Lecture notes for ST329 Topics in Statistics
Markov chain Monte Carlo

Elke Thönnes
Room 65, Ext. 72582
elke@stats.warwick.ac.uk

December 4, 2003

Contents

0 Preliminaries

This course is structured as follows. After the introduction, we discuss some examples of complex distributions. Then, after a revision of some concepts in Markov chain theory we see how to construct a Markov chain whose distribution converges to the desired target distribution. There are different ways of producing such chains, we will introduce the most common ones: Metropolis-Hastings Sampling and Gibbs Sampling. Once we know how to do MCMC, we learn about some of the implementational issues associated with it. For example, how long should the chain be run? When should I start sampling it? How should I use the samples?

- Relevant Literature:

  1. Markov Chain Monte Carlo by Dani Gamerman, Chapman & Hall, 1997.
  2. Monte Carlo Statistical Methods by Christian Robert and George Casella, Springer, 1999.
  3. Markov Chain Monte Carlo by W.R. Gilks, S. Richardson and D.J. Spiegelhalter, Chapman & Hall, 1996.
  4. Markov chains by J.R. Norris, Cambridge, 1997.

Please note that the numbering in these lecture notes does not necessarily coincide with the numbering of the handouts given out in the lecture! Also note that these notes contain some material which was not covered in the lecture and do not contain all of the handouts given out in the lecture!

1 Introduction

As we build more realistic and thus more complex statistical models, simulation has become an important and useful item in the statistical toolbox. When models are too complicated to be examined analytically, we may base inference on simulation. Rather than computing characteristics of a distribution, we simulate from it and then use the obtained samples to estimate these characteristics.

This course will give an overview on one particular class of simulation techniques: *Markov chain Monte Carlo* (MCMC). Some distributions are so complex that we cannot even sample them directly. In these cases MCMC may help out. MCMC essentially constructs a Markov chain whose distribution converges towards the target distribution. Thus, if we simulate the chain for long enough, then we can sample a distribution which approximates the target distribution and base our inference on these samples.

Buzz Task 1: Urn model
Suppose we have two urns and four balls. We repeatedly toss a fair coin. If the coin comes up tails, we pick the left urn, take a ball from it (if there is any) and put it into the right urn. Alternatively, if the coin comes up heads then we pick a ball from the right urn and put it into the left one. If the right urn is empty, then we do nothing. We may describe the number of balls in the left urn as a Markov chain which can take the values 0, 1, 2, 3 or 4. The state-flow diagram of the chain is given by the diagram in Figure 1.

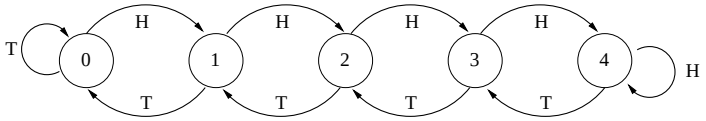


Figure 1: State-flow diagram of the urn model Markov chain. H stands for a coin toss which comes up heads and T for a coin toss that comes up tails.

We can simulate the chain using coin tosses as indicated above. Given a sequence of coin tosses we may produce a path of the Markov chain. Figure 2 illustrates a path that was started in state 3.

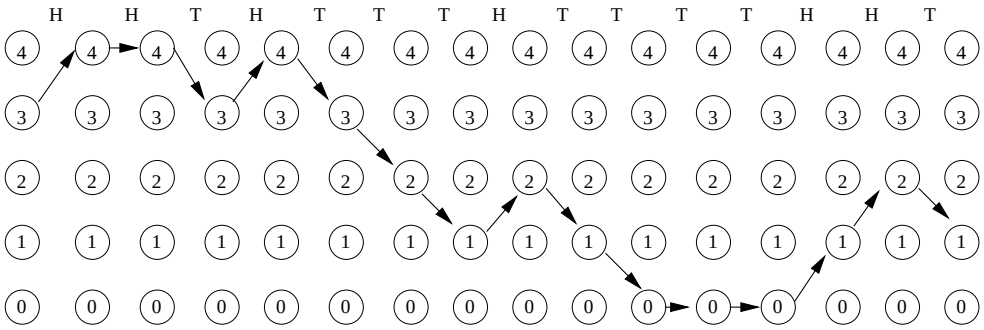


Figure 2: A path of the urn model chain.

The table below shows how many times the chain visits each state (not counting the initial state of the chain.)

State	0	1	2	3	4
Number of times visited	3	4	3	2	3

Thus the average value taken by the Markov chain is 28/15. We can also display the number of times each state was visited in a histogram as in Figure 3. Of course, for the urn model Markov chain we can easily compute

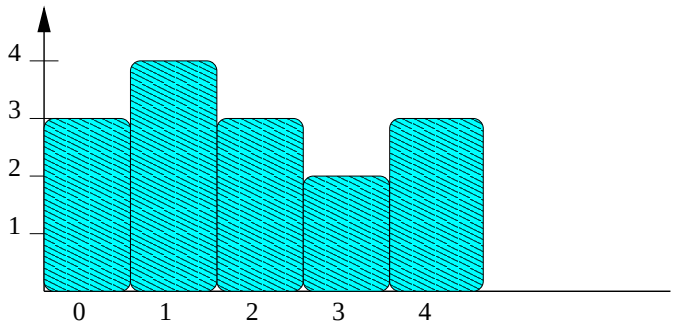


Figure 3: Histogram of states visited by the urn model chain.

and sample the equilibrium distribution which is uniform on the states 0, 1,

2, 3 and 4. However, the buzz task illustrates how we could have used the Markov chain to estimate probabilities of the equilibrium distribution using the histogram. The average value taken by the chain (28/15) is an estimate for the expected value of the stationary distribution which in this case is 2.

1.1 A bit of history (not exam relevant)

MCMC techniques have been used since the early 1950's. Metropolis et al. (1953) were the first to publish a method of simulation based on Markov chains. They used it as a tool to efficiently simulate the energy levels of atoms in a crystalline structure. Whereas physicists then started using MCMC for a variety of problems, it took statistics almost 20 years to catch up. In 1970 Hastings reconsidered the methods by Metropolis et al. and generalized them with a focus on statistical problems. However, it was not before the mid 1980's until MCMC took off in statistics. In 1984 Geman and Geman introduced an MCMC method, called the Gibbs Sampler, to researchers working in image analysis. Then, in the late 1980's, Bayesian statistics finally helped MCMC to establish itself in the statistical community. Nowadays, it is a widely used tool and a vibrant area of research as for example the MCMC preprint server demonstrates.

2 Complex distributions

As mentioned above, MCMC is concerned with sampling complex distributions. Here is a list of complex distributions some of which can be found in the cited literature.

2.1 Examples of complex distributions

1. **Permutations:** Example 4.32 in Ross[1] describes a conditional distribution on permutations. Suppose $\mathcal{S}$ is the space of all permutations $(x^{(1)}, \ldots, x^{(m)})$ of $\{1, 2, \ldots, m\}$ such that $\sum_{j=1}^{m} jx^{(j)} > c$ for some constant $c > 0$. Then let $\pi(x^{(1)}, \ldots, x^{(m)})$ be the uniform distribution on $\mathcal{S}$. (Distributions of this type often occur in coding theory).

[1] S. Ross, *Introduction to Probability Models*, 6th edition, 1997, Academic Press, San Diego

2. **Survival times:** Example 4.34 in Ross[2] describes a conditional distribution using Exponential distributions. For $i = 1, \ldots, m$ let $Y^{(i)}$ be an independent random variable with an Exponential distribution of mean $1/\theta_i$. Let π be the distribution of $(Y^{(1)}, \ldots, Y^{(m)})$ conditional on $\sum_{i=1}^{m} Y^{(i)} > a$ where $a > 0$ is some given constant. (This type of distribution is common in survival analysis).

3. **Points on a line:** Suppose $x^{(1)}, \ldots, x^{(m)}$ are the positions of m points on a line of length 1. Let $\pi(x^{(1)}, \ldots, x^{(m)})$ be the density that distributes the m points uniformly on the line conditional that no points are within distance d of each other. (Distributions of this type often occur in chemical settings where the points are centres of circular molecules of diameter d).

4. **Contingency tables:** Consider a 2×3 contingency table

n_{11}	n_{12}	n_{13}	$n_{1.}$
n_{21}	n_{22}	n_{23}	$n_{2.}$
$n_{.1}$	$n_{.2}$	$n_{.3}$	N

Suppose x is another contingency table which also has row sums $n_{1.}$, $n_{2.}$ and column sums $n_{.1}$, $n_{.2}$, $n_{.3}$, that is x is given by

x_{11}	x_{12}	x_{13}	$n_{1.}$
x_{21}	x_{22}	x_{23}	$n_{2.}$
$n_{.1}$	$n_{.2}$	$n_{.3}$	N

Under the hypothesis of independence a contingency table x with given row and column sums has a multivariate hypergeometric distribution:

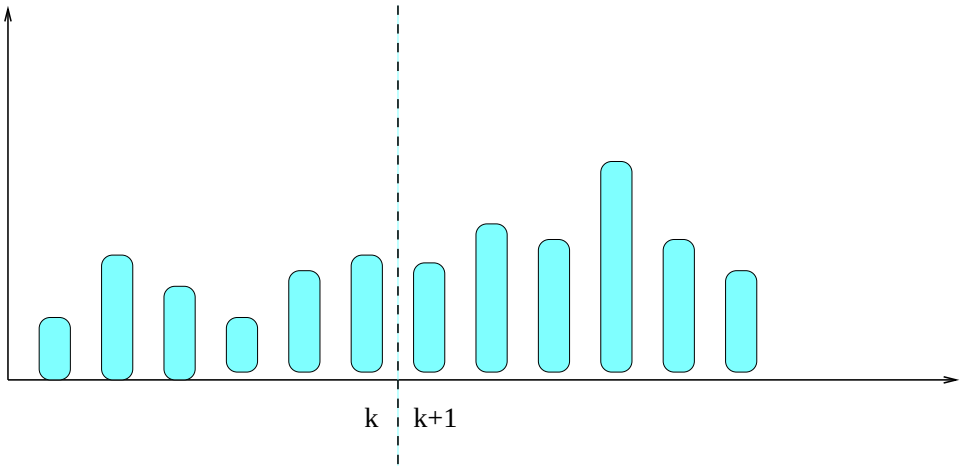
$$\pi(x) = \frac{n_{1.}!\, n_{2.}!\, n_{.1}!\, n_{.2}!\, n_{.3}!}{N! \prod_{i=1}^{2} \prod_{j=1}^{3} x_{ij}!}$$

This distribution is difficult to sample. However, to assess how probable an observed contingency table is under the assumption of independence, sampling may be neccessary because class counts are low.

[2] S. Ross, *Introduction to Probability Models*, 6th edition, 1997, Academic Press, San Diego

5. **Change-point problem:**[3] Suppose $\{y^{(1)}, \ldots, y^{(k)}\}$ is an iid sample from a Poisson distribution of mean λ and $\{y^{(k+1)}, \ldots, y^{(m)}\}$ an iid sample from a Poisson distribution of mean θ. We are interested in the parameters λ and θ and the change-point k. We adopt a Bayesian approach and choose as a prior for k a uniform distribution on $\{1, \ldots, m\}$. The prior distribution for λ is a Gamma$(\alpha, 1/\beta)$ distribution and for θ a Gamma$(\gamma, 1/\delta)$ distribution. Then the posterior distribution is given by

$$\pi(\lambda, \theta, k \mid \underline{y}) = c\lambda^{\alpha-1+\sum_{i=1}^{k} y^{(i)}} \; e^{-\lambda(\beta+k)} \; \theta^{\gamma-1+\sum_{i=k+1}^{m} y^{(i)}} \; e^{-\theta(\delta+m-k)}.$$



6. **Hierarchical model:**[4] Suppose λ is distributed according to a Beta(α, β) distribution where α and β depend on the observation $\underline{x}$. We are interested in the parameter θ which given λ has a Binomial(n, λ) distribution. For the parameter n we choose a prior distribution that is Poisson with mean γ. The posterior distribution is then given by

$$\pi(\theta, \lambda, n) \; = \; c \begin{pmatrix} n \\ \theta \end{pmatrix} \; \lambda^{\theta+\alpha-1}(1-\lambda)^{n-\theta+\beta-1} \; \frac{\gamma^n}{n!} \; e^{-\gamma}.$$

[3] B.P. Carlin, A.E. Gelfand, and A.F.M. Smith (1992) Hierarchical Bayesian analysis of changepoint problems, *Applied Statistics* 41, 389 - 405.

[4] G. Casella and E. George (1992): An introduction to Gibbs Sampling, *American Statistician* 46, 167 - 174.

7. **Bayesian mixture distribution:** Suppose that $\underline{y} = (y^{(1)}, \ldots, y^{(k)})$ is an iid sample from the following mixture density:

$$f(y|\mu_1, \mu_2) \quad = \quad 0.5 \exp\Big(-\frac{1}{2}(y-\mu_1)^2 \Big) + 0.5 \exp\Big(-\frac{1}{2}(y-\mu_2)^2 \Big).$$

Suppose $p(\mu_1, \mu_2)$ is a prior density for the parameters μ_1 and μ_2. Then the posterior density is given by

$$\begin{aligned} \pi(\mu_1, \mu_2|\underline{y}) \quad &\propto \quad p(\mu_1, \mu_2) \\ &\times \quad \prod_{i=1}^{k} \frac{1}{2}\Big[\exp\Big(-\frac{1}{2}(y^{(i)}-\mu_1)^2 \Big) + \exp\Big(-\frac{1}{2}(y^{(i)}-\mu_2)^2 \Big)\Big] \end{aligned}$$

Here the likelihood function is expensive to compute and so it is difficult to determine the constant of proportionality for the posterior density.

8. **Ising model: (not exam relevant)** The Ising model describes a lattice which at each site has a small dipole or spin which is directed upwards or downwards. Thus each site j may take a value $x^{(j)} \in \{-1, 1\}$ representing a downward respectively an upward spin. Figure 4 represents a configuration of an Ising model, where sites are displayed as circles. If the site carries an upwards spin then it is filled in black. If, on the other hand, the site has a downwards spin, it is filled in white. The spin at each site is influenced by the spins of its neighbour sites. Figure 4 shows the neighbours of a site on the boundary of the lattice and the neighbours of a site within the inner of the lattice. The ferro-magnetic Ising model gives a high probability to spin configurations in which many neighbouring sites have the same spin. Physicist tend to speak of energies rather than probabilities. A state that has small energy is easy to maintain and thus has a high probability. The energy function of a simple ferro-magnetic Ising model is given by

$$H(x) \quad = \quad -J \sum_{i \sim j} x^{(i)} x^{(j)}.$$

Here, "$i \sim j$" means that site i is a neighbour of site j. The constant J is positive for the ferro-magnetic Ising model. (If J is negative, then we call it the anti-ferromagnetic Ising model.) Based on the energy function H we can now define the probability mass function:

$$\pi(x) \quad = \quad \frac{1}{Z} \exp(-H(x)).$$

Here x is a spin configuration on a lattice. Thus, if the lattice has m sites, then $x = (x^{(1)}, \ldots, x^{(m)})$ is an m-dimensional vector of 1's and -1's. We call Z the partition function, but it is nothing else but the inverse of the normalizing constant for π.

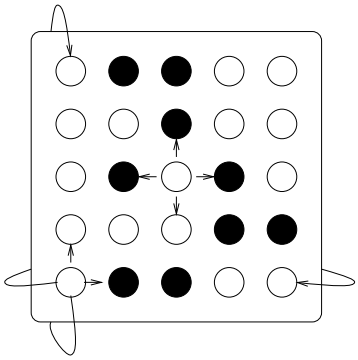


Figure 4: An Ising configuration. Black disks represent sites with an upwards spin and white disks sites with a downwards spin. The arrows show the neighbours for a site in the inner of the lattice and for a site on the boundary of the lattice.

2.2 Types of complex distributions

A common problem of a complex distribution is that we cannot compute in closed form the normalising constant of its density or probability mass function (pmf).

2.2.1 Conditional distributions

Suppose we have a standard pmf or density f but now we condition on the event $\mathcal{E}$. Then the normalising constant c of the conditional pmf or density is given by

$$\frac{1}{c} = \begin{cases} \sum_{x \in \mathcal{E}} f(x) & \text{if } f \text{ is a pmf,} \\ \int_{\mathcal{E}} f(x) dx & \text{if } f \text{ is a density.} \end{cases}$$

The constant c may be difficult to compute if the event $\mathcal{E}$ is very complex. The conditional density or pmf π given the event $\mathcal{E}$ is then

$$\pi(x) \quad = \quad \frac{1}{c}\, f(x)\, \mathbf{1}_{[x \in \mathcal{E}]}$$

where $\mathbf{1}$ is the indicator function.

Example 2.1 (Permutations): Consider the first example in the previous section which describes a distribution on a set of permutations. Suppose $m = 3$. The list of all permutations and the corresponding value of $\sum_{j=1}^{3} j x^{(j)}$ is given in the table below:

permutation	$\sum_{j=1}^{3} j x^{(j)}$
(1,2,3)	14
(1,3,2)	13
(2,1,3)	13
(2,3,1)	11
(3,1,2)	11
(3,2,1)	10

For example, if $c = 12$ then $\mathcal{S} = \Big\{(1,2,3), (1,3,2), (2,1,3)\Big\}$ and $\pi(x)$ gives probability $1/3$ to each element $x \in \mathcal{S}$.

2.2.2 Posterior distribution

MCMC is often used within Bayesian inference. Suppose we have data y from a distribution with likelihood $L(\theta|y)$, where θ is the set of parameters specifying the distribution. Given a prior $p(\theta)$ for the parameters, we can then use Bayes theorem to derive the posterior

$$\pi(\theta \mid y) \quad \propto \quad p(\theta)\, L(\theta|y).$$

The normalizing constant c of π is implicitly defined as

$$c^{-1} \quad = \quad \int p(\theta)\, L(\theta|y)\, d\theta.$$

However, if we do not use conjugate priors then it is often quite hard to compute the above integral.

Example 2.2 (Change point problem): In the change point example in section 2.1 the parameters consist of the discrete change point k and the continuous parameters λ and θ. The prior of the parameter k is given by

$$p(k) \quad = \quad \frac{1}{m}\, \mathbf{1}_{[k\in\{1,\ldots,m\}]}.$$

The parameter λ has prior density

$$p(\lambda) \quad = \quad c_1 \lambda^{\alpha-1} \exp(-\beta\lambda)$$

and the parameter θ has prior density

$$p(\theta) \quad = \quad c_2 \lambda^{\alpha-1} \exp(-\gamma\theta).$$

Moreover, the likelihood for k, λ and θ is given by

$$L(k,\lambda,\theta|y) \quad = \quad \prod_{i=1}^{k} \frac{\lambda^{y^{(i)}}}{y^{(i)}!} \exp(-\lambda) \prod_{i=k+1}^{m} \frac{\theta^{y^{(i)}}}{y^{(i)}!} \exp(-\theta).$$

We can derive the unnormalised posterior density by multiplying the prior pmf and densities with the likelihood. An expression for the posterior density is given in section 2.1.

2.2.3 Models with missing data

Example 2.3 (Bayesian mixture distribution): Suppose we sample y as follows. Toss a fair coin. If the coin comes up heads then we sample from $\mathcal{N}(\mu_1, 1)$ otherwise we sample from $\mathcal{N}(\mu_2, 1)$. The y has the density

$$f(y|\mu_1,\mu_2) \quad = \quad 0.5 \exp\Big(-\frac{1}{2}(y-\mu_1)^2\Big) + 0.5 \exp\Big(-\frac{1}{2}(y-\mu_2)^2\Big).$$

Now consider the data $\underline{y} = (y^{(1)}, \ldots, y^{k)})$ which is iid each with density f. Then the joint density for data $\underline{y}$ is

$$f(\underline{y}|\mu_1,\mu_2) \quad = \quad \prod_{i=1}^{k} 0.5 \exp\Big(-\frac{1}{2}(y^{(i)}-\mu_1)^2\Big) + 0.5 \exp\Big(-\frac{1}{2}(y^{(i)}-\mu_2)^2\Big)$$

Now, if $p(\mu_1, \mu_2)$ is the prior density for μ_1 and μ_2 then the posterior density for μ_1 and μ_2 is given by

$$\pi(\mu_1, \mu_2|\underline{y}) \quad \propto p(\mu_1, \mu_2) \prod_{i=1}^{k} \Big[\exp\Big(-\frac{1}{2}(y^{(i)} - \mu_1)^2 \Big) \exp\Big(-\frac{1}{2}(y^{(i)} - \mu_2)^2 \Big) \Big].$$

Suppose we observed the sequence $I = (I_1, \ldots, I_k$ of coin tosses that produced the data, that is $I_j = 1$ if the jth coin toss came up heads and $I_j = 0$ if it came up tails. Then the density of $\underline{y}$ given I is

$$f(\underline{y}|\mu_1, \mu_2, I) \quad = \quad \prod_{i:I_i=1} \exp\Big(-\frac{1}{2}(y^{(i)} - \mu_1)^2 \Big) \prod_{i:I_i=0} \exp\Big(-\frac{1}{2}(y^{(i)} - \mu_2)^2 \Big).$$

Then, if we assume a uniform prior for I the joint posterior density of μ_1,μ_2 and I is given by

$$\begin{aligned} \pi(\mu_1, \mu_2, I|\underline{y}) \quad &\propto \quad p(\mu_1, \mu_2)(\frac{1}{2})^k \prod_{i:I_i=1} \exp\Big(-\frac{1}{2}(y^{(i)} - \mu_1)^2 \Big) \prod_{i:I_i=0} \exp\Big(-\frac{1}{2}(y^{(i)} - \mu_2)^2 \Big) \\ &= \quad p(\mu_1, \mu_2) \exp\Big(-\frac{1}{2}(\sum_{i:I_i=1} (y^{(i)} - \mu_1)^2 + \sum_{i:I_i=0} (y^{(i)} - \mu_2)^2 \Big). \end{aligned}$$

This is of course a much easier expression than the one given in Section 2.1 as it involves computing much fewer terms. Usually we will not know the missing data I but we will see how we can use MCMC to infer it.

3 Markov chain theory

As the name suggests, MCMC is based on Markov chains. In this section we review some of the concepts in Markov chain theory which are important for MCMC.[5]

Let us start with the basics and first define a Markov chain. We assume that the Markov chain lives on a state space $\mathcal{S}$, for example the non-negative integers, and evolves in discrete time.

[5]Most of this material was discussed in ST202 *Stochastic Processes*!

Definition 3.1 (Markov chain): A random process $X = \{X_n, n = 0, 1, 2, \dots\}$ taking values in $\mathcal{S}$ is a *Markov chain*, if

$$\begin{aligned}\mathbb{P}\Big(X_{n+1} \in A \mid X_n = k_n, X_{n-1} = k_{n-1}, \dots, X_0 = k_0\Big) \\ = \mathbb{P}\Big(X_{n+1} \in A \mid X_n = k_n\Big)\end{aligned}$$

for all $n \geq 0$ and $k_0, \dots, k_n \in \mathcal{S}$.

Thus, given we know the present value of the chain (at time n), the value the chain takes in the *future* (at time $n+1$) does not depend on values in the past (at times before n).

Definition 3.2 (Transition probabilities):

1. Let $\mathcal{S}$ be discrete. The transition probabilities of the Markov chain X are given by

$$\mathbb{P}\Big(X_{n+1} = y \mid X_n = x\Big) \quad = \quad p(x, y).$$

2. Suppose $\mathcal{S}$ is continuous. Then

$$\mathbb{P}\Big(X_{n+1} \in A \mid X_n = x\Big) \quad = \quad \int_A p(x, y)dy$$

where for given $x \in \mathcal{S}$ the function $p(x, y)$ is a density, the transition density.

Definition 3.3 (Distribution at time n):

1. Suppose $\mathcal{S}$ is discrete. The initial distribution $q^{(0)}(\cdot)$ of X is defined as

$$q^{(0)}(x) \quad = \quad \mathbb{P}(X_0 = x), \qquad x \in \mathcal{S}.$$

The distribution of X at time n is given by:

$$q^{(n)}(x) = \mathbb{P}(X_n = x) \quad = \quad \sum_{y \in \mathcal{S}} q^{(n-1)}(y)\, p(y, x).$$

2. Suppose $\mathcal{S}$ is continuous and that the initial distribution of X has density $q^{(0)}(x)$. Then

$$\mathbb{P}(X_0 \in A) \quad = \quad \int_A q^{(0)}(x)dx.$$

We can compute the density of the distribution of X at time n as follows:

$$q^{(n)}(x) \quad = \quad \int_{\mathcal{S}} q^{(n-1)}(y)\ p(y,x)dy$$

(In this course we mostly consider time-homogeneous Markov chains, that is Markov chains whose transition probabilities do not depend on the time.)

Within MCMC we use Markov chains which have a limit distribution. To make sure that the distribution of the Markov chain does actually converge to a limit, the chain needs to be well-behaved. More specifically, we need the chain to be *ergodic*. A discrete time Markov chain on a discrete state space is ergodic, if it is irreducible, aperiodic and positive recurrent.

Theorem 3.4 Equilibrium distribution
The distribution of an *ergodic* Markov chain converges to a limit distribution, that is

$$\lim_{n\to\infty} q^{(n)}(x) \quad = \quad \pi(x),$$

where π is the density or pmf of the limit distribution. We also call the limit distribution the *equilibrium* distribution or the *stationary* distribution.

Standing Assumption:

In the following we assume that the Markov chain X on $\mathcal{S}$ is ergodic and that $\pi(x) > 0$ for all $x \in \mathcal{S}$!

Theorem 3.4 holds for any initial distribution. Thus, if we run an ergodic Markov chain for a long time, then it settles down to a statistical equilibrium, regardless of its starting point. Furthermore, if we start a chain in equilibrium then it remains in equilibrium. This can be easily shown as follows. Suppose $\mathcal{S}$ is discrete (the proof for continuous $\mathcal{S}$ is analogous). For $n \geq 2$

$$q^{(n+1)}(y) \quad = \quad \sum_{x\in\mathcal{S}} q^{(n)}(x)\ p(x,y)$$

and so

$$\lim_{n\to\infty} q^{(n+1)}(y) \quad = \quad \lim_{n\to\infty} \sum_{x\in\mathcal{S}} q^{(n)}(x)\, p(x,y) \quad = \quad \sum_{x\in\mathcal{S}} \lim_{n\to\infty} q^{(n)}(x)\, p(x,y)$$

It follows that

$$\pi(y) \quad = \quad \sum_{x\in\mathcal{S}} \pi(x)\, p(x,y).$$

We call the last equation the *general balance* equation. It shows that the transition probabilities of X preserve the equilibrium. This means that once the chain is in equilibrium it remains in equilibrium. That is why we call a distribution which satisfies the general balance equation the *invariant* distribution of X. If X is ergodic, then we can use the general balance equations to show that π is the equilibrium distribution for the chain X. However, often it is difficult to find transition probabilities/densities which solve the equations given by general balance. Much easier are the *detailed balance* equations.

Lemma 3.5 Detailed balance: Suppose π satisfies

$$\pi(x)\, p(x,y) \quad = \quad \pi(y)\, p(y,x)$$

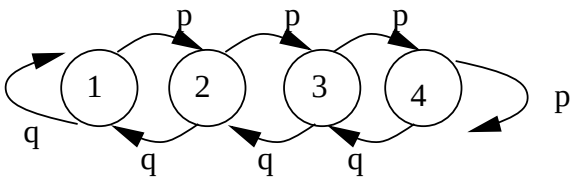
for all $x, y \in \mathcal{S}$, where $p(x,y)$ are the transition probabilities/densities of an ergodic Markov chain X. Then π defines the stationary distribution of X.

Proof:
Suppose $\mathcal{S}$ is continuous. Then, due to detailed balance

$$\int_{\mathcal{S}} \pi(y)p(y,x)dy \quad = \quad \int_{\mathcal{S}} \pi(x)p(x,y)dy \quad = \quad \pi(x)\int_{\mathcal{S}} p(x,y)dy \quad = \quad \pi(x).$$

The proof for discrete $\mathcal{S}$ follows analogously.

Note that detailed balance is not necessary for general balance! (If detailed balance holds, then the chain is time-reversible. This means in equilibrium the chain behaves the same whether it is run forwards or backwards in time. However, there are many ergodic Markov chains which are not time-reversible.)



Example 3.6: Consider the Markov chain with the state-flow diagram shown above. Then the general balance equations of the chain are:

$$\begin{aligned}
\pi(1) &= q\pi(1) + q\pi(2) \\
\pi(2) &= p\pi(1) + q\pi(3) \\
\pi(3) &= p\pi(2) + q\pi(4) \\
\pi(4) &= p\pi(3) + p\pi(4)
\end{aligned}$$

The detailed balance equations are given by

$$\begin{aligned}
p\pi(1) &= q\pi(2) \\
p\pi(2) &= q\pi(3) \\
p\pi(3) &= q\pi(4)
\end{aligned}$$

It follows that

$$\begin{aligned}
\pi(2) &= \frac{p}{q}\pi(1) \\
\pi(3) &= \frac{p}{q}\pi(2) = \left(\frac{p}{q}\right)^2\pi(1) \\
\pi(4) &= \frac{p}{q}\pi(3) = \left(\frac{p}{q}\right)^3\pi(1)
\end{aligned}$$

As $1 = \pi(1) + \pi(2) + \pi(3) + \pi(4)$ we have that

$$\pi(1) = \frac{1 - \frac{p}{q}}{1 - (\frac{p}{q})^4}.$$

Note that detailed balance equations are usually easier to solve than general balance equations.

The usefulness of MCMC is based on the following important theorem for ergodic Markov chains.

Theorem 3.7 Ergodic theorem: Let f be some real function and X an ergodic Markov chain. Consider the ergodic average

$$\overline{f}_N \quad = \quad \frac{1}{N}\sum_{n=1}^{N} f(X_n).$$

Now, suppose that Y has the equilibrium distribution of X. If $\mathbb{E}(|f(Y)|) < \infty$ then, as $N \to \infty$ the ergodic average $\overline{f}_N$ converges to $\mathbb{E}(f(Y))$ with probability one.

We can exploit the ergodic theorem as follows. Suppose we would like to know the expectation of the random variable Y with distribution given by the density π. However, we cannot compute $\mathbb{E}(Y) = \int x\ \pi(x)dx$. Fortunately, we can construct an ergodic Markov chain whose stationary distribution has density π. Then we run X up to some large time N and estimate $\mathbb{E}(Y)$ by $\frac{1}{N}\sum_{n=1}^{N} X_n$. Theorem 3.7 tells us that for sufficiently large N, our estimate will be close to $\mathbb{E}(Y)$. This is the main idea behind MCMC.

There are many implementational issues associated with this construction. For example, how do we know how accurate our estimate is? We will discuss these implementational issues later on, but first let us see how we can construct a Markov chain whose equilibrium distribution is the target distribution given by π.

4 MCMC Algorithms

4.0 Introduction

The common approach in Markov chain theory (see ST202) is to define transition probabilities which lead to an ergodic Markov chain and then to examine which equilibrium it has. MCMC reverses this direction. It starts with an equilibrium distribution and then tries to find an ergodic Markov chain

which has the desired equilibrium distribution. For any given distribution there are usually many Markov chains which have the required equilibrium distribution. Thus there is a variety of ways in which to construct a Markov chain whose distribution converges to the target distribution.

It is actually not very difficult to find a Markov chain whose invariant distribution is the target distribution. There is a set of methods, so-called "samplers", which we can use to define such a Markov chain. If the constructed chain is ergodic then we can proceed by simulating that chain. In the following we will learn about the two most common samplers.

4.1 Metropolis-Hastings Sampler

The transitions of a Metropolis-Hastings chain are produced as follows. First, we choose for each $x \in \mathcal{S}$ a pmf (if $\mathcal{S}$ is discrete) or density (if $\mathcal{S}$ is continuous) $q(x, \cdot)$ on $\mathcal{S}$. Thus $q(x, \cdot)$ specifies the transition probabilities/densities of a Markov chain on the state space $\mathcal{S}$. These transition probabilities/densities $q(x, \cdot)$ should be such that they are easy to sample.

Now suppose the current state of our Markov chain is $X_n = x$. Then we sample a state z according to $q(x, \cdot)$. We propose this state z as the new state of the chain and accept it with probability

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

If the proposed state z is accepted then the Markov chain moves to z, that is $X_{n+1} = z$. Otherwise the chain remains in x, that is $X_{n+1} = x$. We summarize this procedure in the following definition.

Definition 4.1 Metropolis-Hastings Sampler: Choose transition probabilites/densities $q(x, y), x, y \in \mathcal{S}$. These define the proposal distributions. Now suppose $X_n = x \in \mathcal{S}$. Proceed as follows:

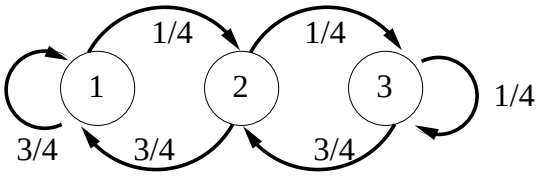
1. Sample $Z = z$ from $q(x, z)$, $z \in \mathcal{S}$.

2. Accept $Z = z$ with probability

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

   If $Z = z$ is accepted set $X_{n+1} = z$. If $Z = z$ is not accepted set $X_{n+1} = x$.

Example 4.2: Suppose $\pi(j) = 1/3$ for $j \in \{1, 2, 3\}$. Of course we can easily sample π directly! However, let us look at how we would produce a Markov chain whose stationary distribution is π using Metropolis-Hastings Sampling. Suppose we choose the following proposal distributions $q(i, j)$ as indicated in the flow-diagram below:



Now suppose that $X_n = 2$. Then we proceed through the following steps:

1. According to our state-flow diagram of proposal distributions we propose $Z = 3$ with probability $1/4$ and $Z = 1$ with probability $3/4$. Suppose we proposed $Z = 1$.

2. Now we need to compute the acceptance probability. We have

$$\alpha(2,1) \quad = \quad \min\left\{1, \frac{\frac{1}{3}\,\frac{1}{4}}{\frac{1}{3}\,\frac{3}{4}}\right\} \quad = \quad 1/3.$$

   Thus we set $X_{n+1} = 1$ with probability $1/3$ and $X_{n+1} = 2$ with probability $2/3$.

Notice that the acceptance probability $\alpha(x, y)$ is based on ratios of π, thus we do not need to know the normalizing constant of π to be able to compute this probability. The acceptance probability is chosen such that detailed balance holds, so let us have a look at the detailed balance equations of the Metropolis-Hastings chain. Suppose $\mathcal{S}$ is discrete. Recall that the chain moves to a new state y if this state was proposed and accepted. This happens with probability $q(x, y)\alpha(x, y)$. This is the probability of going from x to y when $y \neq x$. Now consider the probability of going from x to x. This can happen in two ways. Firstly we may propose x as the new state and then accept it, which happens with probability $q(x, x)\alpha(x, x)$. Secondly, we may propose some state y and reject it, in which case the chain remains in x. This

occurs with probability

$$r(x) \quad = \quad \sum_{y \in \mathcal{S}} q(x,y)\Big(1 - \alpha(x,y)\Big).$$

Thus, in summary, the transition probabilities of the Metropolis-Hastings chain are given by

$$p(x,y) \quad = \quad q(x,y)\alpha(x,y) + \mathbf{1}_{[x=y]} r(x).$$

An analogous proof for continuous $\mathcal{S}$ leads to the following Lemma.

Lemma 4.3 The transition probabilities/densities $p(x,y)$ for the Metropolis-Hastings sampler are given by

$$p(x,y) \quad = \quad q(x,y)\alpha(x,y) + \mathbf{1}_{[x=y]} r(x).$$

where

$$r(x) \quad = \quad \begin{cases} \sum_{y \in \mathcal{S}} q(x,y)\Big(1 - \alpha(x,y)\Big) & \text{if } \mathcal{S} \text{ is discrete} \\ \int_{\mathcal{S}} q(x,y)\Big(1 - \alpha(x,y)\Big) dy & \text{if } \mathcal{S} \text{ is continuous} \end{cases}$$

Now we are in a position to check the detailed balance equations.

Lemma 4.4: The Metropolis-Hastings chain satisfies detailed balance with respect to π.
Proof: We have

$$\begin{aligned} \pi(x)p(x,y) &= \pi(x)q(x,y)\alpha(x,y) + \mathbf{1}_{[x=y]}\pi(x)r(x) \\ &= \min\Big\{\pi(x)q(x,y), \pi(y)q(y,x)\Big\} + \mathbf{1}_{[x=y]}\pi(x)r(x) \\ &= \pi(y)q(y,x)\min\Big\{\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}, 1\Big\} + \mathbf{1}_{[x=y]}\pi(y)r(y) = \pi(y)p(y,x), \end{aligned}$$

where $\mathbf{1}$ denotes the indicator function. Therefore detailed balance holds. Thus if the Metropolis-Hastings chain is ergodic, then its distribution converges towards π. However, we still need to show that the Markov chain is actually ergodic.

Let us now look at some more examples. The first example is about a mixture distribution. Continuous mixture distributions with two components have densities of the form

$$f(x) \quad = \quad pf_1(x) + (1-p)f_2(x)$$

where $0 < p < 1$ and $f_i(x)$ is a density. We can sample mixtures by sampling x from $f_1(\cdot)$ with probability p and from $f_2(\cdot)$ with probability $1-p$. In the following two examples we show how to sample two mixture distributions using MCMC (although these could be sampled directly).

Example 4.5: Bimodal Normal mixture distribution

- The target density is

$$\pi(x) \quad = \quad p\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp(-0.5(x-\mu_1)) + (1-p)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp(-0.5(x-\mu_2))$$

  where $0 < p < 1$. The figure below shows the density above for $\sigma_1 = \sigma_2 = 1$, $\mu_1 = 4$, $\mu_2 = -4$ and $p = 0.8$.

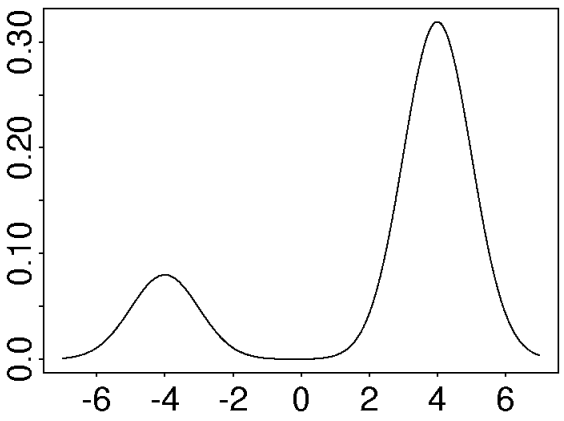


Figure 5: A normal mixture distribution.

- The proposal density:
  We sample z from a standard Normal density and propose $y = x + z$ as our new state. Thus $y \sim \mathcal{N}(x, 1)$ and our proposal density is

$$q(x, y) \quad = \quad \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(y - x)^2 \Big).$$

- The acceptance probability:

$$\begin{aligned} \alpha(x, y) &= \min\Big\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\Big\} \\ &= \min\Big\{1, \frac{\pi(y)\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x - y)^2)}{\pi(x)\frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}(y - x)^2)}\Big\} \\ &= \min\Big\{1, \frac{\pi(y)}{\pi(x)}\Big\} \end{aligned}$$

- The Metropolis-Hastings sampler proceeds as follows:

  1. Choose $X_0 = x_0 \in \mathbb{R}$.
  2. Suppose $X_n = x$. Sample $z \sim \mathcal{N}(0, 1)$ and set $y = x + z$. Accept y with probability $\min\{1, \frac{\pi(y)}{\pi(x)}\}$. If accepted set $X_{n+1} = y$, else set $X_{n+1} = x$.

Example 4.6 (Permutations): Ross[6] describes a Metropolis-Hastings algorithm for the distribution on permutations in section 2.1.

- Here the target distribution is

$$\pi(\underline{x}) \quad = \quad \frac{1}{|S|}\mathbf{1}_{[\underline{x} \in \mathcal{S}]} \quad = \quad \frac{1}{|S|}\mathbf{1}_{[\sum_{j=1}^{m} j x^{(j)} > c]}$$

  where $\underline{x} = (x^{(1)}, \ldots, x^{(m)})$.

[6] S. Ross, *Introduction to Probability Models*, 6th edition, 1997, Academic Press, San Diego, p. 213f

- The proposal distributions:
  We can generate proposals as follows. Let $\underline{x} = (x^{(1)}, \ldots, x^{(m)}) \in \mathcal{S}$. We propose a new state by choosing at random $i \in \{1, \ldots, m\}$ and proposing to swap $x^{(i)}$ with $x^{(i+1)}$ if $i < m$ and $x^{(m)}$ with $x^{(1)}$ otherwise. For example if $\underline{x} = (1, 3, 4, 2)$ and $i = 2$ then the newly proposed state is $\underline{z} = (1, 4, 3, 2)$. If $\underline{x} = (1, 3, 4, 2)$ and $i = 4$ then the newly proposed state is $\underline{z} = (2, 3, 4, 1)$. Let $N(\underline{x})$ be the set of permutations resulting from $\underline{x}$ by swapping one of the coordinates. Then $N(\underline{x})$ has m elements and

$$q(\underline{x}, \underline{z}) \quad = \quad \frac{1}{m} \, \mathbf{1}_{[\underline{z} \in N(\underline{x})]}.$$

- The acceptance probability is then given by

$$\begin{aligned} \alpha(\underline{x}, \underline{z}) &= \min\left\{1, \frac{\pi(\underline{z})\ q(\underline{z}, \underline{x})}{\pi(\underline{x})\ q(\underline{x}, \underline{z})}\right\} \\ &= \min\left\{1, \frac{\frac{1}{|\mathcal{S}|}\mathbf{1}_{[\underline{z} \in \mathcal{S}]}\frac{1}{m}\mathbf{1}_{[\underline{x} \in N(\underline{z})]}}{\frac{1}{|\mathcal{S}|}\mathbf{1}_{[\underline{x} \in \mathcal{S}]}\frac{1}{m}\mathbf{1}_{[\underline{z} \in N(\underline{x})]}}\right\} \\ &= \min\left\{1, \mathbf{1}_{[\underline{z} \in \mathcal{S}]}\right\} \quad = \quad \mathbf{1}_{[\underline{z} \in \mathcal{S}]} \end{aligned}$$

- We can now use this set-up to produce a Metropolis-Hastings chain. Choose $X_0 \in \mathcal{S}$, for example $X_0 = (1, \ldots, m)$. Now suppose $X_n = \underline{x} \in \mathcal{S}$. Proceed as follows:

  1. Choose $i \in \{1, \ldots, m\}$ at random and propose to swap the ith component leading to the permutation $\underline{z}$.
  2. If $\underline{z} \in \mathcal{S}$, that is $\sum_{j=1}^{m} j z^{(j)} > c$, then accept $\underline{z}$ and set $X_{n+1} = \underline{z}$. If $\underline{z} \notin \mathcal{S}$, reject $\underline{z}$ and set $X_{n+1} = \underline{x}$.

Example 4.7 (Contingency tables): Suppose we would like to sample a 2×3 contingency table with given row and column sums under the hypothesis of independence, see the 4th example in section 2.1. Develop a Metropolis-Hastings Sampler to sample the distribution $\pi(x)$ of a contingency table x under the hypothesis of independence using the questions below as guidance.

1. How many cells need to be filled such that all other cells are determined by the row and column sums? Which values can these cells take?

We have a choice for two cells, say x_{11} and x_{12}. The first cell can take any integer value x_{11} between 0 and $\min\{n_{.1}, n_{1.}\}$. Given the value in the first cell, the second cell can then take an integer value x_{12} between 0 and $\min\{n_{.2}, n_{1.} - x_{11}\}$. Once the values x_{11} and x_{12} are fixed, the values of all other cells are determined by the row and the column sums. Therefore, in the following we identify a contingency table x with the two values x_{11} and x_{12}.

2. How would you sample values for the cells which have a choice of value?
The easiest way of sampling these values is to use a uniform distribution. Thus for x_{11} we draw an integer from $I = \{0, \ldots, \min\{n_{.1}, n_{1.}\}\}$, where $\mathbb{P}(x_{11} = j) = 1/(\min\{n_{.1}, n_{1.}\} + 1)$ for each $j \in I$. Given x_{11} we then sample x_{12} uniformly from $\{0, \ldots, \min\{n_{.2}, n_{1.} - x_{11}\}\}$ and so $\mathbb{P}(x_{12} = k|x_{11}) = 1/(\min\{n_{.2}, n_{1.} - x_{11}\} + 1)$. Thus we sample the contingency table

j	k	$n_{1.} - j - k$	$n_{1.}$
$n_{.1} - j$	$n_{.2} - k$	$n_{.3} - (n_{1.} - j - k)$	$n_{2.}$
$n_{.1}$	$n_{.2}$	$n_{.3}$	N

with probability $\Big((\min\{n_{.1}, n_{1.}\} + 1)(\min\{n_{.2}, n_{1.} - x_{11}\} + 1)\Big)^{-1}$.

3. What are the proposal distributions?
Suppose the current state of the Metropolis-Hastings chain is the contingency table y. We make our proposals independent from the current state of the chain according to the procedure outlined above. Thus we propose a new contingency table x with probability

$$q(y, x) \quad = \quad \frac{1}{\min\{n_{.1}, n_{1.}\} + 1} \, \frac{1}{\min\{n_{.2}, n_{1.} - x_{11}\} + 1}$$

4. What is the Metropolis-Hastings acceptance probability?
We simply compute

$$\begin{aligned} \alpha(y, x) &= \min\left\{1, \frac{\pi(x)\ q(x, y)}{\pi(y)\ q(y, x)}\right\} \\ &= \min\left\{1, \frac{\prod_i \prod_j y_{ij}!}{\prod_i \prod_j x_{ij}!} \, \frac{\min\{n_{.2}, n_{1.} - x_{11}\} + 1}{\min\{n_{.2}, n_{1.} - y_{11}\} + 1}\right\} \end{aligned}$$

5. Describe the resulting Metropolis-Hastings sampler!
   Choose $X_0 = x_0 \in \mathcal{S}$. *Now let* $X_n = y \in \mathcal{S}$. *Then we choose* x_{11} *uniformly on* I *and* x_{12} *uniformly from* $I(x_{11})$. *All other entries of* x *are determined by* x_{11} *and* x_{12}. *Propose* x *as the new state and accept it with probability* $\alpha(y, x)$. *If* x *is accepted then set* $X_{n+1} = x$. *Otherwise set* $X_{n+1} = y$.

Example 4.8 (Witch hat distribution)
The witch hat distribution is a two-dimensional mixture distribution with two components: a bivariate normal distribution and the uniform distribution. A density plot of the witch hat distribution looks like a witch's hat with a square brim, see Figure 6 below:

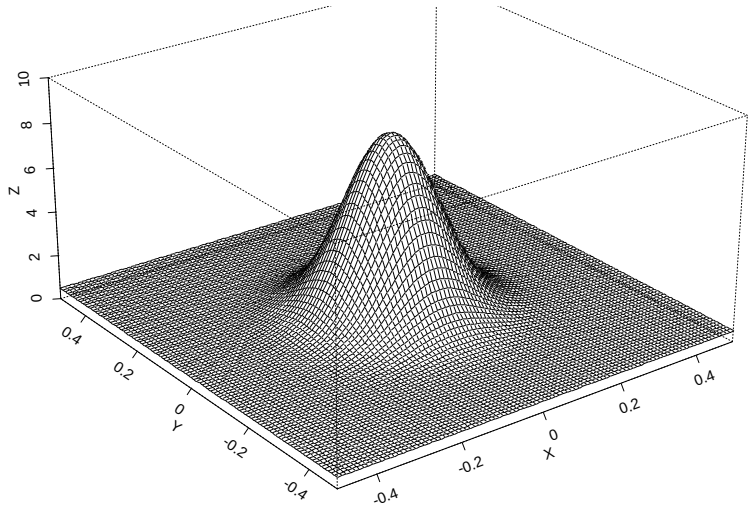


Figure 6: The witch hat density with $\sigma_1 = \sigma_2 = 0.01$.

- The target density is

$$g(x, y) \quad = \quad p\frac{1}{2\pi\sigma_1\sigma_2}\exp\Big(-\frac{1}{2}(\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2})\Big) \; + \; (1-p)\mathbf{1}_{[-\frac{1}{2}\le x\le\frac{1}{2}]}\mathbf{1}_{[-\frac{1}{2}\le y\le\frac{1}{2}]}.$$

- The proposal density:
  Suppose $X_n = (x_n, y_n)$. If n is even then we propose an update for the x-coordinate that is we propose $(\tilde{x}, y_n)$ as a new state, where $\tilde{x}$ is sampled from

$$g(x|y_n) \quad = \quad \frac{g(x, y_n)}{g(y_n)}.$$

  If n is odd we propose an update for the y-coordinate, that is we propose $(x_n, \tilde{y})$ as a new state where $\tilde{y}$ is sampled from

$$g(y|x_n) \quad = \quad \frac{g(x_n, y)}{g(x_n)}.$$

  Here $g(x)$ and $g(y)$ are marginal densities, that is

$$\begin{aligned} g(x) &= \int_{-\infty}^{\infty} g(x,y)dy \quad = \quad \frac{p}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) + (1-p)\mathbf{1}_{[-\frac{1}{2} \le x \le \frac{1}{2}]} \\ g(y) &= \int_{-\infty}^{\infty} g(x,y)dx \quad = \quad \frac{p}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2) + (1-p)\mathbf{1}_{[-\frac{1}{2} \le y \le \frac{1}{2}]} \end{aligned}$$

  Set

$$q_1 \; = \; \frac{p}{\sqrt{2\pi}} \frac{\exp(-\frac{1}{2}y_n^2)}{g(y_n)} q_2 \; = \; \frac{p}{\sqrt{2\pi}} \frac{\exp(-\frac{1}{2}x_n^2)}{g(x_n)}.$$

  Then

$$\begin{aligned} g(x|y_n) &= \frac{q_1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) + (1-q_1)\mathbf{1}_{[-\frac{1}{2} \le x \le \frac{1}{2}]} \\ g(y|x_n) &= \frac{q_2}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2) + (1-q_2)\mathbf{1}_{[-\frac{1}{2} \le y \le \frac{1}{2}]} \end{aligned}$$

  which are mixture densities which can be easily sampled as described above.

- The acceptance probability:
  At even n we have

$$\begin{aligned} \alpha\Big((x_n, y_n), (\tilde{x}, y_n)\Big) &= \min\Big\{1, \frac{g(\tilde{x}, y_n) g(x_n|y_n)}{g(x_n, y_n) g(\tilde{x}|y_n)}\Big\} \\ &= \min\Big\{1, \frac{g(\tilde{x}, y_n) g(x_n, y_n)/g(y_n)}{g(x_n, y_n) g(\tilde{x}, y_n)/g(y_n)}\Big\} \quad = \quad 1. \end{aligned}$$

Similarly at odd times n we have $\alpha\Big((x_n, y_n), (x_n, \tilde{y})\Big) = 1$. Thus a proposed state is always accepted.

- The Metropolis-Hastings sampler proceeds as follows:

  1. Choose $(x_0, y_0) \in \mathbb{R}^2$.
  2. Suppose $X_n = (x_n, y_n)$. If n is even then sample $\tilde{x}$ from $g(x|y_n)$ and set $X_{n+1} = (\tilde{x}, y_n)$. If n is odd then sample $\tilde{y}$ from $g(y|x_n)$ and set $X_{n+1} = (x_n, \tilde{y})$.

4.2 Proposal distributions

Metropolis-Hastings Samplers are often classified according to their proposal distributions.

1. **Independence Sampler**
   As the name suggests the independence sampler proposes states which are independent of the current state of the chain, that is $q(x, y) = f(y)$ for all $x \in \mathcal{S}$, where f is a pmf or density. In this case the acceptance probability reduces to

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

   Example 4.7 is an independence sampler.

2. **The Metropolis sampler**
   Metropolis et al. originally proposed to use symmetric proposal pmf's or densities, that is $q(x, y) = q(y, x)$. The acceptance probability then simplifies to

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

3. **Random walk Metropolis-Hastings sampler**
   Here we choose $q(x, y) = f(y - x)$ for some probability mass function or density f. If f is symmetric around zero then this is a Metropolis sampler. The random walk Metropolis-Hastings sampler derives its

name from the fact that the proposals are made according to a random walk, that is

$$y \quad = \quad x + z$$

where z is drawn from f. The acceptance probability for this proposal distribution is

$$\alpha(x, y) \quad = \quad \min\{1, \frac{\pi(y)f(x-y)}{\pi(x)f(y-x)}\}.$$

4. **Gibbs Sampler**
   The Gibbs Sampler is a popular choice that uses full conditional distributions as proposal distributions. Example 4.8 is a Gibbs Sampler.

4.3 The Gibbs Sampler

Suppose we want to sample a random vector $(Y^{(1)}, \ldots, Y^{(d)})$ with probability mass function or density π. As mentioned earlier, MCMC constructs an ergodic Markov chain whose equilibrium distribution is given by π. The *Gibbs Sampler* uses full conditional distribution to produce a d-dimensional Markov chain $\{(X_n^{(1)}, \ldots, X_n^{(d)}), n = 0, 1, \ldots\}$. Let us first define the full conditional distributions.

Definition 4.9 Full conditional distributions

(a) Let $\pi(y^{(1)}, \ldots, y^{(d)})$ be the p.m.f. of the random vector $Y = (Y^{(1)}, \ldots, Y^{(d)})$. Set

$$\pi\left(y^{(-j)}\right) = \sum_{z \in \mathcal{S}} \pi(y^{(1)}, \ldots, y^{(j-1)}, z, y^{(j+1)}, \ldots, y^{(d)})$$

then the jth full conditional p.m.f. is given by

$$\begin{aligned}
&\pi_j \quad \Big(x \mid y^{(i)}, i \neq j\Big) \\
&= \quad \mathbb{P}\Big(Y^{(j)} = x \;\Big|\; Y^{(i)} = y^{(i)}, i \neq j\Big) \\
&= \quad \frac{\mathbb{P}\Big(Y^{(j)} = x, Y^{(i)} = y^{(i)}, i \neq j\Big)}{\mathbb{P}\Big(Y^{(i)} = y^{(i)}, i \neq j\Big)} \\
&= \quad \frac{\pi(y^{(1)}, \ldots, y^{(j-1)}, x, y^{(j+1)}, \ldots, y^{(d)})}{\pi\Big(y^{(-j)}\Big)}
\end{aligned}$$

(b) continuous case:
Let $\pi(y^{(1)}, \ldots, y^{(d)})$ be the density of the random vector $Y = (Y^{(1)}, \ldots, Y^{(d)})$. Define

$$\pi\Big(y^{(-j)}\Big) = \int_{\mathcal{S}} \pi(y^{(1)}, \ldots, y^{(j-1)}, z, y^{(j+1)}, \ldots, y^{(d)}) dz$$

then the jth full conditional density is defined as

$$\begin{aligned}
&\pi_j \quad \Big(x \mid y^{(i)}, i \neq j\Big) \\
&= \quad \frac{\pi(y^{(1)}, \ldots, y^{(j-1)}, x, y^{(j+1)}, \ldots, y^{(d)})}{\pi\Big(y^{(-j)}\Big)}
\end{aligned}$$

We call the distribution defined by $\pi_j(x|y^{(i)}, i \neq j)$ the jth full conditional distribution of π. Note that we only need to determine the jth full conditional distribution up to a normalizing constant and so terms that do not depend on x can be ignored.

We can now define the Gibbs Sampler.

Definition 4.10 (Gibbs Sampler): Suppose we want to sample a random vector $(Y^{(1)}, \ldots, Y^{(d)})$ whose distribution is given by π. Assume $X_n =$

$\underline{x} = (x^{(1)}, \ldots, x^{(d)}) \in \mathcal{S}$. Choose $j \in \{1, \ldots, d\}$ at random (or sequentially). Sample x from $\pi_j(x|x^{(i)}, i \neq j)$ and set

$$X_{n+1} \quad = \quad \Big(x^{(1)}, \ldots, x^{(j-1)}, x, x^{(j+1)}, \ldots, x^{(d)}\Big).$$

The Gibbs sampler is a Metropolis-Hastings sampler that uses the full conditional distributions as proposal distributions. The acceptance probability for the Gibbs sampler is always accepted as the following computations show. Suppose that the vector $\underline{y}$ is such that $x^{(i)} = y^{(i)}$ for $i \neq j$ and $\mathcal{S}$ is discrete. Then the acceptance probability is

$$\begin{aligned}
\alpha(\underline{x}, \underline{y}) &= \min\left\{1, \frac{\pi(\underline{y})\pi_j(x^{(j)}|x^{(i)}, i \neq j)}{\pi(\underline{x})\pi_j(y^{(j)}|x^{(i)}, i \neq j)}\right\} \\
&= \min\left\{1, \frac{\pi(\underline{y})}{\pi(\underline{x})} \frac{\pi(\underline{x})/\pi(x^{(-j)})}{\pi(\underline{y})/\pi(y^{(-j)})}\right\} \\
&= \min\left\{1, \frac{\pi(y)}{\pi(x)} \frac{\pi(\underline{x})/\pi(x^{(-j)})}{\pi(\underline{y})/\pi(x^{(-j)})}\right\} \quad = \quad 1,
\end{aligned}$$

where the last line uses that $x^{(i)} = y^{(i)}$ for $i \neq j$. The proof is analogous for continuous $\mathcal{S}$.

Example 4.11 (Survival Times): Consider the second distribution in the list of distributions in section 2.1. Ross[7] develops a Gibbs Sampler for this target distribution.

- The target density:
  Unconditionally the random vector $(Y^{(1)}, \ldots, Y^{(m)})$ has density

  $$f(y^{(1)}, \ldots, y^{(m)}) \quad = \quad \prod_{i=1}^{m} \lambda_i \exp(-\lambda_i y^{(i)}) \mathbf{1}_{[y^{(i)}>0]}.$$

  Thus, conditionally on the event $E = \{\sum_{i=1}^{m} Y^{(i)} > a\}$ the vector $(Y^{(1)}, \ldots, Y^{(m)})$ has density

  $$\pi(y^{(1)}, \ldots, y^{(m)}) \quad = \quad \frac{1}{\mathbb{P}(E)} \prod_{i=1}^{m} \lambda_i \exp(-\lambda_i y^{(i)}) \mathbf{1}_{[y^{(i)}>0]} \mathbf{1}_{[\sum_{i=1}^{m} y^{(i)} > a]}$$

[7] S. Ross, *Introduction to Probability Models*, 6th edition, 1997, Academic Press, San Diego, p.216f

- The proposal density:
  Let $q_j = \sum_{i=1}^{j-1} y^{(i)} + \sum_{i=j+1}^{m} y^{(i)}$, then

$$\begin{aligned}\pi_j(x|y^{(i)}, i \neq j) &\propto \pi(y^{(1)}, \ldots, y^{(j-1)}, x, y^{(j+1)}, \ldots, y^{(m)}) \\ &\propto \frac{1}{\mathbb{P}(E)} \prod_{i \neq j} \lambda_i \exp(-\lambda_i y^{(i)}) \lambda_j \exp(-\lambda_j x) \mathbf{1}_{[x>0]} \mathbf{1}_{[q_j + x > a]} \\ &\propto \lambda_j \exp(-\lambda_j x) \mathbf{1}_{[x > (a-q_j)^+]},\end{aligned}$$

  where $(a-q_j)^+ = \max(0, a-q_j)$. Note that we have only determined the jth full conditional density up to a factor of proportionality. However, as we know that it is a density and so needs to integrate to one we can deduce that

$$\pi_j(x|y^{(i)}, i \neq j) \quad = \quad \exp(\lambda_j (a - q_j)^+) \; \lambda_j \exp(-\lambda_j x) \mathbf{1}_{[x > (a-q_j)^+]}$$

- The acceptance probability is one for the Gibbs Sampler.

- Let $Z = X + (a - q_j)^+$ where X is Exponentially distributed with mean $1/\lambda_j$. Due to the memoryless property of the Exponential distribution Z has density $\pi_j(x|y^{(i)}, i \neq j)$. We can now define the Gibbs Sampler to sample π. Let $X_0 = (x_0^{(1)}, \ldots, x_0^{(m)})$ be an initial state such that $\sum_{i=1}^{m} x_0^{(i)} > a$ and $x_0^{(i)} > 0$ for all $i \in \{1, \ldots, m\}$. Now suppose $X_n = (x^{(1)}, \ldots, x^{(m)})$. At random or sequentially we choose j from $\{1, \ldots, m\}$. We then sample x from an Exponential distribution of mean $1/\lambda_j$ and set $z = x + (a - \sum_{i \neq j} x^{(i)})^+$. Then

$$X_{n+1} \quad = \quad (x^{(1)}, \ldots, x^{(j-1)}, z, x^{(j+1)}, \ldots, x^{(m)}).$$

We can use detailed balance to show that the invariant distribution of the Gibbs Sampler is π.

Lemma 4.12: Each transition of the Gibbs Sampler satisfies detailed balance with respect to π.
Proof: Suppose $\mathcal{S}$ is continuous. Let $\underline{x} = (x^{(1)}, \ldots, x^{(d)})$ and the vector $\underline{y} = (x^{(1)}, \ldots, x^{(j-1)}, z, x^{(j+1)}, \ldots, x^{(d)})$. Now, let $p(\underline{x}, \underline{y})$ denote the transition densities for the Gibbs Sampler. Then we have

$$\begin{aligned}\pi(\underline{x}) p(\underline{x}, \underline{y}) &= \pi(\underline{x}) \frac{\pi(\underline{y})}{\pi(x^{(-j)})} \quad = \quad \pi(\underline{y}) \frac{\pi(\underline{x})}{\pi(x^{(-j)})} \\ &= \pi(\underline{y}) \pi_j(x^{(j)}|x^{(i)}, i \neq j) \quad = \quad \pi(\underline{y}) p(\underline{y}, \underline{x}).\end{aligned}$$

and so detailed balance holds. The proof for discrete $\mathcal{S}$ is analogous.

The above theorem shows that the Gibbs Sampler chain has the desired invariant distribution. Thus, if the chain is ergodic then its distribution converges towards π. Note however that we still have to ensure that the Markov chain produced by the Gibbs Sampler is ergodic. This will depend on the individual Markov chain concerned and cannot be shown in the same generality as detailed balance.

Example 4.13 (Bayesian mixture distribution): Consider again the mixture density from Example 2.3 and suppose that

$$p(\mu_1, \mu_2) \quad = \quad p(\mu_1)p(\mu_2) \quad = \quad \frac{1}{2\pi} \exp(-\frac{1}{2}(\mu_1 - \theta_1)^2) \exp(-\frac{1}{2}(\mu_2 - \theta_2)^2)$$

As mentioned earlier, the posterior distribution simplifies if we know I the sequence of coin tosses used to produce the sample $\underline{y}$. We now develop a Gibbs Sampler to sample from the joint posterior distribution

$$\pi(\mu_1, \mu_2, I|\underline{y}) \quad \propto \quad p(\mu_1)p(\mu_2) \exp\Big(-\frac{1}{2}(\sum_{i:I_i=1} (y^{(i)} - \mu_1)^2 + \sum_{i:I_1=0} (y^{(i)} - \mu_2)^2) \Big).$$

- The above density is our target density.
- We now derive the full conditional distributions. To ease notation, let H be the set of indices $i \in \{1, \ldots, k\}$ such that $I_i = 1$ and T the set of indices i such that $I_i = 0$. The first full conditional distribution is

given by

$$\begin{aligned}
\pi_1(\mu_1 \quad | \quad \mu_2, I, \underline{y}) &\propto p(\mu_1)p(\mu_2)\exp\Big(-\frac{1}{2}(\sum_{i\in H}(y^{(i)}-\mu_1)^2+\sum_{i\in T}(y^{(i)}-\mu_2)^2\Big)\\
&\propto p(\mu_1)\exp\Big(-\frac{1}{2}\sum_{i\in H}(y^{(i)}-\mu_1)^2\Big)\\
&\propto \exp\Big(-\frac{1}{2}((\mu_1-\theta_1)^2+\sum_{i\in H}(y^{(i)}-\mu_1)^2\Big)\\
&\propto \exp\Big(-\frac{1}{2}(\mu_1^2-2\mu_1\theta_1-\theta_1^2+\sum_{i\in H}\mu_1^2-2\mu_1 y^{(i)}+(y^{(i)})^2)\Big)\\
&\propto \exp\Big(-\frac{1}{2}(|H|+1)\Big[\mu_1^2-2\mu_1\frac{\theta_1+\sum_{i\in H}y^{(i)}}{|H|+1}+\frac{(\theta_1+\sum_{i\in H}y^{(i)})^2}{(|H|+1)^2}\Big]\Big)\\
&\propto \exp\Big(-\frac{1}{2}(|H|+1)\Big[\mu_1-\frac{\theta_1+\sum_{i\in H}y^{(i)}}{(|H|+1)}\Big]^2\Big)\\
&\sim \mathcal{N}\Big(\frac{\theta_1+\sum_{i\in H}y^{(i)}}{(|H|+1)},\frac{1}{(|H|+1)}\Big).
\end{aligned}$$

Analogously the second full conditional density $\pi_2(\mu_2|\mu_1, I, \underline{y})$ is the Normal density $\mathcal{N}\Big(\frac{\theta_2+\sum_{i\in T}y^{(i)}}{(|T|+1)},\frac{1}{(|T|+1)}\Big)$. Finally the third full conditional p.m.f. is given by

$$\begin{aligned}
\pi_3(I|\mu_1,\mu_2,\underline{y}) &\propto \exp\Big(-\frac{1}{2}(\sum_{i\in H}(y^{(i)}-\mu_1)^2+\sum_{i\in T}(y^{(i)}-\mu_2)^2)\Big)\\
&\propto \prod_{i\in H}\exp\Big(-\frac{1}{2}(y^{(i)}-\mu_1)^2\Big)\prod_{i\in T}\exp\Big(-\frac{1}{2}(y^{(i)}-\mu_2)^2\Big).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{P}(I_i=1|\mu_1,\mu_2,\underline{y}) &\propto \exp\Big(-\frac{1}{2}(y^{(i)}-\mu_1)^2\Big) \quad \text{and}\\
\mathbb{P}(I_i=0|\mu_1,\mu_2,\underline{y}) &\propto \exp\Big(-\frac{1}{2}(y^{(i)}-\mu_2)^2\Big) \quad \text{and so}\\
\mathbb{P}(I_i=1|\mu_1,\mu_2,\underline{y}) &= \frac{\exp(-\frac{1}{2}(y^{(i)}-\mu_1)^2)}{\exp(-\frac{1}{2}(y^{(i)}-\mu_1)^2)+\exp(-\frac{1}{2}(y^{(i)}-\mu_2)^2)}\\
&= 1-\mathbb{P}(I_i=0|\mu_1\mu_2,\underline{y}).
\end{aligned}$$

- The acceptance probability of the Gibbs Sampler is one.
- The Gibbs Sampler performs the following moves in cycles. It first updates μ_1 according to the Normal distribution $\mathcal{N}\left(\frac{\theta_1+\sum_{i\in H} y^{(i)}}{(|H|+1)}, \frac{1}{(|H|+1)}\right)$. It then updates μ_2 according to the Normal distribution $\mathcal{N}\left(\frac{\theta_2+\sum_{i\in T} y^{(i)}}{(|T|+1)}, \frac{1}{(|T|+1)}\right)$. Finally it updates the list of coin tosses I by setting $I_j = 1$ with probability $\mathbb{P}(I_j = 1|\mu_1, \mu_2, \underline{y})$ and $I_j = 0$ with probability $\mathbb{P}(I_j = 0|\mu_1, \mu_2, \underline{y})$.

5 Implementational Issues

5.1 Proposal distributions

We have seen that the Metropolis-Hastings sampler uses proposal distributions. Clearly, the properties of such a sampler will strongly depend on the choice of proposal distribution. But how do we choose a proposal distribution? There are several issues that we may consider.

1. The proposal distribution should be close to π. For example, if $q(x, y) = \pi(y)$, then not only would the acceptance probability be one, but also the chain would be in equilibrium immediately. Proposal distributions which take into account the equilibrium distribution are, for example, the full conditional distributions of π. Thus if we can compute the full conditional distributions and easily sample from them, then the Gibbs Sampler is a popular choice.

2. The proposal distribution should be easy to sample. This will improve the computational efficiency of the MCMC algorithm. For example to sample the contingency tables in Example 4.6 we used uniform distributions which are very easy to sample.

3. The proposal distribution should facilitate fast mixing. The term "mixing" refers to how well the Markov chain explores the state space. Generally, chains that mix fast converge fast to equilibrium. Consider a proposal distribution that proposes large moves. If the moves are too large, then we often propose a state x which lies in the tails of π and so has small probability under the equilibrium distribution. But then this move is likely to be rejected. On the other hand, if we propose very small moves, these moves are frequently accepted. However, with

very small moves it takes the chain a long time to move from one part of the state space to another. Usually, we perform some trial runs of the chain to decide how large the moves should be in order to ensure fast mixing.

Here are some example plots which show the path of a random walk Metropolis-Hastings chain whose equilibrium distribution is Normal with mean -3 and variance 1. The proposal density $f(\cdot)$ is Normal with mean 0 and variance σ^2. Thus given the chain is in x we propose to move to $x + z$ where z is a sample from a Normal distribution with mean 0 and variance σ^2.

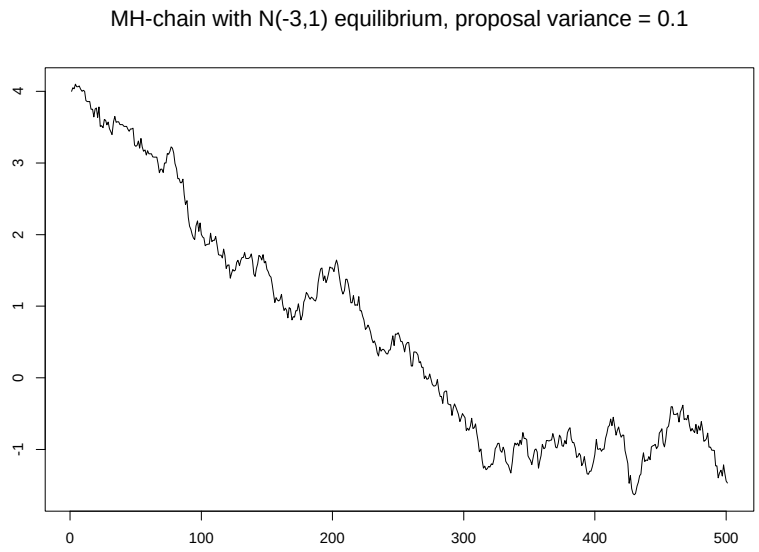


Figure 7: A path of the Metropolis-Hastings chain started in state 4 and run for 500 iterations. The proposals are normally distributed with variance 0.1.

In Figure 7 the proposal variance is only 0.1 and so the proposed moves are rather small. The moves are often accepted but the chain only moves a little at a time. Thus it takes the chain a long time to move from the starting value 4 to the range of values which are typical for

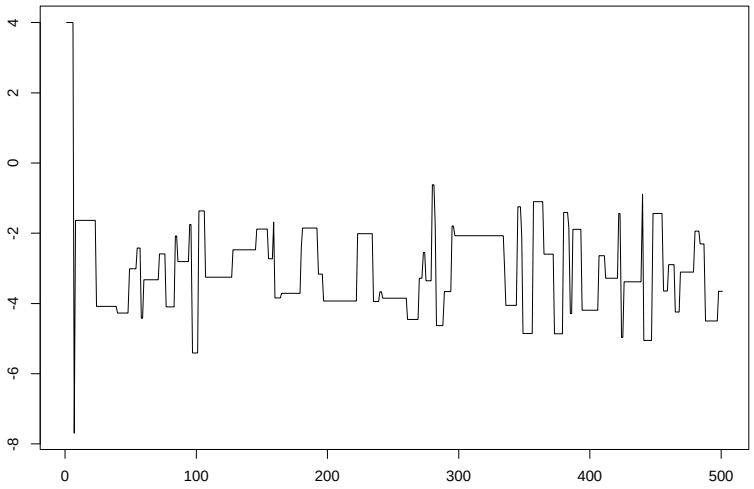


Figure 8: A path of the Metropolis-Hastings chain started in state 4 and run for 500 iterations. The proposals are normally distributed with variance 10.

the equilibrium distribution. The path indicates rather slow mixing. Figure 8 shows proposals which are normally distributed with variance 10. This means that rather large moves are proposed. However, these moves are often rejected which can be seen from the long plateaus in the path of the chain. As in the previous figure mixing is rather slow. Figure 9 uses a proposal distribution that has variance 1. This path seems to be mixing well.

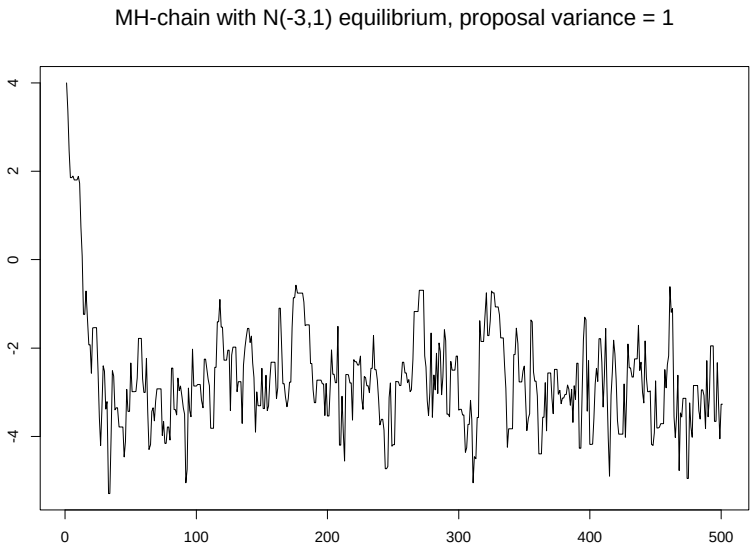


Figure 9: A path of the Metropolis-Hastings chain started in state 4 and run for 500 iterations. The proposals are normally distributed with variance 1.

5.2 Burn in

Suppose we have four balls divided into two urns. With probability 1/2 we pick a ball from the left urn and put it into the right urn. Alternatively, we take a ball from the right urn and put it into the left urn. If we find the chosen urn empty, then we do nothing. We can describe this process as a Markov chain as follows. Let X_n be the number of balls in the left urn at time n. Then this chain can take values 0,1,2,3 or 4. The transition probabilities of the chain are given in the state-flow diagram below:

In the long-run, what is the expected number A of balls in the left urn?

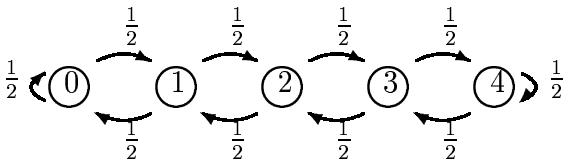


Figure 10: State-flow diagram for the Markov chain counting the number of balls in the left urn.

Markov chain Monte Carlo approaches this problem as follows. It simulates the chain up to some large time N and then takes the average of all the values that the chain has taken as its estimate. Thus our estimate would be $\hat{A} = \frac{1}{N}\sum_{n=1}^{N} X_n$. But how do we choose our initial configuration X_0? Suppose we start in $X_0 = 3$. Then the first few samples $X_1, X_2, X_3 \ldots$ will be slightly larger than expected in the long run and so the ergodic average $\hat{A}$ is likely to overestimate the expectation of the equilibrium distribution. If we choose $X_0 = 3$, then for the given sequence of coin tosses in Figure 11 we would get an ergodic average of 3.6 (not counting X_0.) This is quite a bit higher than the actual expected value of the equilibrium distribution which is 2. On the other hand if we start in $X_0 = 0$, then the first few samples $X_1, X_2, X_3 \ldots$ will be slightly smaller than expected in the long run. In Figure 11 the starting value $X_0 = 0$ leads to an ergodic average of 1.87, which underestimates the expected value of 2.

This phenomenon is often called the *initialisation bias* and is due to the fact that we do not choose the initial state according to the equilibrium distribution. Nevertheless, we know that because X is an ergodic Markov chain, its distribution will converge to the equilibrium distribution for any initial state. Moreover, due to the ergodic theorem, we know that $\hat{A}$ will converge to A with probability one as N tends to infinity. However, to

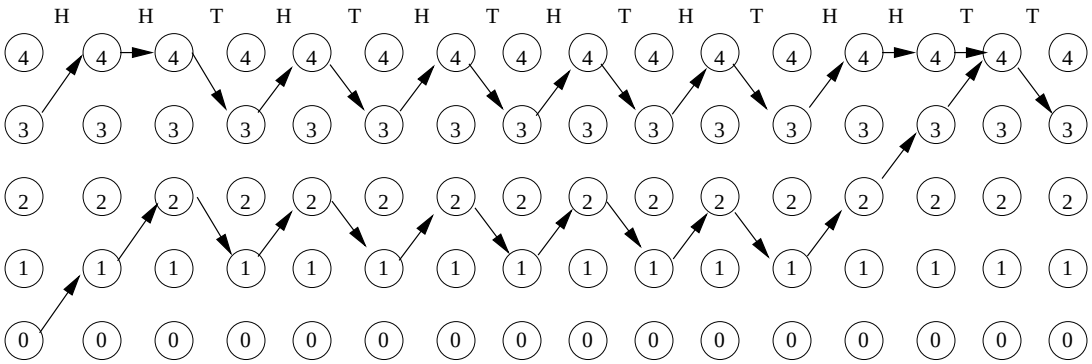


Figure 11: Some sample paths of the urn model Markov chain

reduce the initialisation bias it is common to wait until the chain is closer to equilibrium and to discard the initial samples of the chain. In other words, we choose a time M (usually much smaller than N) and estimate A by $\frac{1}{N-M}\sum_{n=M+1}^{N} X_n$. The time M is often called the *burn-in* time according to a terminology which is borrowed from statistical physics.

This is easy enough, but what is an appropriate value for M? Ideally we would like to wait until the chain is in equilibrium or at least close to it. However, this is not at all easy to determine. Essentially we have two approaches.

1. Convergence rate computations:
   Convergence rate computations compare the distribution $q^{(n)}$ of the chain at time n with the equilibrium distribution π. Then we can compute a value M such that $q^{(M)}$ and π are sufficiently close. The advantage of convergence rate computation is that it is an exact method. However, because it takes into account worst case scenarios which rarely occur, it often chooses a value for M which is pessimistically large. Moreover, convergence rate computations are often very difficult.

2. Convergence diagnostics:
   Convergence diagnostics examine the output of the Markov chain to determine whether it is close to equilibrium. It is the most common method to determine the burn-in. Its advantage is that it is a simple method. However, it has a decisive disadvantage: it may not detect if the chain has not converged yet. We will discuss convergence diagnos-

tics in more detail in the next section.

5.3 Convergence diagnostics

Convergence diagnostics are a heuristic approach in which we examine the output of the chain to determine when it has reached equilibrium. However, these methods usually do not guarantee that the chain is close to equilibrium, but they might warn if equilibrium has not yet been reached. In practice, convergence diagnostics are the most common tool used to determine the burn-in period. Often we use several convergence diagnostics at the same time, hoping that if the chain is far from equilibrium one of the diagnostics will indicate this. We will look at a small selection of convergence diagnostics, but there are many more than is possible to discuss here.

A common diagnostic is to plot the path of the chain. If the Markov chain lives on a high-dimensional space, we may plot some summary statistic instead. For example, for the Ising we may plot the number of sites with an upward spin. It is also common to plot ergodic averages in time. Usually we observe that in the beginning the path of the chain or the value of the summary statistics show an atypical behaviour, for example they may fluctuate a lot. This is often due to the initialisation bias. Then, after some time, the plot becomes more stable, with less extreme fluctuations. It is commonly assumed that the more stable plot indicates that the chain is close to or in equilibrium, and so we now may start sampling.

Let us examine this argument more closely. Once the chain has reached equilibrium, its distribution becomes constant over time. Thus, we would expect the plot to look more stable. However, the reverse is not necessarily true. A stable plot does not necessarily imply that the chain is in equilibrium. For example, consider the ferro-magnetic Ising model with large parameter J. Suppose we update the spin of a site and most of the neighbours of this site have an upward spin. Then the probability of assigning an upward spin to this site is very high. Thus if we start in a configuration with mostly upward spins we will remain in configurations with mostly upward spins for a long time. Hence, if we look at a path of the chain it looks very stable and so we may think that the chain has converged. However, because the distribution of the Ising model is symmetric it gives equally high probability to configurations that have mostly upward spins as it does to those which have mostly downward spins. As we have not yet observed any configurations with mostly downward spins, the chain has probably not yet converged.

If the path of a chain appears to be stable but the chain has not converged yet, then we speak of *"meta-stability"*. This is often caused by multi-modality of the stationary distribution. We may remedy the problem as follows. We sample a range of starting values which are over-dispersed compared to the equilibrium distribution. Then, we simulate a path of the Markov chain from each of these starting values. The idea is that if meta-stability occurs, we can detect it because the different starting values ensure that the corresponding paths will get stuck in different modes of the equilibrium distribution. For example, in the Ising model case, we may start one path in a state with mostly upward spins and one in a state with mostly downward spins. Each path may stabilize, but commonly the path started with mostly upward spins will mainly visit states with mostly upward spins. Similarly, the path started in a state with mostly downward spins will mainly visit states with mostly downward spins. The plots now indicate that the chains are in meta-stability and thus that we should not conclude that convergence has occurred.

1. Plots of the path of the chain or of some suitable summary statistic are often used as convergence diagnostics. What is the reasoning for the use of these methods?
   Once the chain has reached equilibrium its distribution is constant over time. Thus we expect the plot of the chain or some summary statistic to show random fluctuations that follow a more stable pattern.

2. What is meta-stability? What may cause meta-stability? How can we diagnose meta-stability?
   Meta-stability causes plots of the chain or of some summary statistic to look stable although the chain has not converged yet. This is often caused by multi-modality of the stationary distribution. The chain remains close to one mode and does not visit the regions of other modes. To diagnose meta-stability we start several paths of the chain in starting values that are overdispersed compared to the equilibrium distribution. Meta-stability is then indicated by paths which yield estimation results that differ greatly from each other.

3. Now consider Figure 9. Would you assume the chain has converged? What burn-in would you choose?
   The chain looks like it may have converged after 400 iterations, so $M = 400$ *seems a reasonable burn-in time.*

4. Figure 12 shows the evolution of the ergodic average which estimates the mean of the equilibrium distribution without a burn-in. Do you still think your choice of burn-in period is appropriate?
   The ergodic average is still decreasing which may indicate that the chain has not yet converged yet. We are in the fortunate position of actually knowing the mean of the equilibrium distribution which is -3, so we know that the ergodic average is over-estimating. This figure shows that figure 9 may have been deceptive in indicating convergence. Thus it illustrates the importance of using several different convergence diagnostics.

5. Figure 13 shows the path of the same chain up to time 2000. Figure 14 shows the corresponding ergodic average. How would you evaluate these figures?
   Both figures seem to indicate convergence. An appropriate value for the burn in time seems to be $M = 1000$.

6. Consider again the witchhat distribution from example 4.6. Figure 15 is a plot of its density where the mean of the normal component is chosen to be at $(0.99, 0.99)$ and $p = 0.75$. The uniform component is defined on the unit square. Figure 16 is a plot of a path of the Gibbs Sampler for this distribution. What can you say about convergence of the chain?
   At the beginning the chain essentially only samples from the uniform component. Only at the very end is the Normal component sampled. Once the chain gets close to the mean of the Normal component it gets stuck at this mode. This chain is not mixing well. It would have been easy to misdiagnose convergence if one had only observed the first 3000 iterations! This is an example of meta-stability!

7. Finally an example of the bimodal Normal mixture discussed in example 4.5. The density plot in Figure 5 shows that the means of the mixture components are 4 and -4 respectively. The Normal component with mean 4 has a probability of $p = 0.8$. Figure 17 shows two paths of the random walk Metropolis-Hastings sampler for the bimodal Normal mixture distribution. The first path is started at 4 and the second is started at -4. Proposals are Normally distributed with variance 1. What can you say about the paths of the chain?
   Each path seems to be in equilibrium, but one path stays close to the

mode 4 and the other stays close to the mode -4. The two paths would produce very different results if used in an estimation scheme. This again is an example of meta-stability but it was detected by using several paths started in different starting states.

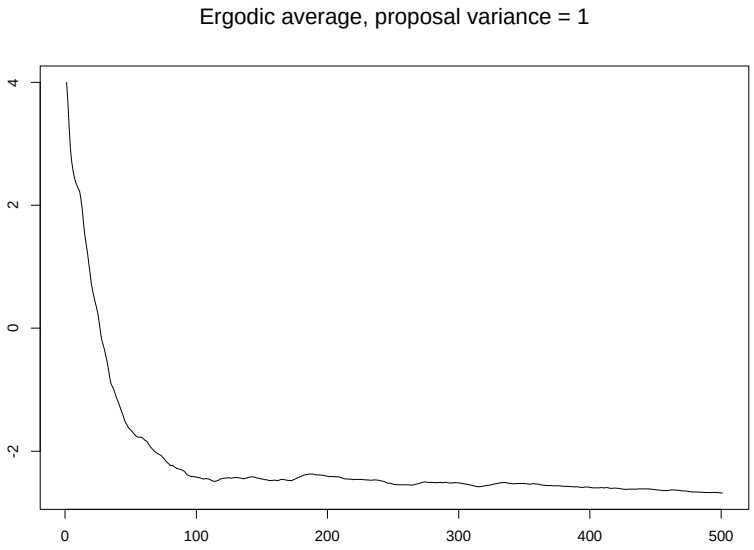


Figure 12: The ergodic average for the path in Figure 9.

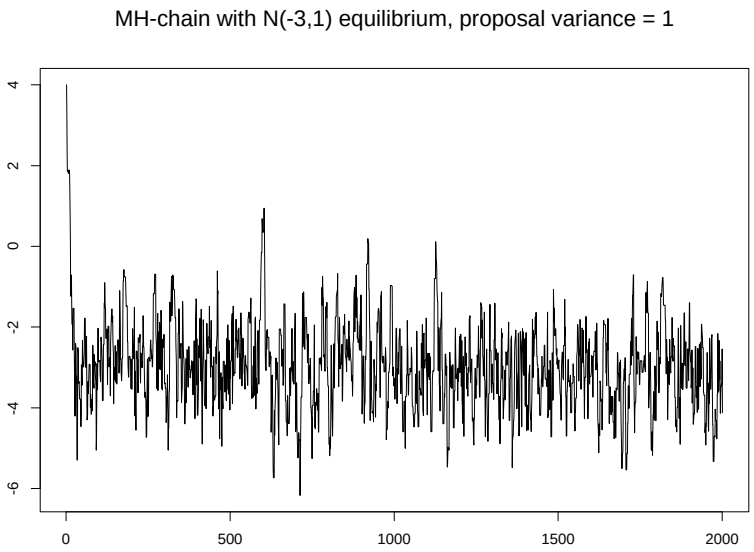


Figure 13: A path of the Metropolis-Hastings chain started in state 4 and run for 2000 iterations. The proposals are normally distributed with variance 1.

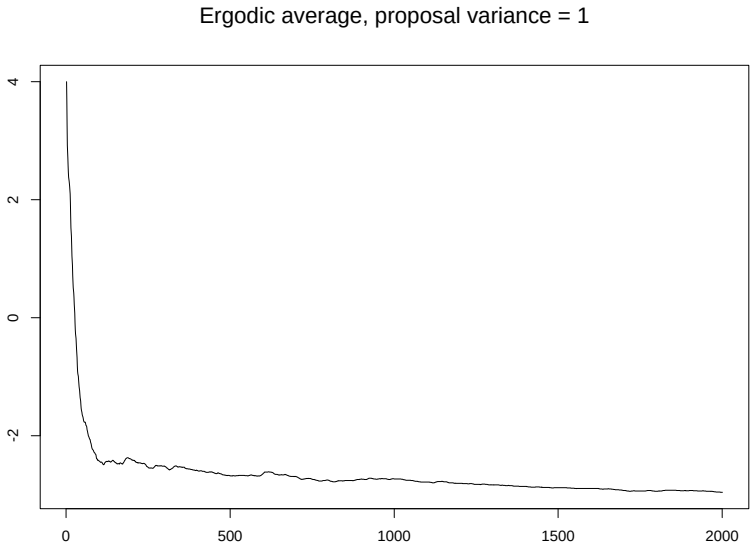


Figure 14: The ergodic average for the path in Figure 13.

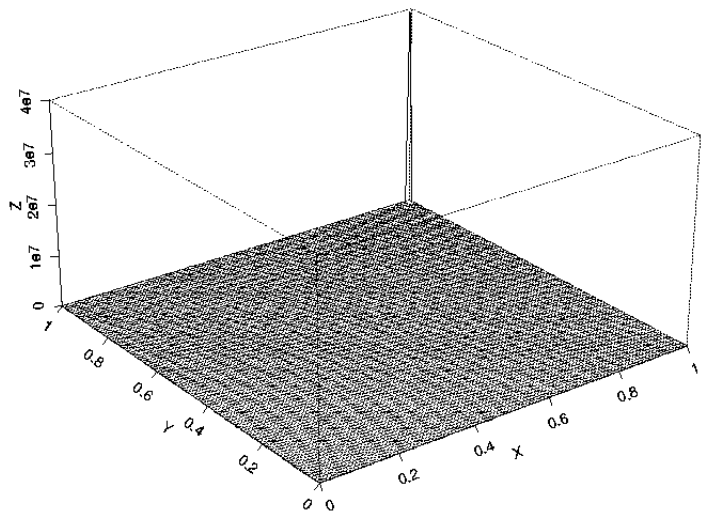


Figure 15: The witchhat distribution. The normal component has a probability of 0.75 and its mean is at (0.99, 0.99)

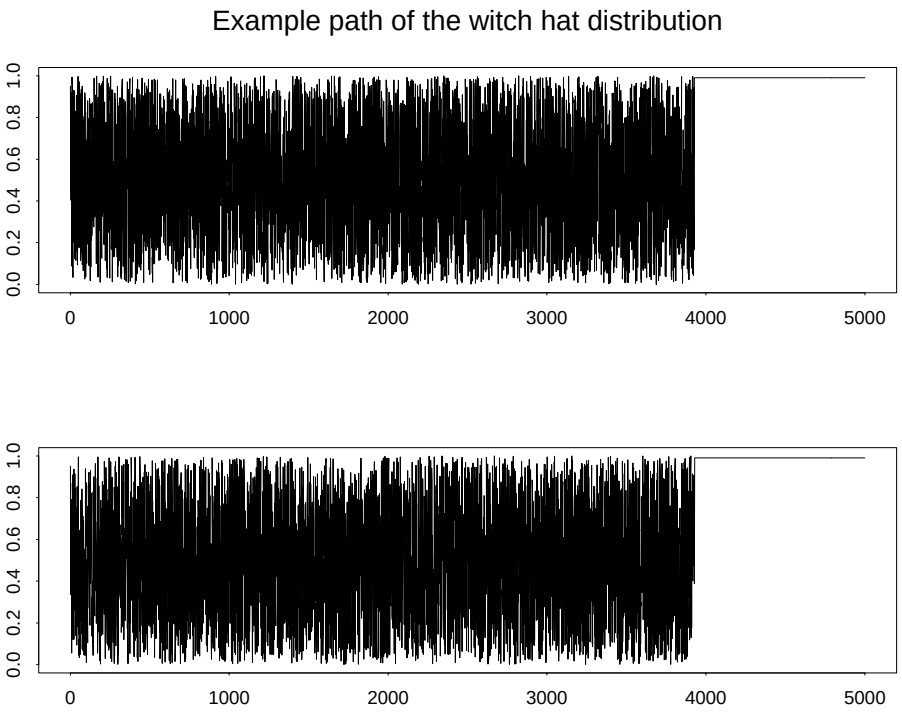


Figure 16: Gibbs Sampling for the witch hat distribution.

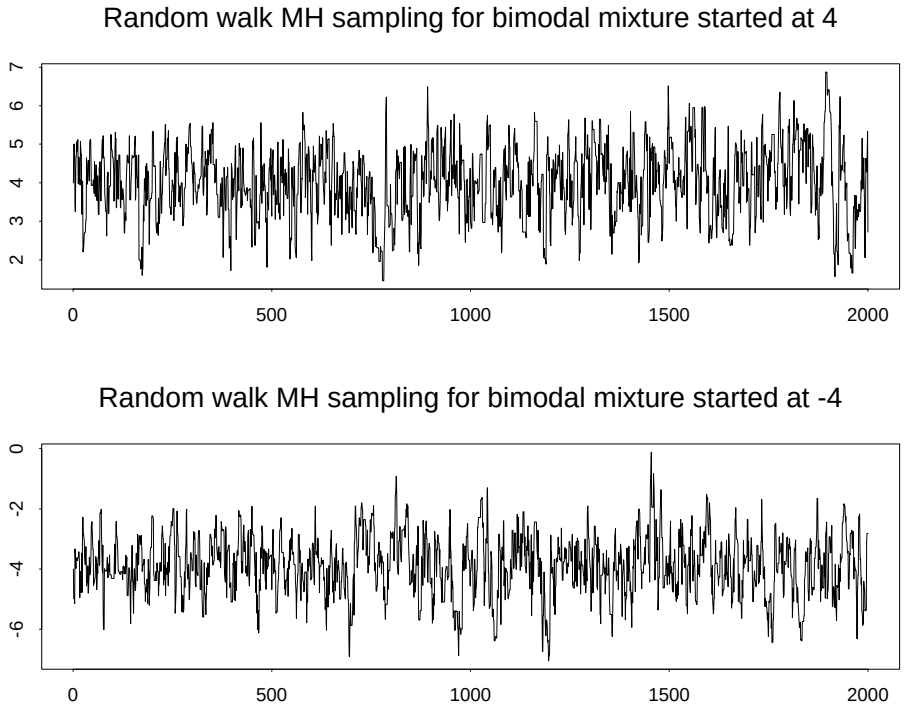


Figure 17: Paths of a random walk Metropolis-Hastings sampler for the bimodal Normal mixture distribution.

5.4 Monte Carlo error

Once we have collected samples from our run of the Markov chain, we use ergodic averages to estimate the quantities of interest. But how accurate are these estimates? How large should I choose my sample size, such that the estimate and the true value differ by less than a small quantity ϵ?

Definition 5.3.1 (Autocovariance/Autocorrelation): The autocovariance of lag k is defined as

$$\begin{aligned} \gamma_k &= \mathrm{Cov}\Big(X_n, X_{n+k}\Big) \qquad n, k \geq 0, \\ \gamma_0 &= \mathrm{Cov}\Big(X_n, X_n\Big) \quad = \quad \mathrm{Var}\Big(X_n\Big) \end{aligned}$$

The autocorrelation of lag k is defined as

$$\rho_k \quad = \quad \frac{\gamma_k}{\gamma_0}, \qquad k \geq 0.$$

Note that if the chain X is in stationarity then $\gamma_0 = \sigma^2$ where σ^2 is the variance of the stationary distribution π. But what we are really interested in is the variance of ergodic averages. Define

$$\frac{\tau_n^2}{n} \quad = \quad \mathrm{Var}\Big(\frac{1}{n}\sum_{k=1}^{n} X_k\Big).$$

Lemma 5.3.2: In stationarity

$$\tau_n^2 = \sigma^2\Big[1 + 2\sum_{k=1}^{n-1}\frac{n-k}{n}\rho_k\Big].$$

Proof:

$$\begin{aligned}
\frac{\tau_n^2}{n} &= \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big) \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n}\mathrm{Var}\Big(X_i\Big) + 2\sum_{i=1}^{n-1}\sum_{j>i}\mathrm{Cov}\Big(X_i, X_j\Big)\Big] \\
&= \frac{1}{n^2}\Big[n\sigma^2 + 2\sum_{i=1}^{n-1}\sum_{k=1}^{n-i}\mathrm{Cov}\Big(X_i, X_{i+k}\Big)\Big] \\
&= \frac{\sigma^2}{n}\Big[1 + \frac{2}{n\sigma^2}\sum_{i=1}^{n-1}\sum_{k=1}^{n-i}\gamma_k\Big] \\
&= \frac{\sigma^2}{n}\Big[1 + \frac{2}{n}\sum_{i=1}^{n-1}\sum_{k=1}^{n-i}\rho_k\Big] \\
&= \frac{\sigma^2}{n}\Big[1 + \frac{2}{n}\sum_{k=1}^{n-1}\sum_{i=1}^{n-k}\rho_k\Big] \\
&= \frac{\sigma^2}{n}\Big[1 + \frac{2}{n}\sum_{k=1}^{n-1}(n-k)\rho_k\Big] \\
&= \frac{\sigma^2}{n}\Big[1 + 2\sum_{k=1}^{n-1}\frac{(n-k)}{n}\rho_k\Big]
\end{aligned}$$

One can show that as $n \to \infty$

$$\tau_n^2 \quad \longrightarrow \quad \tau^2 = \sigma^2\Big[1 + 2\sum_{k=1}^{\infty}\rho_k\Big]$$

The accuracy of MCMC estimates thus depends on the autocorrelation of the Markov chain. Suppose $Y_1, \ldots, Y_N$ are independent and identically distributed according to π. Then the mean of the sample $\frac{1}{N}\sum_{i=1}^{N} Y_i$ has variance

$$\mathrm{Var}\Big(\frac{1}{N}\sum_{i=1}^{N} Y_i\Big) \quad = \quad \frac{\sigma^2}{N}.$$

Now compare this with the variance of the ergodic average $\frac{1}{N}\sum_{i=1}^{N} X_i$ which is $\frac{\tau_N^2}{N}$. From the above lemma it follows that if the autocorrelations of the

chain X are positive then the ergodic average will be less accurate than an estimate based on an independent sample. Informally, this may be explained by the fact that positively correlated variables carry redundant information and so are less informative than independent variables.

On the other hand, if we can introduce negative autocorrelations into our Markov chain, then this will make our estimation procedure more accurate. The use of negatively correlated variables in estimation problems is sometimes referred to as "antithetic variables".