# Stochastic Simulation

## Preface

Mathematical modelling that traditionally contains important elements of mathematics, probability theory and statistics has experienced a drastic development during the last twenty years. Especially the application of computer simulation has been crucial for the development of the field. This course will give an introduction to modern simulation techniques. In addition we aim at giving the participants a better intuitive knowledge of basic concepts in probability theory and statistics. Only basic knowledge of stochastics (probability theory and statistics) is required.

The course covers two quarters. In the first quarter, an exposition will be given of traditional simulation techniques such as inversion, rejection, importance sampling and variance reduction techniques. Also modern techniques such as Markov Chain Monte Carlo simulation will be treated, including the Metropolis-Hastings algorithm. The second quarter of the course will cover applications of simulation in a number of other fields, including operation analysis, insurance and finance.

August 2004
Eva B. Vedel Jensen
Søren Asmussen

# Simulation 1

**Eva B. Vedel Jensen**

## 1. Introduction

### 1.1. Scope of simulation

The term 'computer intensive methods' means different things to different people. It is also a dynamic subject: what requires intensive computing today may be solvable with a pocket calculator tomorrow. Not so long ago, the calculation of normal probabilities to reasonable accuracy would have required considerable CPU time.

An initial classification of computer intensive methods as applied to statistics is the following:

- Computers for graphical data exploration.
- Computers for data modelling.
- Computers for inference.

There is some overlap in these three, but in this course the focus is on the second and the third of the above.

A course in simulation may have two roles. The first is to gain some understanding and knowledge of the techniques and tools which are available. The second is that many of the techniques are themselves clever applications or interpretations of probability and statistics. So, understanding the principles behind the different algorithms can often lead to a better understanding of probability and statistics generally. The simulation techniques have their own intrinsic value as statistical exercises.

This is not a course on computing. We will not get into the details of programming itself. Furthermore, this is not a course which will deal with specialised statistical packages often used in statistical computing. All the examples can be handled using simple S-plus (r) functions - far from the most efficient way of implementing the various techniques. It is important to recognise that high-dimensional complex problems do require efficient programming (commonly in C or Fortran). However the emphasis of this course is to illustrate the various methods and their applications on relatively simple examples.

### 1.2. Computers as inference machines

It is something of a cliché to point out that computers have revolutionized all aspects of statistics. In the context of inference there have really been two substantial impacts: the first has been the freedom to make inferences without the assumptions which standard techniques necessitate in order to obtain analytic solutions - Normality, linearity, independence etc. The second is the ability to apply standard type of models to situations of greater data complexity - missing data, censored data.

### 1.3. References

These notes on simulation are adapted from earlier course notes produced by Coles *et al.* (2001) and Jensen (2001). The notes also use ideas and results from the forthcoming book Asmussen & Glynn (2004). Other important books in the area are:

- *Stochastic Simulation*, B. Ripley.
- *An Introduction to the Bootstrap*, B. Efron and R. Tibshirani.
- *Tools for Statistical Inference*, M. Tanner.
- *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson and D. Spiegelhalter.

## 2. Traditional simulation techniques

In this section we look at different techniques for simulating from distrbutions and stochastic processes. In situations where we study a statistical model, simulating from that model generates realizations which can be analyzed as a means of understanding the properties of that model.

### 2.1. Issues in simulation

Whatever the application, the role of simulation is to generate data which have the statistical properties of some specified model. This generates two questions:

- How to do it; and
- How to do it efficiently.

To some extent, just doing it is the priority, since computers are often sufficiently fast for even inefficient routines to be quick. On the other hand, efficient design of simulation can add insight into the statistical model itself, in addition to CPU savings. We will illustrate the idea with a simple example.

### 2.2. Buffon's needle

Perhaps the most famous simulation experiment is Buffon's needle, originally designed to calculate an estimate of $\pi$. Here, we will use the experiment to calculate an estimate of the length of the needle. There are a number of ways the experiment can be improved on to give better estimates which will highlight the general principle of *designing* simulated experiments to achieve optimal accuracy in the sense of minimizing statistical variability.

Buffon's original experiment is as follows. Imagine a grid of horizontal parallel lines of spacing $d$, on which we randomly drop a needle of unknown length $\ell$, with $\ell \leq d$. We repeat this experiment $n$ times, and count $R$, the number of times the needle intersects a line. An estimate of the needle length $\ell$ is

$$\hat{\ell} = \frac{\pi d}{2} \frac{R}{n}. \tag{1}$$

The rationale behind this is that if we let $X$ be the distance from the centre of the needle to the nearest lower grid line, and $\Theta$ be the angle that the needle makes with the horizontal, then under the assumption of random needle throwing, we have $X \sim U[0, d]$ and $\Theta \sim U[0, \pi]$. (Here, $U[a, b]$ is the notation used for the uniform distribution on the interval $[a, b]$). The needle

intersects the grid if and only if

$$X \leq \frac{\ell}{2}\sin\Theta \text{ or } X \geq d - \frac{\ell}{2}\sin\Theta,$$

cf. Figure 1. The probability $p$ that the needle intersects the grid is therefore

$$\begin{aligned} p &= P(X \leq \frac{\ell}{2}\sin\Theta) + P(X \geq d - \frac{\ell}{2}\sin\Theta) \\ &= 2P(X \leq \frac{\ell}{2}\sin\Theta) \\ &= 2\int_0^{\pi}\int_0^{\frac{\ell}{2}\sin\theta} \frac{1}{d}\frac{1}{\pi} dx d\theta \\ &= \frac{2}{\pi d}\int_0^{\pi} \frac{\ell}{2}\sin\theta d\theta \\ &= \frac{2\ell}{\pi d}. \end{aligned}$$

Therefore,

$$\ell = \frac{\pi d}{2} p.$$

Since the probability of intersection $p$ can be estimated by $\hat{p} = R/n$, we obtain the estimator (1).

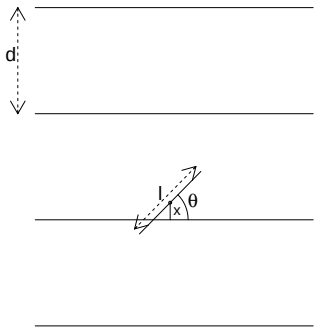


Figure 1: Buffon's needle.

A natural question is how precise is this estimator. To address this we need to consider the variability of the estimator $\hat{\ell}$. Now, $R \sim b(n, p)$, so $\mathrm{Var}(R/n) = p(1-p)/n$. Therefore,

$$\mathrm{Var}(\hat{\ell}) = (\frac{\pi d}{2})^2 \times \frac{p(1-p)}{n}.$$

There are many modifications that may improve the efficiency of this experiment. One example is to use a grid of rectangles and basing the estimate on the number of intersections with either or both horizontal or vertical lines.

### 2.3. Raw ingredients

The raw material for any simulation exercise is random digits. Transformation or other types of manipulation can then be applied to build simulations of more complex distributions or systems. So, how can random digits be generated?

It should be recognised that any algorithmic attempt to mimic randomness is just that: a mimic. By definition, if the sequence generated is deterministic then it is not random. Thus, the trick is to use algorithms which generate sequences of numbers which would pass all the tests of randomness (from the required distribution or process) despite their deterministic derivation.

The most popular such algorithms today are linear congruential generators of the form

$$u_n = \frac{x_n}{M} \text{ where } x_{n+1} = (Ax_n + C) \pmod{M}.$$

The number $x_1$ determines deterministically the whole sequence $\{u_n\}$ and is called the *seed*. One should note that the range of the $u_n$s is not the whole of $[0, 1]$ but only $\{0, 1/M, 2/M, \ldots, 1 - 1/M\}$ (often the value 0 is discarded to avoid problems when using the sequence, say one needs division or to take logarithms). Thus, of course, $M$ should be large for the generator to work well but there are other concerns such as periodicity. Namely, after $d \leq M$ steps, one of the numbers $i/M$ will occur for the second time and the algorithm will then produce replicates of cycles of length $d$ or smaller.

The difficulty is therefore to choose a large $M$ and associated $A, C$ such that the period is large, preferably $M$ (this is denoted *full period*). One difficulty with the generators having short period is that the gaps in the sequence may not be evenly distributed. Fortunately, number-theoretic considerations provide verifiable conditions under which linear congruential generators are of full period. This has led to certain popular parameter choices for $A$, $C$ and $M$. A dominant one in earlier generators of computers and software has $M = 2^{31} - 1 = 2147483647$, $A = 7^5 = 16807$, $C = 0$. This choice has the nice property that its period is (very) close to the number of machine-representable integers in a 32-bit computer.

Ripley (1987) gives details of the number theoretic arguments which support this method, and gives illustrations of the problems which can arise by using inappropriate choices of $A$, $C$ and $M$. We will not worry about

this issue here, as any decent statistics package should have had its random number generator checked pretty thoroughly. The point worth remembering though is that computer generated random numbers are not random at all, but (hopefully) they look random enough for that not to matter.

In subsequent sections then, we assume that we can generate a sequence of numbers which can be regarded as the outcome of $n$ random variables $U_1, U_2, \ldots, U_n$ which are independent and distributed according to the $U[0,1]$ distribution. In the following section we look at ways of simulating data from a specified univariate distribution with distribution function $F$, on the basis of $U_1, U_2, \ldots, U_n$ from the distribution $U[0,1]$.

### 2.4. Inversion

Let us suppose that $F$ is continuous and strictly increasing. If $X$ has distribution function $F$, then $F(X)$ is uniformly distributed on [0,1]. So by inversion if $U$ is uniformly distributed on [0,1], then $F^{-1}(U)$ has distribution function $F$, since

$$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

Thus, defining $X_i = F^{-1}(U_i)$, $i = 1, \ldots, n$, generates a sequence of independent random variables from $F$.

For example, to simulate from the exponential distribution with parameter $\lambda$, we use that

$$F(x) = 1 - \exp(-\lambda x), \quad x > 0,$$

so

$$F^{-1}(u) = -\lambda^{-1} \log(1-u), \quad 0 < u < 1.$$

Since $U \sim U[0,1]$ implies that $1 - U \sim U[0,1]$, we have that

$$-\lambda^{-1} \log U_1, \ldots, -\lambda^{-1} \log U_n$$

is a sequence of independent random variables from the exponential distribution with parameter $\lambda > 0$.

This procedure works equally well for discrete distributions, provided we interpret the inverse distribution function as

$$F^{-1}(u) = \min\{x | F(x) \geq u\}.$$

The procedure then simply amounts to searching through a table of the distribution function. For example, the distribution function of the Poisson distribution with parameter 2 is

| $x$ | $F(x)$ |
|---|---|
| 0 | 0 1353353 |
| 1 | 0.4060058 |
| 2 | 0.6766764 |
| 3 | 0.8571235 |
| 4 | 0.9473470 |
| 5 | 0.9834364 |
| 6 | 0.9954662 |
| 7 | 0.9989033 |
| 8 | 0.9997626 |
| 9 | 0.9999535 |
| 10 | 0.9999917 |

so, we generate a sequence of standard uniforms $U_1, U_2, \ldots, U_n$ and for each $U_i$ obtain a Poisson (2) variate $X_i$ where $F(X_i - 1) < U_i \leq F(X_i)$. So, for example, if $U_1 = 0.7352$ then $X_1 = 3$.

More formally this procedure can be described as follows: Consider a random variable $X$ with a discrete distribution. Let us imagine, for simplicity, that the possible values of $X$ are the non-negative integers and that

$$P(X = j) = p_j, j = 0, 1, \ldots.$$

We then simulate $X$ from $U \sim U[0, 1]$ using

$$X = j \text{ if } \sum_{i=0}^{j-1} p_i < U \leq \sum_{i=0}^{j} p_i.$$

(We use the convention $\sum_{i=0}^{j-1} p_i = 0$ if $j = 0$.)

Let $q_j = p_0 + p_1 + \cdots + p_j$. We can then do the calculations with the following algorithm

**Algorithm 1**

1. *Simulate $U$ from $U[0, 1]$.*
2. *Set $j = 0$.*
3. *Repeat $j = j + 1$ until $U \leq q_j$.*
4. *Set $X = j$.*

This method is not feasible when the number of non-zero probabilities is large. On the average we use $\sum_{j=0}^{\infty} jp_j$ steps to find $X$. If we know that

$p_j = 0$ for $j > k$, a better way is to use the following algorithm (we let $q_{-1} = 0$).

**Algorithm 2**

1. *Simulate $U$ from $U[0,1]$.*
2. *Set $i = -1$ and $j = k$.*
3. *While $i + 1 < j$ do*

   *l=int((i+j)/2)*

   *if $U > q_l$ then $i = l$ else $j = l$.*
4. *Set $X = j$.*

Here, $\text{int}(x)$ is for $x \in \mathbb{R}$ the largest integer smaller than or equal to $x$. It can be shown that this algorithm will use $\log(k)/\log(2)$ steps and so is much quicker than the above method.

Returning to the continuous case, it may seem that the inversion method is sufficiently universal to be the only method required. In fact, there are many situations in which the inversion method is complicated to program or excessively inefficient to run. The inversion method is only really useful if the inverse distribution function is easy to program and compute. This is not the case, for example, with the normal distribution function for which the inverse distribution function, $\Phi^{-1}$, is not available analytically and slow to evaluate numerically. To deal with such cases, we turn to a variety of alternative schemes.

A concrete scheme for simulating normally distributed variables is as follows: Let $X, Y$ be independent and $N(0,1)$-distributed. Then, $X, Y$ is distributed as $R\cos\Theta, R\sin\Theta$, where $R^2$ is exponentially distributed with parameter $1/2$ and $\Theta \sim U[0, 2\pi]$. Furthermore, $R$ and $\Theta$ are independent. Therefore, we can take

$$X = \sqrt{-2\log(U_1)}\cos(2\pi U_2), \quad Y = \sqrt{-2\log(U_1)}\sin(2\pi U_2),$$

where $U_1, U_2$ are independent and $U[0,1]$-distributed.

### 2.5. Rejection sampling

The idea in rejection sampling is to simulate from one distribution which is easy to simulate from, but then only accept a simulated value with some probability. By choosing the probability correctly, we can ensure that the sequence of accepted simulated values are from the desired distribution. This

technique is called *rejection sampling*. We will throughout this section assume that the distribution $F$ to be simulated from has a density function $f$. Recall that the relation between $F$ and $f$ is

$$F(x) = \int_{-\infty}^{x} f(z)dz, \quad x \in \mathbb{R}.$$

Let us start with a concrete example, viz. simulation from the Beta distribution which has density

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha, \beta > 0$ and $B(\cdot, \cdot)$ is the Beta function defined by

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

Here,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

where $\Gamma(\cdot)$ is the gamma function defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx, \quad \alpha > 0.$$

Simulation by inversion is for the Beta distribution difficult because the inverse distribution function is not known explicitly. If $\alpha > 1$ and $\beta > 1$ we can instead bound the density function by a rectangle,

$$\{(x, f(x))|0 < x < 1\} \subseteq [0, 1] \times [0, K],$$

where

$$K = \frac{1}{B(\alpha, \beta)} \frac{(\alpha-1)^{\alpha-1}(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}}$$

and simulate random points $(X_i, Y_i)$ uniformly over the rectangle. We accept $X_i$ as an observation from $f$, if $Y_i \leq f(X_i)$, cf. Figure 2.

This procedure works for the following reason: Let $X, Y$ be independent random variables such that $X$ is uniform in $[0, 1]$ and $Y$ is uniform in $[0, K]$. Accept $X$, if $Y \leq f(X)$. We want to show that

$$P(X \leq x|X \text{ accepted}) = \int_{-\infty}^{x} f(z)dz, \quad x \in \mathbb{R}.$$

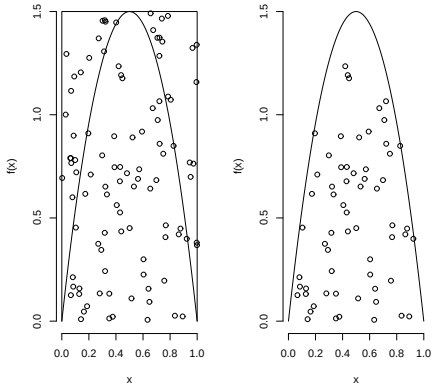


Figure 2: Rejection sampling from the Beta distribution with $\alpha = 2$ and $\beta = 2$.

It is clear that

$$P(X \leq x | X \text{ accepted}) = \begin{cases} 1 & x \geq 1, \\ 0 & x \leq 0. \end{cases}$$

For $0 < x < 1$, we use the following reasoning. First, recall that

$$P(X \leq x | X \text{ accepted}) = \frac{P(X \leq x, X \text{ accepted})}{P(X \text{ accepted})}.$$

Letting $\mathbf{1}\{\cdot\}$ be the indicator function, we find

$$\begin{aligned} P(X \leq x, X \text{ accepted}) &= P(X \leq x, Y \leq f(X)) \\ &= \int_0^x \int_0^{f(z)} \frac{1}{K} dy dz \\ &= \int_0^x \frac{f(z)}{K} dz \\ &= \int_{-\infty}^x \frac{f(z)}{K} dz. \end{aligned}$$

Therefore,

$$\begin{aligned} P(X \leq x | X \text{ accepted}) &= \frac{\int_{-\infty}^x \frac{f(z)}{K} dz}{\int_{-\infty}^{\infty} \frac{f(z)}{K} dz} \\ &= \int_{-\infty}^x f(z) dz. \end{aligned}$$

The efficiency of this method depends on how many points are rejected, which in turn depends on how well the graph of $f$ resembles the bounding rectangle. To improve the efficiency of the procedure and to allow for situations where $f$ may be unbounded or have unbounded support, the technique can be modified to permit the bounding function to take any form $Kg(x)$, where $g$ is the density of a distribution from which it is easy to simulate. If

$$f(x) \leq Kg(x), \quad x \in \mathbb{R},$$

for some $K > 0$, then we simulate from the density $f$ in the following way:

**Algorithm 3**

1. *Simulate $X$ from $g$. Suppose $X = x$.*
2. *Simulate $Y$ from $U[0, Kg(x)]$.*
3. *Accept $X$ if $Y \leq f(X)$.*
4. *Continue.*

The justification of this more general procedure is along the same lines as above. Let $X$ denote a random variable with density $g$. Then,

$$\begin{aligned} P(X \leq x, X \text{ accepted}) &= P(X \leq x, Y \leq f(X)) \\ &= \int_{-\infty}^{x} \int_{0}^{f(z)} g(z) \frac{1}{Kg(z)} \mathbf{1}\{g(z) > 0\} dz \\ &= \int_{-\infty}^{x} \frac{f(z)}{K} \mathbf{1}\{g(z) > 0\} dz \\ &= \int_{-\infty}^{x} \frac{f(z)}{K} dz. \end{aligned} \tag{2}$$

In particular,

$$P(X \text{ accepted}) = \int_{-\infty}^{\infty} \frac{f(z)}{K} dz, \tag{3}$$

so

$$\begin{aligned} P(X \leq x | X \text{ accepted}) &= \frac{\int_{-\infty}^{x} f(z) dz}{\int_{-\infty}^{\infty} f(z) dz} \\ &= \int_{-\infty}^{x} f(z) dz, \end{aligned}$$

so that the accepted values do indeed have density $f$. Note that

$$P(X \text{ accepted}) = P(Y \leq f(X)) = 1/K.$$

Note also that $f$ need only be known up to a constant of proportionality in order for this technique to work. The efficiency of the procedure depends on the degree of agreement between $f$ and the bounding envelope $Kg$ since if a large value of $K$ is necessary, then the acceptance probability is low, so that large numbers of simulations are needed in order to achieve a required sample size.

An adaption of the rejection algorithm which works well for many distributions is *the ratio of uniforms method*. Here a pair of independent uniforms are simulated and the ratio accepted as a simulation from the required distrbution according to a rejection scheme. The method is explained in more detail below.

Suppose we want to simulate from the density $f$ which is known up to a constant of proportionality. Thus,

$$f(x) = Ch(x), \quad x \in \mathbb{R},$$

where the non-negative function $h$ is known and $C > 0$ is an unknown constant. The basis of the technique is the following result. Let

$$C_h = \{(u, v) \in \mathbb{R}^2 : 0 < u < \sqrt{h(v/u)}\}.$$

Then, if $(U, V)$ is uniformly distributed over $C_h$, then $X = V/U$ has density $f$.

So, to simulate from a density proportional to $h$, we simulate uniformly over the region $C_h$, and take ratios of coordinates. In practice, $C_h$ may be complicated in shape, so the only practical solution is to bound it with a rectangle (if possible), simulate within the rectangle (by a pair of uniforms), and apply rejection.

The reason this works is the following. We want to show that

$$P(V/U \leq x) = \int_{-\infty}^{x} f(z)dz, \quad x \in \mathbb{R}. \tag{4}$$

Letting $\Delta_h$ be the area of $C_h$, we find that the density of $(U, V)$ is

$$f_{(U,V)}(u, v) = \frac{1}{\Delta_h} \mathbf{1}\{(u, v) \in C_h\}, \quad (u, v) \in \mathbb{R}^2.$$

Therefore,

$$\begin{aligned}
&P(V/U \le x) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}\{v/u \le x\} f_{(U,V)}(u,v)dvdu \\
&= \frac{1}{\Delta_h}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}\{v/u \le x, 0 < u < \sqrt{h(v/u)}\}dvdu \\
&= \frac{1}{\Delta_h}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}\{z \le x, 0 < u < \sqrt{h(z)}\}udzdu \\
&= \frac{1}{\Delta_h}\int_{-\infty}^{x}\int_{0}^{\sqrt{h(z)}} ududz \\
&= \frac{1}{2\Delta_h}\int_{-\infty}^{x} h(z)dz.
\end{aligned}$$

In particular,

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{(U,V)}(u,v)dvdu = \frac{1}{2\Delta_h}\int_{-\infty}^{\infty} h(z)dz = \frac{1}{2\Delta_h C}.$$

The result (4) now follows immediately.

As discussed above, this is only useful if we can generate uniformly over $C_h$, which is most likely to be achieved by simulating uniformly within a rectangle $[0,a]\times[b_-,b_+]$ which contains $C_h$ (provided such a rectangle exists). If it does, we have the following algorithm.

**Algorithm 4**

1. *Simulate independent* $U \sim U[0,a], V \sim U[b_-,b_+]$.
2. *If* $(U,V) \in C_h$, *accept* $X = V/U$, *otherwise repeat.*
3. *Continue.*

As an example, consider the Cauchy distribution with density

$$f(x) \propto \frac{1}{1+x^2}, \quad x \in \mathbb{R}, \tag{5}$$

cf. Figure 3. Then,

$$\begin{aligned}
C_h &= \{(u,v) : 0 \le u \le \sqrt{h(v/u)}\} \\
&= \{(u,v) : 0 \le u, u^2+v^2 \le 1\},
\end{aligned}$$

a semicircle. Hence, we can take $[0,a]\times[b_-,b_+] = [0,1]\times[-1,1]$ and get the algorithm.

**Algorithm 5**

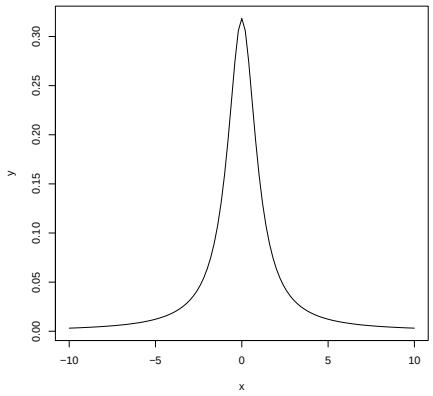


Figure 3: The Cauchy density.

1. *Simulate independent* $U \sim U[0,1], V \sim U[-1,1]$.
2. *If* $U^2 + V^2 \leq 1$, *accept* $X = V/U$, *otherwise repeat.*
3. *Continue.*

A number of modifications have been proposed to improve the efficiency of this procedure, which amount to rescaling and locating distributions before applying the method.

Another method for improving the efficiency is by a process known as 'squeezing' or 'pre-testing'. This applies to both the rejection and ratio of uniform methods. The point is that, in the ratio of uniforms method for example, the slowest part of the algorithm can be to check whether $(u, v) \in C_h$ or not. However, there may be simpler regions $C_1$ and $C_2$ such that $C_1 \subset C_h \subset C_2$, so that if $(u, v)$ is found to lie inside $C_1$ or outside $C_2$ then we immediately know whether it lies inside $C_h$ or not.

### 2.6. Monte Carlo integration

On one form or another, the quantity to be determined by simulation can often be formulated as an integral. This is obviously the case for expectation. Suppose $X$ is a random variable with density $f$ and expectation $E(X)$. Then,

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx,$$

so if $X_1, X_2, \ldots, X_n$ are independent random variables from the distribution of $X$, then

$$n^{-1} \sum_{i=1}^{n} X_i$$

is an unbiased and consistent estimator of $E(X)$. This argument can be generalized. Suppose we wish to calculate

$$\theta = \int_{-\infty}^{\infty} \varphi(x) f(x) dx$$

which is $E(\varphi(X))$, where $X$ has density $f$. Then, if $X_1, X_2, \ldots, X_n$ are independent random variables from this distribution, then

$$\hat{\theta} = n^{-1} \sum_{i=1}^{n} \varphi(X_i) \tag{6}$$

is an unbiased and consistent estimator of $\theta$. This approach is remarkably easy to use, even in high dimensions. The cost for this simplicity is that the variance may be high.

Normally, we state not only $\hat{\theta}$, but also a measure of how close to the true value $\theta$ we expect $\hat{\theta}$ to be. If the variance of $\varphi(X)$ is $\sigma^2$ it follows from the central limit theorem (will be presented in the probability theory course) that

$$\lim_{n\to\infty} P(\frac{1}{\sigma}\sqrt{n}(\hat{\theta} - \theta) \leq x) = \Phi(x).$$

An approximative 95% confidence interval is therefore given by

$$\left[\hat{\theta} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\theta} + \frac{1.96\sigma}{\sqrt{n}}\right]. \tag{7}$$

The choice of 95% is common but other values are, of course, possible. Say 99%, corresponding to $\hat{\theta} \pm 2.58\sigma/\sqrt{n}$. Also, one-sided confidence intervals may sometimes be relevant. Assume, for example, that $\varphi(X)$ is an indicator function telling whether a certain system failure occurs or not. Then, $\theta$ is the corresponding failure probability. An upper 95% confidence limit for $\theta$ is $\hat{\theta} + 1.64\sigma/\sqrt{n}$.

Note that informally phrased, a 4 times increase of $n$ only implies a doubling of our knowledge. This observation is a popular way of expressing that $n$ enters through the square root of $n$ in the confidence interval (7). If the variance $\sigma^2$ is unknown we use the usual estimator

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\varphi(X_i) - \hat{\theta})^2.$$

It can be shown that

$$\lim_{n\to\infty} P(\frac{1}{s}\sqrt{n}(\hat{\theta} - \theta) \leq x) = \Phi(x).$$

and we can still use (7), with $\sigma$ replaced by $s$, as an approximative confidence interval.

We can also use (7) to determine $n$. If we wish that $\theta$ is determined with a precision less than $\varepsilon$ we must choose $n$ such that

$$\frac{1.96\sigma}{\sqrt{n}} = \varepsilon \quad \text{or} \quad n = \frac{1.96^2\sigma^2}{\varepsilon^2}. \tag{8}$$

In addition to Monte Carlo integration as described above, a so-called quasi Monte Carlo method exists. The random variables $X_1, X_2, \ldots, X_n$ are chosen more regularly, resulting in a more precise estimate of $\theta$, but, in contrast to ordinary Monte Carlo integration, it is difficult to calculate a confidence interval.

Finally, let us in this section discuss the problem of simulating the probability of a rare event. As an example, let $\varphi(X) = \mathbf{1}\{X > x\}$, such that $\theta = E\varphi(X) = P(X > x)$. For this case $\sigma^2 = \theta(1-\theta)$. If for instance $\theta = 0.01$ it is of no use if the precision of the simulated value is 0.02. The relevant thing here is to require that the uncertainty is small compared to $\theta$. This is called a small relative error. If, for instance, we want that the precision should be $\frac{1}{10}$ of $\theta$, that is, $\varepsilon = \frac{1}{10}\theta$, we find from (8) that $n$ should be

$$n = \frac{1.96^2\theta(1-\theta)}{(\frac{1}{10}\theta)^2} = \frac{(1.96 \cdot 10)^2}{\theta}(1-\theta) \approx \frac{400}{\theta} \text{ for small } \theta.$$

Taking $\theta = 0.01$ we get $n = 40.000$ and taking $\theta = 10^{-6}$ we get $n = 4 \cdot 10^8$. This means that for very small values of $\theta$ it is necessary to find alternative ways of simulating $\theta$.

## 2.7. Variance reduction

A number of techniques are available for improving the precision of Monte-Carlo integration. We will look at one of these in detail, and describe the idea behind a second one.

*2.7.1. Importance sampling*

Let us suppose that we want to calculate

$$\theta = \int_{-\infty}^{\infty} \varphi(x) f(x) dx, \tag{9}$$

where $f$ is the density of a random variable $X$. Let us suppose that $g$ is another density such that

$$g(x) = 0 \Rightarrow \varphi(x)f(x) = 0.$$

Let $\psi(x) = \varphi(x)f(x)/g(x)$. (For $g(x) = 0$, we let $\psi(x) = 0$, say.) Then, we can rewrite $\theta$ as

$$\begin{aligned} \theta &= \int_{-\infty}^{\infty} \varphi(x)f(x)\mathbf{1}\{g(x) > 0\}dx \\ &= \int_{-\infty}^{\infty} \psi(x)g(x)\mathbf{1}\{g(x) > 0\}dx \\ &= \int_{-\infty}^{\infty} \psi(x)g(x)dx. \end{aligned} \tag{10}$$

Hence, if $X_1, X_2, \ldots, X_n$ are independent random variables from the distribution with density $g$, then we can estimate the integral by the unbiased and consistent estimator

$$\hat{\theta}_g = n^{-1} \sum_{i=1}^{n} \psi(X_i) \tag{11}$$

for which the variance is

$$\begin{aligned} \mathrm{Var}(\hat{\theta}_g) &= n^{-1} \int_{-\infty}^{\infty} \{\psi(x) - \theta\}^2 g(x)dx \\ &= n^{-1}[\int_{-\infty}^{\infty} \psi(x)^2 g(x)dx - \theta^2]. \end{aligned} \tag{12}$$

This variance can be very low, much lower than the variance of the estimator $\hat{\theta}$ given in (6), if $g$ can be chosen so that $\psi$ is nearly constant on the set $\{x \in \mathbb{R} : g(x) > 0\}$. A constant $\psi$ corresponds to choosing $g$ as

$$g(x) = \frac{\varphi(x)f(x)}{\int_{-\infty}^{\infty} \varphi(y)f(y)dy}.$$

Let us here give a very simple example. We want to calculate the probability that an exponentially distributed random variable with parameter 1 exceeds $u$

$$\theta(u) = \int_{u}^{\infty} e^{-x}dx = e^{-u}, \quad u > 0. \tag{13}$$

We suppose that $u$ is very large such that $\theta(u)$ is very small. Here, $\theta(u)$ can be expressed as (9) with

$$\varphi(x) = \mathbf{1}\{u < x\}$$

and

$$f(x) = \begin{cases} e^{-x} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

As alternative density $g$ we will use the density of an exponential distribution with parameter $\lambda < 1$,

$$g(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Note that this distribution has mean $1/\lambda > 1$. We will simulate $X_1, X_2, \ldots, X_n$ from $g$ and estimate $\theta(u)$ by $\hat{\theta}_g$ where

$$\psi(x) = \frac{\varphi(x)f(x)}{g(x)} = \lambda^{-1}e^{(\lambda-1)x}\mathbf{1}\{u < x\}.$$

In order to find the variance of $\hat{\theta}_g$, we need to calculate (14). We find

$$\begin{aligned} &\int_{-\infty}^{\infty} \psi(x)^2 g(x)dx \\ &= \int_u^{\infty} \lambda^{-2}e^{2(\lambda-1)x}\lambda e^{-\lambda x}dx \\ &= \frac{1}{\lambda}\int_u^{\infty} e^{-(2-\lambda)x}dx \\ &= \{\lambda(2-\lambda)\}^{-1}e^{-(2-\lambda)u}, \end{aligned}$$

and therefore,

$$\mathrm{Var}(\hat{\theta}_g) = n^{-1}e^{-2u}\{\frac{e^{\lambda u}}{\lambda(2-\lambda)} - 1\}.$$

The relative variance is

$$\frac{\mathrm{Var}(\hat{\theta}_g)}{\theta(u)^2} = n^{-1}\{\frac{e^{\lambda u}}{\lambda(2-\lambda)} - 1\}. \tag{14}$$

A natural choice is then to take $\lambda$ to minimize (14) for a given value of $u$:

$$\lambda(u) = \frac{2 + 2u - \sqrt{(2+2u)^2 - 8u}}{2u} = 1 + \frac{1}{u} - \sqrt{1 + \frac{1}{u^2}} \approx \frac{1}{u},$$

for large $u$, with the corresponding relative variance $\approx \frac{1}{n}\frac{1}{2}e^1 u$. This should be compared to the variance in the case where we simulate from an exponential distribution with parameter $\lambda = 1$ where the relative variance (14) is $\sim \frac{1}{n}e^u$. Note that the best value $\lambda(u) = \frac{1}{u}$ corresponds to choosing $\lambda$ so that the mean value in the corresponding distribution is $u$.

As another example, suppose we want to estimate the probability $P(X > 2)$ where $X$ follows a Cauchy distribution with the density function

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R},$$

so we require the integral

$$\int_{-\infty}^{\infty} \mathbf{1}\{x \in A\} f(x) dx,$$

where $A = \{x \in \mathbb{R} : x > 2\}$. We could simulate from the Cauchy distribution directly and apply (6) with $\varphi(x) = \mathbf{1}\{x \in A\}$, but the variance of this estimator is substantial.

Alternatively, we observe that for large $x$, $f(x)$ is close to proportional to the density $g$ given by

$$g(x) = \begin{cases} 2/x^2 & x > 2, \\ 0 & \text{otherwise.} \end{cases}$$

By inversion, we can simulate from $g$ by letting $X_i = 2/U_i$ where $U_i \sim U[0, 1]$. Thus, our estimator becomes, cf. (11)

$$\hat{\theta}_g = n^{-1} \sum_{i=1}^{n} \frac{X_i^2}{2\pi(1 + X_i^2)},$$

where $X_i = 2/U_i$.

*2.7.2. Control and antithetic variables*

In general, the idea of control variables is to modify an estimator according to a correlated variable whose mean is known. Thus, let us suppose that we wish to estimate $\theta = E(Z)$ where $Z = \varphi(X)$. Let $W = \psi(X)$ be the control variable with known $E(W)$. We suppose that $\varphi(X)$ and $\psi(X)$ are positively correlated and with variances of similar magnitudes. For a sample $X_1, X_2, \ldots, X_n$, we use the estimator

$$\hat{\theta} = n^{-1} \sum_{i=1}^{n} \{Z_i - W_i\} + E(W),$$

where $Z_i = \varphi(X_i)$ and $W_i = \psi(X_i)$. Clearly, $\hat{\theta}$ is an unbiased and consistent estimator of $\theta$, but since

$$\mathrm{Var}(\hat{\theta}) = n^{-1}[\mathrm{Var}(Z) - 2\mathrm{Cov}(W, Z) + \mathrm{Var}(W)],$$

the variance can be low if $\mathrm{Cov}(W, Z)$ is sufficiently large. A typical choice for $W$ is the first terms of a Taylor series expansion of $\varphi(X)$.

Antithetic variables are almost the converse of control variables: we use a variate $Z^*$ which has the same distribution as $Z$, but is negatively correlated with $Z$. Then,

$$\hat{\theta} = n^{-1} \sum_{i=1}^{n} \frac{Z_i + Z_i^*}{2}$$

is an unbiased and consistent estimator of $\theta$, with variance

$$\mathrm{Var}(\hat{\theta}) = n^{-1}\frac{1}{2}\mathrm{Var}(Z)\{1 + \mathrm{Cor}(Z, Z^*)\},$$

where Cor is the notation used for correlation. This constitutes at least a 2 fold reduction in variance, if the correlation is negative. For simple problems, antithetic variables are easily achieved by inversion, since if $Z = F^{-1}(U)$ then $Z^* = F^{-1}(1 - U)$ has the same distribution as $Z$ and can be shown to be negatively correlated with $Z$ for all choices of $F$. Applying this to the estimation of $\theta = \frac{1}{2} - \int_0^2 [\pi(1 + x^2)]^{-1} dx$ in the Cauchy example leads to the estimator

$$\frac{1}{2} - \frac{1}{n}\sum_{i=1}^{n}\{\frac{1}{\pi(1 + U_i^2)} + \frac{1}{\pi(1 + (2 - U_i)^2)}\}$$

where $U_i \sim U[0, 2]$.

# 3. Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is probably 50 years old, and has been both developed and extensively used in physics for the last four decades. However, the most spectacular increase in its impact and influence in statistics and probability has come since the late 80s.

It has now come to be an all-pervading technique in computational stochastics, in particular for Bayesian inference, and especially in complex stochastic systems. A huge research effort is being expended, in devising new generic techniques, in extending the application of existing techniques, and in investigating the mathematical properties of the methods.

## 3.1. An example

The example described in this section originates from statistical physics. In this field, MCMC is used to simulate models for interaction between particles.

Let $S$ be the region where the particles are living. For simplicity, we assume here that $S$ is the unit square in the plane,

$$S = \{(a, b) \in \mathbb{R}^2 : 0 \leq a \leq 1, 0 \leq b \leq 1\}.$$

Let $\sim$ be a symmetric relation on $S$. For instance, $\sim$ may be the distance relation, defined for a pair of points (particles) $x_1, x_2 \in S$ by

$$x_1 \sim x_2 \Leftrightarrow ||x_1 - x_2|| < R. \tag{15}$$

Two points $x_1, x_2$ in $S$ are said to be neighbours if $x_1 \sim x_2$.

The model describes the interaction between a set of $n$ points (particles) in $S$. This set is denoted $x = \{x_1, \ldots, x_n\}$ where $x_i \in S$. Under the model, the probability density of $x$ is

$$f(x) \propto \pi(x) = \gamma^{s(x)}, \quad x = \{x_1, \ldots, x_n\}, \quad x_i \in S, \tag{16}$$

where $s(x)$ is the number of neighbour pairs in $x$ and $\gamma \geq 0$ is a parameter. If $x = \{x_1, \ldots, x_n\}$, then

$$s(x) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1(x_i \sim x_j).$$

This model is called the Strauss model.

The density $f$ is only specified up to a constant of proportionality. The full specification is

$$f(x) = \alpha(\gamma)\gamma^{s(x)}, \tag{17}$$

where

$$\alpha(\gamma) = [\int_S \cdots \int_S \gamma^{s(\{x_1,\ldots,x_n\})} dx_n \cdots dx_1]^{-1} \tag{18}$$

is the normalizing constant of the density. It is complicated to calculate $\alpha(\gamma)$ when $n$ is large. Also, for $\gamma \neq 1$ simple methods such as those resulting in (6) may lead to very inaccurate estimates of $\alpha(\gamma)$. If possible, one should avoid to try to determine $\alpha(\gamma)$.

Let us now discuss the role of the parameter $\gamma$. It can be regarded as an interaction parameter. For $\gamma = 1$, the density $f$ is constant. Using (18), we find

$$f(x) = \alpha(1) = (\frac{1}{\text{area}(S)})^n = 1, \quad x = \{x_1, \ldots, x_n\}, \quad x_i \in S.$$

For $\gamma > 1$, the point patterns $x$ with high probability density $f(x)$ are those with a high number of neighbours $s(x)$. So, for $\gamma > 1$, the model will typically produce clustered point patterns. For $\gamma < 1$, point patterns with a small value of $s(x)$ are preferred, corresponding to regular point patterns where points do not come too close to each other. In the extreme case where $\gamma = 0$, the density (16) is only positive if $s(x) = 0$ so in this case the model will always generate point patterns $x$ with no neighbour pairs. If the distance relation is used, $s(x) = 0$ means that the distance between any pair of points is at least $R$. If we in such a point pattern $x$ place circular disks of radii $R/2$, centered at each point in $x$, then the disks will not overlap. For $\gamma = 0$, the model is called the *hard-core model*.

Using MCMC, it is possible to simulate from the model and get an impression of how point patterns typically look like. It is also possible to estimate $\gamma$, using MCMC, when an actual point pattern $x$ has been observed. The likelihood function is the density (17) regarded as a function of $\gamma$,

$$L(\gamma) = \alpha(\gamma)\gamma^{s(x)}.$$

The maximum likelihood estimate $\hat{\gamma}$ of $\gamma$ is the value of $\gamma$ that maximizes $L(\gamma)$. It can be shown that if $s(x) > 0$, $\hat{\gamma}$ is the unique solution to

$$E_\gamma s(X) = s(x),$$

where $x$ is the observed point pattern and

$$E_\gamma s(X) = \int_S \cdots \int_S s(x)\alpha(\gamma)\gamma^{s(x)} dx_n \cdots dx_1.$$

The mean value cannot be determined explicitly but as we shall see, it can be found by simulation, using MCMC.

### 3.2. The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a general algorithm that can be used to simulate from a density $f$ of an $m-$dimensional random variable. It is only necessary to know $f$ up to a constant of proportionality. We assume that

$$f(x) \propto \pi(x), \quad x \in \mathbb{R}^m,$$

where $\pi$ is known.

The Metropolis-Hastings algorithm generates a Markov chain

$$X_t, t = 0, 1, \ldots,$$

that has an equilibrium distribution with density $f$. Note that $X_t \in \mathbb{R}^m$. At each step $t$ of the algorithm, a new 'candidate' value $Y$ is proposed according to a proposal density $q(y|X_t)$ that may depend on the actual state $X_t$ of the Markov chain. If $Y$ is accepted, then $X_{t+1} = Y$, otherwise $X_{t+1} = X_t$.

Expressed more precisely, the algorithm can be described as follows:

**Algorithm 6**

1. *Initialize $X_0$; set $t = 0$.*
2. *Simulate $Y$ from the proposal distribution with density $q(y|X_t)$.*
3. *Simulate $U \sim U[0, 1]$.*
4. *If $U \leq \alpha(X_t, Y)$, where*

$$\alpha(x, y) = \min(1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}),$$

   *then set $X_{t+1} = Y$, otherwise set $X_{t+1} = X_t$.*
5. *Increment $t$ and go to 2.*

If $m > 1$ and $m$ is a multiple of $n$ ($m = kn$, say), then we can split $X_t$ into $n$ components

$$X_t = (X_{t1}, \ldots, X_{tn})$$

with $X_{ti} \in \mathbb{R}^k, i = 1, \ldots, n$. The method of generating the proposal $Y$ can then be modified as follows. A uniform integer $i$ amongst $1, \ldots, n$ is chosen with probability $1/n$ and a random proposal $V \in \mathbb{R}^m$ is generated such that

$$Y = (X_{t1}, \ldots, X_{t,i-1}, V, X_{t,i+1}, \ldots, X_{tn}).$$

The density $q(y|X_t)$ is replaced by the density $q(v|X_t)$ of the proposal $V$.

Under mild regularity conditions, the Metropolis-Hastings algorithm will produce a Markov chain that has the distribution with density $f$ as equilibrium distribution. We will argue for this statement in the next sections.

We will finish this section by studying how the Metropolis-Hastings algorithm can be used to simulate the Strauss model.

**Example (continued).** Let $X_t = (X_{t1}, \ldots, X_{tn})$ be the $n$ points in $S$ after $t$ iterations of the algorithm. Since $X_{ti} \in S \subset \mathbb{R}^2$, $m = 2n$. For simplicity, we will omit $t$ and write $X$ instead of $X_t$. Furthermore, let us suppose that the proposal $Y$ is obtained by removing a uniform point $W$ from the point pattern $X$ and adding $V$, which is uniform in $S$. So,

$$Y = (X \backslash \{W\}) \cup \{V\}$$

and

$$q(v|x) = \frac{1}{\text{area}(S)} = 1.$$

The acceptance probability becomes

$$\begin{aligned} \alpha(x, y) &= \min(1, \frac{\pi(y)q(w|y)}{\pi(x)q(v|x)}) \\ &= \min(1, \frac{\gamma^{s(y)}}{\gamma^{s(x)}}) \\ &= \min(1, \gamma^{s(y)-s(x)}). \end{aligned}$$

Notice that for $w \in x$

$$s(x) = s(x \backslash \{w\}) + s(x \backslash \{w\}; w), \tag{19}$$

where

$$s(x \backslash \{w\}; w) = \sum_{z \in x \backslash \{w\}} \mathbf{1}\{w \sim z\}.$$

From (19), we get for a point pattern $x$ with $w \in x$ and $v \notin x$,

$$s((x \backslash \{w\}) \cup \{v\}) = s(x \backslash \{w\}) + s(x \backslash \{w\}; v).$$

Therefore, for $x$ with $w \in x$ and $v \notin x$, we get

$$s((x \backslash \{w\}) \cup \{v\}) - s(x) = s(x \backslash \{w\}; v) - s(x \backslash \{w\}; w).$$

This observation is useful when implementing the Metropolis-Hastings algorithm for the Strauss model.

### 3.3. Markov chains

In this section, we will give the basic concepts and results concerning Markov chains that is needed in order to prove that the Metropolis-Hastings algorithm actually works.

In the course *Mathematical Modelling* 2, Markov chains with finite state space have been treated. Here, we study Markov chains with continuous state space ($\mathbb{R}^m$). We will try to be as comprehensive as needed for understanding the simulation algorithms. It is, however, outside the scope of this simulation course to deal with Markov chains with continuous state space in depth.

Let $\{X_t\}_{t=0}^{\infty}$ be a Markov chain on $\mathbb{R}^m$ such that for any $t$ the conditional distribution of $X_t$ given $X_0, \ldots, X_{t-1}$ is the same as the conditional distribution of $X_t$ given $X_{t-1}$. We suppose that $X_t$ has density $f_{X_t}$.

We say that $\{X_t\}_{t=0}^{\infty}$ has the equilibrium distribution with density $f$ provided that for all $x \in \mathbb{R}^m$ and all $A \in \mathcal{B}(\mathbb{R}^m)$

$$P^t(x, A) \to \int_A f(y)dy, \tag{20}$$

for $t \to \infty$, where $y = (y_1, \ldots, y_m)$ and $dy = dy_1 \cdots dy_m$. Here,

$$P^t(x, A) = P(X_t \in A | X_0 = x)$$

is the $t-$step transition probability.

The density $f$ is called invariant for the Markov chain $\{X_t\}_{t=0}^{\infty}$ if

$$X_t \text{ has density } f \Rightarrow X_{t+1} \text{ has density } f.$$

Note that

$$\begin{aligned} P(X_{t+1} \in A) &= \int_{\mathbb{R}^m} P(X_{t+1} \in A | X_t = x) f_{X_t}(x)dx \\ &= \int_{\mathbb{R}^m} P(x, A) f_{X_t}(x)dx, \end{aligned}$$

where $P(x, A) = P^1(x, A)$. Since we also have

$$P(X_{t+1} \in A) = \int_A f_{X_{t+1}}(x)dx,$$

invariance of $f$ is equivalent to

$$\int_A f(x)dx = \int_{\mathbb{R}^m} P(x, A) f(x)dx \text{ for all } A \in \mathcal{B}(\mathbb{R}^m). \tag{21}$$

It can be shown that if $\{X_t\}_{t=0}^{\infty}$ has the equilibrium distribution with density $f$, then $f$ is invariant. To see this, we use the Chapman-Kolmogorov formula

$$P^{t+1}(x, A) = \int_{\mathbb{R}^m} P(y, A)P^t(x, dy). \tag{22}$$

If $\{X_t\}_{t=0}^{\infty}$ has the equilibrium distribution with density $f$, then as $t \to \infty$, the left-hand side of (22) tends to $\int_A f(y)dy$ while the right-hand side of (22) tends to

$$\int_{\mathbb{R}^m} P(y, A)f(y)dy.$$

It follows that (21) is satisfied and $f$ is therefore invariant.

In practice, any Markov chain Monte Carlo algorithm is therefore constructed so that $f$ becomes invariant. In fact, for most MCMC algorithms (including the Metropolis-Hastings algorithm, as we shall see in the next section) reversibility holds, that is

$$\int_B P(x, A)f(x)dx = \int_A P(x, B)f(x)dx \tag{23}$$

for all $A, B \in \mathcal{B}(\mathbb{R}^m)$. Clearly, reversibility implies invariance.

It can be shown that for a time homogeneous Markov chain with invariant density $f$ the transition probabilities converge if the chain in addition is irreducible and aperiodic. By definition, the chain is irreducible if for all $x \in \mathbb{R}^m$ and all $A \in \mathcal{B}(\mathbb{R}^m)$ with

$$\int_A f(y)dy > 0, \tag{24}$$

there exists $t = t(x, A)$ such that $P^t(x, A) > 0$. Moreover, the chain is said to be aperiodic if there are no disjoint sets $A_0, \ldots, A_{d-1} \in \mathcal{B}(\mathbb{R}^m)$ with $d \geq 2$ such that $P(x, A_{j(i)}) = 1$ for all $x \in A_i$ and $i = 0, \ldots, d-1$ where

$$j(i) = i + 1 \mod d.$$

It can be shown that (20) holds for almost all $x$ and all $A \in \mathcal{B}(\mathbb{R}^m)$ if the chain is irreducible and aperiodic and has $f$ as invariant density.

In order to get rid of the nullset, Harris recurrence is needed. This means that for all $x \in \mathbb{R}^m$ and all $A \in \mathcal{B}(\mathbb{R}^m)$ with (24) satisfied, there exists $t = t(x, A)$ such that

$$P(X_t \in A \text{ for some } t = t(x, A) < \infty | X_0 = x) = P^t(x, A) = 1.$$

Clearly, Harris recurrence implies irreducibility.

### 3.4. The Metropolis-Hastings algorithm (continued)

In this section, we will show that by choosing the acceptance probability as described in Algorithm 6, the resulting Markov chain becomes reversible, i.e. (23) is satisfied for all $A, B \in \mathcal{B}(\mathbb{R}^m)$. This in turn implies that $f$ is an invariant density. Irreducibility and aperiodicity must be checked in each separate case.

First, we show that the transition probabilities are of the form

$$P(x, A) = \int_A \alpha(x, y) q(y|x) dy + (1 - p(x)) \mathbf{1}\{x \in A\}, \tag{25}$$

$x \in \mathbb{R}^m$, $A \in \mathcal{B}(\mathbb{R}^m)$, where

$$p(x) = \int_{\mathbb{R}^m} q(y|x) \alpha(x, y) dy.$$

In order to show (25), we use that

$$X_{t+1} = \mathbf{1}\{0 \le U \le \alpha(X_t, Y)\} \cdot Y + \mathbf{1}\{\alpha(X_t, Y) < U \le 1\} \cdot X_t,$$

cf. Algorithm 6. The proposal $Y$ is accepted if $U \le \alpha(X_t, Y)$. We get

$$\begin{aligned}
&P(X_{t+1} \in A, Y \text{ accepted } | X_t = x) \\
&= P(Y \in A, U \le \alpha(x, Y) | X_t = x) \\
&= \int_A P(U \le \alpha(x, y)) q(y|x) dy \\
&= \int_A \alpha(x, y) q(y|x) dy.
\end{aligned}$$

In particular,

$$P(Y \text{ accepted } | X_t = x) = \int_{\mathbb{R}^m} \alpha(x, y) q(y|x) dy = p(x).$$

Therefore,

$$\begin{aligned}
&P(x, A) \\
&= P(X_{t+1} \in A | X_t = x) \\
&= P(X_{t+1} \in A, Y \text{ accepted } | X_t = x) + P(X_{t+1} \in A, Y \text{ not accepted } | X_t = x) \\
&= \int_A \alpha(x, y) q(y|x) dy + (1 - p(x)) \mathbf{1}\{x \in A\},
\end{aligned}$$

and (25) holds.

In order to show reversibility, we also need the following identity

$$f(x)q(y|x)\alpha(x,y) = f(y)q(x|y)\alpha(y,x), \tag{26}$$

which will be shown as an exercise. Using (25) and (26), we finally get

$$\begin{aligned}
&\int_B P(x,A)f(x)dx \\
&= \int_B [\int_A \alpha(x,y)q(y|x)dy + (1-p(x))\mathbf{1}\{x \in A\}]f(x)dx \\
&= \int_B \int_A \alpha(x,y)q(y|x)f(x)dydx + \int_{\mathbb{R}^m} \mathbf{1}\{x \in A\}\mathbf{1}\{x \in B\}(1-p(x))f(x)dx \\
&= \int_A \int_B \alpha(y,x)q(x|y)f(y)dxdy + \int_{\mathbb{R}^m} \mathbf{1}\{x \in A\}\mathbf{1}\{x \in B\}(1-p(x))f(x)dx \\
&= \int_A P(x,B)f(x)dx,
\end{aligned}$$

and the Markov chain defined in Algorithm 6 is thereby reversible.

### 3.5. Monte Carlo integration, using MCMC

As in Section 2.6, the aim of the MCMC simulations is typically to estimate an integral of the form

$$\theta = E\varphi(X) = \int_{\mathbb{R}^m} \varphi(x)f(x)dx.$$

Here, $f$ is the density of an $m-$dimensional random variable $X$ and

$$\varphi : \mathbb{R}^m \to \mathbb{R}.$$

If the Markov chain $\{X_t\}_{t=0}^{\infty}$ is judged to be in equilibrium at time $t_0$, then $\theta$ is estimated by

$$\hat{\theta} = \frac{1}{N}\sum_{t=t_0+1}^{t_0+N} \varphi(X_t),$$

where $N$ is a suitably chosen integer.

Since the $X_t$s are correlated, it is more complicated to evaluate the variance of $\hat{\theta}$ than in the case of independence. Since the Markov chain is time homogenous, $\mathrm{Cov}(\varphi(X_s), \varphi(X_{s+t}))$ does not depend on $s$. Letting

$$\zeta(t) = \mathrm{Cov}(\varphi(X_s), \varphi(X_{s+t})), \tag{27}$$

the variance becomes

$$\begin{aligned}\mathrm{Var}(\hat{\theta}) &= \frac{1}{N^2}\sum_{t_1=t_0+1}^{t_0+N}\sum_{t_2=t_0+1}^{t_0+N}\mathrm{Cov}(X_{t_1},X_{t_2}) \\ &= \frac{1}{N^2}\sum_{t_1=t_0+1}^{t_0+N}\sum_{t_2=t_0+1}^{t_0+N}\zeta(t_2-t_1) \\ &= \frac{1}{N^2}\sum_{t=-N+1}^{N-1}(N-|t|)\zeta(t).\end{aligned}$$

Note that the covariances (27) depend not only on $f$ and $\varphi$ but also on the transition probabilities.

Several possibilities for estimating $\mathrm{Var}(\hat{\theta})$ have been proposed. A main option is time series methods. Furthermore,

$$\lim_{N\to\infty} P([\sum_{t=-\infty}^{\infty}\zeta(t)]^{-1}\sqrt{N}(\hat{\theta}-\theta)\le x) = \Phi(x), \quad x\in\mathbb{R}, \tag{28}$$

which holds under weak regularity conditions. To be more precise, (28) holds if $E\varphi(X)^2<\infty$ and the Markov chain $\{X_t\}_{t=0}^{\infty}$ is so-called geometric ergodic.

Variance reduction techniques are used in connection with MCMC. This is in particular used in the context where the density $f$ is of the form

$$f(x;\gamma) = \alpha(\gamma)\pi(x;\gamma), \quad x\in\mathbb{R}^m,$$

where $\gamma\in\Gamma\subseteq\mathbb{R}^l$ is an unknown parameter, $\pi$ is a known function parametrized by $\gamma$ and $\alpha(\gamma)$ is the normalization constant

$$\alpha(\gamma) = [\int_{\mathbb{R}^m}\pi(x;\gamma)dx]^{-1}.$$

For statistical analysis, it is sometimes necessary to know $\alpha(\gamma)$ (at least up to a constant). One possibility here is to notice that

$$\frac{\alpha(\gamma_0)}{\alpha(\gamma)} = \int_{\mathbb{R}^m}\frac{\pi(x;\gamma)}{\pi(x;\gamma_0)}f(x;\gamma_0)dx = E_{\gamma_0}\frac{\pi(X;\gamma)}{\pi(X;\gamma_0)},$$

where $E_{\gamma_0}X$ indicates that we take mean value of a random variable with density $f(\cdot;\gamma_0)$. We can therefore estimate $\alpha(\gamma_0)/\alpha(\gamma)$ by

$$\frac{1}{N}\sum_{t_0+1}^{t_0+N}\frac{\pi(X_t;\gamma)}{\pi(X_t;\gamma_0)}, \tag{29}$$

where $\{X_t\}_{t=0}^{\infty}$ is a Markov chain with equilibrium density $f(\cdot;\gamma_0)$. If $\gamma$ is far from $\gamma_0$, it is a good idea to define $\gamma_0, \gamma_1, \ldots, \gamma_K$ where $\gamma_{i-1}$ and $\gamma_i$ are close to each other and $\gamma_K = \gamma$. One then use that

$$\frac{\alpha(\gamma_0)}{\alpha(\gamma)} = \prod_{i=1}^{K} \frac{\alpha(\gamma_{i-1})}{\alpha(\gamma_i)},$$

and estimate each factor $\alpha(\gamma_{i-1})/\alpha(\gamma_i)$ separately by the procedure described above.

**Example (continued).** For the Strauss model defined on the unit square $S$, $m = 2n$, $\Gamma = [0, \infty)$ and

$$\pi(x;\gamma) = \begin{cases} \gamma^{s(x)} & x = \{x_1, \ldots, x_n\}, \quad x_i \in S \\ 0 & \text{otherwise.} \end{cases}$$

Here, (29) becomes

$$\frac{1}{N} \sum_{t_0+1}^{t_0+N} (\frac{\gamma}{\gamma_0})^{s(X_t)}, \tag{30}$$

Clearly, unless $\gamma$ and $\gamma_0$ are close, the estimator (30) may have a very large variance.

# 4. Models for point processes

In this section, we will discuss models for finite point patterns $x$ observed in a bounded subset $S$ of the plane. Models of this type are called point process models.

A point process $X$ on $S$ is a random finite set of points in $S$. We let $\mathcal{S}$ denote the set of finite subsets of $S$. The number $n(X)$ of points in $X$ is not necessarily fixed but a random variable.

A famous theorem is the *void probability theorem*, see e.g. Daley & Vere-Jones (1988). This theorem concerns the void (empty set) probabilities, i.e. the probabilities that there are no points in $A$ where $A$ varies over (essentially) all subsets of $S$. To be more precise, $A$ belongs to the Borel subsets $\mathcal{B}(S)$ of $S$ which is a very rich class a sets.

**Theorem 1**. The distribution of a point process $X$ on $S$ is determined by the void probabilities

$$v(A) = P(n(X \cap A) = 0), \quad A \in \mathcal{B}(S). \qquad \square$$

The void probability theorem is a consequence of a deep result in random set theory, related to so-called capacity functionals. As will be apparent in what follows, this theorem is very useful.

### 4.1. The Poisson point process

The homogenous planar Poisson point process is the cornerstone on which the theory of point processes is built. It represents the simplest possible stochastic mechanism for the generation of point patterns, and in applications the process is used as an idealized standard of complete spatial randomness.

The homogeneous Poisson point process $X$ on $S$ with intensity $\lambda > 0$ is defined by

(**P1**) $n(X \cap A) \sim po(\lambda \text{area}(A)), \quad A \in \mathcal{B}(S)$

(**P2**) For $A_1, \ldots, A_k \in \mathcal{B}(S)$ disjoint,
$n(X \cap A_1), \ldots, n(X \cap A_k)$ are independent

Property (**P2**) can be interpreted as spatial randomness or lack of interaction, since the process behaves independently in disjoint regions.

According to (**P1**), the mean number of points in $A$ only depends on the area of $A$ and not on the position of $A$ inside $S$. This is the reason why the process is called homogeneous.

A class of inhomogenous processes is obtained if the constant intensity $\lambda$ is replaced by a variable intensity function $\lambda$ defined on $S$. A Poisson point process $X$ on $S$ with intensity function $\lambda : S \to [0, \infty)$ is defined by

(**P1'**) $n(X \cap A) \sim po(\int_A \lambda(y)dy), \quad A \in \mathcal{B}(S)$

and (**P2**) above.

A Poisson point process has the following property:

**Theorem 2**. Let $X$ be a Poisson point process on $S$ with intensity function $\lambda : S \to [0, \infty)$. Let $A \in \mathcal{B}(S)$. Then, conditionally on $n(X \cap A) = n$, $X \cap A$ is distributed as $\{X_1, \ldots, X_n\}$, where $X_1, \ldots, X_n$ are independent and identically distributed random points in $A$ with density proportional to $\lambda$.

**Proof**. Since $X \cap A$ is a point process on $A$, the distribution of $X \cap A$ is determined by the void probabilities, cf. Theorem 1. Let us for any $B \in \mathcal{B}(S)$ use the short notation

$$\mu(B) = \int_B \lambda(y)dy. \tag{31}$$

Then, for $B \subseteq A$ we get the following void probability

$$\begin{aligned}
&P(n(X \cap B) = 0 | n(X \cap A) = n) \\
&= \frac{P(n(X \cap B) = 0, n(X \cap A) = n)}{P(n(X \cap A) = n)} \\
&= \frac{P(n(X \cap B) = 0, n(X \cap A \backslash B) = n)}{P(n(X \cap A) = n)} \\
&= \frac{e^{-\mu(B)} \cdot e^{-\mu(A \backslash B)} \frac{\mu(A \backslash B)^n}{n!}}{e^{-\mu(A)} \frac{\mu(A)^n}{n!}} \\
&= (\frac{\mu(A \backslash B)}{\mu(A)})^n.
\end{aligned}$$

This agrees with the void probabilities for $n$ independent random points $X_1, \ldots, X_n$ in $A$ with density proportional to $\lambda$, since for such points we have

$$\begin{aligned}
&P(X_1 \notin B, \cdots, X_n \notin B) \\
&= P(X_1 \in A \backslash B, \cdots, X_n \in A \backslash B) \\
&= (\frac{\mu(A \backslash B)}{\mu(A)})^n.
\end{aligned}$$

□

Using the definition of a Poisson point process, it is possible to derive a formula for probabilities associated with the Poisson point process. Let $F$ be an event for the point process. For instance,

$$F = \{x \in \mathcal{S} : n(x) = k\},$$

i.e. $F$ is the event that the point pattern contains $k$ points. For an event $F$, we have with $\mu$ defined as in (31)

$$P(X \in F) = \sum_{n=0}^{\infty} \exp(-\mu(S)) \frac{1}{n!} \int_{S^n} \mathbf{1}\{\{x_1, \ldots, x_n\} \in F\} \prod_{i=1}^{n} \lambda(x_i) dx_1 \cdots dx_n. \tag{32}$$

The proof of this result uses that $n(X) \sim Po(\mu(S))$ and Theorem 2.

The result can be extended to a result for mean values. Let $g$ be a non-negative function defined on $\mathcal{S}$. Then,

$$Eg(X) = \sum_{n=0}^{\infty} \exp(-\mu(S)) \frac{1}{n!} \int_{S^n} g(\{x_1, \ldots, x_n\}) \prod_{i=1}^{n} \lambda(x_i) dx_1 \cdots dx_n.$$

Note that if we let $g$ be the indicator function of the event $F$, then we again obtain (32).

The intensity function of a Poisson point process may depend on explanatory variables. One simple geometric example is an intensity function of the form

$$\lambda(y) = g(d_C(y)), \quad y \in S,$$

where $d_C(y)$ is the distance from $y$ to a reference structure $C \subset S$. For instance, the reference structure may be a point or a planar curve. For statistical purposes, it is a good idea to model $\lambda$ parametrically, for instance using an exponential expression as

$$\lambda(y) = \alpha e^{\theta \cdot \tau(y)}, \quad y \in S,$$

where $\alpha > 0$, $\theta \in \Theta \subseteq \mathbb{R}^l$ and $\tau(y) \in \mathbb{R}^l$.

### 4.2. Markov point processes

In this section, we will define and study Markov point processes which are finite point processes with a particularly simple interaction structure.

We start by defining the concept of a neighbourhood.

**Definition 1.** Given a symmetric relation $\sim$ on $S$, two points $y_1, y_2 \in S$ are called neighbours if $y_1 \sim y_2$. The neighbourhood of a set $A \subseteq S$ is denoted

$$\partial A = \{y \in S : y \sim a \text{ for some } a \in A\}.$$

In particular for $A = \{a\}$, we use the short notation

$$\partial a = \{y \in S : y \sim a\}.$$

Using the concept of a neighbourhood, we can define a Markov point process $X$. Such a process has a density $f$ with respect to the homogeneous Poisson point process with intensity 1. The density is defined on $\mathcal{S}$, the set of finite subsets of $S$. Probabilities can be calculated as

$$P(X \in F) = \sum_{n=0}^{\infty} \exp(-\text{area}(S)) \frac{1}{n!} \int_{S^n} 1(\{x_1, \ldots, x_n\} \in F) \times f(\{x_1, \ldots, x_n\}) dx_1 \cdots dx_n.$$

Conditionally on $n(X) = n$, the density of $X = \{X_1, \ldots, X_n\}$ is proportional to $f(\{x_1, \ldots, x_n\})$.

The definition of a Markov point process is given below. The requirement (**M2**) in the definition is the essential one which concerns 'the conditional intensity of adding an extra point $u$ to the point pattern $x$'.

**Definition 2.** A point process $X$ with density $f$ is a Markov point process with respect to the relation $\sim$ if for all $x \in \mathcal{S}$

(**M1**) $f(x) > 0 \Rightarrow f(y) > 0$ for all $y \subseteq x$

(**M2**) if $f(x) > 0$, then

$$\lambda(u; x) = f(x \cup \{u\})/f(x), \quad u \in S, \quad x \in \mathcal{S}, \quad u \notin x$$

depends only on $u$ and $\partial u \cap x$.

**Example 1. (The Poisson point process)** A Poisson point process with intensity $\lambda > 0$ has the following density

$$f(x) = e^{(1-\lambda)\text{area}(S)} \lambda^{n(x)}, \quad x \in \mathcal{S}.$$

This process is Markov with respect to any relation $\sim$ since $f(x) > 0$ for all $x \in \mathcal{S}$ and $\lambda(u; x) = \lambda$ is constant for all $u$ and $x$ such that $u \notin x$.

**Example 2. (Hard-core model)** Suppose we want to model a pattern of non-overlapping circular discs with fixed diameter $R > 0$. Then no disc centre can be closer that $R$ to another disc centre. Assuming no other interactions occur, a density could be

$$f(x) = f(\{x_1, \ldots, x_n\}) = \alpha \beta^n \mathbf{1}\{||x_i - x_j|| \geq R, i \neq j\}, \alpha, \beta > 0.$$

This model is called a *hard-core model*.

Let $\sim$ be the symmetric relation on $S$ given by

$$y_1 \sim y_2 \Leftrightarrow ||y_1 - y_2|| < R.$$

The hard-core model is then Markov with respect to this relation.

Thus, suppose that $f(x) > 0$. We then have $||x_i - x_j|| \geq R$, for all $i \neq j$, i.e. $x$ does not contain points closer than $R$ together. If $y \subseteq x$, then also $y$ does not contain points closer than $R$ together, hence $f(y) > 0$.

Also, (**M2**) is fulfilled, since for $u \in S$ and $x = \{x_1, \ldots, x_n\} \in \mathcal{S}$ such that $u \notin x$

$$\begin{aligned}
&\lambda(u; x) \\
&= \frac{\alpha\beta^{n+1}\mathbf{1}\{||x_i - x_j|| \geq R, i \neq j)\}\mathbf{1}\{||x_i - u|| \geq R, i = 1, \ldots, n\}}{\alpha\beta^n\mathbf{1}\{||x_i - x_j|| \geq R, i \neq j\}} \\
&= \beta \cdot \mathbf{1}\{||x_i - u|| \geq R, i = 1, \ldots, n\} \\
&= \beta \cdot \mathbf{1}\{\partial u \cap x = \emptyset\}.
\end{aligned}$$

The density of a Markov point process can be factorized in a simple manner as described in the famous Hammersley-Clifford theorem. An important concept is here the cliques.

**Definition 3.** A pattern $x \in \mathcal{S}$ is called a clique if all members of $x$ are neighbours, i.e. $u \sim v$ for all $u, v \in x$. By convention, sets of 0 and 1 points are cliques. The set of cliques is denoted $\mathcal{C}$.

The Hammersley-Clifford theorem gives a factorization of a Markov density in terms of interactions which are only allowed between elements in cliques.

**Theorem 3. (Hammersley-Clifford)** A density $f$ defines a Markov point process with respect to $\sim$ if and only if there exists a function $\varphi : \mathcal{S} \to [0, \infty)$ such that $\varphi(x) = 1$ unless $x \in \mathcal{C}$ and such that

$$f(x) = \prod_{y \in \mathcal{S}: y \subseteq x} \varphi(y)$$

for all $x \in \mathcal{S}$. The function $\varphi$ is called the clique interaction function.

We will not prove the theorem here, but just mention that a lengthy but rather elementary proof can be constructed, based on induction. The Hammersley-Clifford theorem is important, first of all because it gives a way of breaking up a high-dimensional joint distribution in managable clique interactions that are easier to interpret and have lower dimension. It also provides a natural way to construct parametric models for Markov point processes.

**Example 3. (The Strauss process)** The Strauss process is the Markov

point process with interaction function

$$\varphi(x) = \begin{cases} \alpha & n(x) = 0 \\ \beta & n(x) = 1 \\ \gamma & n(x) = 2, x = \{x_1, x_2\}, x_1 \sim x_2, \end{cases}$$

and $\varphi(x) = 1$ otherwise, where $\alpha, \beta, \gamma > 0$. Using the Hammersley-Clifford theorem, the density of the Strauss process becomes

$$f(x) = \alpha\beta^{n(x)}\gamma^{s(x)}, \quad x \in \mathcal{S}, \tag{33}$$

where $s(x)$ is the number of neighbour pairs in $x$. If we condition on the number of points in $x$, then we get the Strauss model described in Part 1, Section 3.1. The density $f$ is well-defined if

$$\sum_{n=0}^{\infty} \frac{1}{n!} \int_{S^n} \beta^n \gamma^{s(\{x_1,\ldots,x_n\})} dx_1 \cdots dx_n < \infty. \tag{34}$$

It can be shown that (34) holds if $\gamma \leq 1$ while for $\gamma > 1$, the sum in (34) may be infinite. Thus, there does not in general exist a Strauss process for $\gamma > 1$.

**Example 4. (The area-interaction process)** In this example, we consider an alternative to the Strauss process which is called the *area-interaction process*. Suppose that circular discs of radius $R$ are allowed to overlap, and we want the conditional intensity $\lambda(u; x)$ to depend on the area added by the new circular disc with centre $u \in S$. A natural choice for the intensity is then

$$\lambda(u; x) = \beta\gamma^{-\text{area}(B(u,R)\setminus U_x)}, \quad u \in S,$$

where $B(u, R)$ is a circular disc with centre $u \in S$ and radius $R$, and $U_x$ is a short notation for $\cup_{v\in x} B(v, R)$. The model parameters satisfy $\beta, \gamma > 0$.

If $\gamma < 1$, $\lambda(u; x)$ is large when the added area is large, resulting in regular patterns. Similarly, for $\gamma > 1$, realizations tend to be clustered. For $\gamma = 1$, we reobtain a Poisson process.

From the actual form of the conditional intensity, we can derive the form of the corresponding density. For $x = \{x_1, \ldots, x_n\}$, we get

$$\begin{aligned} f(x) &= f(\emptyset)\lambda(x_1; \emptyset)\lambda(x_2; \{x_1\}) \cdots \lambda(x_n; x\setminus\{x_n\}) \\ &= f(\emptyset)\beta^n\gamma^{-\sum_{i=1}^n \text{area}(B(x_i,R)\setminus U_{x_1,\ldots,x_{i-1}})} \\ &= \alpha\beta^n\gamma^{-\text{area}(U_x)}, \quad \alpha = f(\emptyset). \end{aligned} \tag{35}$$

It can be shown that (35) really defines a density, because

$$\sum_{n=0}^{\infty} \frac{1}{n!} \int_{S^n} \beta^n \gamma^{-\text{area}(U_{\{x_1,\ldots,x_n\}})} dx_1 \ldots dx_n < \infty.$$

The process in (35) is Markov with respect to the relation

$$y_1 \sim y_2 \Leftrightarrow ||y_1 - y_2|| < 2R.$$

To see this, note first that $f(x) > 0$ for all $x \in \mathcal{S}$, so (**M1**) is satisfied. Next, note that $\lambda(u; x)$ depends only on $u$ and the points in $x$ closer than $2R$ to $u$. So (**M2**) is satisfied.

## References


1. Coles, S., Roberts, G. and Jarner. S. (2001) *Computer Intensive Methods.* Lecture Notes.

2. Jensen, J.L. (2001) *Stochastic Simulations: Concepts and Applications.* Department of Theoretical Statistics, University of Aarhus.