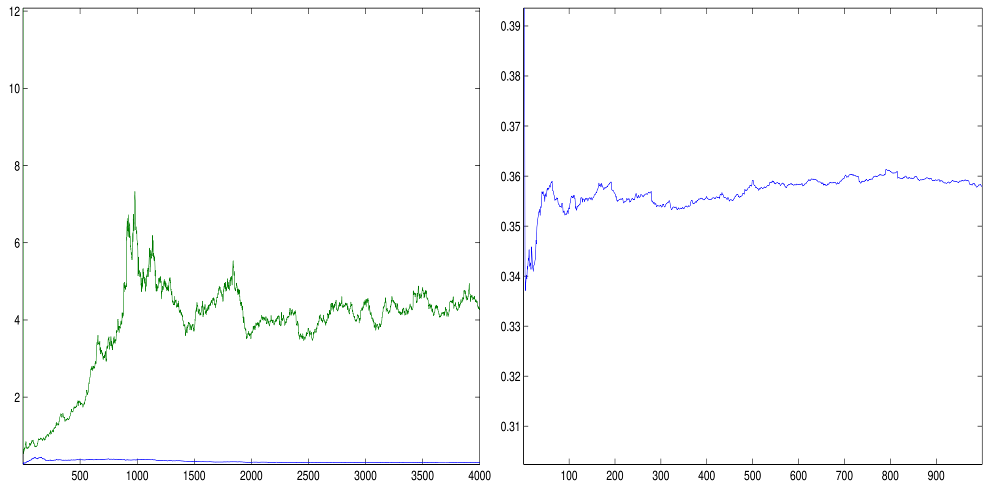

Fig. 12. Convergence of parameters a and b of the bivariate Gaussian proposal distribution, subsampled $(1/50)$.

kernels that maximize the truncated efficiency. Results are presented on Fig. 18 and 19 for $100,000$ iterations. As expected the proposal with principal axis $-\pi/4$ is attributed the lowest weight, and for symmetry reasons 0 and $\pi/2$ eventually recover identical weights.

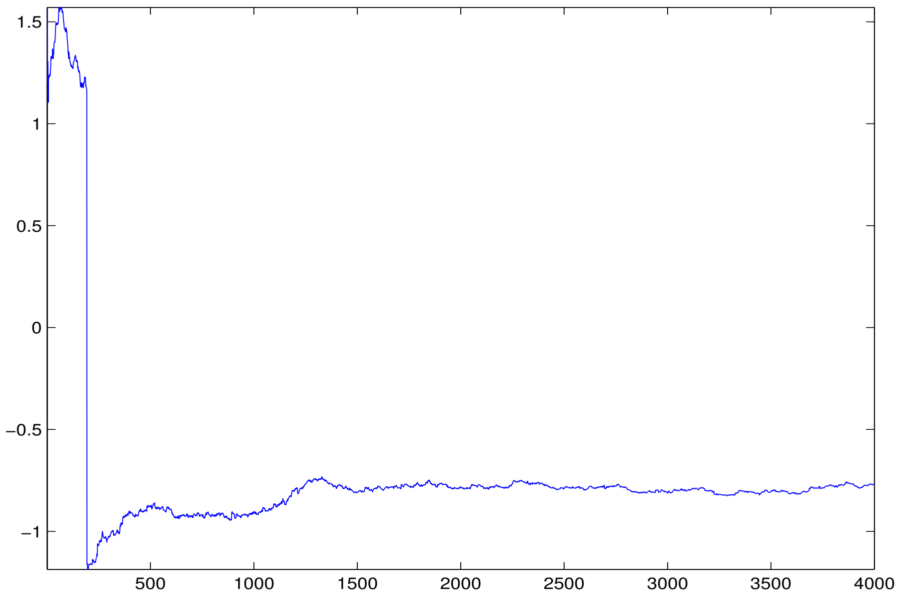

Fig. 13. Convergence of parameter α of the bivariate Gaussian proposal distribution, subsampled $(1/50)$.

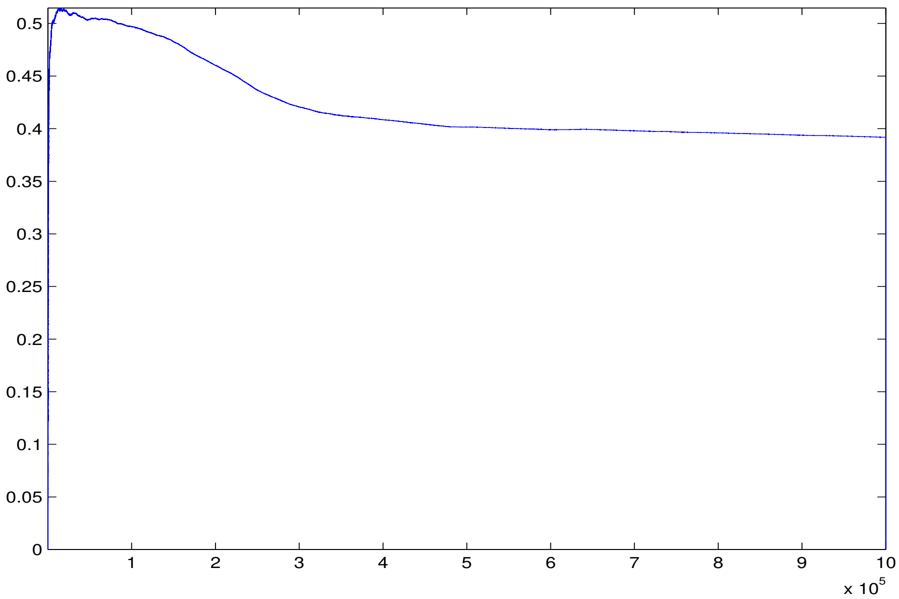

Fig. 14. Convergence of the empirical acceptance probability for the bivariate Gaussian target and proposal distributions.

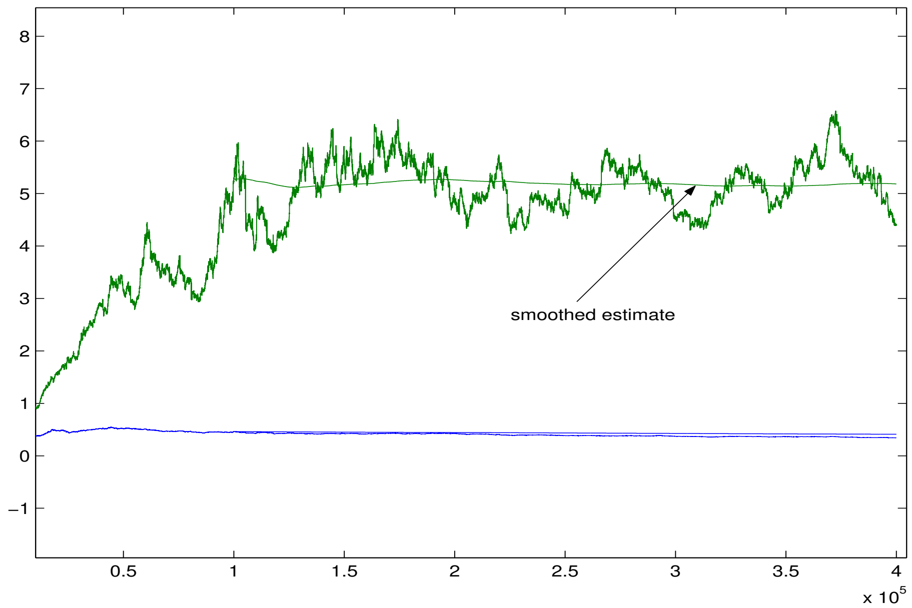

Fig. 15. Convergence of parameters a and b (and their smoothed estimates, calculated from 100,000) of the bivariate Gaussian target and proposal distributions.

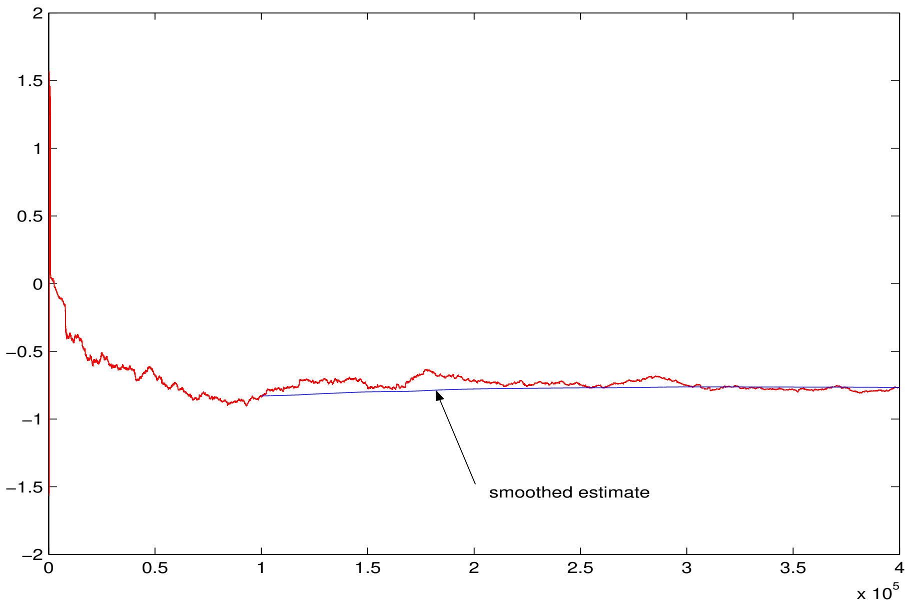

Fig. 16. Convergence of α (and its smoothed estimate calculated from 100,000) of the bivariate Gaussian target and proposal distributions.

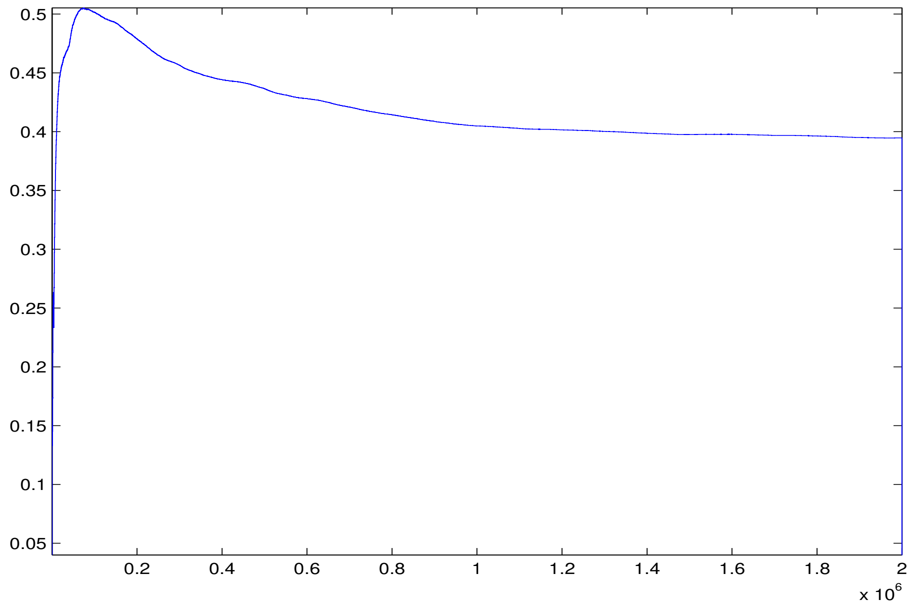

Fig. 17. Convergence of the empirical acceptance probability for the bivariate target and proposal distributions.

8. Discussion

We focused on the optimum scaling of random walks, the Langevin algorithm and mixtures of transition kernels so as to reduce the variance of integral estimators, but there are other possible extensions of this work.

- In the spirit of variational methods that have recently received attention in the computer science literature, we suggest the use of the framework developed here. In many cases these approximations depend on moments of the distribution π. It would be advantageous to use the approach presented in Section 2 (Example 2) to improve these estimates.
- The design of complex reversible jump moves requires many parameters to be tuned in order to lead to acceptable acceptance probabilities. These parameters could be chosen in order to optimize some criteria.
- In the framework of so-called parallel MCMC algorithms, it could be interesting to optimize the parameters that influence their interaction/combination.
- It is believed that adaptation of our framework to non-homogeneous Markov chains is possible either when the sequence of invariant distributions evolves slowly or by

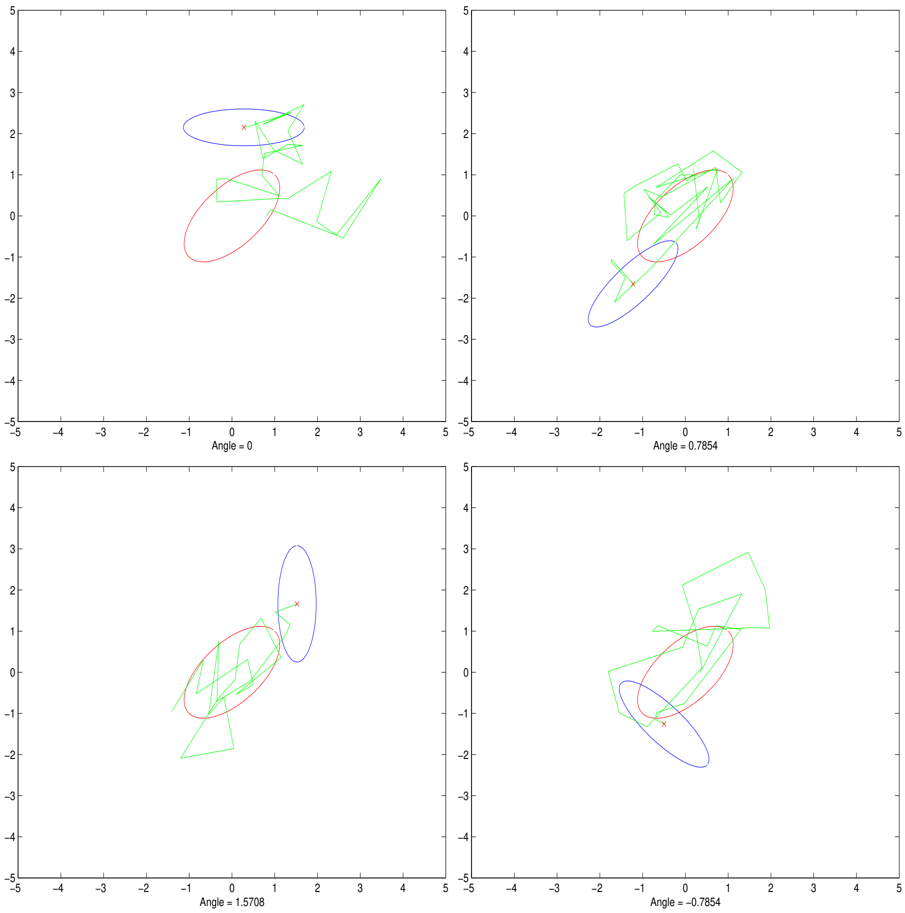

Fig. 18. The target distribution and the four possible proposal densities for the mixture of strategies example, along with 50 steps of the corresponding Markov chain.

using parallel MCMC as suggested above. Application in the context of simulated annealing could help designing efficient optimization algorithms.

- Inspired by our approach and adaptive importance sampling strategies, work on the development of similar techniques adapted to the difficult context of particle filtering is under investigation. This includes the solution to the question "when shall we perform selection?" (Arnaud Doucet, personal communication).

- Finally, in the context of learning algorithms, one can envisage the development of algorithms that learn to take decisions according to the current state of the Markov chain in order to achieve a best infinite horizon reward (Sutton and Barto 1998). Note that this adaptation to general state space would certainly require aggregation techniques *e.g.* the points x belong to the same level set $\{x; a \leq \pi(x) < b\}$, which is beyond the scope of this paper.

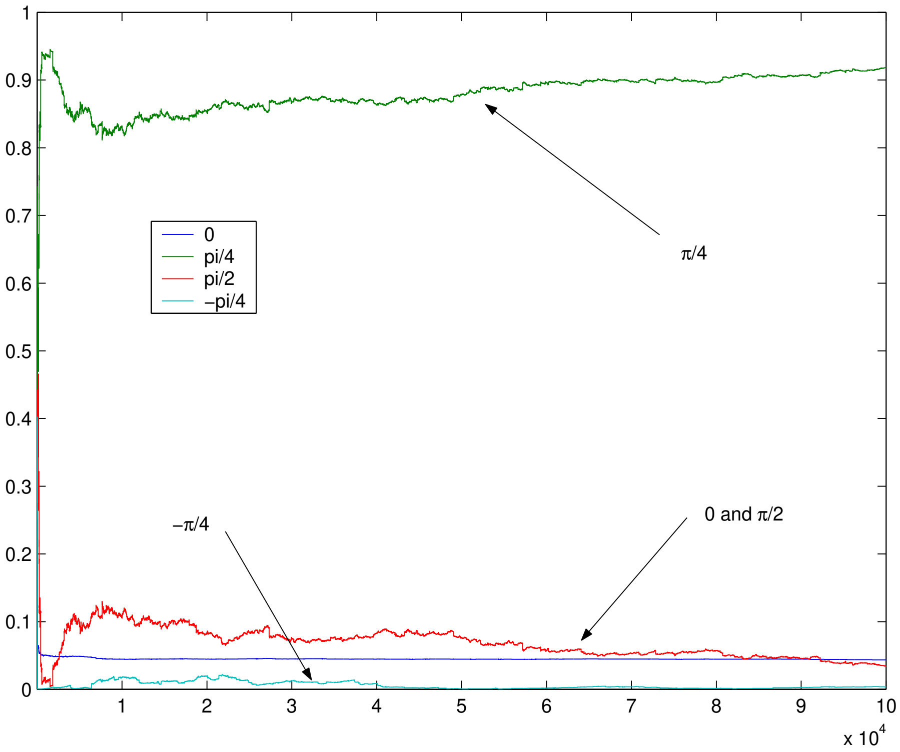

Fig. 19. Evolution of the proportions of the mixture of strategies.

9. Acknowledgments

The authors would like to thank Arnaud Doucet, Nando de Freitas and Merrilee Hurn for helpful comments on an earlier version of this paper.

A. Gradient of α_θ

A.1. Function $\min\{1, z\}$

We examine here the differential of the function

$$f(z) = \min\{1, z\} \text{ for } z \geq 0.$$

Consider $h > 0$

$$f(z+h) - f(z) = \min\{1, z+h\} - \min\{1, z\}$$

- If $z < 1$ then for h small enough

$$f(z+h) - f(z) = z + h - z = h$$

- If $z = 1$ then f is not differentiable.

- If $z > 1$ then again
$$f(z+h) - f(z) = 1 - 1 = 0$$
Consequently for $z \neq 1$
$$d_z \{f(\cdot)\}(h) = h\mathbb{I}_{\{y;y<1\}}(z).$$
Now if $g \geq 0$ is a differentiable function then
$$\begin{aligned} d_z \{\min\{1, g(\cdot)\}\}(h) &= d_x \{f \circ g(\cdot)\}(h) \\ &= \mathbb{I}_{\{y;y<1\}}(g(z))\, d_z \{g(\cdot)\}(h) \end{aligned}$$
when $g(z) \neq 1$ from which we can calculate the gradient.

A.2. Integral Differentiation

Now consider for a scalar θ (extension to the multivariate case is direct),
$$\phi(\theta) = \int_{\mathcal{Z}} f \circ g_\theta(z)\, \lambda(dz),$$
for some measure λ. From the Dominated Convergence Theorem (see Th. 1) under the following conditions

- $g_\theta(z)$ is differentiable in a neighborhood v of θ_0 for all $z \in \mathcal{Z}$,.
- The following inequality holds in the neighborhood v of θ_0,
$$\mathbb{I}_{\{g;g<1\}}(g_\theta(z)) \frac{\partial g_\theta(z)}{\partial \theta} \leq \gamma(z)$$
where $\gamma(z)$ is summable.
- In v, $\lambda(\{z; g_\theta(z) = 1\}) = 0$.

Then
$$d_{\theta_0}\{\phi(\cdot)\}(h) = \int_{\mathcal{Z}} \mathbb{I}_{\{y;y<1\}}(g_{\theta_0}(z))\, d_{\theta_0}\{g_\theta(z)\}(h)\, \lambda(dz).$$
We see that the only difficult condition to check in practice is the last one.

B. Existence of Derivatives

B.1. Random Walk

In the scalar case, for a neighborhood $(\theta_{\min}, \theta_{\max})$ of $\theta_0 > 0$,
$$\begin{aligned} \frac{\partial q_\theta(x,y)}{\partial \theta} &= \left(\frac{-1}{\theta} + \frac{(x-y)^2}{\theta^3}\right) q_\theta(x,y) \\ &\leq \left(\frac{-1}{\theta_{\min}} + \frac{(x-y)^2}{\theta_{\min}^3}\right) \frac{1}{\sqrt{2\pi}\theta_{\min}} \exp\left(-\frac{1}{2\theta_{\max}^2}(x-y)^2\right) \end{aligned}$$
which is obviously summable in y. The sum is independent of x. The multidimensional case can be treated in the same way by introducing the largest eigenvalue of the covariance matrix.

B.2. Langevin Algorithm

We restrict ourselves here to the scalar case, the adaptation to the multivariate case is direct. We need to check here that

$$\frac{\partial \rho_\theta(x,y)\, q_\theta(x,y)}{\partial \theta}\pi(x) = \left[\frac{\partial \rho_\theta(x,y)}{\partial \theta} q_\theta(x,y) + \rho_\theta(x,y)\frac{\partial q_\theta(x,y)}{\partial \theta}\right]\pi(x)$$

is upper bounded by a summable function. We start first checking that in the scalar case

$$q_\theta(x,y)\,\pi(x) = \frac{1}{\sqrt{2\pi}\theta}\exp\left(-\frac{1}{2\theta^2}\left(y-x-\frac{\theta^2}{2}\nabla_x \log \pi(x)\right)^2\right)\pi(x)$$

is upper bounded by a summable function. First it is easy to establish the following bound for a neighborhood $(\theta_{\min},\theta_{\max})$ of $\theta_0>0$

$$\begin{aligned}
q_\theta(x,y) &\leq \frac{1}{\sqrt{2\pi}\theta_{\min}}\exp\left(-\frac{1}{2\theta_{\max}^2}\left(y-x-\frac{\theta^2}{2}\nabla_x \log\pi(x)\right)^2\right)\\
&\leq \frac{1}{\sqrt{2\pi}\theta_{\min}}\exp\left(-\frac{1}{2\theta_{\max}^2}\left(y-x-\frac{\theta_{\min}^2}{2}\nabla_x \log\pi(x)\right)^2\right)\\
&\quad+\frac{1}{\sqrt{2\pi}\theta_{\min}}\exp\left(-\frac{1}{2\theta_{\max}^2}(y-x)^2\right)\\
&\quad+\frac{1}{\sqrt{2\pi}\theta_{\min}}\exp\left(-\frac{1}{2\theta_{\max}^2}\left(y-x-\frac{\theta_{\max}^2}{2}\nabla_x \log\pi(x)\right)^2\right),
\end{aligned}$$

which is summable. Now if we look at the derivative of the quantity (for $\rho_\theta(x,y)<1$),

$$\begin{aligned}
\frac{\partial q_\theta(x,y)}{\partial\theta} &= \left\{\frac{-1}{\theta}+\frac{\left(y-x-\frac{\theta^2}{2}\nabla_x\log\pi(x)\right)^2}{\theta^3}\right.\\
&\quad\left.+\frac{\nabla_x\log\pi(x)\left(y-x-\frac{\theta^2}{2}\nabla_x\log\pi(x)\right)}{\theta}\right\}q_\theta(x,y),
\end{aligned}$$

we can use the bound of q_θ on $(\theta_{\min},\theta_{\max})$ and the triangle inequality on $\left|\frac{\partial \log q_\theta(x,y)}{\partial\theta}\right|$. Integration of the bound with respect to y is routine and leads to

$$\begin{aligned}
&\frac{3\theta_{\max}^2}{\theta_{\min}^2}+\sum_{\theta'\in\{\theta_{\min},\theta_{\max},0\}}\frac{\theta_{\max}^2+\left(\theta'^2-\theta^2\right)^2\left(\frac{\nabla_x\log\pi(x)}{2}\right)^2}{\theta_{\min}^3}\\
&+\frac{|\nabla_x\log\pi(x)|\sqrt{\theta_{\max}^2+\left(\theta'^2-\theta^2\right)^2\left(\frac{\nabla_x\log\pi(x)}{2}\right)^2}}{\theta_{\min}}
\end{aligned}$$

where we have used Jensen's inequality. Now this quantity can easily be uniformly bounded in θ on $(\theta_{\min},\theta_{\max})$. Then we require the existence of some moments of the target distribution, namely

$$\mathbb{E}_\pi\left(\left[\nabla_x\log\pi(x)\right]^2\right)<+\infty.$$

Now we inspect the summability (for $\rho_\theta(x,y) < 1$) of

$$\begin{aligned}
\frac{\partial \log \rho_\theta(x,y)}{\partial\theta}\rho_\theta(x,y)\, q_\theta(x,y)\,\pi(x) &= \left[\frac{\partial \log q_\theta(y,x)}{\partial\theta} - \frac{\partial \log q_\theta(x,y)}{\partial\theta}\right]\rho_\theta(x,y)\, q_\theta(x,y)\,\pi(x) \\
&\leq \left|\frac{\partial \log q_\theta(y,x)}{\partial\theta}\right| q_\theta(y,x)\,\pi(y) \\
&\quad + \left|\frac{\partial \log q_\theta(x,y)}{\partial\theta}\right| q_\theta(x,y)\,\pi(x),
\end{aligned}$$

which are both summable in a neighborhood of θ_0 as shown above.

References

[Andrieu *et al.* 2001] Andrieu, C., Moulines, É. and Robert C.P. (2001) On the Convergence of Controlled MCMC, *in preparation.*

[Benveniste *et al.* 1990] Benveniste, A, Métivier M. and Priouret P. (1990) *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag.

[Bennet *et al.* 1996] Bennet, J.E., Racine-Poon, A. and Wakefield, J.C. (1996) MCMC for Nonlinear Hierarchical Models, in *MCMC in Practice*, Chapman & Hall London.

[Besag and Green 1992] Besag J., Green P.J. (1993) Spatial Statistics and Bayesian Computation, *Journal of the Royal Statistical Society, Series B, Methodological*, vol. 55, pp. 25-37.

[Browne and Draper 2000] Browne, W.J. and Draper, D. (2000) Implementation and Performance Issues in the Bayesian and Likelihood Fitting of Multilevel Models, *Computational Statistics*, vol. 15, pp. 391-420.

[Casella and Robert 1996] Casella, G. and Robert, C.P. (1996) Rao-Blackwellization of Sampling Schemes, *Biometrika*, vol. 83, pp. 81-94.

[Chauveau and Vandekerkhove 2001] Chauveau, D. and Vandekerkhove, P. (2001) Improving Convergence of the Hastings-Metropolis Algorithm with an Adaptive Proposal, *Scandinavian Journal of Statistics*, *to appear.*

[Chen *et al.* 1988] Chen, H.F., Guo, L. and Gao, A.J. (1988) Convergence and Robustness of the Robbins-Monro Algorithm Truncated at Randomly Varying Bounds, *Stochastic Processes and their Applications*, vol. 27, no. 2, pp. 217-231.

[Chib *et al.* 1998] Chib, S., Greenberg, E. and Winkelmann, R. (1998) Posterior Simulation and Bayes Factors in Panel Count Data Models, *Journal of Econometric*, vol. 86, pp. 33-54.

[Delyon 1996] Delyon, B. (1996) General Results on the Convergence of Stochastic Algorithms, *IEEE Trans. Automatic Control*, vol. 41, no. 9, pp. 1245-1256.

[de Freitas *et al.* 2001] de Freitas, N., Højen-Sørensen, P., Jordan, M. and Russell, S. (2001) Variational MCMC, in *Uncertainty in Artificial Intelligence,* J. Breese and D. Koller (Editors), pp. 120-127.

[Gåsemyr 2000] Gåsemyr, J. (2000) On an Adaptive Metropolis-Hastings Algorithm with Independent Proposal, *submitted.*

[Gelfand and Sahu 1994] Gelfand, A.E. and Sahu, S.K. (1994) On Markov Chain Monte Carlo Acceleration, *Journal of Computational and Graphical Statistics*, vol. 3, no. 3, pp. 261-276.

[Gelman *et al.* 1995] Gelman, A., Roberts, G., and Gilks, W. (1995) Efficient Metropolis Jumping Rules, In Bayesian Statistics 5. Oxford University Press, New York.

[Geyer and Thompson 1995] Geyer, C.J. and Thompson, E.A. (1995) Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference, *Journal of the American Statistical Association*, vol. 90, pp. 909-920.

[Gilks *et al.* 1998] Gilks, W. R., Roberts, G. O. and Sahu, S. K. (1998) Adaptive Markov Chain Monte Carlo Through Regeneration, *Journal of the American Statistical Association*, vol. 93, pp. 1045–1054.

[Gilks *et al.* 1994] Gilks, W.R., Roberts, G.O., and George, E.I. (1994) Adaptive Direction Sampling, *The Statistician* vol. 43, pp. 179-189.

[Green 1995] Green, P.J. (1995) Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination, *Biometrika*, vol. 82, pp. 711–732.

[Green and Mira 2000] Green, P.J. and Mira, A. (2001) Delayed Rejection in Reversible Jump Metropolis-Hastings, *submitted.*

[Haario *et al.* 1999] Haario, H., Saksman, E. and Tamminen, J. (1999) Adaptive Proposal Distribution for Random Walk Metropolis Algorithm, *Computational Statistics*, vol. 14, no. 3, pp 375-395.

[Haario *et al.* 2001] Haario, H., Saksman, E. and Tamminen, J. (2001) An Adaptive Metropolis Algorithm, *Bernoulli*, vol. 7, no. 2.

[Kim *et al.* 1998] Kim, S., Shephard, N. and Chib, S. (1998) Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models, *Review of Economic Studies*, vol. 65, pp. 361-393.

[Liu *et al.* 2000] Liu, J., Liang, F. and Wong, W.H. (2000) The Use of Multiple-Try Method and Local Optimization in Metropolis Sampling , *Journal of the American Statistical Association*, vol. 95, pp. 121-134.

[Laskey and Myers 2001] Laskey, K.B., Myers J. (2001) Population Markov Chain Monte Carlo, *to appear in Machine Learning.*

[Mykland *et al.* 1994] Mykland, P., Tierney, L. and Yu, B. (1994) Regeneration in Markov Chain Samplers, *Journal of the American Statistical Association*, vol. 90, pp. 233-241.

[Robbins and Monro 1951] Robbins, H. and Monro, S. (1951) A Stochastic Approximation Method, *Annals of Mathematical Statistics*, vol. 22, pp. 400-407.

[Robert and Casella 1999] Robert, C.P. and Casella, G. (1999) *Monte Carlo Statistical Methods*, Springer-Verlag.

[Roberts *et al.* 1997] Roberts, G.O., Gelman, A. and Gilks, W. (1997) Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms, *Annals of Applied Probability*, vol. 7, pp. 110-120.

[Roberts *et al.*1998] Roberts, G.O. and Rosenthal, J. (1998) Optimal Scaling of Discrete Approximation to Langevin Diffusion, *Journal of the Royal Statistical Society*, Series B, vol. 60, pp. 255-268.

[Sahu and Zhigljavsky 1998] Sahu, S.K. and Zhigljavsky A.A. (1998) Adaptation for Self Regenerative MCMC, *available from* http://www.maths.soton.ac.uk/staff/Sahu/research/papers/self.html.

[Spall 2000] Spall, J.C. (2000) Adaptive Stochastic Approximation by the Simultaneous Perturbation Method, *IEEE Transactions on Automatic Control*, vol. 45, no. 10, pp. 1839-1853.

[Sutton and Barto 1998] Sutton R. S. and Barto A. G. (1998) *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA.

[Tierney and Mira 1999] Tierney, L. and Mira, A. (1999) Some Adaptive Monte Carlo Methods for Bayesian Inference, *Statistics in Medicine*, vol. 18, pp. 2507-2515.