Notes on
Markov Chain Monte Carlo Methods

Dina Research Summer School, August 16 – 27,1999

Computer Intensive Statistical Methods
– with Applications in Agriculture

Jesper Møller (jm@math.auc.dk)
Department of Mathematical Sciences, Aalborg University, Denmark

Preface: These notes has been prepared as a part of an intensive course on 'Computer Intensive Statistical Methods – with Applications in Agriculture'. The material on 'Markov chain Monte Carlo Methods' was presented within three days: The present notes was used the first and third day, while Heidi Hundborg was presenting some material on 'Markov chain Monte Carlo in Practice' (including some material on WinBugs) the second day. Various theoretical and practical exercises were used as well during the course.

Contents

1 Introduction

1.1 Monte Carlo Methods

A Monte Carlo method is a particular form of numerical integration which uses randomness. Suppose X is a random variable with probability density $\pi(x)$ and a statistic $g(X)$ of interest and we wish to estimate the expectation $\mathbb{E}_\pi g(X)$, assuming this is not analytically tractable. Recall that

$$\begin{aligned}\mathbb{E}_\pi g(X) &= \sum_x g(x)\pi(x), \qquad \text{discrete case}\\ &= \int g(x)\pi(x)\,dx, \qquad \text{continuous case}\\ &= \int g(x)\pi(x)\,\nu(dx), \qquad \text{general case}\end{aligned}$$

($\pi(x)$ density w.r.t. measure ν). Typically, X is multidimensional and its distribution

$$\Pi(dx) = \pi(x)\nu(dx) \qquad \text{is complicated!}$$

We refer to Π as the *target distribution* and π as the *target density*.

REMARK 1
We are often working with an unnormalized density π_u,

$$\pi(x) \propto \pi_u(x)$$

i.e.

$$\pi(x) = \frac{1}{c}\pi_u(x),$$

where

$$c = \int \pi_u(x)\nu(dx)$$

is the normalizing constant. *For 'complicated models' c is often not known on closed form. Sometimes we forget the subscript u and write $\pi(x)$ for $\pi_u(x)$.*

EXAMPLE 1 (EXPONENTIAL FAMILY MODELS)
Many parametric models $\pi_u(x) = \pi_\theta(x)$ are of the form

$$\pi_\theta(x) = b(x)\exp(\theta \cdot t(x)), \tag{1}$$

where $\theta \in \mathbf{\Theta}$ is the unknown parameter, $\mathbf{\Theta} \subseteq \mathbb{R}^d$ is the parameter space, $t(x) \in \mathbb{R}^d$ and $b(x) \geq 0$ are functions, and $\cdot$ denotes the usual inner product in $\mathbb{R}^d$:

$$\theta \cdot t(x) = (\theta_1, \dots, \theta_d) \cdot (t_1(x), \dots, t_d(x)) = \sum_{i=1}^{d} \theta_i t_i(x).$$

Then $c = c(\theta)$ depends on θ and

$$L(\theta; x) = \frac{1}{c(\theta)}\pi_\theta(x)$$

is the *likelihood function*.

EXAMPLE 2 (BAYES STATISTICS)
Bayesians use probability to express all types of uncertainty – replacing things like p-values and confidence intervals – assigning probabilities to all possible situations before looking at the data. So they treat θ as a random variable with some *prior density*

$$\theta \sim p(\theta)$$

while the (conditional) distribution of X (given θ) is specified by the likelihood. Hence (θ, X) has joint density

$$\pi(\theta, x) = L(\theta; x)p(\theta) = \frac{1}{c(\theta)}\pi_\theta(x)p(\theta).$$

Suppose we observe $X = x$ (the data). According to the Bayesian paradigm inference about θ should be based on the *posterior distribution* of θ given $X = x$ with (conditional) density

$$\pi(\theta \mid x) = \pi(\theta, x) \Big/ \int_\Theta \pi(\theta, x)\, d\theta \propto \pi(\theta, x)$$

(as x is fixed). So

$$\pi_u(x) = \pi(\theta, x)$$

is an *unnormalized posterior density*; its normalizing constant is often unknown (unless, as in 'classical Bayesian statistics', one uses 'conjugated priors').

REMARK 2
A more general point of view, which is also useful for 'missing data' problems: Joint densities lead to unnormalized conditional densities.

1.1.1 Simple Monte Carlo

Simple Monte Carlo methods use a sequence of random variables

$$X_1, X_2, \dots, X_n$$

all having distribution Π. The *Monte Carlo estimate*

$$\mathbb{E}_\pi g(x) \simeq \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

is unbiased.

Special cases:

(i)

$$g(x) = \mathbb{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

$$\Pi(A) = \mathbb{P}(X \in A) = \mathbb{E}_\pi \mathbb{1}_A(X) \simeq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_A(X_i)$$

so we can estimate probabilities.

(ii) We can also estimate quantiles of $g(X)$:

q'th quantile of $g(X) \simeq g^*$ so that $g^* \times n$ of the $g(X_1), \dots, g(X_n)$ are less than g^*.

We can exploit the *classical strong law of large numbers* (SLLN): Suppose

$$X_1, X_2, \dots \quad \text{IID} \sim \Pi.$$

Then

$$\overline{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_\pi g(X) \quad \text{a.s.} \tag{2}$$

('a.s.' reads as 'almost surely'; it means that the convergence in (2) holds with probability 1).

If $\sigma_g^2 = \mathrm{Var}_\pi(g(X))$ exists (i.e. $\sigma_g^2 < \infty$) then by the *classical central limit theorem* (CLT),

$$\overline{g}_n \underset{n \text{ large}}{\simeq} \mathcal{N}(\mathbb{E}_\pi g(X), \sigma_g^2/n)$$

or more precisely

$$\sqrt{n}\,\frac{\overline{g}_n - \mathbb{E}_\pi g(X)}{\sqrt{\sigma_g^2}} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1) \tag{3}$$

(meaning convergence in distribution).

Some methods for generating IID samples:

A) **Inversion:**

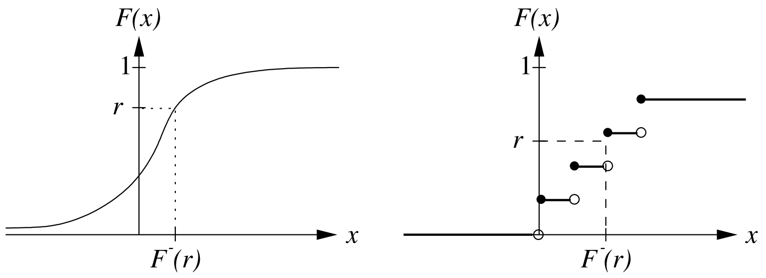
Suppose X is one-dimensional with c.d.f.

$$F(x) = \mathbb{P}(X \le x), \quad x \in \mathbb{R}.$$

Define the inverse

$$F^{-}(r) = \min\{s : F(s) \ge r\}, \quad 0 \le r \le 1.$$

Then



$$F^{-}(U) \sim \Pi \quad \text{if } U \sim \text{Uniform}[0,1]$$

So generating

$$U_1, U_2, \ldots \quad \text{IID} \sim \text{Uniform}[0,1]$$

then

$$X_1 = F^{-}(U_1), X_2 = F^{-}(U_2), \ldots \quad \text{IID} \sim \Pi.$$

PROBLEM:

This technique is essentially 1-dimensional. It is mainly used for discrete distributions. There are alternative methods - in particular for continuous distributions; see, for example, Ripley (1987).

B) **Sequential simulation:**

Suppose $X = (X^{(1)}, \ldots, X^{(d)})$ is multidimensional with blocks (e.g. coordinates) $X^{(1)}, \ldots, X^{(d)}$. Generate first $X^{(1)}$ from its marginal distribution, then $X^{(2)}$ from its conditional distribution given $X^{(1)}, \ldots$, finally $X^{(d)}$ from its conditional distribution given $(X^{(1)}, \ldots, X^{(d-1)})$. Repeat this to obtain IID samples from Π.

PROBLEM:

We often don't know this marginal and these conditional distributions. One exception is the Gaussian case (but here many other methods like Choleski decomposition and fast Fourier transform of circulant embeddings of covariance matrices may do better. . .)

REMARK 3

Recall Remarks 1 and 2: we wish to find simulation procedures not depending on unknown normalizing constants.

C) **Rejection sampling:**

Consider again an unnormalized target density $\pi_u(x)$. Suppose there is another unnormalized density $f_u(x)$ so that it is easier to simulate from f_u and

$$\exists\, K > 0 \quad \forall\, x : \ \pi_u(x) \le K f_u(x).$$

ALGORITHM 1 (REJECTION)

a) draw $U \sim$ Uniform$[0,1]$ and independently $X \sim f_u$,

b) return X if $U \le \dfrac{\pi_u(X)}{K f_u(X)}$ (setting $0/0 = 0$),

repeat (a)-(b) otherwise (using a new (U, X) independently drawn from the previous ones).

The output follows Π, and so repeating the algorithm we obtain IID simulations from Π. Note that we do not need know any normalizing constants.

PROBLEM:

This can be extremely slow if $K f_u$ is not close enough to π_u. For high dimensional densities it is usually hard to find a good 'enveloping function' f_u (but for low dimensional log-concave densities one can use a technique called adaptive rejection sampling. . .).

REMARK 4

A), B), C), and many other 'direct simulation procedures' are not feasible for our purpose (simulating 'complicated distributions'). A fourth Monte Carlo method is importance sampling, *which uses variables generated from another distribution, reweighted to give estimates with respect to Π. However, this typically only works (for our purpose) if it is combined with MCMC.*

1.1.2 Markov chain Monte Carlo (MCMC) methods

MCMC methods enables us to draw a sequence $X_0, X_1, \ldots$ of random variables that converges to Π. These variables need not to be (and are typically not) neither independent nor identically distributed. But they satisfies a Markov property: $\{X_n\}$ is a *time-homogeneous Markov chain*, that is the conditional distribution

$$\mathcal{D}(X_{n+1} \,|\, X_0, \ldots, X_n) = \mathcal{D}(X_{n+1} \,|\, X_n) \qquad (n \in \mathbb{N}_0) \tag{4}$$

depends not on $X_0, \ldots, X_{n-1}$ ('the future depends only on the present and not on the past'), and

$$\mathcal{D}(X_1 \,|\, X_0), \mathcal{D}(X_2 \,|\, X_1), \ldots \quad \text{are all identical conditional distributions} \tag{5}$$

This Markov chain can be defined in many ways, but it must leave Π invariant and be ergodic — adding further conditions we obtain a SLLN and a CLT as discussed later.

1.2 Statistical applications of Monte Carlo

- p-values for *hypothesis tests* (e.g. Besag and Clifford (1989, 1991)).
- Imputing missing data (*prediction*): What are the likely values of variables that we could not or did not observe ?
- *Maximum likelihood estimation*, often with missing data.
- *Bayesian inference* using MCMC has grown enormously over the last then years.

A few examples of these applications will be sketched in the lectures. They would all take too long time to explain in details (references are given below). Simple (but less interesting) examples will follow later. Heidi Hundborg will treat a larger example in detail.

The examples presented in the lectures are the following.

- Weed example: prediction in time-space multitype point processes modelled by log Gaussian Cox processes (Brix and Møller, 1998).
- Agricultural weed experiments and Bayesians analysis (Besag and Higdon, 1999).
- Confocal data and time series of ultrasound images: Bayesian object recognition using template models (Rue and Hurn, 1997, Hansen, Møller, Sloth and Tøgersen, 1999).
- Forestry example: likelihood inference for inhomogeneous point processes with missing data (Baddeley, Møller and Waagepetersen, 1998).

1.3 Further reading

Apart from the references given elsewhere in this text I would like to mention a few others. Much of the mathematics behind Markov chains and stability is well-covered in the book by Meyn and Tweedie (1993). See also Tierney (1994). Statistical aspects of MCMC are treated in the survey papers by Besag and Green (1993), Besag, Geeen, Higdon and Mengersen (1995) and Smith and Roberts (1993). Practical MCMC is discussed in the contributions collected in Gilks, Richardson and Spiegelhalter (1996).

When preparing these lectures I have also been inspired by a forthcoming book by Gareth Roberts and Richard Tweedie (easier to access than Meyn and Tweedie (1993)) and a paper by Peter Green ('A primer on MCMC'), which is submitted for publication.

Note that many preprints on MCMC are available from the MCMC Preprint Service at

http://www.maths.surrey.ac.uk/personal/st/S.Brooks/MCMC/

2 Markov chains: Part 1

Let $X_0, X_1, \ldots$ be a time-homogeneous Markov chain. Recall the conditions (4) and (5):

$$\begin{aligned}\mathbb{P}(X_{n+1} \in A | X_n = x, X_{n-1}, \ldots, X_0) &= \mathbb{P}(X_{n+1} \in A | X_n = x) \\ &= \mathbb{P}(X_1 \in A | X_0 = x)\end{aligned}$$

for all events A, states x, and times $n \in \mathbb{N}_0$. Hence the *transition kernel* defined by

$$P(x, A) = \mathbb{P}(X_1 \in A \,|\, X_0 = x)$$

and the *initial distribution* $\mu = \mathcal{D}(X_0)$ determines the distribution of $(X_0, \dots, X_n)$ for any time n:

$$P(X_0 \in A_0, X_1 \in A_1) = \int_{A_0} P(x_0, A_1)\,\mu(dx_0),$$
$$P(X_0 \in A_0, X_1 \in A_1, X_2 \in A_2) = \int_{A_0} \Big[\int_{A_1} P(x_1, A_2)\,\mu(dx_1) \Big]\,\mu(dx_0),$$

and so on. For example, in the discrete state space case,

$$\int_A h(y) P(x, dy) = \sum_{y \in A} h(y) P(x, \{y\}),$$

while in the continuous case where $\mathcal{D}(X_1 \,|\, X_0 = x)$ has a conditional density $f(x_1 \,|\, x_0)$,

$$\int_A h(y) P(x, dy) = \int_A h(y) f(y \,|\, x)\, dy.$$

REMARK 5
Much more complicated state spaces than discrete (such as $\{0, 1, \dots\}$) or continuous (such as $\mathbb{R}^d$) may be of relevance. For example, for finite point processes *defined on a bounded region $S \subset \mathbb{R}^2$, the state space is*

$$\Omega = \{x \subset S : n(x) < \infty\} \tag{6}$$

where $n(x)$ denotes the cardinality of x. Sets $x \in \Omega$ represents point patterns of e.g. trees, weeds or nests (see Figure 1 on page 14).

REMARK 6
We'll construct $P = P(\cdot, \cdot)$ so that no matter which initial distribution we use,

$$X_n \xrightarrow[n \to \infty]{\mathcal{D}} \Pi$$

(where Π is out target distribution). That is, the n step probabilities

$$P^n(x, A) = P(X_n \in A \,|\, X_0 = x)$$

satisfies

$$P^n(x, A) \underset{n \to \infty}{\longrightarrow} \Pi(A). \tag{7}$$

EXAMPLE 3
In the discrete case, (7) implies that $\pi(y) = \Pi(\{y\})$ must satisfy

$$\pi(y) = \lim_{n \to \infty} P^n(x, \{y\}) = \lim_{n \to \infty} \sum_z P^{n-1}(x, \{z\}) P(z, \{y\})$$
$$= \sum_z \lim_{n \to \infty} P^{n-1}(x, \{z\}) P(z, \{y\}) = \sum_z \pi(z) P(z, \{y\}),$$

i.e.

$$\pi(y) = \sum_z \pi(z) P(z, \{y\}).$$

Hence

$$\Pi(A) = \sum_{y \in A} \pi(y) = \sum_z \pi(z) P(z, A).$$

This motivates the next definition.

DEFINITION 1
A distribution Π *is* invariant *or* stationary *for* P *if, for any event* A,

$$\Pi(A) = \int P(x, A)\,\pi(dx).$$

REMARK 7
A stationary distribution is also called an equilibrium distribution *for* P*, since*

$$X_0 \sim \Pi \quad \Rightarrow \quad X_1 \sim \Pi \quad \Rightarrow \quad X_2 \sim \Pi \quad \Rightarrow \quad \ldots$$

Example 3 shows that we must construct P *so that* Π *(our target distribution) is invariant for* P.

There is a case where it is trivial to verify invariance:

DEFINITION 2
P *is* reversible w.r.t. Π *if*

$$(\Pi \times P)(A \times B) = (\Pi \times P)(B \times A) \tag{8}$$

for all events A *and* B*, where*

$$(\Pi \times P)(A \times B) = \int_A P(x, B)\,\pi(dx)$$

(= joint distribution of (X_n, X_{n+1}) *if* $X_n \sim \Pi$*).*

THEOREM 1

Reversibility $\Rightarrow$ *invariance.*

EXAMPLE 3 (CONTINUED)
In the discrete case,

$$(\Pi \times P)(A \times B) = \sum_{x \in A} P(x, B)\pi(dx),$$

so (8) is equivalent to the *detailed balance condition*

$$\pi(x)P(x, \{y\}) = \pi(y)P(y, \{x\}). \tag{9}$$

REMARK 8
Reversibility can equivalently be expressed as follows. If $X_n \sim \Pi$ *and* $X_{n+1} \,|\, X_n = x \sim P(x, \cdot)$ *then* $\mathcal{D}(X_n, X_{n+1}) = \mathcal{D}(X_{n+1}, X_n)$*. Combining this with the Markov property (4) we obtain that*

$$\mathcal{D}(X_0, X_1, \ldots, X_n) = \mathcal{D}(X_n, X_{n-1}, \ldots, X_0).$$

REMARK 9
Detailed balance conditions (DBCs) can be given for many other cases than just discrete state spaces. In general, DBC means that

$$\Pi(dx)P(x, dy) = \Pi(dy)P(y, dx).$$

This can usually be expressed in terms of an unnormalized target density. For instance, we can replace $\pi(\cdot)$ *with* $\pi_u(\cdot)$ *in (9).*

3 Algorithms

3.1 Gibbs sampler

This name was coined by Geman and Geman (1984). In statistical physics it is known as the *heat bath algorithm* (Creutz, 1979). Possibly it first appeared in Suomela's (1976) thesis.

EXAMPLE 4 (BAYESIAN ANALYSIS OF A NORMAL SAMPLE)
Consider the following Bayesian model.

Likelihood:

$$Y_1, \dots, Y_n \qquad \text{IID} \qquad \mathcal{N}(\mu, \sigma^2).$$

Prior:

$$\mu \sim \mathcal{N}(\xi, \kappa^{-1}) \quad \text{and} \quad \sigma^{-2} \sim \Gamma(\alpha, \beta)$$

independent (here ξ, κ, α, and β are known numbers; $\Gamma(\alpha, \beta)$ denotes the gamma distribution with mean α/β and variance α/β^2).

Posterior:

$$\pi(\mu, \sigma^{-2} \,|\, y) \propto (\sigma^{-2})^{\alpha+1+n/2} \exp\left\{ -\frac{\kappa(\mu-\xi)^2}{2} - \Big(\frac{\sum_{i=1}^n (y_i-\mu)^2}{2} + \beta\Big)\sigma^{-2} \right\}$$

if $(Y_1, \dots, Y_n) = y = (y_1, \dots, y_n)$ is observed.

Note that μ and σ^{-2} are independent a priori, but dependent a posteriori. However, the 'full conditionals' are easy to handle:

$$\text{(a)} \quad \mu \,|\, \sigma^{-2}, y \sim \mathcal{N}\Big(\frac{\sigma^{-2}\sum_{i=1}^n y_i + \kappa\xi}{\sigma^{-2}n+\kappa}, \frac{1}{\sigma^{-2}n+\kappa}\Big)$$
$$\text{(b)} \quad \sigma^{-2} \,|\, \mu, y \sim \Gamma\Big(\alpha + n/2, \beta + \sum_{i=1}^n (y_i-\mu)^2/2\Big)$$

By alternately drawing μ and σ^{-2} from the full conditionals we obtain a Markov chain.

This was a simple example of a Gibbs sampler with cyclic updating and two components. In general we consider $k \geq 1$ components

$$X = (X^{(1)}, \dots, X^{(k)})$$

and k full conditionals

$$X^{(i)} \,|\, X^{(-i)} = x^{(-i)}, \qquad i = 1, \dots, k, \tag{10}$$

where

$$X^{(-i)} = (X^{(1)}, \dots, X^{(i-1)}, X^{(i+1)}, \dots, X^{(k)}).$$

Let $P^{(i)}$ denote the transition kernel given by (10) when $X \sim \Pi$, i.e. for $x^{(j)} \in A_j, j \neq i$:

$$P^{(i)}(x, A_1 \times \dots \times A_k) = G_i(x^{(-i)}, A_i) = \Pi(X^{(i)} \in A_i \,|\, X^{(-i)} = x^{(-i)})$$

depends only on $x^{(-i)}$ and A_i. In Example 4, $k=2, X^{(1)}=\mu, X^{(2)}=\sigma^{-2}$, (a) specifies G_1, (b) specifies G_2, and the transition kernel $P=P^{(1)}P^{(2)}$ is a *product kernel*. In general we define $P^{(1)}P^{(2)}$ by

$$(P^{(1)}P^{(2)})(x,A)=\int P^{(2)}(y,A)P^{(1)}(x,dy)$$

so in the special case of Example 4,

$$P((\mu,\sigma^{-2}),A_1\times A_2)=\int_{A_1} G_2(\mu',A_2)G_1(\sigma^{-2},d\mu')$$

does not depend on μ. Similarly, by induction we define the product kernel

$$P^{(1)}P^{(2)}\cdots P^{(k)}$$

corresponding to *cyclic Gibbs sampling* where we first update the first component given the rest, then the second component given the rest (which may differ from the previous 'rest'), and so on.

Suppose that

$$\Pi \text{ has density } \pi(x^{(1)},\ldots,x^{(k)}) \text{ w.r.t. a product measure } \nu(dx)=\nu_1(dx^{(1)})\times\cdots\times\nu_k(x^{(k)}) \qquad (11)$$

and

$$G_i(x^{(-i)},A_i)=\int_{A_i}\pi(x^{(i)}\,|\,x^{(-i)})\,\nu_i(dx^{(i)}),\qquad i=1,\ldots,k.$$

(In Example 4, $\pi(x^{(1)}\,|\,x^{(2)})$ is a normal density and $\pi(x^{(2)}\,|\,x^{(1)})$ is a Gamma density). Then detailed balance holds for each update of components, since

$$\pi(x)\pi(y^{(i)}\,|\,x^{(-i)})\mathbb{1}[x^{(-i)}=y^{(-i)}]$$

is seen to be symmetric in x and y. Thereby we obtain reversibility at each update of components:

$$(\Pi\times P^{(i)})(A\times B)=(\Pi\times P^{(i)})(B\times A),$$

and so invariance of Π after a cycle.

REMARK 10
Cyclic Gibbs sampling leaves Π invariant, but it is in general not reversible.

Cyclic Gibbs sampling is one example of a *systematic updating scheme*. A *reversible systematic Gibbs sampler* is e.g. provided by doing a forwards-backwards scan corresponding to the product kernel

$$P^{(1)}P^{(2)}\cdots P^{(k)}P^{(k)}\cdots P^{(2)}P^{(k)}.$$

Another reversible algorithm is a *random updating Gibbs sampler*, where we select the i'th component with probability $q_i>0$ so that

$$P=\sum_{i=1}^{k}q_iP^{(i)}$$

is a *mixture kernel* (usually, $q_i=1/k$ is uniform).

REMARK 11
Similar statements apply for mixture and product kernels obtained by using other algorithms than Gibbs sampling.

REMARK 12
All we need is invariance, not reversibility (though in practice invariance is usually obtained by 'reversibility at each update of components'). However, reversibility may be easier to use for 'backwards constructions'.

3.2 Metropolis et al. (1953) algorithm

Let the situation be as in Section 3.1, but the i'th component $X^{(i)}$ given the rest $X^{(-i)}$ is now updated by

a) a *proposal* y drawn from
$$p_i(x, y) = \mathbb{1}[y^{(-i)} = x^{(-i)}] q_i(y^{(i)}; x)$$
where $q_i(\cdot; x)$ is a density w.r.t. ν_i in (11) so that
$$p_i(x, y) = p_i(y, x) \qquad \text{(for all } x, y\text{);} \tag{12}$$

b) return y with probability
$$\alpha_i(x, y) = \min\{1, \frac{\pi(y)}{\pi(x)}\} \tag{13}$$
else return x.

3.3 Hastings (1970) algorithm (Metropolis-Hastings sampler)

As above but without the symmetry condition (12), and extending the definition of the *acceptance probability* in b) to
$$\alpha_i(x, y) = \min\{1, H_i(x, y)\}$$
where
$$H_i(x, y) = \frac{\pi(y) p_i(y, x)}{\pi(x) p_i(x, y)} \tag{14}$$
is the *Hastings ratio* (setting $0/0 = 0$). As in Section 3.1 we can verify that detailed balance holds for each update of a component, so that Π is invariant for the Metropolis-Hastings algorithm. If we are using a reversible systematic updating scheme or a random updating scheme, then the Metropolis-Hastings algorithm is reversible w.r.t. Π (see Remark 11).

Clearly, (12) and (14) imply (13).

The Gibbs sampler is the special case where each $q_i(y^{(i)}; x) = \pi(y^{(i)} \mid x^{(-i)})$is the full conditional, and so $\alpha_i = 1$.

Note that we do not need to know any normalizing constants, since
$$\begin{aligned}\pi(y)/\pi(x) &= \pi_u(y)/\pi_u(x) \\ &= \pi(y^{(i)} \mid y^{(-i)})/\pi(x^{(i)} \mid x^{(-i)}) \\ &= \pi_u(y^{(i)} \mid y^{(-i)})/\pi_u(x^{(i)} \mid x^{(-i)})\end{aligned}$$
as $x^{(-i)} = y^{(-i)}$.

Compared to rejection sampling we have now a Markov chain, not an independent sequence; we do not repeat, if the proposal is rejected; and we do not need a bound on $\pi(y)/p_i(x, y)$.

3.4 How to choose the blocking and proposal distribution

This depends of course on the particular application. Some rules:

- It is usually a good idea to explore some conditional independence properties and use small components.
- On the other hand it is sometimes better to update many coordinates at once (the extreme case being $k = 1$ components). See the MALA (Section 3.5.4 on page 13).
- Use different types of updates for different kinds of full conditionals: Use a Gibbs update if the full conditional is easy to sample from (i.e. when it is a known distribution or (adaptive) rejection sampling is feasible) — *but Gibbs updates are not always easy or the right thing to do* — so use Metropolis-Hastings type updates for others (see Section 3.5.1).
- If the coordinates in $X^{(i)}$ given $X^{(-i)}$ are independent, then they can be updated in parallel.

3.5 Common choices for proposal distributions

3.5.1 Independence Metropolis-Hastings

Here

$$q_i(y^{(i)}; x) = q_i(y^{(i)}) \qquad \text{does not depend on } x.$$

This algorithm is of little use in practice (but it is useful for theoretical considerations).

3.5.2 Random walk Metropolis

This is a Metropolis algorithm as we assume

$$q_i(y^{(i)}; x) = q_i(y^{(i)} - x^{(i)}) = q_i(x^{(i)} - y^{(i)}).$$

The proposal amounts to adding a random walk increment to the current $x^{(i)}$. For example, $q_i(\cdot)$ could be a normal density with mean 0.

3.5.3 Random walk Metropolis-Hastings on the log scale

This is used for one-dimensional components which are necessarily positive. A multiplicative rather than additive update is proposed: If, conditional on current x, $\log y^{(i)} - \log x^{(i)}$ has a symmetric density around 0, then

$$H_i(x, y) = \frac{\pi(y) y^{(i)}}{\pi(x) x^{(i)}}.$$

3.5.4 Metropolis-adjusted Langevin algorithm (MALA)

This was introduced in statistics by Besag (1994) (he called it the Langevin-Hastings algorithm) and later studied in detail by Roberts and Tweedie (1996) among others. It is mainly used for higher-dimensional continuous components; so let us say that $k = 1$ and the state space is $\mathbb{R}^d$ in the sequel. Letting

$$\nabla(x) = \frac{\partial}{\partial x} \log \pi_u(x)$$

the proposal y given the current x is generated by

$$y = x + \frac{h}{2}\nabla(x) + \sqrt{h}z \tag{15}$$

where the d coordinates in z are IID $\mathcal{N}(0,1)$ variates and $h > 0$ is a user specified parameter. (Equation (15) is inspired by the time-discretization of the stochastic differential equation of a Langevin diffusion). So the Hastings ratio becomes

$$H(x,y) = \frac{\pi(y)}{\pi(x)} \frac{\exp(-\frac{1}{2}\|x - y - \frac{h}{2}\nabla(y)\|^2)}{\exp(-\frac{1}{2}\|y - x - \frac{h}{2}\nabla(x)\|^2)}.$$

The optimal value of h seems to be obtained if the acceptance rate is about 0.574. The use of gradient information may lead to much better algorithms than e.g. a random walk Metropolis algorithm (Møller, Syversveen and Waagepetersen, 1998; Hansen, Møller, Sloth and Tøgersen, 1999). If $\pi_u(\cdot)$ has too light tails, one may need to truncate in the gradient (Roberts and Rosenthal, 1998).

3.6 Metropolis-Hastings algorithm for point processes

In order to discuss a situation with a state space of varying dimension, consider (as in Remark 5) a bounded region $S \subset \mathbb{R}^2$ of area $|S| > 0$. The state space is

$$\Omega = \cup_{n=0}^{\infty} \Omega_n \qquad \text{with } \Omega_n = \{x \subset S : n(x) = n\}$$

and $\pi(x)$ is assumed to be a density w.r.t. the unit rate Poisson process ν on S (thus under ν, the number of points is Poisson($\nu|S|$) and the positions of points are IID with a uniform distribution on S).

EXAMPLE 5

$$\pi_u(x) = \beta^{n(x)}\gamma^{s(x)} \tag{16}$$

where

$$s(x) = \sum_{i<j} \mathbb{1}[\|x_i - x_j\| \le R] \qquad \text{for } x = \{x_1, \ldots, x_n\}$$

and $\beta > 0$, $0 \le \gamma \le 1$, $R > 0$ are parameters (setting $0^0 = 1$). For $\gamma = 0$ we have a hard core process. If $\gamma = 1$ we obtain a Poisson process of rate β on S. For $0 \le \gamma \le 1$ we observe an increasing regularity in simulated point patterns as γ decreases, see Figure 1 below.

ALGORITHM 2 (GEYER AND MØLLER (1994), GEYER (1999) AND MØLLER (1999))
Suppose x is the current state:

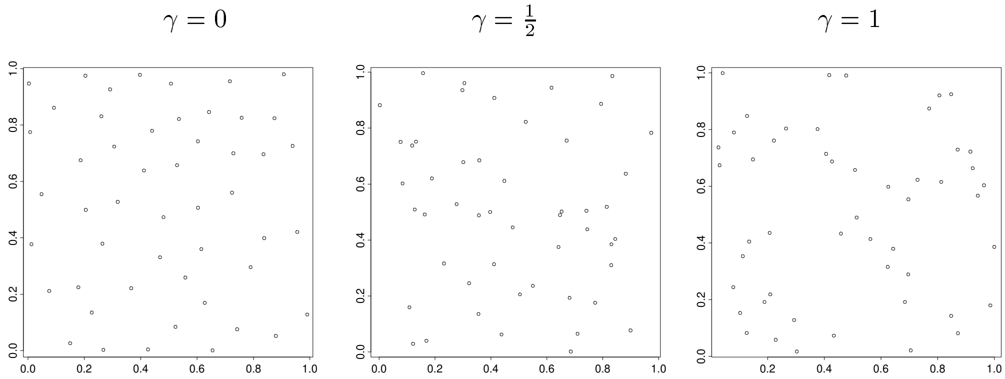


Figure 1: Simulations of the Strauss process with 50 points in the unit square, $R = 0.1$ and for different values of the interaction parameter, γ.

With probability $1/2$ *do a birth proposal:*

generate $\xi \sim$ *Uniform(S)*

return $y = x \cup \{\xi\}$ *with probability* $\min\{1, r(x,\xi)\}$

else return $y = x$*;*

else do a death proposal

if $x \neq \emptyset$

then generate $\eta \sim$ *FiniteUniform*(x)

return $y = x \setminus \{\eta\}$ *with probability* $\min\{1, r(x,\eta)^{-1}\}$

else return $y = x$.

else return $y = \emptyset$.

Here

$$r(x,\xi) = \begin{cases} \dfrac{\pi(x \cup \{\xi\})|S|}{\pi(x)(n(x)+1)} & \text{if } \pi(x) > 0 \\ 0 & \text{else.} \end{cases}$$

For example, for the Strauss process,

$$r(x,\xi) = \beta\gamma^{\sum_{\eta\in x} 1\!\!1[\|\eta-\xi\|\leq R]} \frac{|S|}{n(x)+1}.$$

The algorithm is a particular simple example of Green's (1995) *reversible MCMC jump algorithm.* However, it is simpler to show that a detailed balance condition holds directly:

$$\frac{1}{2}\frac{1}{|S|}\min\{1, r(x,\xi)\}\pi(x) = \frac{1}{2}\frac{1}{n(x\cup\{\xi\})}\min\{1, r(x,\xi)^{-1}\}\pi(x\cup\{\xi\})$$

provided

$$\pi(x\cup\{\xi\}) > 0 \Rightarrow \pi(x) > 0 \tag{17}$$

Condition (17) is a so-called *hereditary condition* which we assume is satisfied. From the detailed balance condition one obtains reversibility w.r.t. π.

In the sequel we assume (17) to hold and redefine the state space as

$$\Omega = \{x \subset S : n(x) < \infty, \pi(x) > 0\}.$$

Our Metropolis-Hastings sampler remains within Ω whenever $X_0 = x$ is started in $x \in \Omega$.

4 Markov chains: Part 2

Consider again a time-homogeneous Markov chain $\{X_n\}_{n=0}^{\infty}$ with transition kernel P and invariant distribution Π. Let Ω denote the state space of each X_n. Formally Ω is equipped with a σ-field $\mathcal{F}$ containing all the 'interesting' events. We assume that $\mathcal{F}$ is countable generated; if Ω is finite or at least countable (the discrete case) then $\mathcal{F} = \{A : A \subseteq \Omega\}$; if Ω admits a topology then $\mathcal{F}$ is the Borel σ-field.

In the sequel we investigate the *stability* properties of the Markov chain under various conditions. Recall, that we want at least that

$$X_n \xrightarrow[n\to\infty]{\mathcal{D}} \Pi,$$

that is by (7),

$$P^n(x, A) \underset{n\to\infty}{\longrightarrow} \Pi(A) \quad \text{for all } x \in \Omega, A \in \mathcal{F} \tag{18}$$

where $P^n(x, A) = \mathbb{P}(X_n \in A \mid X_0 = x)$ is the n-step transition kernel.

4.1 Irreducibility

For (18) to hold we need at least the following property.

DEFINITION 3
*The Markov chain $\{X_n\}$ is φ-*irreducible *if there exists a non-zero measure φ on $\mathcal{F}$ such that for all $x \in \Omega$ and $A \in \mathcal{F}$ with $\varphi(A) > 0$,*

$$\mathbb{P}\big(\cup_{n=1}^{\infty} \{X_n \in A\} \mid X_0 = x\big) > 0.$$

In other words

$$\forall\, x \in \Omega \quad \forall A \in \mathcal{F} \text{ with } \varphi(A) > 0 \quad \exists\, n \in \mathbb{N} :\ P^n(x, A) > 0. \tag{19}$$

PROPOSITION 1
If $\{X_n\}$ is φ-irreducible, then

(i) Π *is the only invariant measure for P,*

(ii) $\{X_n\}$ *is* Π*-irreducible,*

(iii) $\Pi(A) = 0 \ \Rightarrow\ \varphi(A) = 0.$

REMARK 13
We have earlier noticed that (18) implies that Π is invariant, so (i) shows that no other measure than Π can be a limiting measure for the Markov chain. By (ii), φ-irreducibility implies Π-irreducibility and by (iii), Π dominates φ. The point in using another measure φ than Π is simply that we can consider fewer events A in order to check (19) (see Example 5 below). By irreducibility *we simply mean Π-irreducibility.*

EXAMPLE 5 (CONTINUED)
Take

$$\varphi(A) = \mathbb{1}[\emptyset \in A]$$

where $\emptyset$ denotes the empty point configuration. Let $x = \{x_1, \dots, x_n\} \in \Omega$ and $A = \{\emptyset\}$. For $\xi \in S \setminus x$,

$$r(x, \xi) = \beta\gamma^{\sum_{i=1}^{n} \mathbb{1}[\|x_i - \xi\| \leq R]} \frac{|S|}{n+1} \leq \beta \frac{|S|}{n+1}$$

so

$$\prod_{i=0}^{n-1} r(\{x_1, \dots, x_i\}, x_{i+1})^{-1} \geq \prod_{i=0}^{n-1} \frac{i+1}{\beta|S|}.$$

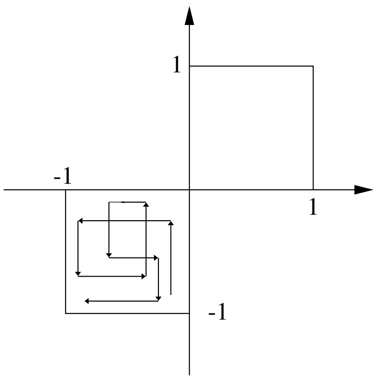
Hence

$$\begin{aligned} P^n(x, A) &\geq \prod_{j=1}^{n} P\Big(\{x_1, \dots, x_j\}, \{\{x_1, \dots, x_{j-1}\}\}\Big) \\ &= \prod_{i=0}^{n-1} \frac{1}{2} \min\{1, r(\{x_1, \dots, x_i\}, x_{i+1})^{-1}\} \\ &\geq \prod_{i=0}^{n-1} \frac{i+1}{2\beta|S|} \\ &> 0. \end{aligned}$$

so the chain is φ-irreducible.

REMARK 14
In general one has to check for irreducibility in each separate case of a MCMC algorithm. For the Gibbs sampler, irreducibility is closely related to connectness of Ω as indicated below.

EXAMPLE 6
The Gibbs sampler for $\Pi = \text{Uniform}([-1, 0]^2 \cup [0, 1]^2)$ is reducible (i.e. not irreducible) as shown in the figure.



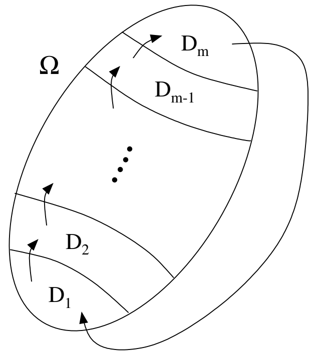
4.2 Aperiodicity

Apart from irreducibility, (18) implies aperiodicity as defined below.

DEFINITION 4
Suppose the Markov chain is irreducible. We call $D_1, \dots, D_m \in \mathcal{F}$ *an* m*-cycle if*

$$\begin{array}{ccc} x \in D_1 & \Rightarrow & P(x, D_2) = 1 \\ x \in D_2 & \Rightarrow & P(x, D_3) = 1 \\ & \vdots & \\ x \in D_{m-1} & \Rightarrow & P(x, D_m) = 1 \\ x \in D_m & \Rightarrow & P(x, D_1) = 1 \end{array}$$

and $\Pi(\cup_{i=1}^m D_i) = 1$*. The largest* $d \in \mathbb{N}$ *for which a* d*-cycle occurs is called the* period *of the Markov chain. When* $d = 1$*, the Markov chain is* aperiodic*.*



EXAMPLE 7
If $P(x, \{x\}) > 0$ for some $x \in \Omega$, we say the chain is *strongly aperiodic*. For example, for the Metropolis-Hastings sampler for point processes

$$P(\emptyset, \{\emptyset\}) \geq \frac{1}{2}.$$

The periodic case can essentially be reduced to the aperiodic case:

PROPOSITION 2
Suppose $\{X_n\}$ *is irreducible with period* d *and* d*-cycle* $D_1, \dots, D_d$*. For any* $i \in \{1, \dots, d\}$*, the Markov chain with state space* D_i *and transition kernel* $P_i = P^d = P \cdots P$ *(d times) is* Π_i*-irreducible and aperiodic, where*

$$\Pi_i(A) = \Pi(A)/\Pi(D_i) \quad \forall\, A \in \mathcal{F} \text{ with } A \subseteq D_i$$

is the normalized restriction of Π *to* D_i*.*

4.3 Ergodicity

The following is a somewhat stronger result than (18).

THEOREM 2
Irreducibility and aperiodicity is equivalent to the existence of a Π-null set N (i.e. $N \in \mathcal{F}$ and $\Pi(N) = 0$) so that

$$\|P^n(x,\cdot) - \Pi\| \underset{n\to\infty}{\longrightarrow} 0 \qquad \text{for all } x \in \Omega \setminus N \tag{20}$$

where

$$\|P^n(x,\cdot) - \Pi\| = \sup_{A\in\mathcal{F}} |P^n(x,\cdot) - \Pi|$$

is the total variation norm *of $P^n(x,\cdot) - \Pi$.*

In the discrete case, $N = \emptyset$ is vanishing. In the general case, in order to get rid of the null set N in (20), we need the following concept.

DEFINITION 5
The Markov chain $\{X_n\}$ is Harris (recurrent) *if*

$$\forall\, x \in \Omega \quad \forall\, A \in \mathcal{F} \text{ with } \Pi(A) > 0 : \ \mathbb{P}\Big(\bigcup_{n=1}^{\infty} \{X_n \in A\} \,\Big|\, X_0 = x\Big) = 1.$$

It is ergodic *if it is both Harris and aperiodic.*

Clearly, Harris recurrence implies irreducibility. If Ω is finite, then irreducibility is the same as Harris recurrence.

THEOREM 3
Ergodicity is equivalent to that

$$\|P^n(x,\cdot) - \Pi\| \underset{n\to\infty}{\longrightarrow} 0 \qquad \forall\, x \in \Omega.$$

In fact, in order to establish a SLLN, we do not need aperiodicity. Let $L^1(\Pi)$ be the set of measurable functions $f : \Omega \to [-\infty, \infty]$ so that $\int |f(x)|\Pi(dx) < \infty$, and set

$$\mathbb{E}_\pi(f) = \int f(x)\,\Pi(dx) \qquad \text{for } f \in L^1(\Pi),$$
$$\overline{f}_n = \frac{1}{n}\sum_{k=1}^{n} f(X_k).$$

THEOREM 4 (STRONG LAW OF LARGE NUMBERS)
Harris recurrence is equivalent to that, for any function $f \in L^1(\Pi)$,

$$P\big(\overline{f}_n \underset{n\to\infty}{\longrightarrow} \mathbb{E}_\pi(f)\big) = 1.$$

How do we check for Harris recurrence ? For this the concept of small sets is useful.

DEFINITION 6
A set $C \in \mathcal{F}$ is small if $\exists\, m \in \mathbb{N}\, \exists\, \varepsilon > 0\, \exists\, \kappa$ probability measure on $\mathcal{F}$ $\forall\, A \in \mathcal{F}\, \forall x \in C$:

$$P^m(x, A) \geq \varepsilon\kappa(A).$$

REMARK 15
In the discrete case of an irreducible Markov chain, $\{x\}$ is small for any $x \in \Omega$. Typically, if Ω admits a topology, compact sets are small sets.

REMARK 16
Consider again the Metropolis-Hastings sampler for point processes. Assume local stability *in the sense that*

$$\exists\, K > 0 \quad \forall\, x \in \Omega \quad \forall\, \xi \in S \setminus x : \quad \pi(x \cup \{\xi\}) \leq K\pi(x). \tag{21}$$

Clearly, (21) implies the hereditary condition (17). The Strauss process satisfies (21) (with $K = \beta$), and as in Example 5 we can show that (21) implies φ-irreducibility. By Example 7, we have aperiodicity. Moreover, one can show that any set of the form $\{x \subset S : \ n(x) \leq m, \pi(x) > 0\}$ is small ($m \in \mathbb{N}_0$).

PROPOSITION 3
Suppose the Markov chain $\{X_n\}$ is irreducible and aperiodic, and there exists a small set $C \in \mathcal{F}$, a function $V : \Omega \to \mathbb{R}$ such that $\{x : V(x) \leq n\}$ is a small set for any $n \in \mathbb{N}$ and

$$\forall\, x \in \Omega \setminus C : \ \mathbb{E}[V(X_1) \,|\, X_0 = x] \leq V(x) \tag{22}$$

(this is called the drift condition for recurrence*). Then the chain is Harris.*

REMARK 16 (CONTINUED)
The drift condition of recurrence can be verified by taking

$$V(x) = M^{n(x)} \quad \textit{and} \quad C = \{x \subset S : \pi(x) > 0,\ n(x) \leq m\} \tag{23}$$

where $\max\{1, K|S|\} < M < \infty$ and $m \in \mathbb{N}$ is sufficiently large (Geyer, 1999, Proposition 3.3).

4.4 Geometric ergodicity

For functions $V : \Omega \to [0, \infty]$ we define the V*-norm* of $P^n(x, \cdot) - \Pi$ by

$$\|P^n(x, \cdot) - \Pi\|_V = \sup_{f \in L^1(\Pi):\ |f| \leq V} \Big| \int f(y) P^n(x, dy) - \mathbb{E}_\pi(f) \Big|.$$

In the special case where $V = 1$ the V-norm is just the total variation norm, while

$$\|P^n(x, \cdot) - \Pi\| \leq \|P^n(x, \cdot) - \Pi\|_V \quad \text{if } V \geq 1.$$

DEFINITION 7
The Markov chain $\{X_n\}$ is V-uniformly ergodic *if*

$$\exists\, V \in L^1(\Pi) \text{ with } V \geq 1 \quad \exists\, \varrho < 1 \quad \exists\, R < \infty \quad \forall\, n \in \mathbb{N} \quad \forall\, x \in \Omega : \ \|P^n(x, \cdot) - \Pi\|_V \leq RV(x)\varrho^n.$$

If V is finite, the chain is geometrically ergodic.

REMARK 17
Note that ϱ does not depend on x. If V is finite, V-uniform ergodicity implies ergodicity. Geometric ergodicity is a so-called 'qualitative property' and its importance is mainly that it implies a CLT to hold (see below).

Qualitative bounds for convergence rates are established by Rosenthal (1993) and others, but so far this seems to be of rather limited use (see, e.g., Møller, 1999, Appendix B).

Now let $g(X)$ be a statistic of interest so that $g \in L^1(\Pi)$. Consider the *MCMC estimate* $\overline{g}_n$ of $\mu_g = \mathbb{E}_\pi(g)$ and assume that the *asymptotic variance*

$$\sigma_g^2 = \lim_{n\to\infty} n\,\mathrm{Var}_\pi(\overline{g}_n) \tag{24}$$

exists and is given by

$$\sigma_g^2 = \mathrm{Var}_\pi(g(X_0)) + 2\sum_{k=1}^{\infty} \mathrm{Cov}_\pi(g(X_0), g(X_k)) \tag{25}$$

(the subscript in Var_π and Cov_π means that variances and covariances are calculated in the case where X_0 follows Π).

THEOREM 5 (CENTRAL LIMIT THEOREM)
Under any of the conditions (C1), (C2), and (C3) we have that σ_g^2 exists and is given by (25), and if $\sigma_g^2 > 0$ then

$$\sqrt{n}\,\frac{\overline{g}_n - \mu_g}{\sqrt{\sigma_g^2}} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1).$$

(C1) $\int |g(x)|^{2+\varepsilon}\pi(dx) < \infty$ *for some* $\varepsilon > 0$ *and the chain is geometrically ergodic (Chan and Geyer, 1994).*

(C2) $\int |g(x)|^2\pi(dx) < \infty$ *and the chain is geometrically ergodic and reversible (Roberts and Rosenthal, 1997).*

(C3) *The geometric drift condition below is satisfied for a finite function V and $g^2 \le V$ (Meyn and Tweedie, 1993).*

The *geometric drift condition* states that

$$\exists\, C \in \mathcal{F} \text{ small set } \exists\, b < \infty\ \exists\, \lambda < 1\ \exists\, V : \Omega \to [1,\infty] \text{ finite at some } x_0 \in \Omega\ \forall\, x \in \Omega :$$

$$\mathbb{E}[V(X_1) \mid X_0 = x] \le \lambda V(x) + b\mathbb{1}[x \in C].$$

THEOREM 6
The geometric drift condition implies V-uniform ergodicity.

REMARK 16 (CONTINUED)
The Metropolis-Hastings sampler for a locally stable point process is geometrically ergodic (Geyer and Møller, 1994; Geyer, 1999). In fact the geometric drift condition is satisfied with V and C given by (23).

4.5 Uniform ergodicity

This is the strongest ergodicity property we shall consider — it is equivalent to V-uniform ergodicity with $V = 1$.

DEFINITION 8
The Markov chain is uniformly ergodic *if*

$$\sup_{x \in \Omega} \|P^n(x, \cdot) - \Pi\| \underset{n \to \infty}{\longrightarrow} 0.$$

THEOREM 7
Uniform ergodicity holds if and only if the state space Ω is small.

REMARK 18
By Definition 6, if Ω is small, $P^m(x, A) \geq \varepsilon\kappa(A)$ where $m \in \mathbb{N}$ and $0 < \varepsilon < 1$ are constants and κ is a probability measure. Then, it can be shown that

$$\|P^n(x, \cdot) - \Pi\| \leq (1 - \varepsilon)^{n/m}$$

REMARK 19
For a finite state space, uniform ergodicity is just the same as irreducibility and aperiodicity. Typically, compactness of Ω and certain continuity conditions give uniform ergodicity. However, in many applications, only geometric and not uniform ergodicity can be established.

4.6 What is the use of all this theory ?

Qualitative results like geometric ergodicity or at least Harris recurrence tell us something. In the latter case it is meaningful to use ergodic averages as Monte Carlo estimates due to the SLLN, and in the former case a CLT applies so that Monte Carlo errors can be calculated (we return to this later). But apart from this, they do not tell us much. Statistical methods investigating the performance of MCMC methods, as discussed in the sequel, play therefore an important role.

5 Performance of MCMC methods

Two main issues to consider are

(i) **Convergence:** how quickly is the chain (effectively) in equilibrium ?

(ii) **Efficiency:** how well are mean values w.r.t. Π estimated from the chain ?

Other questions are

(iii) One long run or many short runs ?

(iv) Simplicity ?

We turn first to (iii). Using one long run of a Harris recurrent Markov chain, by the SLLN, the initial effect of the initial variables on the *ergodic average* $\overline{g}_n$ will be overcome. If we instead draw K independent chains $\{X_n^j\}_{n=1}^M$ each of length M, we know by the classical SLLN that

$$\frac{1}{K}\sum_{j=1}^{K} g(X_M^j) \simeq \mathbb{E}_\pi(g).$$

Neither of the two methods can be advocated as preferable to the other in general — usually a strategy somewhere in between these two works best in practice.

A commonly used method is to discard an initial '*burn-in*' part of the Markov chain sample path, computing the ergodic average using the remaining values — or perhaps using a subset of the remaining values obtained by a *spacing* m (i.e. each m'th value is used) where m is chosen to reduce the correlation of adjoining values (not necessarily a good idea!) or maybe m is used because of storage problems.

Regarding (iv), we can ask how convenient is a MCMC method to code reliably and to use ? Should we reuse existing code in order to save time ? Programming your own MCMC method from scratch is not always so time-consuming as you may think; it provides flexibility and run-time efficiency that cannot really be matched by package software. The code for a sampler may be written in e.g. C (to speed things up), while a high-level statistical language such as S may be used for statistical and graphical procedures. Note that it is absolutely essential to check and double-check MCMC code — it is not foolproof.

5.1 Monitoring convergence

Returning to the issue (i) of convergence, what is an appropriate burn-in ? Usually one compares *time series plots* of various statistics reflecting different aspects of target distribution; perhaps with chains started from a number of distinct starting values in order to assess the effect of the initial values by comparison.

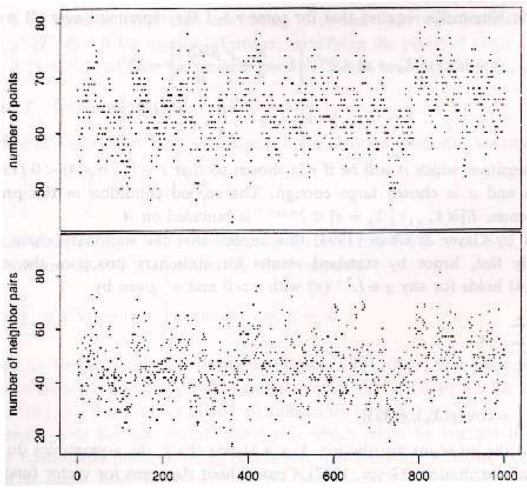
If we are just interested in finding e.g. maximum likelihood estimates, we may restrict attention to '*sufficient statistics*'. For example, in a Strauss process, $t_1(x) = n(x)$ and $t_2(x) = s(x)$ are the sufficient statistics and the MLE is the solution to the likelihood equation

$$\mathbb{E}_\theta(t_1) = n(x) \quad \text{and} \quad \mathbb{E}_\theta(t_2) = s(x),$$

with x being an observed point pattern, $\theta = (\beta, \gamma)$ the unknown parameter, and $\mathbb{E}_\theta$ denoting expectation w.r.t. the Strauss process.

There exists many other kinds of *convergence diagnostics*, see e.g. Brooks and Gelman (1998). Suppose, for example, that we know $\mu = \mathbb{E}_\pi(t)$ for some statistic t — one example could be $t(x) = \frac{\partial}{\partial x}\log \pi_u(x)$ and $\Omega \subseteq \mathbb{R}^d$ open, where under fairly general conditions $\mathbb{E}_\pi(t) = 0$ (this is of course something we have to check for specific models). Consider K independent sample paths $\{X_n^j\}_{n\in\mathbb{N}_0}, j = 1, \ldots, K$ (i.e. K independent realizations of our Markov chain). By the classical CLT, provided $\sigma^2 = \mathrm{Var}_\pi(t)$ exists and K is large,

$$\frac{1}{K}\sum_{j=1}^{K} t(X_n^j) \simeq \mathcal{N}(\mu, \sigma^2/K)$$



Iterations

Figure 2: Time series for the canonical statistics $n(x)$ and $s(x)$ for an unconditional Strauss process on the unit square with canonical parameters $\theta_1 = \log(100)$ and $\theta_2 = \log(0.75)$ and interaction radius $r = 0.1$. The Metropolis algorithm had a warm-up of 200 iterations followed by a run of 1000 iterations. Each comprised 200 basic update steps. Figure taken from Geyer and Møller (1994).

if the chain is (effectively) in equilibrium at time t. Since μ is known and σ^2 can be estimated, a plot of $\frac{1}{K}\sum_{j=1}^{K} t(X_n^j)$ versus n, with, for example, 95%-envelopes around μ indicated, would show us something about how large a burn-in we need (at least with respect to that particular statistic t).

A problem with convergence diagnostics is of course that we can never be absolutely sure if we use a long enough burn-in. If Π has two or more high probability regions which are separated by low probability regions, the Markov chain can be very slowly converging and apparently convergence (as observed from time series plots or similar) may just be caused by *meta-stability*. This can be related to what physicists call '*critical slowing down*', where a solution could be to use a *heating strategy* like in simulated tempering (see later). Recently exact/perfect MCMC methods that give guaranteed convergence have been developed (we return also to this later). These can be rather time consuming to implement. On the other hand we have no problem with convergence at all — if one is very careful in monitoring convergence of a regular MCMC method (as one should, but people are not always!), this may be rather time consuming as well.

5.2 Monte Carlo standard errors

We turn now to the issue (ii) of how well $\mathbb{E}_\pi(g)$ is estimated by $\overline{g}_n$ (after using an appropriate burn-in).

The *Monte Carlo error* can be expressed by

$$\mathrm{Var}_\pi(\overline{g}_n) = \frac{1}{n}\Big[\mathrm{Var}_\pi(g(X_0)) + 2\sum_{k=1}^{n-1}\Big(1-\frac{k}{n}\Big)\mathrm{Cov}_\pi(g(X_0), g(X_k))\Big],$$

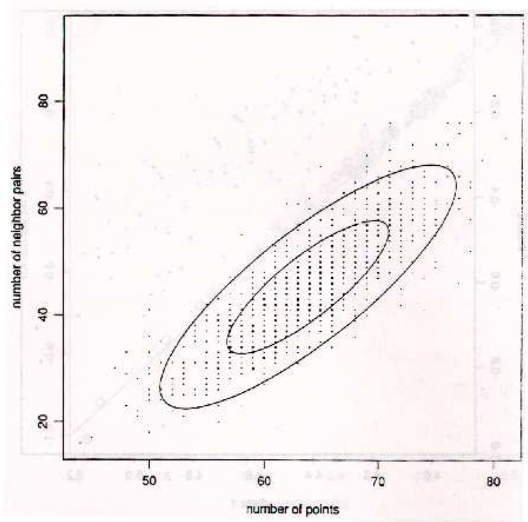


Figure 3: Scatter plot for the canonical statistics $n(x)$ and $s(x)$ for an unconditional Strauss process of Figure 2. The ellipses are the 50% and 90% contours of the normal approximation. Multiple points in the plot are indicated by the variable size dots. Figure taken from Geyer and Møller (1994).

or if the CLT holds

$$\mathrm{Var}_\pi(\overline{g}_n) \simeq \sigma_g^2/n$$

where

$$|\overline{g}_n - \mathbb{E}_\pi(g)| \leq 1.96\sqrt{\sigma_g^2/n}$$

with (approximately) probability 0.95. The asymptotic variance σ_g^2 can be estimated (see below). It can be written as

$$\sigma_g^2 = \mathrm{Var}_\pi\, g(X_0)\tau$$

where

$$\tau = 1 + 2\sum_{k=1}^{\infty} \mathrm{Corr}_\pi(g(X_0), g(X_k))$$

is called the *integrated autocorrelation time*.

Note that negative correlations 'helps' in the sense that σ_g^2 is a decreasing function of τ. On the other hand, numerical high correlations lead to small changes from one iteration to the next and therefore to a slow convergence rate. This illustrates a somewhat conflicting matter between the issues (i) and (ii). So it is sometimes suggested to use one MCMC algorithm (with good convergence properties) for obtaining an equilibrium state, and then switching to another MCMC algorithm with *good mixing properties* (in the sense of issue (ii)).

By the *mixing properties* of a Markov chain we usually mean how $\tau = \tau_g$ performs for various statistics g of interest. Note that in the special IID case, $\tau = 1$. In general

$$\tau^{-1} = \mathrm{Var}_\pi(g(X_0))/\sigma_g^2$$

can be thought of as the relative frequency of 'effective independent samples' in $\{g(X_k)\}$. Let M denote the number of 'updates in total' used to generate the sample path $X_1, \ldots, X_n$; the definition of M depends on the context: In a MCMC algorithm one may take $M = n+$burn-in (other definitions may be appropriate if e.g. a perfect simulation procedure is involved, see Møller and Nicholls (1999). In rejection sampling, an update is one rejection or acceptance operation (so M is not the length of the sequence of accepted samples). Alternatively, M may be the total run-time measured in CPU-seconds (on fixed hardware), but we should of course be aware that two rival methods may not been coded comparably efficiently. Recall also the issue (iv) of simplicity.

Having defined M appropriately, the *efficiency of the algorithm* (generating $g(X_1), \dots, g(X_n)$) can be estimated by

$$\xi = \frac{1}{\hat{\tau} M}$$

where $\hat{\tau}$ is an estimate of τ (see below). So ξ is 'the estimated relative frequency of effective independent samples generated per update of work done'.

Several possibilities for estimating σ_g^2 and τ_g have been proposed. We estimate the *lag-k covariance*

$$\gamma_k = \mathrm{Cov}_\pi(g(X_0), g(X_k))$$

by

$$\hat{\gamma}_k = \frac{1}{n} \sum_{i=1}^{n-k} (g(X_i) - \overline{g}_n)(g(X_{i+k}) - \overline{g}_n),$$

$k = 0, 1, \dots, n-1$. A reason for dividing with n and not $n-k$ is that the estimated covariance matrix $\{\hat{\gamma}_{|i-j|}\}_{i,j=0,\dots,n-1}$ is positive semi-definite. It is well-known that the obvious estimate for σ_g^2, $\hat{\gamma}_0 + 2\sum_{k=1}^{n} \hat{\gamma}_k$ is not consistent as $n \to \infty$. A consistent estimator is the *truncated periodogram* of Sokal (1989):

$$\hat{\sigma}_g^2 = \hat{\gamma}_0 + 2\sum_{k=1}^{K} \hat{\gamma}_k \quad \text{and} \quad \hat{\tau}_g = \hat{\sigma}_g^2 / \hat{\gamma}_0$$

when K is the smallest integer so that $K \geq 3\hat{\tau}_g$ (with $\hat{\tau}_g$ depending on K). Another is the *initial series estimator* of Geyer (1992) obtained by observing that for an irreducible and reversible Markov chain,

$$\Gamma_k = \gamma_{2k} + \gamma_{2k+1}$$

is a strictly positive/decreasing/convex function of $k = 0, 1, \dots$ Let L be the smallest integer so that $\widehat{\Gamma}_0, \dots, \widehat{\Gamma}_{L+1}$ is not strictly positive, decreasing or convex (with $\widehat{\Gamma}_k = \hat{\gamma}_{2k} + \hat{\gamma}_{2k+1}$). Then Geyer's estimator

$$\hat{\sigma}_g^2 = \hat{\gamma}_0 + 2\sum_{k=1}^{L} \hat{\gamma}_k$$

is in fact a conservative estimator of σ_g^2.

Geyer and Sokal's estimators are so-called window estimators. Other types of estimators can be derived using so-called regeneration techniques (these apply if the chain can be divided into different IID parts (called tours)).

6 Some tools for improving performance

Some tools are often introduced for 'hard simulation problems'. Here we consider only some examples of *enlarging the state space* using *auxiliary variables*.

6.1 Slice sampler

This is perhaps the simplest example of an auxiliary variable method. Suppose $\pi_u(x)$ is an unnormalized density on $\mathbb{R}^d$. Introduce an auxiliary variable $z > 0$ and the (unnormalized) joint density

$$\pi_u(x,z) = \mathbb{1}[\pi_u(x) > z].$$

Clearly, under $\pi_u(x,z)$, x has marginal density

$$\int_0^\infty \pi_u(x,z)\,dz = \pi_u(x)$$

as it should be. The slice sampler consists in alternately updating from (i) — (ii) as follows:

(i) $\pi(x \mid z) \propto \pi_u(x,z)$ specifies that x given z follows a uniform distribution on the horizontal slice $H(z) = \{x : \pi_u(x) > z\}$;

(ii) $\pi(z \mid x) \propto \pi_u(x,z)$ specifies that z given x follows a uniform distribution on the vertical slice $V(z) = (0, \pi_u(x))$.

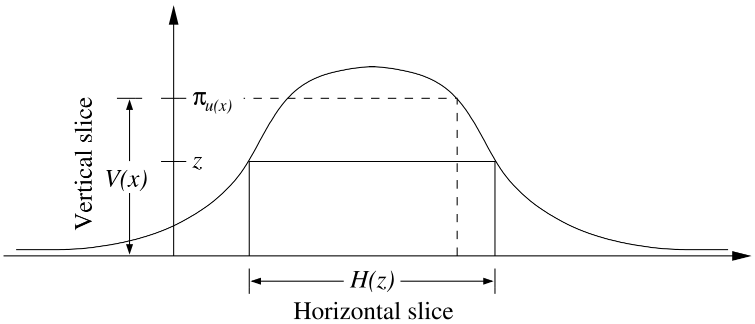


Figure 4: The slice sampler.

So this is a two component Gibbs sampler.

6.2 Swendsen and Wang (1987) algorithm

This is probably the first example of an auxiliary variable method. Consider the joint density

$$\pi_u(x,z) = \prod_{e=\{i,j\}\in E} \mathbb{1}[x_i \neq x_j \Rightarrow z_e = 0]\, p^{1-z_e}(1-p)^{z_e}$$

defined for

$$x = (x_i)_{i\in V} \in \{\pm 1\}^V, \quad z = (z_e)_{e\in E} \in \{0,1\}^E,$$

where $0 < p < 1$ is a parameter and $G = (V, E)$ is an undirected finite graph, e.g. a rectangular nearest-neighbour graph. The marginal density of x is an *Ising model* (with no external field),

$$\pi_u(x) = \prod_{\{i,j\} \in E} p^{\mathbb{1}[x_i \neq x_j]}$$

which is of fundamental interest in statistical physics, in particular at the *critical temperature* $\beta = -\ln p \simeq 0.88$ (where phase transition happens). The other variable z is in this context the auxiliary variable; its marginal distribution is called a *random cluster model*.

Regular MCMC such as Metropolis-Hastings and Gibbs samplers based on $\pi_u(x)$ alone perform badly compared to Gibbs sampling based on $\pi_u(x, z)$ with components x and z:

(i) $\pi(z \mid x)$ specifies that the bonds $z_e, e \in E$, are conditionally independent given x with

$$z_{\{i,j\}} = 0 \quad \text{whenever } x_i \neq x_j$$

and

$$z_{\{i,j\}} = 0 \quad \text{with probability } p \text{ if } x_i = x_j;$$

(ii) $\pi(z \mid x)$ specifies that the $x_C = (x_i)_{i \in C}$ with $C \in \mathcal{C}(z)$ are conditionally independent given x, where

$$\text{all } x_i \text{ equal for } i \in C$$

and

$$\text{the common value of } x_i \text{ being } +1 \text{ or } -1 \text{ with probability } \frac{1}{2} \text{ each.}$$

Here $\mathcal{C}(z)$ denotes a partition of V into maximal connected z-components defined as follows. We say that $i, j \in V$ are *connected by* z if either $i = j$ or there exists $e_1 = \{i_1, i_2\} \in E, \ldots, e_k = \{i_1, i_k\} \in E$ such that $i_1 = i, i_{k+1} = j$, and $z_{e_1} = \cdots = z_{e_k} = 1$. A set $C \subseteq V$ is a *connected* z*-component* if any $i, j \in C$ are connected by z. It is maximal if there exists no other connected z-component D so that $C \nsubseteq D$. As a matter of fact there exists efficient algorithms for finding such maximal connected components.

6.3 Simulated tempering

This technique has been independently invented by Marinari and Parisi (1992) and Geyer and Thompson (1995). The idea it to put the target density $\pi_u(x)$ into a mixture of unnormalized densities

$$g_1(x), g_2(x), \ldots, g_m(x) = \pi_u(x)$$

of 'increasing complexity'. The original example of this is temperature based:

$$g_i(x) = \pi_u(x)^{1/t_i}$$

where

$$t_1 > t_2 > \cdots > t_m = 1$$

so that $g_m = \pi_u$ is the *'cold' density* and g_1 is the *'hot' density*. But other constructions may be of relevance, including cases where the dimension of x depends on the index i — see Møller and Nicholls (1999).

Now consider the (unnormalized) joint density

$$\pi_u(x,i) = g_i(x)\pi_i, \quad x \in \Omega, i \in \{1,\dots,m\}$$

where each $\pi_i > 0$ are constants; Geyer and Thompson (1995) call $\pi = (\pi_1,\dots,\pi_m)$ for the *pseudo-prior*. Suppose that $(X,I) \sim \pi_u(x,i)$. Then X follows the mixture density

$$\tilde{\pi}_u(x) = \sum_{i=1}^{m} g_i(x)\pi_i$$

and X given $I = i$ has (conditional) density g_i. Particularly, X given $I = i$ follows the cold distribution. Moreover,

$$\mathbb{P}(I = i) \propto c_i\pi_i$$

where c_i is the normalizing constant of g_i. By choosing $c_i = 1/\pi_i$ a uniform mixture is obtained. Estimates $\hat{c}_i$ of $c_i, i = 1,\dots,m$ can be obtained in different ways as described in Geyer and Thompson (1995). Thereby an approximate uniform mixture can be obtained.

Furthermore, let P_i denote a transition kernel which is invariant with respect to $g_i, i = 1,\dots,m$, and let Q be the transition kernel on the state space $\{1,2,\dots,m\}$ defined by

$$Q(i,\{j\}) = \begin{cases} p & \text{for } |i-j| = 1 \\ 1-2p & \text{for } 1 < i = j < m \\ 1-p & \text{for } i = j = 1 \text{ or } i = j = m \end{cases}$$

and $Q(i,\{j\}) = 0$ else. The kernels P_i are usually given by some related Metropolis-Hastings algorithms. The parameter $0 < p \leq 1/2$ is user-specified; common choices are $p = 1/2$ and $p = 1/3$.

ALGORITHM 3 (SIMULATED TEMPERING ALGORITHM)
Given a current state (x,i)

(i) update $x \to x'$ using $P_i(x,\cdot)$;

(ii) propose $i \to i'$ using $Q(i,\cdot)$;

(iii) return (x',i') with probability $\min\{1, H(i,i';x)\}$, retain (x',i) otherwise.

Here

$$H(i,i';x) = \frac{\pi_u(x',i')}{\pi_u(x',i)} \frac{Q(i',\{i\})}{Q(i,\{i'\})}$$

is a Hastings ratio so that the algorithm has invariant density $\pi_u(x,i)$. It remains to check for irreducibility, etc.

A *regeneration step* can be done at the hot temperature: if g_1 is 'flat enough', choose $P_1(x,\cdot) = P_1(\cdot)$ independently of x; then we obtain IID tours: when the simulated tempering chain (X_n, I_n) satisfies $I_n = 1$, it regenerates.

The following figures, taken from Mase, Møller, Stoyan, Waagepetersen and Döge (1999) relate to the case where $pi_u(x) = g_m(x)$ is a hard core point process ($\gamma = 0$ in the Strauss process) and

$$g_i(x) = \beta_i^{n(x)} \exp\left(-\theta_i \sum_{1 \leq i < j \leq n} \left[\mathbb{1}(\|x_i - x_j\| \leq R) + c\frac{|b(x_i,R/2) \cap b(x_j,R/2)|}{b(0,R/2)} \right] \right)$$

for $x = \{x_1, \ldots, x_n\}$ and $i = 1, 2, \ldots, m-1$, with $0 = \theta_1 < \theta_2 < \cdots < \theta_{m-1} < \infty, \beta_i > 0$ and $c > 0$ being user-specified parameters. Here $b(x_i, R/2)$ is the disc with center x_i and diameter R.

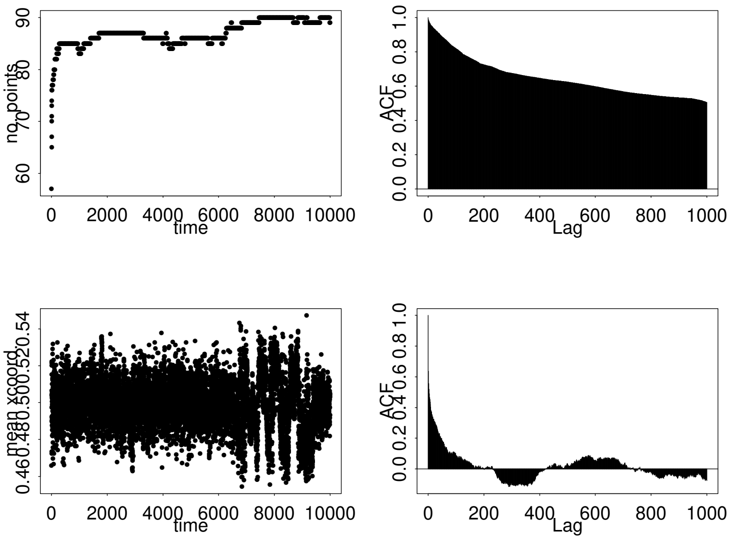


Figure 5: Time series and estimated autocorrelations obtained with the Metropolis-Hastings algorithm. Upper row: number of points. Lower row: statistic given by the empirical mean of the first coordinates in the point pattern (Figure taken from Mase, Møller, Stoyan, Waagepetersen and Döge (1999)).

7 MCMC likelihood inference

7.1 Importance sampling

Originally this was introduced as a means to reduce the asymptotic variance when making IID samples from another distribution (the importance distribution) than the target (adjusting samples as described below); see, e.g., Ripley (1987). We shall use importance sampling for estimating ratios of normalizing constants.

Let, as usual, $\pi_u(\cdot)$ be the unnormalized target density with normalizing constant

$$c = \int \pi_u(x)\, \nu(dx).$$

Suppose $f_u(\cdot)$ is another unnormalized density with normalizing constant

$$\tilde{c} = \int f_u(x)\, \nu(dx)$$

and so that

$$f_u(x) = 0 \;\Rightarrow\; \pi_u(x) = 0.$$

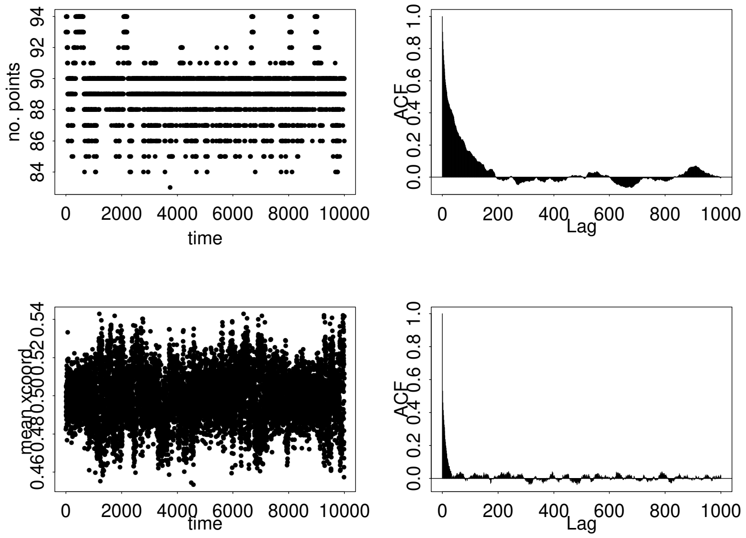


Figure 6: Time series and estimated autocorrelations for the coldest chain in simulated tempering. Upper row: number of points. Lower row: statistic given by the empirical mean of the first coordinates in the point pattern (Figure taken from Mase, Møller, Stoyan, Waagepetersen and Döge (1999)).

Then, taking $0/0 = 0$,

$$\mathbb{E}_\pi(g) = \frac{\tilde{c}}{c}\int g(x)\frac{1}{\tilde{c}}\frac{\pi_u(x)}{f_u(x)}f_u(x)\,\nu(dx) = \frac{\tilde{c}}{c}\mathbb{E}_f\left[g\frac{\pi_u}{f_u}\right]$$

for any function $g \in L^1(\Pi)$. Taking $g = 1$ we obtain

$$\frac{c}{\tilde{c}} = \mathbb{E}_f\left[\frac{\pi_u}{f_u}\right]. \tag{26}$$

Thus, for any $g \in L^1(\Pi)$,

$$\mathbb{E}_\pi(g) = \mathbb{E}_f\left[g\frac{\pi_u}{f_u}\right]\Big/\mathbb{E}_f\left[\frac{\pi_u}{f_u}\right]. \tag{27}$$

If $X_1, \dots, X_n$ is a MCMC sample from f, we estimate $\mathbb{E}_f\left[g\frac{\pi_u}{f_u}\right]$ by

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{\pi_u(X_i)}{f_u(X_i)} \tag{28}$$

and similarly (26) by

$$\widehat{c/\tilde{c}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi_u(X_i)}{f_u(X_i)}. \tag{29}$$

Combining (27) — (29) we obtain an estimator of $\mathbb{E}_\pi(g)$.

Importance sampling works only if the *importance weights*

$$\omega(X_i) = \frac{\pi_u(X_i)}{f_u(X_i)}, \quad i = 1, \ldots, n$$

are all of the same magnitude. Note that

$$\widehat{\mathbb{E}_\pi(g)} = \mathbb{E}_{n,\pi,f}(g) = \sum_{i=1}^{n} g(X_i)\overline{\omega}(X_i)$$

where the normalized importance weights

$$\overline{\omega}(X_i) = \frac{\omega(X_i)}{\sum_{j=1}^{n} \omega(X_j)}, \quad i = 1, \ldots, n$$

specifying a distribution on $\{1, \ldots, n\}$; the mean $\mathbb{E}_{n,\pi,f}(g)$ is w.r.t. this distribution.

7.2 MCMC MLE

We follow Geyer (1999). Consider a parametric family of unnormalized densities

$$\pi_\theta(x), \quad \theta \in \Theta$$

where $\Theta \subseteq \mathbb{R}^d$ is the parameter space. Suppose all densities are defined w.r.t. the same measure ν, and that the normalizing constants

$$c(\theta) = \int \pi_\theta(x)\, \nu(dx), \quad \theta \in \Theta$$

are analytical intractable. Assume the MLE $\hat{\theta}$ exists, that is $\hat{\theta}$ is the unique $\theta \in \Theta$ which maximizes the likelihood function

$$L(\theta; x) = \frac{1}{c(\theta)} \pi_\theta(x). \tag{30}$$

However, since $c(\theta)$ is not completely known, we do not know the value of $\hat{\theta}$.

We shall use importance sampling to approximate $c(\theta)$ (up to a constant). The importance density could be

(i) $f_u(x) = \pi_\psi(x)$ for some fixed $\psi \in \Theta$,

(ii) a mixture $f_u(x) = \sum_{i=1}^{k} \delta_i \pi_{\psi_i}(x), \quad (\delta_i > 0, \psi_i \in \Theta)$

or another choice so that importance sampling works well — see comments later. For specificity, we consider the case (i) in the following.

It makes no difference if we instead of (30) use the log likelihood ratio

$$\ell(\theta; x) = \log \frac{L(\theta; x)}{L(\psi; x)} = \log \frac{\pi_\theta(x)}{\pi_\psi(x)} - \log \frac{c(\theta)}{c(\psi)}.$$

Let $X_1, \ldots, X_n$ be a MCMC sample from π_ψ. Then we obtain the approximation

$$\ell(\theta; x) \simeq \ell_n(\theta; x) = \log \frac{\pi_\theta(x)}{\pi_\psi(x)} - \log \frac{1}{n} \sum_{i=1}^{n} \frac{\pi_\theta(X_i)}{\pi_\psi(X_i)}.$$

Note that for fixed θ, x and ψ, we have that $\ell_n(\theta; x)$ is a random variable depending on $X_1, \ldots, X_n$. Now, an *approximate MLE* (MCMC MLE) is a value $\hat{\theta}_n(x) \in \Theta$ which maximizes $\ell_n(\theta; x)$ uniquely.

Very often in applications, π_θ is of exponential family form:

$$\pi_\theta(x) = b(x) \exp(\theta \cdot t(x)).$$

Then

$$\ell_n(\theta; x) = (\theta - \psi) \cdot t(x) - \log \frac{1}{n} \sum_{i=1}^{n} \exp((\theta - \psi) \cdot t(X_i))$$

Taking derivatives w.r.t. θ we obtain that $\hat{\theta}_n(x)$ (if it exists) is a solution to the *MCMC approximation of the likelihood equation:*

$$\mathbb{E}_{n,\theta,\psi}(t) = t(x)$$

where $\mathbb{E}_{n,\theta,\psi}(t) = \mathbb{E}_{n,\pi_\theta,\pi_\psi}(t)$ with normalized importance weights

$$\overline{\omega}(X_i) = \frac{\exp((\theta - \psi) \cdot t(X_i))}{\sum_{j=1}^{n} \exp((\theta - \psi) \cdot t(X_j))}.$$

Differentiating once more, we obtain

$$\frac{\partial^2}{\partial\theta\partial\theta^\top} \ell_n(\theta; x) = -\operatorname{Var}_{n,\theta,\psi}(t).$$

Hence $\ell_n(\theta; x)$ is concave and $\hat{\theta}_n(x)$ may be found using e.g. Newton-Raphson.

THEOREM 8
Suppose that we have an exponential family as above, where $\Theta \subseteq \mathbb{R}^d$ is convex. If the true MLE $\hat{\theta}(x)$ exists, then

$$\mathbb{P}(\exists\, N > 0\; \forall\, n \geq N: \; \hat{\theta}_n(x) \text{ exists}) = 1$$

and

$$\mathbb{P}(\lim_{n \to \infty} \hat{\theta}_n(x) = \hat{\theta}(x)) = 1.$$

Comments:

(i) The theorem establishes consistency of the approximate MLE towards the true MLE as the length of the Markov chain increases — it does not tell anything about consistency towards the true (unknown) θ.

(ii) In practice we need ψ to be close enough to $\hat{\theta}(x)$ (which we don't know). One may try with another estimator than the MLE, for example some kind of maximum pseudo-likelihood estimator (MPLE), a moment estimator,. . .

(iii) Typically, $\ell_n(\theta; x)$ approximates $\ell(\theta; x)$ well in a neighbourhood to ψ, so an iterative procedure like the following one can be used .

ALGORITHM 4 (TRUST REGION ALGORITHM)

1 $\psi_1 =$ *appropriate initial value of* ψ *(e.g. the MPLE);*

2 $i = 0$*;*

3 repeat

a) $i = i + 1$

b) generate a Markov chain sample from π_{ψ_i}

c) find maximizer $\hat{\theta}_{n,i}(x) \in \Theta$ *in a small trust region around* ψ_i

d) $\psi_{i+1} = \hat{\theta}_{n,i}(x)$

until ψ_{i+1} *is indistinguishable from* psi_i*;*

4 return $\hat{\theta}_n(x) = \psi_{i+1}$

(iv) If we do not have an exponential family model, we need to be even more careful. The parameters may be estimated as briefly indicated below.

EXAMPLE 5 (STRAUSS PROCESS (CONTINUED))

If $R > 0$ is fixed we have an exponential family model with

$$\theta = (\theta_1, \theta_2) = (\log \beta, \log \gamma),$$
$$\Theta = \mathbb{R} \times (-\infty, 0]$$

(excluding the hard core case where $\theta_2 = -\infty$), and

$$t(x) = (t_1(x), t_2(x)) = (n(x), s(x)).$$

Often the parameter of interest is the interaction parameter θ_2. If we condition on $n(x) = k$, the (unnormalized) conditional density

$$\pi_{\theta_2}(x \mid n(x) = k) = \mathbb{1}[n(x) = k] \exp(\theta_2 s(x))$$

depends only on θ_2. The 'conditional MLE' obtained from this conditional density is compared with the 'unconditional MLE' obtained from

$$\pi_\theta(x) = \exp(\theta_1 n(x) + \theta_2 s(x))$$

in the following figures (taken from Geyer and Møller (1994); note that a mixture is used for the importance density).

Suppose now that R is unknown also, so

$$\theta = (\log \beta, \log \gamma, R) \in \mathbb{R} \times (-\infty, 0] \times (0, \infty).$$

Write $s_R(x)$ for $s(x)$. It is easily seen that

$$\begin{gathered} s_{R_1}(x) = s_{R_2}(x), \quad R_1 < R_2 \\ \Updownarrow \\ \pi_{(\log \beta, \log \gamma, R_1)}(x) \le \pi_{(\log \beta, \log \gamma, R_2)}(x) \end{gathered}$$

(here considered as normalized densities). So in principle we need only to consider the inter-event distances $R_1, R_2, \ldots, R_q$ between point pairs. However, $q = n(x)(n(x) - 1)/2$ is usually very large. Instead we may consider a smaller number of R values, maximize the approximate likelihood function for each fixed R value (then we are back in an exponential family model), and compare these in order to find the MCMC MLE — see Baddeley, Møller and Waagepetersen (1998) for an example.

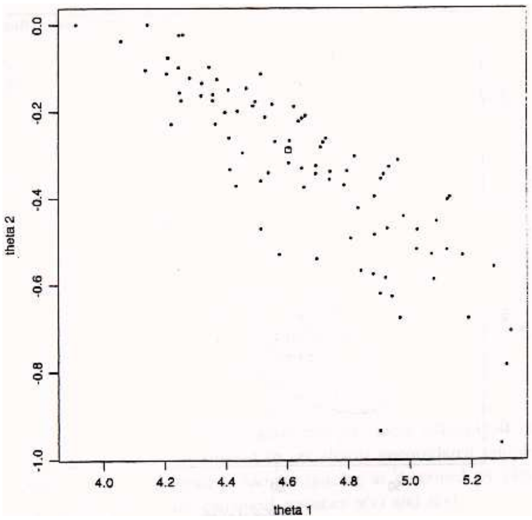


Figure 7: Scatter plot for the unconditional MLE for the unconditional process of Figure 2. The square marks the true parameter value. Figure taken from Geyer and Møller (1994).

7.3 Missing data

An approximate MLE may still be found if we have incomplete observations — see Geyer (1999) and Baddeley, Møller and Waagepetersen (1998) for theory and examples of applications. Briefly the idea is the following. Suppose (X, Y) has unnormalized density $\pi_\theta(x, y)$ parameterized by $\theta \in \Theta$, where $Y = y$ is observed while X is unobserved. Then, if $\theta \in \Theta$ is a fixed point in the parameter space,

$$\frac{L(\theta; y)}{L(\psi; y)} = \frac{\mathbb{E}_\psi\left[\frac{\pi_\theta(X,y)}{\pi_\psi(X,y)}\Big| Y = y\right]}{\mathbb{E}_\psi\left[\frac{\pi_\theta(X,Y)}{\pi_\psi(X,Y)}\right]} \simeq \frac{\sum_{i=1}^{n} \frac{\pi_\theta(X_i^*, y)}{\pi_\psi(X_i^*, y)}}{\sum_{i=1}^{n} \frac{\pi_\theta(X_i, y)}{\pi_\psi(X_i, y)}},$$

where $(X_i, Y_i), i = 1, \ldots, n$ is a sample from an ergodic Markov chain with invariant density $\pi_\theta(x, y)$, and $X_i^*, i = 1, \ldots, n$ is a sample from another ergodic Markov chain with invariant density $\pi_\theta(\cdot|y) \propto \pi_\theta(\cdot, y)$ (see Remark 2).

8 Perfect simulation

First I planned to say a bit more on this topic - but time is running out and the topic is perhaps a little too advanced for this summer school. Two seminal works are Propp and Wilson (1996) and Fill (1998). The paper by Møller and Nicholls (1999) concerns *perfect simulated tempering* applied on a number of applications —- further references on perfect simulation can be found in that paper.

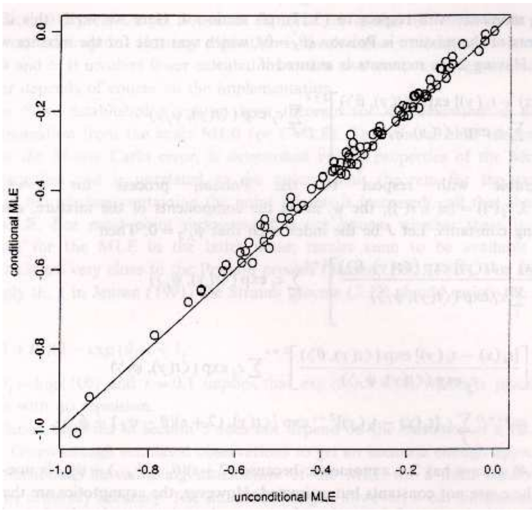


Figure 8: Scatter plot for the conditional and unconditional MLEs of θ_2 for the unconditional Strauss process of 2. The line is the locus of points where the two estimators agree. Figure taken from Geyer and Møller (1994).

References

[1] A.J. Baddeley, J. Møller, and R. Waagepetersen. Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. Technical Report R-98-2019, Department of Mathematics, Aalborg University, 1998.

[2] J. Besag and P. Clifford. Generalized Monte Carlo significance tests. *Biometrika*, 76:633–642, 1989.

[3] J. Besag and P. Clifford. Sequential Monte Carlo p-values. *Biometrika*, 78:301–304, 1991.

[4] J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10:3–66, 1995.

[5] J.E. Besag. Discussion on the paper by Grenander and Miller. *Journal of the Royal Statistical Society*, B 56:591–592, 1994.

[6] J.E. Besag and P.J. Green. Spatial statistics and Bayesian computation (with discussion). *Journal of the Royal Statistical Society*, B 55:25–37, 1993.

[7] A. Brix and J. Møller. Space-time multitype log Gaussian Cox processes with a view to modelling weed data. Technical Report R-98-2012, Department of Mathematical Sciences, Aalborg University, 1998.

[8] S. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1998.

[9] M. Creutz. Confinement and the critical dimensionality of space-time. *Physical Review Letters*, 43:553–556, 1979.

[10] J.A. Fill. An interruptible algorithm for perfect sampling via Markov chains. *Annals of Applied Probability*, 8:131–162, 1998.

[11] A.E. Gelfand and A.F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[12] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[13] C. J. Geyer. Likelihood inference for spatial point processes. In O.E. Barndorff-Nielsen, W.S. Kendall, and M.N.M. van Lieshout, editors, *Stochastic Geometry: Likelihood and Computation*, pages 79–140, Boca Raton, 1999. Chapman and Hall/CRC.

[14] C.J. Geyer. Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, 8:473–483, 1992.

[15] C.J. Geyer and J. Møller. Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, 21:359–373, 1994.

[16] C.J. Geyer and E.A. Thompson. Annealing Markov chain Monte Carlo with applications to pedigree analysis. *Journal of American Statistical Association*, 90:909–920, 1995.

[17] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, 1996.

[18] P. Green. Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[19] M.B. Hansen, J. Møller, and F.Aa. Tøgersen. Bayesian image analysis and the Metropolis adjusted Langevin algorithm with an application to time series of ultrasound images. (In preparation).

[20] E. Marinari and G. Parisi. Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters*, 19:451–458, 1992.

[21] S. Mase, J. Møller, D. Stoyan, R.P. Waagepetersen, and G. Döge. Packing densities and simulated tempering for hard core Gibbs point processes. Technical Report R-99-2002, Department of Mathematical Sciences, Aalborg University, 1999.

[22] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[23] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

[24] J. Møller. Markov chain Monte Carlo and spatial point processes. In O.E. Barndorff-Nielsen, W.S. Kendall, and M.N.M. van Lieshout, editors, *Stochastic Geometry: Likelihood and Computation*, number 80 in Monographs on Statistics and Applied Probability, pages 141–172, Boca Raton, 1999. Chapman and Hall/CRC.

[25] J. Møller and G. Nicholls. Perfect simulation for sample-based inference. Technical Report R-99-2011, Department of Mathematical Sciences, Aalborg University, 1999.

[26] J. Møller, A.R. Syversveen, and R.P. Waagepetersen. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25:451–482, 1998.

[27] J.G. Propp and D.B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9:223–252, 1996.

[28] B.D. Ripley. *Stochastic Simulation*. Wiley, New York, 1987.

[29] G.O. Roberts and J.S. Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997.

[30] G.O. Roberts and J.S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of Royal Statistical Society*, B 60:255–268, 1998.

[31] G.O. Roberts and R.L. Tweedie. Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli*, 2:341–363, 1996.

[32] J.S. Rosenthal. Minorization conditions and covergence rates for Markov chain Monte Carlo. Technical Report 9321, Department of Statistics, University of Toronto, 1993.

[33] H. Rue and M.A. Hurn. Bayesian object identification. Technical Report Statistics 6/1997, Department of Mathematical Sciences, Norwegian University of Science and Technology, 1997.

[34] A.D. Sokal. Monte Carlo methods in statistical mechanics: foundations and new algorithms. In *Cours de Troisième Cycle de la Physique en Suisse Romande*, Lausanne, 1989.

[35] P. Suomela. *Construction of Nearest Neighbour Systems*. Disertation 10, Department of Mathematics, University of Helsinki, 1976.

[36] R.H. Swendsen and J.-S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58:86–88, 1987.

[37] L. Tierney. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22:1701–1728, 1994.