A quick introduction to Markov chains and Markov chain Monte Carlo (revised version)

Rasmus Waagepetersen
Institute of Mathematical Sciences
Aalborg University

1 Introduction

These notes are intended to provide the reader with knowledge of basic concepts of Markov chain Monte Carlo (MCMC) and hopefully also some intuition about how MCMC works. For more thorough accounts of MCMC the reader is referred to e.g. Gilks *et al.* (1996), Gamerman (1997), or Robert and Casella (1999).

Suppose that we are interested in generating samples from a target probability distribution π on $\mathbb{R}^n$ and that π is so complex that we can not use direct methods for simulation. Using Markov chain Monte Carlo methods it is, however, often feasible to generate an ergodic Markov chain $X_1, X_2, \ldots$ which has π as equilibrium distribution, i.e. after a suitable burn-in period m, $X_{m+1}, X_{m+2}, \ldots$ provides a (correlated) sample from π which can be used e.g. for Monte Carlo computations.

Before we turn to MCMC methods we briefly consider in the next section the concepts of Markov chains and equilibrium distributions.

2 Examples of Markov chains and convergence to a stationary distribution

A Markov chain X is a sequence $X_0, X_1, X_2, \ldots$ of stochastic variables (in short notation $X = (X_l)_{l \geq 0}$) which for all $n > 0$ and all events $A_0, A_1, \ldots, A_n$

satisfies the following conditional independence property:

$$P(X_n \in A_n | X_{n-1} \in A_{n-1}, X_{n-2} \in A_{n-2}, \ldots, X_0 \in A_0) = P(X_n \in A_n | X_{n-1} \in A_{n-1}). \quad (1)$$

That is, the value of the nth variable depends on the past variables only through the immediate predecessor one timestep ahead. The variable X_l, $l \geq 0$ could e.g. designate the average temperature in Denmark on the lth day in 1998, and A_l could be the event that X_l is greater than a temperature T, so that $A = \{x \in \mathbb{R} | x \geq T\}$.

An independent sequence is one example of a Markov chain since in this case

$$P(X_n \in A_n | X_{n-1} \in A_{n-1}, X_{n-2} \in A_{n-2}, \ldots, X_0 \in A_0) = P(X_n \in A_n) \quad (2)$$

which does not depend on the past at all, due to the independence.

As a first specific example we consider a Markov chain on the discrete state space $E = \{0, 1\}$.

Example 1 A Markov chain X on $E = \{0, 1\}$ is determined by the initial distribution given by

$$p_o = P(X_0 = 0) \text{ and } p_1 = P(X_0 = 1),$$

and the one-step transition probabilities given by

$$p_{00} = P(X_{n+1} = 0 | X_n = 0),\ p_{10} = P(X_{n+1} = 0 | X_n = 1),$$
$$p_{01} = 1 - p_{00} \text{ and } p_{11} = 1 - p_{10}.$$

Often the one-step transition probabilities are gathered in a matrix

$$P = \begin{bmatrix} p_{00} & p_{10} \\ p_{10} & p_{11} \end{bmatrix}.$$

□

The next example is a Markov chain with the continuous state space $E = \mathbb{R}$, the real numbers.

Example 2 In this example we consider an autoregressive Markov chain of order 1 (an AR(1)). The chain is given by

a) $X_0 = \mu_0$ (a fixed value).

b) $X_l = \beta X_{l-1} + \epsilon_l$, $l \geq 1$,

where $(\epsilon_l)_{l\geq 1}$ is a sequence of independent and normally distributed "innovations" with $\epsilon \sim N(0, \sigma^2)$. That is, $X_1 = \beta X_0 + \epsilon_1$, $X_2 = \beta X_1 + \epsilon_2$ etc. It is clear that $(X_l)_{l\geq 0}$ forms a Markov chain, since in order to compute X_n the only knowledge required about the past is the value of the predecessor X_{n-1}.

It is easy to calculate the distribution of the variables X_l, $l \geq 0$. They are given as sums of the normal variables ϵ_l:

$$X_n = \beta X_{n-1} + \epsilon_n = \beta^2 X_{n-2} + \beta\epsilon_{n-1} + \epsilon_n = \cdots = \\ \beta^n \mu_0 + \beta^{n-1}\epsilon_1 + \cdots + \beta\epsilon_{n-1} + \epsilon_n = \beta^n \mu_0 + \sum_{l=1}^{n} \beta^{n-l}\epsilon_l, \quad (3)$$

and are therefore normal themselves. It thus only remains to calculate the means and variances:

$$EX_n = \beta^n \mu_0 + \sum_{l=1}^{n} \beta^{n-l} E\epsilon_l = \beta^n \mu_0 \quad (4)$$

$$VarX_n = \sum_{l=1}^{n} (\beta^{n-l})^2 Var\epsilon_l = \sum_{l=1}^{n} (\beta^2)^{n-l}\sigma^2 = \sigma^2 \frac{(\beta^2)^n - 1}{\beta^2 - 1}. \quad (5)$$

(it is here required that $\beta^2 \neq 1$). For the calculation of the variance the formula $\sum_{l=0}^{n} z^l = (z^{n+1} - 1)/(z - 1)$, $z \neq 1$ was used.

From these expressions we see that an AR(1) behaves very differently depending on whether $-1 < \beta < 1$ or $|\beta| > 1$. If $|\beta| < 1$ then EX_n tends to 0 and $Var(X_n)$ tends to $\sigma^2/(1 - \beta^2)$ as n tends to infinity. For large n, X_n thus approaches an $N(0, \sigma^2/(1 - \beta^2))$ distribution. If on the other hand $\beta > 1$, then both EX_n and $Var(X_n)$ tends to infinity as n tends to infinity. This behaviour is reflected by the simulations in Figure 1.

For $-1 < \beta < 1$ the normal distribution $N(0, \sigma^2/(1 - \beta^2))$ is a socalled *invariant* or *stationary* distribution. That is, if X_{n-1} is $N(0, \sigma^2/(1 - \beta^2))$ then this implies that also X_n is normal distributed with mean 0 and variance $\sigma^2/(1 - \beta^2)$. This is seen as follows:

$$E(X_n) = \beta E(X_{n-1}) = 0 = E(X_n),$$

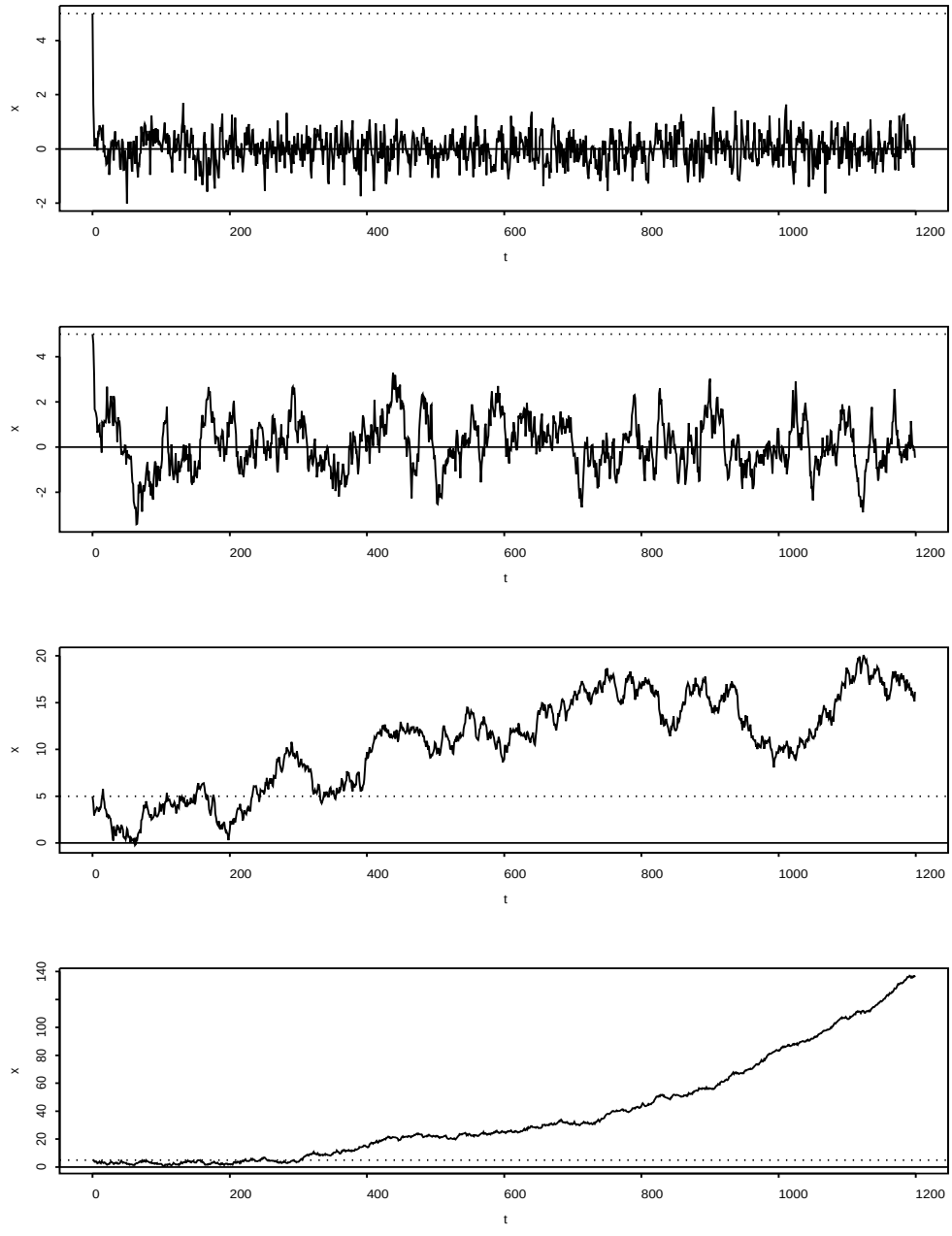

Figure 1: Simulations of AR(1)-chains. For all simulations, $\mu_0 = 5$ and $\sigma^2 = 0.25$. The values of β are, from top to bottom, 0.4, 0.9, 1.0, and 1.0025.

and

$$Var(X_n) = \beta^2 Var(X_{n-1}) + Var(\epsilon_n) = \frac{\beta^2\sigma^2}{1-\beta^2} + \sigma^2 =$$
$$\sigma^2(\frac{\beta^2}{1-\beta^2} + \frac{1-\beta^2}{1-\beta^2}) = \sigma^2\frac{1}{1-\beta^2} = Var(X_{n-1}).$$

An AR(1) with $|\beta| < 1$ thus illustrates the important properties of "lack of memory" and stochastic stability since the chain converges to an uniquely determined stationary distribution, regardless of the chosen initial value. □

The idea of MCMC is that for an arbitrary distribution π of interest, one can generate a Markov chain whose invariant distribution is given by π, which have the "lack of memory" property as for the AR(1) in Example 2, and which converges to the invariant distribution, so that samples of π can be obtained from the Markov chain.

3 Some theoretical concepts for Markov chains

Let $X = (X_l)_{l\geq 0}$ denote a Markov chain with state space E and let π denote a probability distribution on E. Then

- π is an *invariant distribution* for X if

$$X_l \sim \pi \Rightarrow X_{l+1} \sim \pi,\ l \geq 0$$

- X is *aperiodic* if there does not exist a disjoint subdivision of E into subsets $A_0, \ldots, A_{d-1},\ d \geq 2$, such that

$$P(X_l \in A_{(k+1) \bmod d} | X_{l-1} \in A_k) = 1,\ k = 0, \ldots, d-1,\ l \geq 1$$

- X is *irreducible* if for all $x \in E$ and all events A with $\pi(A) > 0$ there exist an $n \geq 1$ such that $P(X_n \in A | X_0 = x) > 0$.

- X is Harris recurrent if

$$P(\exists n : X_n \in A \mid X_0 = x) = 1$$

for all $x \in E$ and all events A with $\pi(A) > 0$.

Example 1 continued Suppose that π is a distribution on $\{0,1\}$ given by probabilities π_0 and π_1. Suppose that $P(X_{n-1}=0)=\pi_0$. Then π is an invariant distribution for the Markov chain given by the one-step transition probabilities p_{00} and p_{10} provided $P(X_{n-1}=0)=\pi_0$ implies $P(X_n=0)=\pi_0$, i.e.

$$\pi_0 = P(X_n=0) = P(X_n=0|X_{n-1}=0)\pi_0 + P(X_n=0|X_{n-1}=1)\pi_1 = p_{00}\pi_0 + p_{10}\pi_1 \Leftrightarrow \pi_0 = \frac{p_{10}}{p_{10}+p_{01}}.$$

In matrix form the condition is

$$\pi = P\pi.$$

where $\pi = \binom{\pi_0}{\pi_1}$. The Markov chain is periodic if $p_{01}=p_{10}=1$. □

Example 2 cntd. Let $\pi = N(0,\sigma^2/(1-\beta^2))$. An AR(1) is then π-irreducible: if $X_0=x$ then $X_1 \sim N(\beta x,\sigma^2)$ so that $P(X_1\in A|X_0=x)>0$ for any event A with $\pi(A)>0$. It is also aperiodic: assume first that the Markov chain is periodic and that $X_n=x\in A_k$ (where A_k is one of the subsets in the "periodic" splitting of E above). But $P(X_n\in A_k|X_{n-1}\in A_{k-1})=1>0$ implies $P(X_{n+1}\in A_k|X_n=x)>0$ so that $P(X_{n+1}\in A_{k+1 \bmod d}|X_n\in A_k)<1$. □

3.1 Convergence towards a stationary distribution

Suppose now that π is an invariant distribution for X, and that X is π-irreducible and aperiodic. Then under the further assumption of Harris recurrency, X_n converges in distribution to π for any chosen starting condition for X_0. This means, that for any event A,

$$P(X_n\in A)\to\pi(A),$$

so that X_n can be considered a simulation from the distribution π for large n. Harris recurrence holds under mild conditions for the Metropolis-Hastings samplers described in section 4, provided that irreducibility is fulfilled, see Tierney (1994) and Chan and Geyer (1994).

It is intuitively clear that irreducibility is required for convergence to the stationary distribution, since the chain must be able to reach all parts A of

the state space for which $\pi(A) > 0$. If the chain is periodic and started in A_0, say, then $P(X_n \in A_k) = 1$ whenever $n = 0, k, 2k, \ldots$, and zero otherwise, so that the distribution of X_n can never converge to π.

Convergence to a stationary distribution can e.g. be observed in the two upper plots in Figure 1.

3.2 Convergence of Markov chain Monte Carlo estimates

Let g be a function on E where we wish to estimate $E(g(Z))$, where $Z \sim \pi$. If X is Harris recurrent (irreducible) with stationary distribution π then the empirical average $\sum_{l=0}^{n} g(X_l)/n$ converges:

$$\frac{1}{n}\sum_{l=0}^{n} g(X_l) \to E(g(Z)), \text{ with probability one,}$$

as n tends to infinity, regardless of the chosen initial value for X_0. Expectations can thus be approximated by empirical averages just as for ordinary Monte Carlo. The correlation in the Markov chain however implies that the size n of the Markov chain sample needs to be greater than when independent simulations are used in ordinary Monte Carlo, in order to obtain a given level of accuracy.

4 The Metropolis-Hastings algorithm

Let π denote a complex, multivariate target distribution for a stochastic vector $Z = (Z_1, \ldots, Z_m)$, $m \geq 1$, and let f be the density of π. We shall now consider how one constructs a Markov chain $X = (X_l)_{l \geq 0}$ for which π is the invariant distribution. The constructed chain will then produce samples of π provided that the chain is irreducible and aperiodic. Note that the statespace of X is multidimensional, so that a state $X_l = (X_1^l, \ldots, X_m^l)$, $l \geq 0$ of the Markov chain has components $X_1^l, \ldots, X_m^l$. The initial state X_0 can be chosen rather arbitrarily, but it is advantageous if it belongs to the center of the target distribution π since convergence to the stationary distribution is then faster.

4.1 Metropolis-Hastings algorithm with simultaneous updating of all components

The Metropolis-Hastings algorithm (Hastings, 1971) is given in terms of a proposal kernel q. That is, for any $x \in E$, $q(x, \cdot)$ is a probability density on E. The Metropolis-Hastings algorithm now iterates the following steps:

1. Let $X_l = x = (x_1, \ldots, x_m)$ be the current state of the chain and generate a proposal Y from the proposal density $q(x, \cdot)$.

2. Generate a uniform number U on $[0, 1]$.

3. If
$$U < a(x, Y) = \min\left\{1, \frac{f(Y)q(Y, x)}{f(x)q(x, Y)}\right\}$$
then $X_{l+1} = Y$. Otherwise $X_{l+1} = X_l$.

Note that all components in the current state X_l are updated if the proposal Y is accepted. Note also that if the density $f(Y)$ is small compared to the density $f(x)$ of the current state, then Y will tend to be rejected, so that the chain stays in the stationary distribution.

This is perhaps clearer if one considers the Metropolis algorithm (Metropolis et al., 1953) which is the special case where q is symmetric (i.e. $q(x, z) = q(z, x)$) so that $a(x, y) = \min\{1, f(y)/f(x)\}$

Example 2 cntd. Remember that the AR(1) chain was not convergent when $|\beta| > 1$. We will now "Metropolize" the chain corresponding to the bottom plot in Figure 1. Given the current state $X_l = x$, the next state for the AR(1)-chain was generated from $N(\beta x, \sigma^2)$ with density

$$q(x, z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(z - \beta x)^2).$$

Let $\pi = N(-5, 1)$ with density $f(z) = \exp(-(z - (-5))^2/2)/\sqrt{2\pi}$. Instead of just accepting $X_{l+1} = Y$ where Y is generated from $N(\beta x, \sigma^2)$ we now accept or reject Y according to the acceptance probability $a(x, Y)$ (this is of course not exactly a natural way to simulate $N(-5, 1)$). A simulation of the "Metropolized" chain is given in Figure 2. □

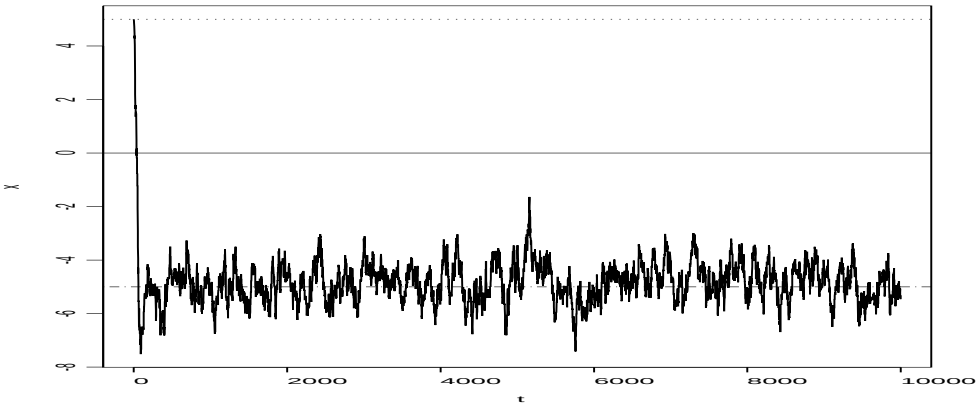

Figure 2: Simulation of "Metropolized" AR(1) chain with stationary distribution given by $N(-5, 1)$.

4.2 Single-site updating Metropolis-Hastings

We now assume that $m > 1$. For a single-site updating Metropolis-Hastings algorithm, the components $X_1^l, \ldots, X_m^l$ are updated individually in a random or systematic order. Often a systematic scheme is chosen so that the components are updated in turn starting with X_1^l, then X_2^l, and so forth.

Assume that the current state is $X_l = (X_1^l, \ldots, X_m^l) = x$ and that the jth component is to be updated. Then the next state X_{l+1} of the chain only differs from X_l on the jth component, i.e. $X_i^{l+1} = X_i^l$, $i \neq j$, and X_{l+1} is generated as follows:

1. A proposal Y_j is generated from a proposal density $q_j(x, \cdot)$. Let $Y = (X_1^l, \ldots, X_{j-1}^l, Y_j, X_{j+1}^l, \ldots, X_m^l)$
2. A uniform variable U on $[0, 1]$ is generated.
3. If
$$U < a_j(x, Y) = \min\left\{1, \frac{f(Y)q_j(Y, x_j)}{f(x)q_j(x, Y_j)}\right\}$$
then $X_{l+1} = Y$. Otherwise $X_{l+1} = X_l$.

Example (Gibbs sampler) The Gibbs sampler is an important example of the single-site Metropolis-Hastings sampler. In this case, $q_j(x, \cdot)$ is simply

the conditional density of Z_j given Z_i, $i \neq j$, i.e.

$$q_j(x, y_j) = f_j(y_j|x_i, i \neq j) = \frac{f(x_1, \ldots, x_{j-1}, y_j, x_{j+1}, \ldots, x_m)}{f_{-j}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)},$$

where $f_{-j}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$ is the marginal density of Z_j, $j \neq i$. It is easy to realize that in this case, $a_j(x, y_j) = 1$ so that the proposals are always accepted. □

4.3 Choice of proposal kernel

The target distribution π may be very complex and impossible to sample directly. The advantage of the Metropolis-Hastings algorithm is that it is only required to sample from the proposal kernel Q and we are free to choose any proposal kernel which is easy to sample from, as long as the resulting chain becomes irreducible and aperiodic. Aperiodicity is usually not a problem. It is e.g. enough that there is a positive probability of rejecting generated proposals.

When irreducibility cause problems, it is usually in cases where f, the density of π, is not positive everywhere. Consider e.g. a distribution on $\{0, 1\}^7$, say, where $\pi(x) = 0$ if $\sum_{i=1}^{8} x_i = 4$. Then a single-site updating algorithm is reducible because it can not move from an x with $\sum_{i=1}^{8} x_i < 4$ to an x' with $\sum_{i=1}^{8} x'_i > 4$.

On $\mathbb{R}^2$ one may consider a π whose density is zero outside the balls $b\big((1, 1), 0.5\big)$ and $b\big((2, 2), 0.5\big)$. In this case the Gibbs sampler becomes reducible (make a drawing) .

5 Some practical issues

The theory of MCMC tells us that the Markov chain eventually will produce samples from the target distribution if we run the chain for sufficiently long time. The difficult question is how long is enough. A hot research topic is "perfect simulation" which in fact answers this question in some cases, but perfect simulation is still a technique for specialists.

A useful way to assess whether convergence is achieved is to consider timeseries of various statistics derived from the Markov chain. Figure 2 e.g. shows a timeseries given by the "Metropolized" AR(1) itself. It seems that convergence has been reached after a burn-in of approximately 100 iterations.

In the Monte Carlo calculations one may feel inclined to discard the first steps of the Markov chain where the chain still have not reached the stationary distribution.

Ordinary timeseries methods are useful for analyzing output from Markov chains. Figure 3 shows autocorrelations estimated from the simulated chain in Figure 2. The chain is in fact highly autocorrelated and the estimate of

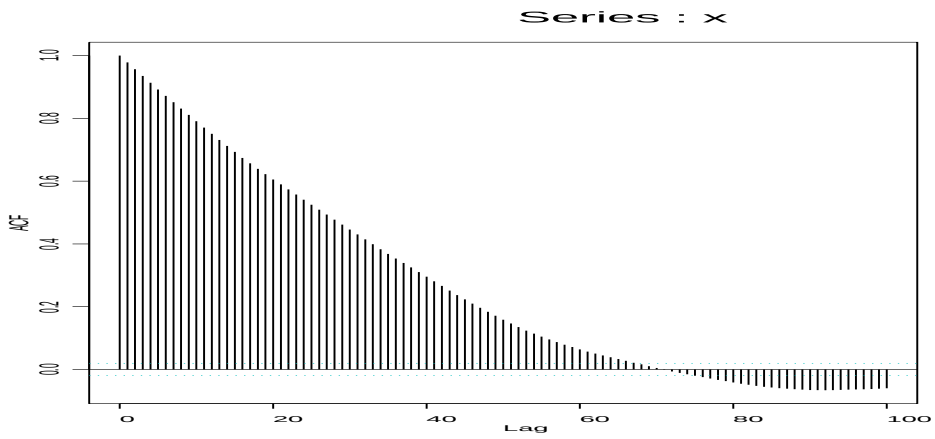

Figure 3: Autocorrelations calculated from the simulation in Figure 2.

the mean (-5) of the stationary distribution accordingly converges slowly, see Figure 4. Compare also with the convergence of the estimate based on independent simulations in Figure 5.

If the MCMC sample is highly autocorrelated, and a way to improve the sampler can not be found, then one may may sometimes wish to subsample the chain, i.e. to create a less correlated sample by retaining only every 10th, say, observation in the original sample. Subsampling throws information away and should in general not be used, but other factors may sometimes render subsampling advantageous. This may e.g. be the case if storage space in the computer is a problem or if the expense of calculating $g(X_l)$ is high, where $g(\cdot)$ is the function whose mean $E(g(Z))$ is to be estimated.

References

Chan, K. S. & Geyer, C. J. (1994). Discussion of the paper 'Markov chains for exploring posterior distributions' by Luke Tierney. *Annals of Statistics*

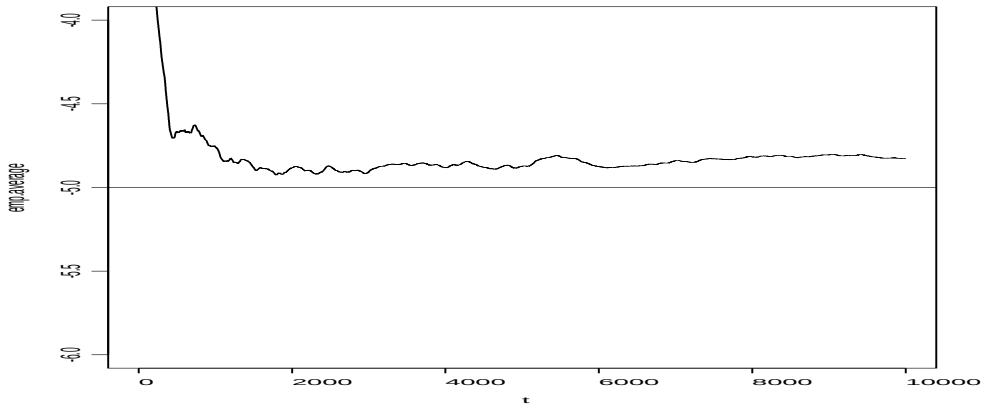

Figure 4: Convergence of empirical average as a function of sample size ("Metropolized" AR(1)).

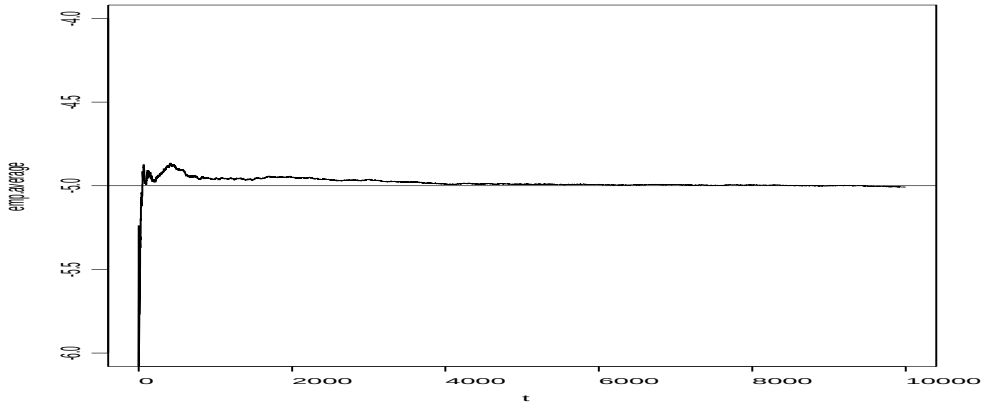

Figure 5: Convergence of empirical average as a function of sample size (independent simulations).

22, 1747–1747.

Gamerman, D. (1997). *Markov Chain Monte Carlo*. Chapman and Hall, London.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Robert, C. P. & Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics* **22**, 1701–1728.