An Introduction to Markov Chain Monte Carlo

Supervised Reading at the University of Toronto

Fall 2005

Supervisor: Professor Jeffrey S. Rosenthal[†]

Author: Johannes M. Hohendorff[‡]

[†] Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Email: jeff@math.toronto.edu. Web: http://probability.ca/jeff/

[‡] Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Email: johannes.hohendorff@utoronto.ca

Preface

The following text is an overview and summary of the supervised reading course I took with Professor Rosenthal at the University of Toronto in fall 2005. The first part should be considered as an introduction to MCMC on finite state spaces since I hadn't worked on MCMC before. My studies on this part were largely based on a book by Häggström [3] and lecture notes from Schmidt [7].

The second part summarizes my work on more advanced topic in MCMC on general state spaces. I focused on papers by Rosenthal [4],[6] and Tierney [8]. Finally, the reader will find a couple of simulation algorithms I implemented using the free statistical software R.

Outlines and references for important proofs or proofs using techniques that are worthwhile to be studied are included. Nevertheless, many propositions are stated without proof since I already went through them with my supervisor Professor Rosenthal and anything else would - more or less directly - be reusing material from the sources I cited above.

Throughout the whole term I had regular meetings with Professor Rosenthal. He encouraged me to study the theoretical background necessary for MCMC as well as to implement several examples. My first exposure to probability theory and statistics were courses with Professor Schmidt at the University of Ulm who also encouraged me to focus on MCMC.

Johannes M. Hohendorff
Toronto, December 2005

Contents

Part I

MCMC on Finite State Spaces

1 Introduction

Markov chains are a general class of stochastic models. In combination with computer simulation methods they are widely used in various scientific areas such as finance and insurance or even in physics, chemistry or biology where one might wouldn't expect it at the first place. Since the resulting models are often too difficult to be analyzed analytically, computers are used for inference. In this first part we will introduce the concept of Markov chains and common algorithms for their simulation. Since finite models are easier to follow, we will limit the discussion to this case and postpone the introduction of more general concepts to the second part.

1.1 Definition and Basic Properties

Definition 1 *Let $\{X_0, X_1, X_2, \dots\} : \Omega \to S$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and mapping to a finite state space $S = \{s_1, \dots, s_k\}$. The sequence is said to be a (homogeneous)* **Markov chain** *with initial distribution $\mu = (\mu_1, \dots, \mu_k)^T$ and transition matrix $P = (p_{i,j})$ if*

$$P(X_0 = s_0, X_1 = s_1, \dots, X_n = s_n) = \mu_{s_0} p_{s_0,s_1} \dots p_{s_{n-1},s_n}$$

holds for all $n = 0, 1, \dots$ and all $s_0, s_1, \dots, s_n \in S$.

Proposition 1 *The above sequence of random variables $\{X_0, X_1, X_2, \dots\} : \Omega \to$*

S is a Markov chain if and only if there exists a stochastic matrix[1] $P = (p_{i,j})$ *such that*

$$P(X_n = s_n | X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) = p_{s_{n-1},s_n}$$

for all $n = 0, 1, \ldots$ and all $s_0, s_1, \ldots, s_n \in S$ with $P(X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) > 0$.

Corollary 1 *Let $\{X_0, X_1, X_2, \ldots\}$ be a Markov chain. Then it holds:*

$$P(X_n = s_n | X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) = P(X_n = s_n | X_{n-1} = s_{n-1})$$

where $P(X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) > 0$. This is frequently referred to as the **Markov property***.*

1.2 A First Example

To demonstrate the concept of Markov chains we consider the classical example of random walks. The setup is as follows:

- Let $Z, Z_1, Z_2, \cdots : \Omega \to \mathbb{Z}$ be a sequence of *iid* random variables with values in the set of all integers.

- Assume that $X_0 : \Omega \to \mathbb{Z}$ is a random variable that is independent from $Z_1, Z_2, \ldots$ and set

$$X_n := X_{n-1} + Z_n \quad \forall \quad n \geq 1.$$

Then the random variables $\{X_0, X_1, X_2, \ldots\}$ form a Markov chain with state space S=$\mathbb{Z}$, initial distribution $\mu = (\mu_1, \ldots, \mu_k)^T$ (where $\mu_i = P(X_0 = i)$) and transition probability $p_{i,j} = P(Z = j - i)$. Such a model could for example

[1] i.e. $p_{i,j} \geq 0 \quad \forall \quad i, j \in \{1, \ldots, k\}$ and the sum of each of the matrix's row equals 1

describe the risk reserves of an insurance firm where X_0 is the initial reserve and Z is the difference between the insurance premium and damage payments (this is why I used the set of all integers as state space).

2 Important Properties of Markov Chains

In the following section we will summarize some of the most common properties of Markov chains that are used in the context of MCMC. We always refer to a Markov chain $\{X_0, X_1, X_2, \dots\}$ with transition matrix P on a finite state space $S = \{s_1, \dots, s_k\}$.

Assume we are interested in the distribution of the Markov chain after n steps. The following proposition tells us that we can receive this information by simple matrix multiplication.

Proposition 2 *Consider a Markov chain with transition matrix P, initial distribution μ and denote the chain's distribution after the n^{th} transition with $\mu^{(n)}$. Then it holds:*

$$\mu^{(n),T} = \mu^T P^n.$$

Note: We will also consider the transition matrix $P^{(n)}$ for n transitions. In analogy to the above proposition it holds: $P^{(n)} = P^n$ which is also referred to as the *Chapman-Kolmogorov-Equation.*

Some further definitions have to follow:

Definition 2

- *A matrix A is called* **non-negative** *if all its elements $a_{i,j}$ are non-negative.*
- *A non-negative matrix A is called* **quasi-positive** *if there exists an $n_0 \geq 1$ such that all elements of A^{n_0} are positive.*

Definition 3 *A Markov chain $\{X_0, X_1, X_2, \dots\}$ is called* **ergodic** *if the limit*

$$\pi_j = \lim_{n\to\infty} p_{i,j}^{(n)}$$

1. *exists for all $j \in \{1, \dots, k\}$*

2. *is positive and does not depend on i*

3. *$\pi = (\pi_1, \dots, \pi_k)^T$ is a probability distribution on S.*

Proposition 3 *A Markov chain with transition matrix P is ergodic if and only if P is quasi-positive.*

Theorem 1 *Consider an ergodic Markov chain. Then the vector $\pi = (\pi_1, \dots, \pi_k)^T$ where $\pi_j = \lim_{n\to\infty} p_{i,j}^{(n)}$ it the unique solution of*

$$\pi^T = \pi^T P$$

and π is a probability distribution on S.

At this point, we just state the theorem but will return to this important property of the distribution π later on this chapter because it is the foundation for MCMC algorithms. But before we can characterize π in a different way, we need to introduce some more concepts.

Definition 4 *A Markov chain is said to be* **irreducible** *if all states $s_i, s_j \in S$* communicate*, that is, there exists an n such that*

$$P(X_{m+n} = s_j | X_m = s_i) > 0$$

(due to the homogeneity independent of m).

Definition 5 *The* **period** *of a state $s_i \in S$ is defined as*

$$d(s_i) := \gcd\{n \geq 1 : (P^n)_{i,i} > 0\}.$$

A Markov chain is **aperiodic** *if all its states have period 1.*

Theorem 2 *A transition matrix P is irreducible and aperiodic if and only if P is quasi-positive.*

Note: On general state spaces, a irreducible and aperiodic Markov chain is not necessarily ergodic.

Since it is used in proofs, we note the following property:

Proposition 4 *Suppose we have an aperiodic Markov chain. Then there exists an $N < \infty$ such that*

$$(P^n)_{i,i} > 0$$

for all $i \in \{1, \dots, k\}$ and all $n \geq N$.

Proposition 5 *Now suppose our Markov chain is aperiodic and irreducible. Then there exists an $M < \infty$ such that*

$$(P^n)_{i,j} > 0$$

for all $i, j \in \{1, \dots, k\}$ and all $n \geq M$.

In the context of MCMC a question of particular interest is the question of the long-term behavior of a Markov chain. Given certain conditions, can we hope that the distribution of the chain converges to a well defined and unique limit? The concept of irreducibility and aperiodicity will provide an answer.

Definition 6 *A vector $\pi = (\pi_1, \dots, \pi_k)^T$ is said to be a* **stationary distribution** *for the Markov chain $\{X_0, X_1, X_2, \dots\}$ if:*

1. $\pi \geq 0 \quad \forall \quad i \in \{1, \dots, k\}$ *and* $\sum_{i=1}^{k} \pi_i = 1$

2. $\pi^T P = \pi^T$

So if we use π as initial distribution, the distribution of X_1 will also be π (and of course of any $X_n, n \geq 1$).

3 Convergence Theorems

Using the above concepts, we can formulate important convergence theorems. We will combine this with expressing the result of the first theorem in a different way. This helps to understand the main concepts.

3.1 A Markov Chain Convergence Theorem

Theorem 3 *For any irreducible and aperiodic Markov chain, there exists at least one stationary distribution.*

Proof: We will only give the main ideas of the proof here. A complete proof can be found in [3], pp. 29-33. The first step in proofing this theorem is to show that for any irreducible and aperiodic Markov chain we have

$$P(T_{i,j} < \infty) = 1$$

and

$$E[T_{i,j}] < \infty$$

where T is a hitting time for a Markov chain $\{X_0, X_1, X_2, \dots\}$ starting in $s_i \in S$ and

$$T_{i,j} := \min\{n \geq 1 : X_n = s_j\}.$$

We then propose a candidate for the stationary distribution. For a Markov chain starting in s_1 and for $i = 1, \dots, k$ set

$$\rho_i := \sum_{n=0}^{\infty} P(X_n = s_i, T_{1,1} > n).$$

With $\tau_{1,1} := E[T_{1,1}]$ our candidate is:

$$\pi = (\pi_1, \dots, \pi_k) = \left(\frac{\rho_1}{\tau_{1,1}}, \dots, \frac{\rho_k}{\tau_{1,1}} \right).$$

We would then simply need to check that π is a probability distribution on $S = \{s_1, \dots, s_k\}$ with the desired property. □

To achieve our final goal - a Markov chain convergence theorem - we need to introduce a metric on probability distributions.

Definition 7 *Let $\mu = (\mu_1, \dots, \mu_k)^T$ and $\nu = (\nu_1, \dots, \nu_k)^T$ be two probability distributions on the state space $S = \{s_1, \dots, s_k\}$. The* **total variation distance** *between μ and ν is defined as*[2]

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{i=1}^{k} |\mu_i - \nu_i|.$$

A sequence of probability distributions $\pi^{(i)}$ **converges in total variation** *to a distribution π if*

$$\lim_{i \to \infty} d_{TV}(\pi^{(i)}, \pi) = 0.$$

Shorthand we write $\pi^{(i)} \xrightarrow{TV} \pi$.

Theorem 4 (Markov chain convergence theorem) *Consider an irreducible and aperiodic Markov chain $\{X_0, X_1, X_2, \dots\}$. If we denote the chains distribution after the n^{th} transition by $\mu^{(n)}$ we have for any initial distribution $\mu^{(0)}$ and*

[2] This is equivalent to $\max |\mu(A) - \nu(A)|$ where $A \subset S$.

a stationary distribution π:

$$\mu^{(n)} \xrightarrow{TV} \pi.$$

In words: If we run the Markov chain for a long time, its distribution will be very close to the stationary distribution π.

Proof: A complete proof of the theorem is rather long and wouldn't fit in a summary like this. A proof in full length can be found in [3] pp. 34-37. Nevertheless, it is very important to outline the main idea that could be applied to achieve an elegant proof. The following *coupling* technique is widely used. We compare the sequence of unknown distributions $\mu^{(n)}$ to the distribution of interest which is of course π. To do so, we consider a Markov chain $\{X_0, X_1, X_2, \dots\}$ who's initial distribution already is the stationary distribution π and an arbitrary other Markov chain $\{X'_0, X'_1, X'_2, \dots\}$. It would then be possible to show that the hitting time

$$T := \min_{n \geq 1} \{X_n = X'_n\}$$

is finite with probability 1. Since we started the first chain according to π, we can conclude that $\mu^{(n)}$ is the same distribution after the chains met. □

Theorem 5 *Any irreducible and aperiodic Markov chain has exactly one stationary distribution.*

Proof: We just learned that an irreducible and aperiodic Markov chain has at least one stationary distribution. So assume the chain $\{X_0, X_1, X_2, \dots\}$ has more than one stationary distribution, say π and π'. Then the chains distribution after n transitions is $\mu^{(n)} = \pi'$. On the other hand we have got

$$\mu^{(n)} \xrightarrow{TV} \pi.$$

Since $\mu^{(n)} = \pi'$

$$\lim_{n\to\infty} d_{TV}(\pi', \pi) = 0.$$

As this does not depend on n at all we conclude

$$\pi' = \pi.$$

□

Definition 8 *A probability distribution π on the state space $S = \{s_1, \ldots, s_k\}$ is* **reversible** *for the Markov chain $\{X_0, X_1, X_2, \ldots\}$ with transition matrix P if for all $i, j \in 1, \ldots, k$ we have*

$$\pi_i P_{i,j} = \pi_j P_{j,i}.$$

The next simple property will help us with constructing MCMC algorithms that (approximately) sample from a given distribution π.

Proposition 6 *If the probability distribution π is reversible for a Markov chain, then it is also a stationary distribution for the chain.*

Proof: The proof is very short and a nice illustration of the properties we introduced on the last few pages and though it is worthwhile to have quick look at it. We just have to proof the second property of 6 since the first one is really obvious. We compute:

$$\pi_j = \pi_j \sum_{i=1}^{k} P_{j,i} = \sum_{i=1}^{k} \pi_j P_{j,i} = \sum_{i=1}^{k} \pi_j P_{i,j} \quad \forall \quad j \in \{1, \ldots, k\}.$$

□

We finish this section with some notes about Markov chains and eigenvalues. This is important in the context of the convergence speed of a Markov chain to a limiting distribution (if it exists, of course).

3.2 Markov Chains and Eigenvalues

Consider the following setup:

- A quasi-positive transition matrix P with k different eigenvalues
- A stationary distribution $\pi = (\pi_1, \ldots, \pi_k)^T$ on S
- An initial distribution μ on S
- Order the eigenvalues of the transition matrix according to their absolute value, starting with the larges one: $|\theta_1| \geq |\theta_2| \geq \cdots \geq |\theta_k|$

We can then formulate:

Proposition 7 (Perron-Frobenius)

$$\sup_{j \in \{1,\ldots,k\}} |\mu_j^{(n)} - \pi_j| = O(|\theta_2|^n)$$

Note: If π is reversible, the basis of the second largest eigenvalue cannot be improved. But we can state more precisely:

$$\sup_{j \in \{1,\ldots,k\}} |\mu_j^{(n)} - \pi_j| \leq \frac{1}{\sqrt{\min_{i \in \{1,\ldots,k\}} \pi_i}} |\theta_2|^n.$$

For several reasons, this bound is of limited practical use since:

- given interesting cases of large sample spaces, it can become infeasible to determine θ_2
- the Markov chain needs to be reversible
- the bound does not depend on the initial distribution μ.

An alternative is to use so called χ^2-contrasts. To outline this concept we need the following definitions:

Definition 9 *The matrix $M := P\tilde{P}$ is called the multiplicative reversible version of the transition matrix P if we set*

$$\tilde{p_{i,j}} := \frac{\pi_j p_{i,j}}{\pi_i}$$

As one can check easily, M is indeed reversible.

Definition 10 *The χ^2-**contrast** of μ given ν is then defined as*

$$\chi^2(\mu, \nu) := \sum_{i \in S} \frac{(\mu_i - \nu_i)^2}{\nu_i}$$

where we require $\nu_i > 0 \quad \forall i \in S$.

It would then be possible to proof the following property:

Proposition 8 *Using the notation we just introduced it holds:*

$$d_{TV}^2((\mu^T P^n)^T, \pi) \leq \frac{\chi^2(\mu, \pi)}{4} \theta_{M,2}^n$$

4 Markov Chain Monte Carlo

In this chapter we will consider algorithms that help us with sampling from potentially complicated and high dimensional distributions. We start by looking at the more classical approaches. The last section then deals with an alternative, very different approach.

4.1 The Hard-Core Model

Let us start by discussing a common example. Consider a graph G=(V,E). We randomly assign 0 or 1 to every vertex $v_i \in V = \{v_1, \ldots, v_k\}$ in such a way that no two neighbors, i.e. two vertices that share an edge $e_i \in E = \{e_1, \ldots, e_k\}$ both take the value 1. If a configuration fulfills this condition, it is called *feasible*. We now pick a feasible configuration from the set of all feasible configurations uniformly at random. What number of 1's should we expect? We can formulate this problem in a more precise way: Let $\xi \in \{0,1\}^V$ be any configuration, set Z_G to the total number of feasible configurations and define a probability function μ_G on $\{0,1\}^V$ by

$$\mu_G(\xi) = \begin{cases} \frac{1}{Z_G} & , \quad \xi \text{ is feasible} \\ 0 & , \quad \text{otherwise} \end{cases}$$

Using this notation and denoting the number of 1's in a configuration by $n(\xi)$ we are interested in

$$E_{\mu_g}(n(X)) = \sum_{\xi \in \{0,1\}^V} n(\xi)\mu_G(\xi) = \frac{1}{Z_G} \sum_{\xi \in \{0,1\}^V} n(\xi) I_{\{\xi \text{ is feasible}\}}$$

Even for moderately sized graphs it is obviously impossible to evaluate this sum. But using an MCMC algorithm we will be able to generate samples from μ_g. We can then apply the law of large numbers and estimate the expected number of 1's in a configuration. We approach the problem by constructing an irreducible and aperiodic Markov chain $\{X_0, X_1, X_2, \dots\}$ with reversible distribution μ_G on the state space $S = \{\xi \in \{0,1\}^V : \xi \text{ is feasible}\}$. A Markov chain with the desired properties can be obtained using the following algorithm.

1. Pick a vertex $v \in V$ uniformly at random
2. Toss a fair coin
3. If the coin heads and if all neighbors of v take the value 0 in X_n, then set $X_{n+1} = 1$. Otherwise, $X_{n+1} = 0$.
4. Leave the value for all other vertices unchanged, that is: $X_{n+1}(w) = X_n(w) \quad \forall \quad w \neq v$.

The resulting Markov chain is irreducible for the following reason: Given a feasible configuration ξ, we can reach the "all 0" configuration in a finite number of steps. From there on we can reach any other feasible configuration ξ'. Moreover, the chain is aperiodic since we can go from ξ to ξ in one step. It remains to show that μ_G is reversible (and hence stationary for the chain). For two feasible configurations ξ and ξ' we need to check that

$$\mu_G(\xi)P_{\xi,\xi'} = \mu_g(\xi')P(\xi', \xi).$$

We split the problem into three different cases. If the configurations are exactly the same, the equation is trivial. Secondly, if the configurations differ in more than two vertices, the equation is also obvious since the algorithm changes only one vertex at a time. Finally, assume the configurations differ at a single vertex v. All neighbors of ξ and ξ' must then be 0 because we deal with feasible configurations. We then get:

$$\mu_G(\xi)P_{\xi,\xi'} = \frac{1}{Z_G}\frac{1}{2k} = \mu_g(\xi')P_{\xi',\xi}.$$

Hence our chain is reversible with respect to μ_G and μ_G is also the stationary distribution. □

This example is one of the models I implemented using R. The source code can be found in the appendix, page 32.

4.2 The Metropolis-Hastings-Algorithm

The method used in the above example can be generalized. Again, suppose we would like to sample from a potentially difficult distribution π that lives on a sample space of high dimension. If we could find a Markov chain whose unique stationary distribution is π we could run this chain long enough and then take the result as an approximate sample form π.

A very general way to construct such a Markov chain is the Metropolis-Hastings-Algorithm. Alternatively, one could use the so called Gibbs-Sampler which turns out to be a special case of the Metropolis-Hastings-Algorithm. The above example of the hard core model uses such a Gibbs sampler.

The algorithm now works as follows. It constructs - as the next theorem will show - a Markov chain that has π even as its reversible distribution:

1. Consider the distribution of interest π and an arbitrary other Markov chain on S with transition matrix $Q = (q_{i,j})$.

2. Choose a starting value X_0.

3. Given the current state X_n generate a *proposal* Z_{n+1} from $q_{X_n,\cdot}$

4. Perform a Bernoulli experiment with probability of success $\alpha(X_n, Y_{n+1})$ where
$$\alpha(i,j) = \min\left\{1, \frac{\pi_j q_{j,i}}{\pi_i q_{i,j}}\right\}$$
and set $\alpha = 1$ in case of $\pi(i)q_{i,j} = 0$.

5. If the experiment is successful, set $X_{n+1} = Y_{n+1}$. Otherwise leave X_n unchanged, i.e. $X_{n+1} = X_n$.

6. Continue the procedure with n+1.

Theorem 6 *The Metropolis-Hastings-Algorithm produces a Markov chain* $\{X_0, X_1, X_2, \dots\}$ *which is reversible with respect to* π.

Proof: First recall that we have to show $\pi_i p_{i,j} = \pi_j p_{j,i}$. We assume $i \neq j$

(the other case is obvious) and can write:

$$\begin{aligned}
\pi_i p_{i,j} &= \pi_i q_{i,j} \alpha_{i,j} \\
&= \pi_i q_{i,j} \min\left\{1, \frac{\pi_j q_{j,i}}{\pi_i q_{i,j}}\right\} \\
&= \min\{\pi_i q_{i,j}, \pi_j q_{j,i}\}
\end{aligned}$$

The fact that this equation is symmetric in i and j completes the proof. □

So this approach is indeed very general. Some choices for the transition probabilities $q_{i,j}$ are common:

- Original Metropolis algorithm: $q_{i,j} = q_{j,i}$
- Random walk Metropolis-Hastings: $q(i,j) = q(j-i)$
- Independence sampler: $q(i,j) = q(j)$ independent of i

Finally, let us remark that we receive the $(k^(th)$ component)Gibbs-Sampler if we set $\alpha(i,j) \equiv 1$.

To demonstrate the algorithm, I implemented an example (random walk Metropolis-Hastings). The complete source code for this example can be found in the appendix on page 32.

4.3 Convergence rates of MCMC algorithms

Let us continue the current section by demonstrating how difficult and labor-intensive it can be to give useful bounds on convergence rates of Markov chains. Consider a graph $G = (V, E)$. Use k to denote the number of vertices in G and suppose that a vertex $v \in V$ has at most d neighbors. The problem of *q-colorings* is well-known from graph theory. In short, the problem is to assign one of q colors (or one of q integers) to every vertex v such that no two adjacent vertices have the same color. For our setup assume that $q > 2d^2$. If we would use a Gibbs-sampler to solve this problem and if we expect from this algorithm to produce a solution with total variation distance less than ϵ from the distribution that puts equal probability mass on all valid q-colorings, we need at most

$$k\left(\frac{\log(k) - log(\epsilon) - \log(d)}{\log(q) - \log(2d^2)} + 1\right)$$

iterations. This is indeed a quite useful bound since in essence it stats that the number of iterations needed behaves as $O(k(\log(k) - \log(\epsilon)))$. Note that it is possible to improve this result in the sense that the bound holds for $q > 2d$.

Proof: As done earlier, we consider the main ideas of the proof (in this case a coupling argument) and leave the technical details to [3] pp. 57-62 (indeed almost six pages!). The proof considers two coupled Markov chains $\{X_0, X_1, X_2, \dots\}$ and $\{X_0', X_1', X_2', \dots\}$. The first one is started in a fixed state and the second one according to the (stationary) distribution which gives equal weight to all valid q-colorings. The chains are linked in such a way that if they coincide at a certain time T they will stay the same for all times $t \geq T$. The proof continues by showing that the total variation distance between the distribution $\mu^{(n)}$ of the chain $\{X_0, X_1, X_2, \dots\}$ after n transitions and the stationary distribution $\rho_{G,q}$ gets smaller as $P(X_n = X_n')$ gets closer to 1. The probability $P(X_n = X_n')$ is examined in several steps. First, the probability that the configurations X_n and X_n' coincide at a certain vertex v is discussed. This is then generalized to the case of interest where all vertices coincide. It turns out that

$$P(X_n \neq X_n') \leq \frac{k}{d}\left(\frac{2d^2}{q}\right)^m$$

where $m = n/k$ (use $\lfloor m \rfloor$ if not an integer). The total variation distance $d_{TV}(\mu^n, \rho_{G,q})$ can then be derived using the remark on page 6. □

4.4 The Propp-Wilson-Algorithm

The algorithms we considered so far have two major drawbacks in common. First of all, the samples generated were only *close* to the real distribution but in general never *exact*. Furthermore, we had difficulties in determining what the phrase "close" means and when we have actually reached that level of accuracy. The following algorithm was developed by Jim Propp and David Wilson at the MIT in the mid 1990's. It's main idea is running not just one Markov chain at a time but several copies of it. The algorithm stops as soon as a perfect sample is achieved. Let us also note, that the algorithm starts in the past, that is, it doesn't run the Markov chain from time 0 on into the future but rather starts in the past and stops at time 0. For a more specific description of how the Propp-Wilson-Algorithm works, consider the following setup:

- A finite state space $S = \{s_1, \dots, s_k\}$
- We want to sample from a probability distribution π.
- For the algorithm itself we need a sequence of *iid* random numbers that are uniformly distributed on $[0, 1]$.
- Let $N_1, N_2, \dots$ be an increasing sequence of positive integers. A broadly used sequence is $\{1, 2, 4, 8, \dots\}$.

Now here is how the algorithm works:

1. Set m=1
2. For each possible state $s_i, i = 1, \dots, k$ simulate a Markov chain starting in state s_i at time $-N_m$ and run it up to time 0. To do so use an update

function[3] $\phi : S \times [0,1] \longrightarrow S$ and the random numbers $U_{-N_m+1}, \ldots, U_{-1}, U_0$. It is important to note that these random numbers are the same for all k chains! We also have to reuse the random numbers $U_{-N_{m-1}}, \ldots, U_{-1}, U_0$ for the m^{th} step. Otherwise the algorithm will not produce correct results.

3. If at time 0 all k chains end in the same state $\hat{s}$ stop the algorithm and use $\hat{s}$ as a sample. Otherwise increase m by 1 and continue with step 2.

Two interesting questions arise immediately: Will the algorithm terminate and if it does so, will it give a correct, unbiased sample? As far as the first part of this question is concerned, we will only note that there are cases when the algorithm doesn't stop. But there is a 0-1 law: Either the algorithm terminates almost surely or the probability that it stops is 0. An obvious consequence is that it suffices to show that P(algorithm terminates)>0 holds.

The second part of the question will be answered by the next theorem:

Theorem 7 *Consider an irreducible and aperiodic Markov chain with state space $S = \{s_1, \ldots, s_k\}$ and stationary distribution $\pi = (\pi_1, \ldots, \pi_k)$. As above, let ϕ be an update function and $N_1, N_2, \ldots$ an increasing sequence of positive integers. If the Propp-Wilson algorithm terminates, then we have*

$$P(Y = s_i) = \pi_i \qquad \forall \quad i \in 1, \ldots, k$$

where Y represents the algorithms output.

Proof:

Let us fix $\epsilon > 0$ and $s_i \in S$. We have to show that

$$|P(Y = s_i) - \pi_i| < \epsilon$$

[3] An update function helps us to get form the current state X_n into the next state X_{n+1}. For $s_i \in S$ it's a piecewise constant function of x and for $s_i, s_j \in S$ the length of the interval where $\phi(s_i, x) = s_j$ equals $p_{i,j}$.

holds. By assumption the algorithm terminates with probability 1 and by picking M large enough we can achieve:

$$P(\text{algorithm doesn't need starting times earlier than} - N_M) \geq 1 - \epsilon$$

If we fix such an M we can apply a coupling argument: We run a first chain from time $-N_M$ up to 0 and, using the same update function and random numbers U_i, we also run a second chain with the initial state chosen according to the stationary distribution π. Let us denote the state of this second, imaginary chain at time 0 with $\tilde{Y}$. It is important to keep in mind that $\tilde{Y}$ has distribution π since this is the stationary distribution. Furthermore,

$$P(Y \neq \tilde{Y}) \leq \epsilon$$

since we chose M large. Hence we can write:

$$\begin{aligned} P(Y = s_i) - \pi_i &= P(Y = s_i) - P(\tilde{Y} = s_i) \\ &\leq P(Y = s_i, \tilde{Y} \neq s_i) \\ &\leq P(Y \neq \tilde{Y}) \leq \epsilon \end{aligned}$$

Similarly, we get $\pi_i - P(Y = s_i) \leq \epsilon$ and by combining these two results we obtain

$$|P(Y = s_i) - \pi_i| < \epsilon$$

□

Before we work through a detailed example, we should discuss one more point. In general, we would use MCMC for complicated models that make an analytical analysis infeasible; for example, if we need to analyze a model with a large sample space. Using the above Propp-Wilson-Algorithm, this inevitably means running a large number of Markov chains. An important technique for considerably reducing the number of chains that need to be simulated is the *sandwiching* technique. It can be used for Markov chains that have certain monotonicity properties. Assume we can order the state space and we could also find a transition function that preserves this ordering. Then it would be enough to run two Markov chains. One starting in the smallest possible state and one in the largest. If the two chains coincide at time 0, we can stop the algorithm since all other chains remain between the smallest and largest possible value.

4.5 The Ising Model

Let $G = (V, E)$ be a graph. The Ising model is a certain way of picking a random element of $\{-1, 1\}^V$. It's physical interpretation might be to think of the vertices as atoms in a ferromagnetic material and of -1 and 1 as possible spin orientations of the atoms. To describe the probability distribution of all possible configurations, we introduce two parameters:

- The **inverse temperature** $\beta \geq 0$ which is a fixed non-negative number
- The **energy** $H(\xi$ where $\xi \in \{-1, 1\}^V$ is a spin configuration and H is defined as:

  $$H(\xi) := -\sum_{(x,y)\ inE} \xi(x)\xi(y)$$

  Here, $\xi(x)$ is the orientation of the chosen configuration at vertex x.

A certain spin configuration $X \in \{-1, 1\}^V$ is chosen according to the probability distribution $\pi_{G,\beta}$ with

$$\pi_{G,\beta}(\xi) := \frac{1}{Z_{G,\beta}} exp(-\beta H(\xi))$$

$Z_{G,\beta}$ is nothing but a normalizing constant to ensure that we end up with a probability measure.

In order to make the Propp-Wilson-Algorithm work for this problem, we need to introduce an ordering. We can then use the sandwiching technique and thereby reduce the number of Markov chains necessary from 2^k to 2 (where k is the number of vertices). For two spin configurations ξ and η we shall write $\xi \preceq \eta$ if $\xi(x) \leq \eta(x) \quad \forall \quad x \in V$. Of course, this is not a total ordering but it allows us to define a smallest and largest spin configuration (-1 everywhere and 1 everywhere respectively).

As a last step, we need to specify how to simulate the two remaining Markov chains. We will use a simple Gibbs sampler. Given $X_n \in -1, 1^V$ we obtain X_{n+1} by picking a vertex x at at random. We leave all vertices except x unchanged and the spin of $X_{n+1}(x)$ is determined using a random number U_{n+1} where U_{n+1} is uniformly distributed on [0,1]. We set:

$$X_{n+1} := \begin{cases} 1 & , \quad U_{n+1} < \frac{\exp[2\beta(k_+(x,\xi)-k_-(x,\xi))]}{\exp[2\beta(k_+(x,\xi)-k_-(x,\xi))]+1} \\ -1 & , \quad \text{otherwise} \end{cases}$$

Here, $k_+(x, X_n)$ denotes the neighbors of x having positive spin and $k_-(x, X_n)$ the number of neighbors having negative spin.

We can now implement the algorithm. The reader will find the complete source code using R in the appendix on page 32. Figure 1 on page 22 shows a possible sample from the model. It was computed using a quadratic grid of 100 times 100 and $\beta = 0.3$. Note the significant clustering.

Figure 1: A sample from the Ising model with $\beta = 0.3$. Note the significant clustering

Part II

MCMC on General State Spaces

Most of the concepts and ideas we have studied so far can be applied to the case of a *general state space* $E \subset R^k$, E measurable. Nevertheless, we need some measure theory to generalize the definitions.

5 Definitions and Basic Properties

5.1 Definition

Definition 11 *From measure theory we recall the concept of a* **kernel**. *Consider two measure spaces* $(\Omega, \mathcal{A})$ *and* $(\Omega', \mathcal{A}')$. *The function*

$$K : \Omega \times \mathcal{A}' \quad \rightarrow \quad [0, \infty]$$

is a **kernel** *from* $(\Omega, \mathcal{A})$ *to* $(\Omega', \mathcal{A}')$ *if*

- $\omega \mapsto K(\omega, A')$ *is* $\mathcal{A}$*-measurable* $\quad \forall \quad A' \in \mathcal{A}'$
- $A' \mapsto K(\omega, A')$ *is a measure on* $\mathcal{A}'$ $\quad \forall \quad \omega \in \Omega.$

If $K(\omega, \Omega') = 1$, *the kernel is a also called* **Markov kernel**.

Loosely spoken, a kernel gives the probability of ending up somewhere in A' if starting in ω.

Definition 12 *A sequence of random variables* $\{X_0, X_1, X_2, \dots\}$ *with values in E is called (homogeneous)* **Markov chain** *if*

$$P(X_{n+1} \in A | X_n = x) = P(x, A) \quad \forall \quad n \geq 1.$$

We will denote $P(\dots | X_0 = x)$ *with* $P_x(\dots)$.

5.2 Important Properties

Definition 13 *A Markov chain is $\varphi-$ irreducible for a probability measure φ on E if for all measurable sets $A \subset E$ with $\varphi > 0$ we have the possibility of finite return times, that is for the time of first return to A*

$$\tau_A := \inf\{n \in N : X_n \in A\}$$

it holds

$$P_x(\tau_A < \infty) > 0 \quad \forall \quad x \in E.$$

A Markov Chain is **irreducible** *if it is $\varphi-$ irreducible for a probability distribution φ.*

We could say that irreducibility means that all “interesting“ sets can be reached. Note that a Markov chain might have many irreducibility distributions; but there exists a maximal irreducibility distribution φ_0 in the sense that all other irreducible distributions φ are absolutely continuous with respect to φ_0.

The way we define aperiodicity in case of the general state space is slightly more technical than in part 1 but is in essence still the same:

Definition 14 *A* m-cycle *for an irreducible Markov chain with transition kernel P is a collection $\{E_0, \ldots, E_{m-1}\}$ of disjoint sets such that*

$$P(x, E_j) = 1$$

for $i = 0, \ldots, m-1; \quad j = (i+1) mod\ d \quad$ and $\quad \forall \quad x \in E_i$. The **period** *$d$ of the chain is the largest m for which an m-cycle exists. The chain is* **aperiodic** *if $d = 1$.*

Definition 15 *A probability distribution π on E is a* **stationary distribution** *for the Markov chain $\{X_0, X_1, X_2, \dots\}$ with transition kernel $P(x, A)$ if*

$$\pi(A) = \int_E \pi(dx) P(x, A) \quad \forall \quad A \subset E \textit{ measurable.}$$

Again, before we can consider the behavior of a Markov chain after many transitions, we need to introduce a metric.

Definition 16 *Given two probability distributions μ and ν on E we define the* **total variation distance** *of the two measures by*

$$\sup_{A \subset E} |\mu(A) - \nu(A)|$$

where we require A to be measurable and write $||\mu - \nu||$ short-hand.

6 Markov Chain Convergence Theorem

Theorem 8 *Suppose $\{X_0, X_1, X_2, \dots\}$ is an irreducible, aperiodic Markov chain on the state space E with transition kernel P and stationary distribution π. Denote the n^{th} transition probabilities*[4] *by $P^n(x, \cdot)$. Then we have*

$$||P^n(x, \cdot) - \pi(\cdot)|| \to 0$$

for $\pi-$ a.e. $x \in E$.

Proof: The proof of this theorem is based on several lemmas, that describe properties of Markov chains related to irreducibility and aperiodicity. As in the discrete case, we would use a coupling technique and return times to complete the proof. A proof in full length can be found in [5]

[4] To be precise: $P^1(x, A) := P(x, A)$ and $P^{n+1} := \int_E P^n(x, dy) P(y, A) \quad n = 1, 2, \dots$

7 The Metropolis-Hastings-Algorithm

In this last section, let us consider how our theoretical results could be used for simulation purposes. As in the discrete case we turn to the general ideas of the Metropolis-Hastings-Algorithm. As it turns out, the algorithm works in almost the same way and even the proof of convergence is closely related to the discrete case.

7.1 The Algorithm

Suppose we are interested in properties of a distribution π with density π_d and want to sample from this distribution. Such problems arise for example in Bayesian statistical inference.

As we already learned, the bottom-line of MCMC algorithms is to construct a Markov chain with reversible distribution π. Now in the given case of a general state space the Metropolis-Hastings-algorithm works as follows:

- Choose an arbitrary Markov chain with Markov kernel of the form $Q(x, dy) = q(x, y)dy$.
- Choose a starting point X_0.
- Given the state X_n, generate a proposal Y_{n+1} from $Q(X_n, \cdot)$
- Perform a Bernoulli experiment with probability of success $\alpha(X_n, Y_{n+1})$ where

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

  and set $\alpha = 1$ in case of $\pi_d(x)q(x, y) = 0$.

- If the experiment is successful, set $X_{n+1} = Y_{n+1}$. Otherwise leave X_n unchanged, i.e. $X_{n+1} = X_n$.
- Continue the procedure with n+1.

Theorem 9 *This algorithm produces a Markov chain $\{X_0, X_1, X_2, \dots\}$ which is reversible with respect to π.*

Proof: First recall that we have to show $\pi(dx)P(x, dy) = \pi(dy)P(y, dx)$. We assume $i \neq j$ (the other case is obvious) and can write:

$$\begin{aligned}\pi(dx)P(x,dy) &= \pi_d(x)dxq(x,y)dy\alpha(x,y)\\ &= \pi_d(x)q(x,y)dxdx\min\left\{1, \frac{\pi_d(y)q(y,x)}{\pi_d(x)q(x,y)}\right\}\\ &= \min\{\pi_d(x)q(x,y), \pi_d(y)q(y,x)\}dxdy\end{aligned}$$

The fact that this equation is symmetric in i and j completes the proof. □

For special choices of $\alpha(x, y)$ see the discrete version on page 15.

7.2 Convergence Bounds

As one of the last points of this notes let us state a theorem dealing with convergence bounds. I studied the theorem and its proof since it uses a coupling technique and also allowed me to revise my knowledge about martingales. Now here is the theorem:

Theorem 10 *For a Markov chain on a state space E with transition kernel P consider the following objects:*

- *Two copies of the Markov chain: $\{X_0, X_1, X_2, \dots\}$ and $\{X'_0, X'_1, X'_2, \dots\}$ started from initial distributions $\mathcal{L}(X_0)$ and $\mathcal{L}(X'_0)$*

- *A probability measure $\nu(\cdot)$ on E*
- *A* minorisation condition[5]

  $$P(x,A) \geq \epsilon\nu(A) \quad \forall \quad x \in C$$

  for some subset $C \subset E$ and a $\epsilon > 0$

- *A* drift condition

  $$\bar{P}h(x,y) \leq \frac{h(x,y)}{\alpha} \quad \forall \quad (x,y) \notin C \times C$$

  where $\alpha > 1$ and h is a function $H : E \times E \to [1,\infty)$ with

  $$\bar{P}h(x,y) \equiv \int_E \int_E h(z,w)P(x,dz)P(y,dw).$$

- *Finally, define*

  $$B = max\{1, \alpha(1-\epsilon)\sup_{C \times C}\{(\bar{R})h(x,y)\}\}$$

  where

  $$(\bar{R})h(x,y) = \int_E \int_E (1-\epsilon)^{-2} h(z,w)(P(x,dz) - \epsilon\nu(dz)))(P(y,dw) - \epsilon\nu(dw))$$

  for $(x,y) \in C \times C$.

  For our Markov chains $\{X_0, X_1, X_2, \dots\}$ and $\{X'_0, X'_1, X'_2, \dots\}$ started according to a joint initial distribution $\mathcal{L}(X_0, X'_0)$ and integers $1 \leq j \leq k$ we can then say that:

  $$||\mathcal{L}(X_k) - \mathcal{L}(X'_k)|| \leq (1-\epsilon)^j + \alpha^{-k}B^{j-1}E[h(X_0, X'_0)].$$

[5] See Definition 19 for a detailed and formal definition.

8 Outlook and a CLT for Markov Chains

I'd like to close the notes with a short outlook on some further topics I've already been looking at but without following the proofs to a deeper extend:

Definition 17 *An irreducible Markov chain with maximal irreducibility distribution ψ is* **recurrent** *if $\forall \quad A \subset E$ measurable with $\psi(A) > 0$ we have:*

- $P_x(X_n \in A \text{ infinitely often}) > 0 \quad \forall \quad x$
- $P_x(X_n \in A \text{ infinitely often}) = 1 \quad \psi\text{-a.s.}.$

If we recall that irreducibility *meant that all "interesting" sets can be reached,* recurrence *means that all such sets will be reached infinitely often from at least almost all starting points.*

In addition we can be slightly stricter and say that a Markov chain is **Harris recurrent** *if $\forall \quad A \subset E$ measurable with $\psi(A) > 0$ we have*

$$P_x(X_n \in A \text{ infinitely often}) = 1 \quad \forall \quad x \in E.$$

As in the discrete case, we can also define *ergodicity* (compare Proposition 3 and Theorem 2) but need to add one more condition.

Definition 18 *A Markov chain $\{X_0, X_1, X_2, \dots\}$ with stationary distribution π is* **ergodic** *if it is irreducible, aperiodic and Harris recurrent. It is* **geometrically ergodic** *if there exists $r < 1$ and a non-negative (possibly extended real-valued) function M such that $M(x)$ is Lebesgue-integrable with respect to π and*

$$||P^n(x,\cdot) - \pi(\cdot)|| \leq M(x)r^n \quad \forall \quad x \quad \forall n \geq 1.$$

The chain is **uniformly ergodic** *if there exist constants $M > 0$ and $r < 1$ such that $\forall \quad x \quad \forall n \geq 1$*

$$||P^n(x,\cdot) - \pi(\cdot)|| \leq Mr^n.$$

So the total variation distance is of order $O(r^n)$.

Easier to verify is the following condition:

Definition 19 *A π-irreducible Markov chain with transition kernel P satisfies a* **minorisation condition** *if we find a measure ν on $\sigma(E)$, $m \geq 1$, $\beta \geq 0$ and a measurable set $C \subset E$ of positive ν-measure such that*

$$P(x, A) \geq \beta\nu(A) \quad \forall \quad x \in C \text{ and } \forall \quad A \subset E \text{ measurable.}$$

Finally, we can close the notes with a nice central limit theorem.

Theorem 11 *Consider an uniformly ergodic Markov chain $\{X_0, X_1, X_2, \dots\}$ with stationary distribution π and a real valued function f with f^2 is integrable with respect to π. Using*

$$\overline{f_n} := \frac{1}{n+1} \sum_{i=0}^{n} f(X_i).$$

Then

$$\sqrt(n)(\overline{f_n} - E(f)) \xrightarrow{d} Y \sim N(0, \sigma_f^2).$$

Proof: A proof can be found in [2].

A The Hard-Core Model

```
# Supervised reading in MCMC with Professor Rosenthal
# October 2005

# This function creates a single feasible configuration
# and returns the number of vertices that are marked "1"

# The function takes the size of the graph and the number
# of simulation steps as arguments.
# Furthermore, the current configuration can be plotted

getConfiguration<-function(size,steps,plotResult=FALSE) {
  graph<-matrix(0,size,size)  # Construct initial graph

  for(i in 1:steps) {
    x<-trunc(runif(1,1,size+1))  # Choose one vertex uniformly at random
    y<-trunc(runif(1,1,size+1))
    coin<-trunc(runif(1,0,2))  # Toss a fair coin

    if(coin) {      # Coin comes up heads?
      change<-TRUE    # Then see if we change our configuration
      if(x+1<=size) {
        if(graph[x+1,y]==1) change<-FALSE  # Check right neighbor
      }
      if(x>=2 & change) {
        if(graph[x-1,y]==1) change<-FALSE  # Check left neighbor
      }
      if(y+1<=size & change) {
        if(graph[x,y+1]==1) change<-FALSE  # Check upper neighbor
      }
      if(y>=2 & change) {
        if(graph[x,y-1]==1) change<-FALSE  # Check lower neighbor
      }
      if(change) graph[x,y]<-1  # All neighbors are still marked "0"?
    }
    else graph[x,y]<-0
  }

  # Plot configuration?
  if(plotResult) {
    print(graph)

    plot(0:size,0:size,type="n")
    for(i in 1:size) {
      for(j in 1:size) {
        if(graph[i,j]==0) points(j,size-i+1,pch=21)
        else points(j,size-i+1,pch=19,col="red")
      }
    }
  }

  return(sum(graph))  # Return number of vertices marked "1"
}
```

B Example: The Metropolis-Hastings-Algorithm

```
# Supervised Reading in MCMC with Prof. Rosenthal; fall 2005

# The following code is a small application of the Metropolis-Hastings algorithm
# MCMC is used to sample from a inverse gamma distribution


# The densitiy function of an inverse gamma distribution with parameter alpha and beta at x
dinvgamma<-function(alpha,beta,x){
  if(x<=0) return(0)    # Support is x > 0
  return(beta^alpha/gamma(alpha)*x^(-alpha-1)*exp(-beta/x))
}


# The function does the simulation and takes the number of steps as its first argument
# The second and third arguments are the parameters for the inverse gamma distribution
MCMC<-function(n,alpha,beta) {
  state<-1      # Start simulation by defining a starting value
  for(i in 1:n) {      # Simulate n steps
    proposal<-runif(1,state-1,state+1)    # Make a suggestion for the next state
    pi.y<-dinvgamma(alpha,beta,proposal)    # Evaluate density of destination distriubtion
    pi.x<-dinvgamma(alpha,beta,state)
    if(pi.x==0) p.success<-1
    else p.success<-min(1,pi.y/pi.x)    # Probability of accepting the suggestion
    if(runif(1)<=p.success) state<-proposal    # Accept suggestion?
  }
  return(state)      # Return the last state, i.e. an approximate inverse gamma sample
}
```

C The Ising Model

```
# Supervised reading in MCMC with Professor Rosenthal
# November 2005
# The Ising model using the Propp-Wilson algorithm
# to obtain spin-configurations.


# The function receives a spin configuration and a specified element
# of this configuration. It then returns the number of neigbours
# with positiv and negativ spin as well as the difference
# between these two numbers. Finally, size is the size of the
# quadratic grid.

neighbours<-function(spin.conf,x,y,size) {
  counter<-vector("numeric",2)      # Count positive and negativ neighbours
  if(x+1<=size) {          # Check right neighbour
    if(spin.conf[x+1,y]==1) counter[1]<-1 else counter[2]<-1
  }
  if(x>=2) {          # Check left neighbor
    if(spin.conf[x-1,y]==1) counter[1]<-counter[1]+1 else counter[2]<-counter[2]+1
  }
  if(y+1<=size) {          # Check lower neighbor
    if(spin.conf[x,y+1]==1) counter[1]<-counter[1]+1 else counter[2]<-counter[2]+1
  }
  if(y>=2) {          # Check upper neighbor
    if(spin.conf[x,y-1]==1) counter[1]<-counter[1]+1 else counter[2]<-counter[2]+1
  }

  return(list(counter=counter,diff=(counter[1]-counter[2])))
}


# The function takes the size of the (quadratic) graph
# and the inverse temperature beta as parameters.
# Furthermore, the current configuration can be plotted.

getSpinConfiguration<-function(size,beta,plotResult=FALSE) {
  conf.bottom<-matrix(-1,size,size)  # Configutaion with -1 everywhere
  conf.top<-matrix(1,size,size)    # Configutaion with +1 everywhere
  m<-1          # Counter
  x.cord<-trunc(runif(1,1,size+1))  # x-coordinates of vertices chosen
  y.cord<-trunc(runif(1,1,size+1))  # y-coordinates of vertices chosen
  uniformRV<-runif(1)      # Uniform RVs drawn

  repeat {      # Loop until configurations coalesce
    for(i in 2^(m-1):1) {  # Perform 2^(m-1) steps on the spin configurations
      exp.bottom<-exp(2*beta*neighbours(conf.bottom,x.cord[i],y.cord[i],size)$diff)
      exp.top<-exp(2*beta*neighbours(conf.top,x.cord[i],y.cord[i],size)$diff)
      if(uniformRV[i]<exp.bottom/(exp.bottom+1)) conf.bottom[x.cord[i],y.cord[i]]<-1
```

```
        else conf.bottom[x.cord[i],y.cord[i]]<-(-1)  # Update bottom chain
      if(uniformRV[i]<exp.top/(exp.top+1)) conf.top[x.cord[i],y.cord[i]]<-1
        else conf.top[x.cord[i],y.cord[i]]<-(-1)  # Update top chain
    }

    # Exit loop if both configurations coalesce
    if(sum(conf.bottom==conf.top)==size*size) break

    # Generate new random variables and append to list.
    # Choose vertices uniformly at random.
    x.cord<-c(x.cord,trunc(runif(2^(m-1),1,size+1)))
    y.cord<-c(y.cord,trunc(runif(2^(m-1),1,size+1)))
    uniformRV<-c(uniformRV,runif(2^(m-1)))

    # Reset both configurations
    conf.bottom[]<-(-1)
    conf.top[]<-1

    m<-m+1
  }

  # Plot configuration?
  if(plotResult) {
    # print(conf.top)

    plot(0:size,0:size,type="n")
    for(i in 1:size) {
      for(j in 1:size) {
        if(conf.top[i,j]==-1) points(j,size-i+1,pch=21,cex=1.1)
        else points(j,size-i+1,pch=19,cex=1.1)
      }
    }
  }

  return(list(conf=conf.top,steps=2^m))
}
```

References

[1] Bauer, H. (2002), Wahrscheinlichkeitstheorie, 5., durchgesehene und verbesserte Auflage. Walter de Gruyter, Berlin, New York

[2] Cogburn, R. (1970), The central limit theorem for Markov processes. Proc. Sixth Berkeley Sympos. Math. Statist. Probab., Vol.2, pp. 485-512. University of California, Berkeley, CA

[3] Häggström, O. (2002), Finite Markov Chains and Algorithmic Applications. Cambridge University Press, Cambridge

[4] Roberts, G. O. and Rosenthal, J. S. (2004), General state space Markov chains and MCMC algorithms. Technical Report No. 0402, University of Toronto, Toronto

[5] Jeffrey S. Rosenthal (2002), A review of asymptotic convergence for general state space Markov chains. http://probability.ca/jeff/research.html, University of Toronto, Toronto

[6] Jeffrey S. Rosenthal (2002), Quantitative convergence rates of Markov chains: A simple account. http://probability.ca/jeff/research.html, University of Toronto, Toronto

[7] Schmidt, V. (2003), Vorlesungsskript Markov-Ketten und Monte-Carlo-Simulation. www.mathematik.uni-ulm.de/stochastik, Universität Ulm, Ulm

[8] Tierney, L. (1994), Markov Chains for Exploring Posterior Distributions. Technical Report No. 560 (Revised), University of Minnesota, Minnesota