Stat Comput (2008) 18: 343–373
DOI 10.1007/s11222-008-9110-y

A tutorial on adaptive MCMC

Christophe Andrieu · Johannes Thoms

Received: 23 January 2008 / Accepted: 19 November 2008 / Published online: 3 December 2008

Abstract We review adaptive Markov chain Monte Carlo algorithms (MCMC) as a mean to optimise their performance. Using simple toy examples we review their theoretical underpinnings, and in particular show why adaptive MCMC algorithms might fail when some fundamental properties are not satisfied. This leads to guidelines concerning the design of correct algorithms. We then review criteria and the useful framework of stochastic approximation, which allows one to systematically optimise generally used criteria, but also analyse the properties of adaptive MCMC algorithms. We then propose a series of novel adaptive algorithms which prove to be robust and reliable in practice. These algorithms are applied to artificial and high dimensional scenarios, but also to the classic mine disaster dataset inference problem.

Keywords MCMC · Adaptive MCMC · Controlled Markov chain · Stochastic approximation

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 a space $\mathsf{X} \subset \mathbb{R}^{n_x}$ (assumed for simplicity to have a density with respect to the Lebesgue measure, also denoted π), from which integrals of the type

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

for some π-integrable functions $\mathsf{X} \to \mathbb{R}^{n_f}$ can be approximated using the estimator

$$\hat{I}_N(f) := \frac{1}{N}\sum_{i=1}^{N} f(X_i), \tag{1}$$

provided that the Markov chain generated with, say, transition P is ergodic *i.e.* it is guaranteed to eventually produce samples $\{X_i\}$ distributed according to π. Throughout this review we will refer, in broad terms, to the consistency of such estimates and the convergence of the distribution of X_i to π as π-ergodicity. The main building block of this class of algorithms is the Metropolis-Hastings (MH) algorithm. It requires the definition of a family of proposal distributions $\{q(x,\cdot), x \in \mathsf{X}\}$ 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. Indeed, the choice of the proposal distribution is crucial: the statistical properties of the Markov chain heavily depend upon this choice, an inadequate choice resulting in possibly poor performance of the Monte Carlo estimators. For example, in the toy case where $n_x = 1$ and the normal symmetric random walk Metropolis algorithm (N-SRWM) is used to produce transitions, the

C. Andrieu (✉)
School of Mathematics, University of Bristol,
Bristol BS8 1TW, UK
e-mail: c.andrieu@bristol.ac.uk
url: http://www.stats.bris.ac.uk/~maxca

J. Thoms
Chairs of Statistics, École Polytechnique Fédérale de Lausanne,
1015 Lausanne, Switzerland

density of the proposal distribution is of the form

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

where θ^2 is the variance of the proposed increments, hence defining a Markov transition probability P_θ. The variance of the corresponding estimator $\hat{I}_N^\theta(f)$, which we wish to be as small as possible for the purpose of efficiency, is well known to be typically unsatisfactory for values of θ^2 that are either "too small or too large" in comparison to optimal or suboptimal value(s). In more realistic scenarios, MCMC algorithms are in general combinations of several MH updates $\{P_{k,\theta},\ k = 1, \ldots, n,\ \theta \in \Theta\}$ for some set Θ, with each having its own parametrised proposal distribution $q_{k,\theta}$ for $k = 1, \ldots, n$ and sharing π as common invariant distribution. These transition probabilities are usually designed in order to capture various features of the target distribution π and in general chosen to complement one another. Such a combination can for example take the form of a mixture of different strategies, *i.e.*

$$P_\theta(x, dy) = \sum_{k=1}^{n} w_k(\theta) P_{k,\theta}(x, dy), \tag{2}$$

where for any $\theta \in \Theta$, $\sum_{k=1}^n w_k(\theta) = 1$, $w_k(\theta) \geq 0$, but can also, for example, take the form of combinations (*i.e.* products of transition matrices in the discrete case) such as

$$P_\theta(x, dy) = P_{1,\theta} P_{2,\theta} \cdots P_{n,\theta}(x, dy).$$

Both examples are particular cases of the class of Markov transition probabilities P_θ on which we shall focus in this paper: they are characterised by the fact that they (a) belong to a family of parametrised transition probabilities $\{P_\theta, \theta \in \Theta\}$ (for some problem dependent set Θ, $\Theta = (0, +\infty)$ in the toy example above) (b) for all $\theta \in \Theta$ π is an invariant distribution for P_θ, which is assumed to be ergodic (c) the performance of P_θ, for example the variance of $\hat{I}_N^\theta(f)$ above, is sensitive to the choice of θ.

Our aim in this paper is to review the theoretical underpinnings and recent methodological advances in the area of computer algorithms that aim to "optimise" such parametrised MCMC transition probabilities in order to lead to computationally efficient and reliable procedures. As we shall see we also suggest new algorithms. One should note at this point that in some situations of interest, such as tempering type algorithms (Geyer and Thompson 1995), property (b) above might be violated and instead the invariant distribution of P_θ might depend on $\theta \in \Theta$ (although only a non θ-dependent feature of this distribution π_θ might be of interest to us for practical purposes). We will not consider this case in depth here, but simply note that most of the arguments and ideas presented hereafter generally carry on to this slightly more complex scenario *e.g.* (Benveniste et al. 1990; Atchadé and Rosenthal 2005).

The choice of a criterion to optimise is clearly the first decision that needs to be made in practice. We discuss this issue in Sect. 4.1 where we point out that most sensible optimality or suboptimality criteria can be expressed in terms of expectations with respect to the steady state-distributions of Markov chains generated by P_θ for $\theta \in \Theta$ fixed, and make new suggestions in Sect. 5 which are subsequently illustrated on examples in Sect. 6. We will denote by θ^* a generic optimal value for our criteria, which is always assumed to exist hereafter.

In order to optimise such criteria, or even simply find suboptimal values for θ, one could suggest to sequentially run a standard MCMC algorithm with transition P_θ for a set of values of θ (either predefined or defined sequentially) and compute the criterion of interest (or its derivative etc.) once we have evidence that equilibrium has been reached. This can naturally be wasteful and we will rather focus here on a technique which belongs to the well known class of processes called controlled Markov chains (Borkar 1990) in the engineering literature, which we will refer to as controlled MCMC (Andrieu and Robert 2001), due to their natural filiation. More precisely we will assume that the algorithm proceeds as follows. Given a family of transition probabilities $\{P_\theta, \theta \in \Theta\}$ defined on X such that for any $\theta \in \Theta$, $\pi P_\theta = \pi$ (meaning that if $X_i \sim \pi$, then $X_{i+1} \sim \pi, X_{i+2} \sim \pi, \ldots$) and given a family of (possibly random) mappings $\{\theta_i : \Theta \times \mathsf{X}^{i+1} \to \Theta, i = 1, \ldots\}$, which encodes what is meant by optimality by the user, the most general form of a controlled MCMC proceeds as follows:

Algorithm 1 Controlled Markov chain Monte Carlo

- Sample initial values θ_0, $X_0 \in \Theta \times \mathsf{X}$.
- Iteration $i + 1$ ($i \geq 0$), given $\theta_i = \theta_i(\theta_0, X_0, \ldots, X_i)$ from iteration i
  1. Sample $X_{i+1} | (\theta_0, X_0, \ldots, X_i) \sim P_{\theta_i}(X_i, \cdot)$.
  2. Compute $\theta_{i+1} = \theta_{i+1}(\theta_0, X_0, \ldots, X_{i+1})$.

In Sect. 4.2 we will focus our results to particular mappings well suited to our purpose of computationally efficient sequential updating of $\{\theta_i\}$ for MCMC algorithms, which rely on the Robbins-Monro update and more generally on the stochastic approximation framework (Benveniste et al. 1990). However, before embarking on the description of practical procedures to optimise MCMC transition probabilities we will first investigate, using mostly elementary undergraduate level tools, some of the theoretical ergodicity properties of controlled MCMC algorithms.

Indeed, as we shall see, despite the assumption that for any $\theta \in \Theta$, $\pi P_\theta = \pi$, adaptation in the context of MCMC