Stochastic Simulation
Script for the course in spring 2012

Hansruedi Künsch
Seminar für Statistik, ETH Zurich

June 2012

Contents

Chapter 1

Introduction and Examples

1.1 What is Stochastic Simulation?

(stochastic) Simulation = Implementation of a (stochastic) system on a computer for investigating the properties of the system.

Simulation is therefore different from both a mathematical analysis of a system and a real world experiment or observation study based on data.

The common point with a real experiment is

the empirical approach (measuring or counting something)

The commom point with a mathematical analysis is

the use of a mathematical model to represent reality

The advantages of simulations over a real experiment are the savings of time and money and the possibility to change the parameters of a system easily.

The advantage of simulations over a mathematical analysis is the possibility to use complex and thus more realistic models which cannot be handled with current mathematical techniques. In particular, one has not to rely on asymptotic approximations.

In the following sections of this chapter, we will illustrate the range of problems which can be solved with the help of stochastic simulations. But let us first give a general mathematical description of a stochastic simulation. We consider a system which consists of random input variables $\mathbf{X} = (X_1, \ldots, X_p)$ and a deterministic function h which maps inputs into outputs $\mathbf{Y} = (Y_1, \ldots, Y_q) = h(\mathbf{X})$. Here, p may be large and the function h can be quite complicated, but the distribution of $\mathbf{X}$ and h are supposed to be known. The goal is to obtain information about the distribution of the output $\mathbf{Y}$. If we can draw samples of $\mathbf{X}$ on a computer (something you will learn in this course), then we can easily obtain draws from the output distribution. We simply draw a sample of size N of $\mathbf{X}$, i.e.

$$\mathbf{X}_i = (X_{i1}, \ldots, X_{ip}) \quad i = 1, \ldots, N$$

and compute the corresponding output

$$\mathbf{Y}_i = h(X_{i1}, \ldots, X_{ip}) \quad i = 1, \ldots, N.$$

The law of large numbers justifies then the use of approximations like

$$\mathbb{E}(Y_1) \approx \frac{1}{N} \sum_{i=1}^{N} Y_{i1}, \quad \mathbb{P}(Y_1 \leq c) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{Y_{i1} \leq c\}}.$$

The technique of approximating expected values by sample averages of simulated random variables is also called the *Monte Carlo method.*

1.2 Distribution of Estimators and Test Statistics

1.2.1 Precision of the trimmed mean

Van Zwet (Statistica Netherlandica, 1985) reports a historical example of a simulation before the age of computers:

"In the issue of May 20, 1942, of the Bulletin of the Astronomical Institutes of the Netherlands, E. Hertzsprung, director of the Observatory at Leiden, describes a sampling experiment to determine the variance of the trimmed mean. In connection with the determination of relative proper motions of stars in the Pleiades, Hertzsprung discusses how one should assign weights to the observed values to account for differences in quality of the observations. He writes: "The simplest way to deal with exorbitant observations is to reject them. In order to avoid special rules for onesided rejection the easy way of symmetrical rejection of the largest deviations to each side may be considered. The first question is then: How much is, in the case of Gaussian distribution of errors, the weight of the result diminished by a priori symmetrical rejection of outstanding observations? As the mathematical treatment of this question appears to be laborious beyond the needs mentioned above I gave preference to an empirical answer. On each of 12534 slips of paper was written with two decimals a deviation from zero in units of the mean error, in such a way that these deviations showed a Gaussian distribution. Thus 50 slips were marked with .00, 50 with +.01, 50 with -.01 etc.. Of these slips somewhat more than 1000 times 24 were picked out arbitrarily. Such 24 slips were in each case arranged according to the size of the deviation and the mean squares of the sums of $24 - x$ deviations calculated after symmetrical rejection of $x = 0, 2, 4, \ldots, 22$ extreme values."

This paragraph should warm a statistician's heart, except that he may feel slightly uneasy about "somewhat more than 1000" replications. And he has reason to feel uneasy: "Of all these samples of 24 exactly 1000 were picked out in such a way that the sum of all 24 deviations ($x = 0$) fairly well showed a Gaussian distribution with a mean square of 24." From a theoretical point of view, this ruins a perfectly good sampling experiment, as Van Dantzig was quick to point out, especially since no further information is supplied. There is no way of assessing the accuracy of the estimated variances any more. On the other hand, if we assume that this data cleaning was done sensibly, there seems to be no reason, a priori, why the estimates should be much worse than they would have been otherwise."

Let us formulate this problem and the solution by Hertzsprung mathematically. We consider the trimmed mean

$$\bar{X}_n^{(k)} = \frac{1}{n-2k} \sum_{j=k+1}^{n-k} X_{(j)},$$

where the X_i are i.i.d. $\sim \mathcal{N}\left(\mu, \sigma^2\right)$ and $X_{(j)}$ denotes the j-th element of the ordered sample. One would like to know how much larger the variance $\sigma(n,k)^2$ of the trimmed mean is compared to the variance $\sigma(n,0)^2 = \sigma^2/n$ for $k = 1, 2, \ldots$ and $n = 24$. Without loss of generality, we can assume $\mu = 0$ und $\sigma = 1$.

This fits into our general framework: The inputs are $(X_1, \ldots, X_n)$, they are i.i.d. standard normal, and the function h is the trimmed mean. We would thus generate a $N \times n$ matrix of independent standard normal random variables X_{ij}, compute the trimmed mean $\bar{X}_{n,i}^{(k)}$ of each row and then use

$$n\sigma(n,k)^2 \approx \frac{n}{N} \sum_{i=1}^{N} (\bar{X}_{n,i}^{(k)})^2.$$

At the time of Hertzsprung, they had to generate the X_{ij} from a discretization of the normal distribution: The real line was partitioned into equispaced intervals centered at $x_k = k \cdot 0.01$. Then an urn was made which contained M slips, each value x_k appearing approximately

$$M(\Phi(x_k + 0.005) - \Phi(x_k - 0.005)) \approx 0.01 \cdot M\phi(x_k)$$

times on a slip. Random draws from this urn have then approximately a standard normal distribution. If one wants the value 0 to appear 50 times, one should take $M \approx 12'533.2$. In order to have a symmetric composition of the urn, they chose an even M.

Note that Hertzsprung used a slightly different approximation for the relative increase in variance due to trimming, namely

$$\frac{\sum_{i=1}^{N} (\bar{X}_{n,i}^{(k)})^2}{\sum_{i=1}^{N} (\bar{X}_{n,i}^{(0)})^2}.$$

This estimate does not use the well-known fact that $\sigma(n,0)^2 = 1/n$ which seems strange at first glance. One can however show that this improves the precision of the approximation because numerator and denominator are strongly correlated. We will discuss this in more detail in section 3.9.2 below.

Hertzsprung's procedure to eliminate certain rows of X_{ij} in order to make the untrimmed mean more normally distributed is doubtful. In particular, it destroys the possibility to estimate the precision of the approximation, and it is hard to decide whether one has done too much adjustment.

Nowadays, this problem can be solved easily by asymptotic arguments. In fact, even at Hertzsprung's time, one student of him, Van de Hulst hated the labor involved in the simulation by hand and looked instead for an analytic answer. He succeeded, and showed his result to van Dantzig, who was the pioneer of statistics in Holland at that time. But van Dantzig mainly criticized the lack of rigor in the proof and failed to do justice to van de Hulst's achievement.

Formulated in modern mathematical language, the result is the following

Theorem 1.1. *If X_i i.i.d. $\sim f(x)dx$ and $f(x) = f(-x)$, then for $n \to \infty$ and $\frac{k}{n} \to \alpha$*

$$\mathbb{P}(\sqrt{n}\bar{X}_n^{(k)} \leq x) \to \Phi\left(\frac{x}{\sigma_\alpha}\right)$$

where $\sigma_\alpha^2 = \frac{1}{(1-2\alpha)^2} \int \min\left(x^2, a^2\right) f(x)dx$ and $a = F^{-1}(1-\alpha)$.

Proof. See any textbook on mathematical statistics. □

This means that $\sigma(n,k)^2 \approx \sigma^2_{k/n}/n$, and one can compute σ^2_α easily in the case f is the standard normal density.

Even nowadays, simulations and asymptotics are the two main tools to approximate the distribution of any kind of statistical estimator and thus to compare the properties of competing estimators of the same parameter. Similarly, one can use simulations and asymptotics in order to derive critical values or power functions of statistical tests. Both methods have advantages and disadvantages. Asymptotics usually shows how results depend for instance on the assumed distribution of the observations whereas simulations can only cover a few cases which have to be carefully chosen. On the other hand, asymptotics relies on the consideration of limits which may have little to do with the sample size of interest. Analytical error bounds are typically both very difficult and too pessimistic. Because of this, simulations usually complement asymptotic arguments in order to obtain an idea how different the distribution for a given sample size is from the asymptotic limit distribution.

1.2.2 Bootstrap

In the previous subsection we assumed 1.2.1 the distribution of X_i to be known. This makes sense if one wants to understand the advantages and disadvantages of the trimmed mean compared to the arithmetic mean. Because it is known that the latter is optimal under the assumption of normality, one wants to know how much precision the trimmed mean looses in this case. Because the answer is “only little” and because the trimmed mean offers in addition protection against outliers, the conclusion is that one should rather use the trimmed mean.

Things are different if one wants to use the standarad deviation $\sqrt{\mathrm{Var}\,(T_n)}$ of an estimator T_n to assess uncertainty about the true value, for instance in the form of an approximate 95% confidence interval a $T_n \pm 2\sqrt{\mathrm{Var}\,(T_n)}$. In this case, one would like to avoid assuming the distribution of the X_i to be known. Instead, we estimate also the distribution of the X_i from the same data which were used to compute the estimator T_n. Such a procedure is called *bootstrap* because of the double use of the data.

The empirical distribution $\widehat{F}_n$ of the observations is the natural estimator of the underlying distribution if one does not assume that it belongs to some parametric family like Normal or Gamma. The empirical distribution is a discrete distribution which puts mass $\frac{1}{n}$ on every observed value x_i $(i = 1, \ldots, n)$. If X_i is univariate, the distribution function of $\widehat{F}_n$ is a step function In any case, if $X^* \ \sim \ \widehat{F}_n$, then

$$\mathbb{P}(X^* = x_j) = \frac{1}{n} \quad (j = 1, \ldots, n).$$

(We write X^* instead of X to distinguish it from the random variable with distribution F).

The situation that we consider here can be summarized as follows

- We assume $X_1, \ldots, X_n$ to be i.i.d. random vectors with unknown distribution F.
- We use an estimator $T_n = t_n(X_1, \ldots, X_n)$, $t_n : \mathbb{R}^{np} \to \mathbb{R}$ to estimate a parameter of the underlying distribution F (e.g. we estimate the true correlation matrix of X_i by the empirical correlation matrix of the data).

- We estimate the standard error $\sigma(T_n; F) = \sqrt{\mathrm{Var_F}\left(t_n(X_1, \ldots, X_n)\right)}$ of T_n by $\sigma(T_n; \widehat{F}_n)$

Since we usually cannot compute $\sigma(T_n; \widehat{F}_n)$ analytically, we use simulation instead. The procedure is like in the Hertzsprung example, but we draw our samples from the empirical distribution $\widehat{F}_n$ instead of some assumed distribution F. If we had no computer, then our urn would contain n slips with the n values x_i which were actually observed.

The bootstrap algorithm is thus the following:

Algorithm 1.1.

1. *Draw $n \cdot N$ times with replacement from the observations $(x_1, \ldots x_n)$ and put the draws into a matrix (X^*_{ij}; $1 \leq i \leq N$, $1 \leq j \leq n$)*
2. *Compute the function t_n for every row $T^*_{n,i} = t_n(X^*_{i1}, \ldots, X^*_{in})$*
3. *Approximate $\sigma^2(T_n; \widehat{F}_n)$ and thus also $\sigma(T_n; F)^2$ by*

$$\frac{1}{N-1} \sum_{i=1}^{N} (T^*_{n,i} - \bar{T}^*_n)^2 \text{ where } \bar{T}^*_n = \frac{1}{N} \sum_{k=1}^{N} T^*_{n,k}.$$

1.3 Simulation in Bayesian Statistics

We need here the concept of conditional distributions in the (absolutely) continuous case which we recall first briefly.

1.3.1 Conditional distributions of continuous random vectors

Let $\mathbf{X} = (X_1, X_2)$ be a two-dimensional random vector with joint density f, i.e. for any (measurable) subset $A \subseteq \mathbb{R}^2$ we have

$$\mathbb{P}((X_1, X_2) \in A) = \int_A f(x_1, x_2) dx_1 dx_2$$

Heuristically, the density is the probability of a small rectangle divided by the area of the rectangle

$$\mathbb{P}(x_1 \leq X_1 \leq x_1 + dx_1, x_2 \leq X_2 \leq x_2 + dx_2) = f(x_1, x_2) dx_1 dx_2.$$

If such a joint density exists, we call $\mathbf{X}$ absolutely continuous.

The *marginal* densities are then

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2, \quad f_{X_2}(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1.$$

We thus accumulate the mass of the density along one of the two coordinate axes. We call X_1 and X_2 *independent* if

$$f(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2).$$

Under independence, the two marginal densities thus determine the joint density. In general this is not true. There are many joint densities which have the same marginal densities.

If $\mathbf{X}$ is absolutely continuous, $\mathbb{P}(X_i = x) = 0$ for any x. Hence we cannot define conditional probabilities given $X_i = x$ by the well known formula from discrete probability theory. Nevertheless, we call

$$f_{X_2|X_1}(x_2 \mid x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)}$$

the *conditional* density of X_2 at x_2 given $X_1 = x_1$. Heuristically, the left hand side is

$$\frac{\mathbb{P}(x_2 \leq X_2 \leq x_2 + dx_2 \mid x_1 \leq X_1 \leq x_1 + dx_1)}{dx_2},$$

and this can be justified rigorously as a limit if f and f_{X_1} are continuous. Note that the conditional density of X_2 given $X_1 = x_1$ is obtained as the restriction of the joint density along the line $X_1 = x_1$, normalized such that the total mass is one. Thus we can also write

$$f_{X_2|X_1}(x_2 \mid x_1) \propto f(x_1, x_2)$$

(proportional means up to a factor which depends on x_1, but not on x_2 – x_1 is considered fixed because it is the value on which we condtion).

The above formula can be used in two directions: If we know the joint density, then we can compute the marginal and the conditional densities. On the other hand, we can choose one marginal and one conditional density arbitrarily up to positivity and total mass one and then compute the joint density as the product. Using each of the two directions once, we obtain Bayes formula

$$f_{X_1|X_2}(x_1 \mid x_2) = \frac{f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1)}{\int_{-\infty}^{\infty} f_{X_2|X_1}(x_2 \mid x_1') f_{X_1}(x_1') dx_1'} \propto f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1).$$

This is *Bayes formula* in the absolutely continuous case.

1.3.2 Introduction to Bayesian statistics

Statistics usually assumes that the observations were obtained as realizations of random variables whose distribution is not fully known, but depend on an unknown parameter θ. The parameter could be infinite dimensional, but we consider here only cases where θ belongs to an open subset $\Theta \subseteq \mathbb{R}^p$. Moreover, we assume that for any θ the distributions have densities with respect to some measure μ which for our purposes here is just the Lebesgue measure.

$$\mathbf{X} = (X_1, \ldots, X_n) \sim p_\theta(\mathbf{x}) dx.$$

The parameter θ is unknown to us, and we want to obtain information about it from the observations, In the frequentist approach to statistics, θ has an unknown, but fixed value, whereas in the Bayesian approach θ is also considered as a random variable. Although we usually cannot sample different values of θ, the interpretation as a random variable makes sense if we interprete probabilities as expressions of (subjective) beliefs how likely different values are.

If we accept the idea of putting probabilities on the possible values of θ, then there are two distributions for θ: The prior α which describes our beliefs about possible values of θ before we have seen the data, and the posterior which describes the beliefs after we have seen the data. The two distributions are then connected through Bayes formula: We interprete $p_\theta(\mathbf{x})$ as the conditional density of $\mathbf{X}$ given θ and the prior density $\alpha(\theta)$ as the marginal density. Thus the joint density of $(\theta, \mathbf{X})$ is

$$\alpha(\theta)p_\theta(\mathbf{x}).$$

Bayes formula then tells us that the posterior density of θ which is nothing else than the conditional density of θ given $\mathbf{X} = \mathbf{x}$ is equal to

$$\alpha(\theta|\mathbf{x}) = \frac{\alpha(\theta)p_\theta(\mathbf{x})}{\int_\Theta \alpha(\theta')p_{\theta'}(\mathbf{x})d\theta'} \propto \alpha(\theta)p_\theta(\mathbf{x})$$

In words: The posterior is proportional to the product of prior and likelihood (where proportional means up to factors which may depend on $\mathbf{x}$, but not on θ).

Example 1.1. *Let $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}\left(\theta, \sigma^2\right)$, with σ^2 known and θ unknown. Thus the likelihood is*

$$p_\theta(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right).$$

As our prior for θ we choose a $\mathcal{N}\left(\xi, \kappa^2\right)$-distribution, that is

$$\alpha(\theta) = \frac{1}{\sqrt{2\pi}\kappa} \exp\left(-\frac{1}{(2\kappa^2)}(\theta - \xi)^2\right)$$

How we should choose the "hyperparameters" ξ and κ is one of the main difficulties in Bayesian statistics, but we do not discuss it here.

The joint density is then

$$\alpha(\theta, \mathbf{x}) = \frac{1}{(2\pi)^{\frac{n+1}{2}}\sigma^n\kappa} \exp\left(-\frac{1}{2\kappa^2}(\theta - \xi)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right)$$

and this is proportional to the posterior. In order to derive the posterior, we have to consider only those factors which contain θ. By completing the square, we obtain

$$\begin{aligned} \alpha(\theta|\mathbf{x}) &\propto \exp\left(-\frac{1}{2\kappa^2}(\theta - \xi)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right) \\ &\propto \exp\left(-\frac{1}{2\nu^2}\left(\theta - \frac{\nu^2}{\kappa^2}\xi - \frac{\nu^2}{\sigma^2}n\bar{\mathbf{x}}\right)^2\right) \end{aligned}$$

where ν^2 is the harmonic mean of κ^2 and σ^2/n:

$$\frac{1}{\nu^2} = \frac{1}{\kappa^2} + \frac{n}{\sigma^2} \quad \Leftrightarrow \quad \nu^2 = \frac{\kappa^2\,\sigma^2}{n\kappa^2 + \sigma^2}.$$

Since the last expression is up to a constant a normal density, we have found that the posterior is again a normal density, but with expectation

$$\mu(\mathbf{x}) = \frac{\nu^2}{\kappa^2}\,\xi + \frac{\nu^2 n}{\sigma^2}\,\bar{\mathbf{x}} = \frac{\sigma^2}{\sigma^2 + n\kappa^2}\,\xi + \frac{n\kappa^2}{\sigma^2 + n\kappa^2}\,\bar{\mathbf{x}}$$

and standard deviation ν. Note that the expectation of the posterior is a convex combination of the prior expectation and the maximum likelihood estimate (MLE) of θ. It thus represents a compromise between these two pieces of information. The weighting relates the prior variance to the variance of the MLE.

What can we do with the posterior ? As a point estimate of a function $g(\theta)$ we can use for instance the posterior mean

$$\mathbb{E}[g(\theta) \mid \mathbf{X} = \mathbf{x}] = \int_{\Theta} g(\theta)\alpha(\theta|\mathbf{x})d\theta$$

or the posterior median of $g(\theta)$. Instead of a frequentist confidence interval a Bayesian would use a credible interval, that is any interval $I(\mathbf{x})$ such that

$$\mathbb{P}(g(\theta) \in I(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}) = 1 - \gamma.$$

In the example above, $\mu(\mathbf{x}) \pm \Phi^{-1}(1-\gamma/2)\nu$ would be such an interval. Note however the different interpretation: $\mathbf{x}$ is fixed, and $1-\gamma$ is our belief on the basis of the data and the prior that the invterval contains $g(\theta)$.

Another use of the posterior concerns prediction of new observations. If Y is independent of $\mathbf{X}$, but its distribution has the same value of the parameter θ, then

$$\mathbb{P}(Y \in B \mid \mathbf{X} = \mathbf{x}) = \int \mathbb{P}(Y \in B \mid \theta)\alpha(\theta \mid \mathbf{x})d\theta = \int_B \int_{\Theta} p_{Y|\theta}(y)\alpha(\theta \mid \mathbf{x})d\theta d\mathbf{y}.$$

In contrast to a simple plug-in rule $\mathbb{P}(Y \in B \mid \widehat{\theta})$ where $\widehat{\theta}$ is a point estimate like the MLE, the above prediction interval takes the uncertainty about the parameter θ into account.

In the above example, let us assume that Y is also $\mathcal{N}\left(\theta, \sigma^2\right)$-distributed. Then

$$\mathbb{P}(Y \le b \mid \mathbf{X} = \mathbf{x}) = \int_{-\infty}^{b} \int_{\mathbb{R}} \frac{1}{\sigma^2}\phi\left(\frac{y-\theta}{\sigma}\right)\frac{1}{\nu}\phi\left(\frac{\theta-\mu(\mathbf{x})}{\nu}\right) d\theta dy.$$

Because the convolution of two normal densities is again normal, we obtain

$$\mathbb{P}(Y \le b \mid \mathbf{X} = \mathbf{x}) = \Phi\left(\frac{b-\mu(\mathbf{x})}{\sqrt{\sigma^2+\nu^2}}\right).$$

In the example above, we could do all computations analytically. The reason for this is that we have chosen the prior such that the posterior belongs to a standard distribution family where moments or the distribution function are readily available. If we had used as prior for instance a Cauchy distribution, then the posterior would not belong to any of the standard distribution families. Because the parameter is one-dimensional, we still could study the shape of the posterior by plotting $\alpha(\theta)p_\theta(\mathbf{x})$ and do the integrations numerically. In many applications, the parameter θ is however high-dimensional and the posterior does not belong to a standard family of distributions. In high-dimensions numerical integration is difficult. Therefore, in such cases simulation is usually the easiest way to approximate marginal posterior distributions, posterior expectations, credible intervals and prediction intervals. We will see examples later on.

1.4 Simulations in Statistical Mechanics

We consider a so-called spin system which consists of magnetic particles arranged on a lattice $L = \{1, 2, \ldots, n\}^d$ with two possible values ±1 for the spin. The possible states (configurations) $\mathbf{x}$ of the system are therefore elements of $\Omega = \{\pm 1\}^L$. For physical reasons, we use a so-called *Gibbs distribution* on Ω:

$$\pi(\mathbf{x}) \;=\; \frac{1}{Z(T)} \exp\left(-\frac{1}{T}\sum_{i \neq k} J_{ik}x_i x_k\right).$$

Here $Z(T)$ is a normalizing constant, T is the absolute temperature and $J_{ik} = J_{ki}$ denotes the interaction between particles at sites i and k. If $x_i = x_k$, the probability p contains the term $\exp(-J_{ik})$, and if $x_i \neq x_k$ it contains the term $\exp(J_{ik})$, that is the sign of J_{ik} tells whether there is a preference for equal or opposite spins at sites i and k, and the absolute value of J_{ik} expresses the strength of this preference. If all $J_{ik} \leq 0$, we speak of *ferromagnetic interaction*. In any case, the distribution π is symmetric with respect to an interchange of the spin at all sites.

In physics $\sum_{i \neq k} J_{ik} x_i x_k$ is the energy of the configuration. Hence $\log p(x)$ is a constant minus energy divided by temperature. As $T \to 0$, π converges to the uniform distribution on the configurations with minimal energy. As $T \to \infty$ $\pi(\mathbf{x})$ converges to a constant, that is to the uniform distribution on all configurations.

The simplest special case which already shows interesting features is the so-called Ising model (studied first by Ising in 1924) where

$$J_{ik} = \begin{cases} 0 & \text{falls } \|i-k\| \neq 1 \\ -1 & \text{falls } \|i-k\| = 1. \end{cases}$$

Ising was interested in the distribution of the so-called average magnetization

$$M_n = \frac{1}{n^d} \sum_{i \in L} X_i$$

for large n. It is always symmetric about zero. For large fixed n and $T = \infty$, $n^{d/2} M_n$ is approximately standard normal by the central limit theorem. Also in the limit $T \to 0$, for fixed n, $M_n = \pm 1$ with probability $\frac{1}{2}$ each. For some values T in between, the distribution will thus be bimodal. The question is what happens if we keep T fixed and let n go to infinity: Will the two modes disappear, or will there be a bimodal and thus non-normal limit distribution ? If the two modes remain, we call this *spontaneous magnetization.* For $d = 1$ one can show rather easily that there is no spontaneous magnetization for any $T > 0$, a result also obtained by Ising. Moreover, he gave a plausibility argument that this is the case also for $d = 2$. Unfortunately, this is wrong: In 1936, Peierls showed that for $d \geq 2$, there is a so-called phase transition: There is some critical temperature T_c such that for $T > T_c$ there is no spontaneous magnetization whereas for $T < T_c$ spontaneous magnetization occurs. For $d = 2$, Onsager in 1944 was also able to compute T_c.

For more complicated systems, analytical results are very difficult to obtain, and physicists often use simulation instead in order to formulate conjectures and to verify plausibility arguments.

How to simulate from a Gibbs distribution is immediately obvious because the space Ω is huge (although finite !). The easiest simulation algorithms is based on the fact that the conditional distribution of a spin at a single site given all the other spins can be easily computed. We can write for any $i \in L$:

$$\sum_{k \neq l} J_{kl} x_k x_l = 2x_i \sum_{k \neq i} J_{ik} x_k + \sum_{\ell \neq k \neq i} J_{\ell k} x_\ell x_k =: 2A_i x_i + B_i$$

where A_i and B_i do not contain x_i. Then

$$\begin{aligned} \mathbb{P}(X_i = +1 | X_k,\, k \neq i) &= \frac{\exp(-\frac{1}{T}(2A_i + B_i))}{\exp(-\frac{1}{T}(2A_i + B_i)) + \exp(-\frac{1}{T}(-2A_i + B_i))} \\ &= \frac{\exp(-\frac{1}{T}2A_i)}{\exp(-\frac{1}{T}2A_i) + \exp(+\frac{1}{T}A_i)} \end{aligned}$$

If most $J_{ik} = 0$, then we can compute A_i and therefore also these conditional probabilities easily (in particular, we do not need the normalizing constant $Z(T)$).

From the definition of the conditional probability it follows immediately, that if $\mathbf{X}$ has distribution $\pi(\mathbf{x})$, then the configuration $\mathbf{X}'$ which has $X'_k = X_k$ for all $k \neq i$ whereas X'_i is drawn according the conditional probabilities above, also has the distribution $\pi(\mathbf{x})$. We will see later in Chapter 4, that repeated sampling according to the correct conditional distribution not only leaves the Gibbs distribution invariant, but we even have convergence to the Gibbs distribution from any starting distribution. We thus can choose an arbitrary initial configuration and sweep through the lattice many times, always adjusting one spin according to the correct conditional distribution. After a suitable burn-in period, we will have samples from the Gibbs distribution. This is called the *Gibbs-sampler*.

The Gibbs sampler has found applications also in Bayesian statistics. In most cases it is not possible to sample directly from the posterior, but often the so-called full conditionals, that is the conditional distribution of one component of *theta* given all the other components $\alpha(\theta_i \mid \theta_j, j \neq i, \mathbf{x})$ are simple enough so that we can simulate from them.

1.5 Simulations in Operations Research

We discuss here queueing systems as an example. Such a system consists of different servers and customers which wait until a server is available to meet their service requests. Input variables are the arrival times and service requests (type and time required for its execution) of the customers. They are usually modeled as stochastic. Output variables are for instance the time a customer spends in the system until his service requests are fulfilled, the total amount of work the system has still to provide at a given time, the length of time a server is idle etc. The function which links output variables to input variables depends on the number and the specialization of servers and the queueing discipline.

In this example the time component is important. The state of the system changes in time in a random way due to the randomness of the inputs; it is thus a stochastic process. The state of the system does however not change continuously, but at discrete random time points. Because of this, one speaks of “discrete event simulation”. Output variables also change with time. Often, the system becomes stationary asymptotically, which means that the distribution of the state of the system is the same at different time points. In such a case one can simulate the system only once, but over a long time and then take time averages of output variables instead of averages over independent realizations.

In contrast to such jump type processes there are also examples where the state changes continuously. In deterministic systems, this is almost always the case because they are usually modeled by ordinary or partial differential equations. If one perturbs such differential equations by a stochastic noise term, one obtains stochastic differential equations where simulations are also widely used. We will give a brief introduction to this topic in Section ?? below.

1.6 Simulations in Financial Mathematics

Financial mathematics is another area where simulations are wide spread, both for pricing of financial instruments and for risk assessment. We consider a simple model for loss in a portfolio with investments in J creditors: l_j is the loss if the j-th creditor defaults and Y_j is the indicator for this event (in a given time period). The total loss is then

$$L = \sum_{j=1}^{J} Y_j \ell_j.$$

We assume that l_j is deterministic with integer values (in a suitable unit) whereas the Y_J's are stochastic. The joint distribution of the Y_j's is of the following form: There are latent variables $W = (W_1, \ldots, W_p) \sim F$ such that given W the Y_j's are conditionally independent with

$$\mathbb{P}(Y_j = 1 \mid W) = f_j(W).$$

The W_i represent the economic conditions of different countries and different industrial sectors which influence the risk of default. The functions f_j describe how much the j-th investment is influenced by these economic conditions and are assumed to be known. A possible form is

$$f_j(W) = \frac{1}{1 + \exp(-\sum_{i=1}^{p} a_{ji} W_i)}$$

with suitable weights a_{ji}. The W_i finally are assumed to be independent with a known distribution, e.g. a Gamma or Lognormal distribution.

The task is to compute the distribution of L. In principle, setting up a simulation which generates replicates of $(Y_1, Y_2, \ldots, Y_n)$ is straightforward. However, this is usually not sufficiently accurate, in particular for the right tails which are of interest here. We will see in the following that straightforward simulation is not suitable for approximating small probabilities, and we will learn alternatives like importance sampling.

In the example here, another possible solution is to use the characteristic function of L which is defined as

$$\chi_L(\lambda) = \mathbb{E}(\exp(i\lambda L)) = \sum_{n=0}^{N} \exp(i\lambda n)\mathbb{P}(L = n)$$

(we use that L takes positive integer values and is bounded). If we know χ_L for $\lambda = 2\pi k/(N+1)$ ($k = 0, 1, \ldots, N$), then we can use Fourier inversion to compute

$$\mathbb{P}(L = n) = \frac{1}{N+1} \sum_{k=0}^{N} \exp(-i2\pi nk/(N+1))\chi_L(2\pi k/(N+1)).$$

By choosing $N+1$ as a power of 2, this can be evaluated easily with the Fast Fourier Transform even when $N+1$ is large. In order to approximate χ_L, we use

$$\chi_L(\lambda) = \mathbb{E}(\mathbb{E}(\exp(i\lambda \sum_{j=1}^{J} Y_j \ell_j) \mid W)).$$

The conditional expectation on the right-hand side can be evaluated analytically because conditionally on W, we have a product of independent random variables with only two possible values

$$\mathbb{E}(\exp(i\lambda \sum_{j=1}^{J} Y_j \ell_j) \mid W) = \prod_{j=1}^{J} \left(1 + (\exp(i\lambda \ell_j) - 1) f_j(W)\right).$$

Simulations is then used to approximate the expectation over the distribution of W. This means we generate independent replicates of W, compute the right-hand side above for each replicate and average the results for all values $\lambda = 2\pi k/(N+1)$. More details can be found in S. Merino und M. Nyfeler, "Calculating portfolio loss", RISK, August 2002, 82–86.

1.7 The accuracy of Monte Carlo methods

The use of sample averages to approximate expected values is justified by the law of large numbers. The central limit theorem provides additional information on the accuracy.

Let us assume that the input variables belong to a space $\mathbb{X}$ and denote their distribution of the input variables by π. The output variable is assumed to be one-dimensional and has the form $Y = h(\mathbf{X})$ where h is a deterministic function. Moreover, we are interested in the expectation of Y, that is

$$\theta = \mathbb{E}(h(\mathbf{X})) = \int h(\mathbf{x})\pi(d\mathbf{x}).$$

We use the simulation approximation

$$\hat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} h(\mathbf{X}_i), \quad \mathbf{X}_i \sim \pi$$

The following theorem gives a bound on the approximation error.

Theorem 1.2. *Assume* $\int h(\mathbf{x})^2\pi(d\mathbf{x}) < \infty$ *and let*

$$\sigma^2 = \sigma^2(h) = \int (h(\mathbf{x}) - \theta)^2\pi(d\mathbf{x}).$$

Then for all $t \in \mathbb{R}$

$$\mathbb{P}(\sqrt{N}(\hat{\theta}_N - \theta) \leq \sigma t) \to \Phi(t)$$

Moreover, if

$$S_N^2 = \frac{1}{N}\sum_{i=1}^{N}(h(\mathbf{X}_i) - \hat{\theta}_N)^2$$

denotes the sample variance, then the probability that the interval

$$I_N = \hat{\theta}_N \pm \Phi^{-1}(1 - \frac{\alpha}{2})\frac{S_N}{\sqrt{N}},$$

contains the true value θ*, converges to* $1-\alpha$ *for* $N \to \infty$*.*

Remarks:

1. The accuracy is not known in advance and has uncertainty α.
2. The rate $\frac{1}{\sqrt{N}}$ is slow! For twice the accuracy, we need four times as many replicates.
3. Since N is always large , it does not matter whether we use N or $N-1$ in the denominator of S_N^2.

Proof. The first statement is simply the central limit theorem. The second statement is a part of the Slutsky theorem, see Mathematical Statistics for details. To sketch the idea, we fix a $\delta > 0$ and we set $W_N = \sqrt{N}|\hat{\theta}_N - \theta|$ and $z = \Phi^{-1}(1 - \frac{\alpha}{2})$. Then $\mathbb{P}(W_N \leq zS_N)$ differs from $\mathbb{P}(W_N \leq zS_N, |\frac{S_N}{\sigma} - 1| \leq \delta)$ by at most $\mathbb{P}(|\frac{S_N}{\sigma} - 1| \geq \delta)$ which by the law of large numbers for S_N converges to zero as $N \to \infty$. Moreover, if $|\frac{S_N}{\sigma} - 1| \leq \delta$, then $W_N \leq zS_N$ implies that $W_N \leq z\sigma(1+\delta)$ and on the other hand $W_N \leq z\sigma(1-\delta)$ implies $W_N \leq zS_N$. Combining all this, we obtain

$$2\Phi(z(1-\delta)) - 1 \leq \liminf \mathbb{P}(W_N \leq zS_N) \leq \limsup \mathbb{P}(W_N \leq zS_N) \leq 2\Phi(z(1+\delta)) - 1$$

As $\delta \to 0$, the two bounds converge to $1 - \alpha$. □

In the special case where one wants to approximate the probability $\theta = \pi(A)$ (i.e. $h(x) = 1_A(x)$), there is also an a priori estimation for the required number of replications because in this case $\sigma^2 = \pi(A)(1 - \pi(A))$. Assume for instance that the error should be with probability 95% at most $0.1 \times \pi(A)$. Then

$$1.96 \times \sqrt{\frac{\pi(A)(1-\pi(A))}{N}} \leq 0.1 \times \pi(A)$$

which is equivalent to

$$N \geq 385 \times \frac{(1 - \pi(A))}{\pi(A)}.$$

When $\pi(A)$ is very small, then N must be very large. Different methods to increase the accuracy are discussed in Section 3.10.

Similarly, we can obtain information on the Monte Carlo error for more complicated functionals of π than expected values. We consider specifically the quantiles and refer to a course in Mathematical Statistics for the proofs.

Let $Y_i := h(\mathbf{X}_i)$ and let q_α be the α-quantile of Y_i, $q_\alpha = \inf\{y; \mathbb{P}(Y_i \leq y] \leq \alpha\}$. The empirical quantile is $\hat{q}_\alpha = Y_{([(N+1)\alpha])}$, where $Y_{(1)} \leq Y_{(2)} \leq ... \leq Y_{(N)}$ denotes the ordered observations, and $[x]$ the integer part of x. Then we have:

Theorem 1.3. *If* $\mathbb{P}(Y_i \leq q_\alpha) = \mathbb{P}(Y_i < q_\alpha) = \alpha$, *then* $[Y_{(k_1)}, Y_{(k_2)}]$ *where*

$$\begin{aligned} k_1 &= [N\alpha + 0.5 - \sqrt{N\alpha(1-\alpha)}\Phi^{-1}(1 - \frac{\gamma}{2})], \\ k_2 &= [N\alpha + 0.5 + \sqrt{N\alpha(1-\alpha)}\Phi^{-1}(1 - \frac{\gamma}{2})] + 1, \end{aligned}$$

is an approximate $(1-\gamma)$*-confidence interval for* q_α.

Proof. This follows from the Central Limit Theorem for binomial random variables. We have

$$\mathbb{P}\left(q_\alpha \notin [Y_{(k_1)}, Y_{(k_2)}]\right) = \mathbb{P}(Y_{(k_1)} > q_\alpha) + \mathbb{P}(Y_{(k_2)} < q_\alpha).$$

The event $Y_{(k_1)} > q_\alpha$ is equivalent to saying that there are at most $k_1 - 1$ observations $\leq q_\alpha$. It follows

$$\mathbb{P}(Y_{(k_1)} > q_\alpha) = \sum_{j=0}^{k_1-1} \binom{N}{j} \alpha^j (1-\alpha)^{N-j} \to \Phi(\frac{k_1 - 1 - N\alpha + 0.5}{\sqrt{N\alpha(1-\alpha)}}) = \frac{\gamma}{2}$$

for $N \to \infty$ (0.5 is a continuity correction).

For the second term we proceed similarly. □

For a concrete example, we take $\alpha = 0.9$, $N = 1000$ and $\gamma = 0.05$. Then we get $k_1 = 881$ and $k_2 = 920$.

1.8 Other applications of stochastic simulation

We mention briefly two other applications of stochastic simulations: randomized algorithms and teaching.

The most straightforward example of a randomized algorithm is Monte Carlo integration. Assume we have a function $h : A \subseteq \mathbb{R}^p \to R$ and we want to compute the integral $\int_A h(\mathbf{x})d\mathbf{x}$. By a suitable transformation of variables we can always achieve $A = [0,1]^p$. Monte Carlo integration generates then Np uniform random variables U_{ij} on $[0,1]$ and uses the approximation

$$\frac{1}{N}\sum_{i=1}^{N} h(U_{i1},\ldots,U_{ip}).$$

A numerical integration method chooses N deterministic points $\mathbf{x}_i \in [0,1]^p$ and uses the approximation $\sum_{i=1}^N w_i h(\mathbf{x}_i)$ with given weights w_i. In the simplest case, the $\mathbf{x}_i$ are on a cubic lattice

$$\mathbf{x}_i \in \left\{\frac{1}{2K}, \frac{3}{2K}, \ldots, \frac{2K-1}{2K}\right\}^p,$$

with $N = K^p$ and the weights are constant. For smooth functions h, one can show that in this case the convergence rate is equal to $N^{-1/p}$, which is very slow in high dimensions. In contrast, Monte Carlo integration has convergence rate $N^{-1/2}$ independent of the dimension p and without any smoothness assumptions.

It may be surprising that a random choice leads to a better algorithm than a systematic, deterministic choice, even though the problem that one wants to solve is purely deterministic. This is however not the only example of such a phenomenon, there are many randomized algorithms which perform better than deterministic ones. We refer to the literature for such examples.

Simulation is also a very useful tool for teaching. It allows to experience random variability since one obtains slightly different results each time an experiment is repeated, and one can observe cases where randomness behaves differently that one would naively expect (consider for instance the lenght of runs in a coin toss). Using the computer instead of throwing dice or tossing a coin also has the advantage that one can do many replicates without getting tired. Finally, simulations allow to quickly see interesting phenomena which are difficult to prove analytically, e.g. arc sine laws for random walks.

Chapter 2

Generating Uniform Random Variables

Practically all simulations use “random” numbers, which are not really random, but generated by a deterministic algorithm. They are thus call pseudo-random numbers. This seems to be a contradiction, but from a practical point of view what matters are the properties of the numbers, not how they were generated. Pseudo-random numbers should behave in as many ways as possible like realizations of i.i.d. uniform random variables. There have been attempts to exploit the randomness of some physical systems like radioactive decay. However, for these physical random numbers it is usually easier than for good pseudo-random numbers to detect that they don’t behave like i.i.d. uniform random variables . Moreover, reproducibility of results is also important which is easier to achieve with pseudo random numbers.

The question what it means to behave like i.i.d. uniform random numbers leads to deep mathematical theories developed by Kolmogorov and Martin-Löf. We take here a more pragmatic approach based on visual properties and some numerical indices for the distribution of d-tupels.

All algorithms that we discuss here, generate pseudo random numbers (u_n) according to the following rules:

$$z_{n+1} = f(z_n), \quad u_n = h(z_n), \tag{2.1}$$

with given functions f and h. This means that u_n is a in general not invertible function of a sequence z_n defined recursively from its predecessor. Such an algorithm needs a starting or seed value z_0.

If the set of possible values of z_n is finite, then it is easy to see that (z_n) and (u_n) are periodic except for maybe a finite number of values at the beginning. (Consider the first n such that $z_n \in \{z_0, z_1, \ldots z_{n-1}\}$). Clearly, a good generator should have a period length which is substantially longer than the number of values used in a given simulation experiment. This is however only necessary, but not sufficient for a good generator. Important is also that the sequence of successive d-tupels fills out the d-dimensional unit cube well.

We discuss first a few simple generators and then show how one can combine them in order to satisfy more stringent requirements.

2.1 Linear Congruential Generators

A linear congruence generator has the form $u_n = x_n/M$ where (x_n) satisfies the recursion

$$x_{n+1} = (ax_n + c) \mod M$$

where $x_0, a, c, M \in \mathbb{N}$.

The issue of the period length of such generators is covered in the following theorem.

Theorem 2.1. *It holds*

1. *If $c \neq 0$, the period is equal to M for all x_0 iff c and M are relatively prime and if $a \equiv 1 \mod p$ for all prime divisors p of M and also for $p = 4$ if M is a multiple of 4.*
2. *If $c = 0$, the period is equal to $M - 1$ for all $x_0 \neq 0$ iff M prim and $a^{(M-1)/p} \neq 1 \mod M$ for all prime divisors p of $M - 1$.*
3. *If $c = 0$ and $M = 2^k \geq 16$, then the period is $\frac{M}{4}$ iff x_0 is odd and $a \mod 8 \in \{3, 5\}$.*
4. *If $c = 0$, $M = 2^k \geq 16$ and $a \mod 8 = 5$, then $x_n \mod 4$ is constant $=: b$, and if $b \in \{1, 3\}$, then $\frac{1}{4}(x_n - b)$ is identical to the sequence produced by the generator with $a' = a$, $c' = b\frac{a-1}{4}$, $M' = \frac{M}{4}$. (This means we should simply ignore the last two bits which are constant anyhow).*

Proof. The proof uses basic results of number theory, see siehe Ripley (1987) Section 2.7, or Knuth (1998) Theorems A and C in Section 3.2. □

Thus the case $c \neq 0$ leads to the easiest criterion for a maximal period, but it has the disadvantage that 0 appears in the sequence (u_n). The choice $M = 2^k$ is prefered because $\mod M$ is then particularly easy to implement on a computer. However, low order bits have then always a very short period. For $M = 2^k - r$ with r small, $\mod M$ is also easy to compute, but the condition for a in case 2 of the above theorem is then more involved.

For any M there are many choices of c and a which guarantee a long period. As the figures show, these choices differ in the distribution of successive pairs. We study therefore the distribution of d-tupels. We denote by Λ_d the set of all d-tupels produced by the generator and in the case $c = 0$ we add the origin

$$\Lambda_d = \{(x_n, x_{n+1}, \ldots, x_{n+d-1}),\ 0 \leq n < M\} \quad (\cup\{(0, \ldots, 0)\} \text{ if } c = 0).$$

If the period of the generator is maximal, Λ_d contains for any d only M points from the M^d possible points of the set $\{0, \ldots, M - 1\}^d$. Hence with increasing d, the d-tupels are necessarily farther apart from each other. Whereas this is unavoidable, for a good generator distances should increase in all directions equally. We are going to formulate this more precisely in the following.

The figures show that Λ_d has the regular structure of a so-called *lattice*. A lattice $L \subset \mathbb{R}^d$ is the set of all integer linear combinations of a set of d linear independent vectors $g_i \in \mathbb{R}^d$:

$$L = \{x = t_1 g_1 + \ldots t_d g_d;\ t_i \in \mathbb{Z}\}.$$

The set of the g_i's are called a basis of L. Note that there are many bases of the same lattice. A lattice is a subgroup of $(\mathbb{R}^d, +)$.

We now show that Λ_d consists of all points of a lattice shifted by a fixed vector which belong to the the integer cube $\{0, \ldots, M - 1\}^d$.

Theorem 2.2. *Let L_d denote the lattice with generating vectors*

$$\begin{aligned} g_1 &= (1, a, a^2, \ldots, a^{d-1})^T, \\ g_j &= (0, \ldots, \underbrace{M}_{j}, \ldots, 0)^T \quad (j = 2, 3, \ldots, d). \end{aligned}$$

If $c > 0$ and the period is equal to M or if $c = 0$ and the period is $M - 1$, then

$$\Lambda_d = \Big(c(0, 1, 1 + a, \ldots, (1 + a + \cdots + a^{d-2}))^T + L_d\Big) \cap \{0, \ldots, M - 1\}^d.$$

Proof. First one shows by induction that

$$x_{n+j} = (a^j x_n + c(a^{j-1} + \cdots + 1) + M \cdot \mathbb{Z}) \cap \{0, 1, \ldots, M - 1\}.$$

Hence Λ_d is a subset of the lattice L_d shifted by $c(0, 1, \ldots, (1+\cdots+a^{d-2}))^T$ and intersected with $\{0, \ldots, M-1\}^d$. In order to show the opposite inclusion, we prove that the two sets have the same cardinality. Because we have assumed that the period is maximal, Λ_d contains M points. For any $t_1 \in \{0, 1, \ldots, M-1\}$ there is exactly one choice for $t_2, \ldots, t_d$ such that $t_1 g_1 + \ldots + t_d g_d$ lies in $\{0, \ldots, M-1\}^d$ shifted by $-c(0, 1, \ldots, (1+\cdots+a^{d-2}))^T$. Hence the claim follows. □

In particular, the theorem shows that the quality of a generator is not affected by the value c because c simply shifts the lattice.

The points of any lattice lie on parallel equidistant hyperplanes as the figures clearly indicate. Mathematically, this can be described with the so-called dual (or reciprocal) lattice

$$L^\perp = \{v \in \mathbb{R}^d;\ v^T x \in \mathbb{Z} \ \ \forall x \in L\}.$$

Let us show first that $L^\perp$ is a lattice and find a basis.

Lemma 2.1. *If L is a lattice with basis vectors g_i, then also $L^\perp$ is a lattice and the set of vectors h_i which satisfy $h_i^T g_k = 0$ for $i \neq k$ and $h_i^T g_i = 1$ is a basis of $L^\perp$. In other words, if G is the matrix with column vectors g_i, then the columns of $H = G^{-T}$ from a basis of $L^\perp$.*

Proof. It is clear that $v \in L^\perp$ iff $v^T g_k \in \mathbb{Z}$ for all k. Hence the lattice generated by the vectors h_i is contained in $L^\perp$. On the other hand, if $v \in L^\perp$, then the vector $\sum_i (g_i^T v) h_i$ is in the lattice generated by the the vectors h_i. But $\sum_i (g_i^T v) h_i$ is in matrix notation nothing else that $HG^T v = v$. □

For $v \in L^\perp$, the hyperplanes $v^T x = 0, \pm 1, \pm 2, \ldots$ are parallel, they contain by definition all points in L, and the distance between two neighboring such hyperplanes is equal to $\frac{1}{\|v\|}$. Hence points $v \in L^\perp$, $v \neq 0$ with small norm give those directions where the points of the lattice are far apart.

For a good generator, we therefore want that the smallest possible value $\|v\| > 0$ with $v \in L^\perp$ is large. By the Lemma above, we can easily find a basis of $L^\perp$, namely

$$h_1 = (1, 0, \ldots, 0)^T, \quad h_j = \frac{1}{M}(-a^{j-1}, 0, \ldots, 1, \ldots, 0)^T \ \ (j = 2, 3, \ldots, d).$$

Moreover, any integer combination of the h_i's has the form

$$\begin{aligned} v &= \frac{1}{M}(Mt_1 - at_2 - a^2t_3 + \cdots - a^{d-1}t_d, t_2, \ldots, t_d)^T \quad (t_i \in \mathbb{Z}) \\ &= \frac{1}{M}(u_1, \ldots, u_d)^T \quad (u_i \in \mathbb{Z}, u_1 + au_2 + a^2u_3 + \cdots + a^{d-1}u_d = 0 \mod M). \end{aligned}$$

Hence we take as a measure of the quality of a generator in d dimensions the value

$$\nu_d := M \min\{\|t\|\,;\, \|t\| \neq 0,\ t \in \mathbb{Z}^d,\ t_1 + at_2 + a^2t_3 + \cdots + a^{d-1}t_d = 0 \mod M\}.$$

(large values of ν_d are good).

Unfortunately, there is no simple formula to express ν_d as a function of a and M, but there are algorithms to compute ν_d efficiently, see Ripley (1987) or Knuth (1998). Moreover, ν_d can decrease drastically if d is increased by one, which makes choosing a difficult (typically M is chosen large and in such a way that the recursion can be computed quickly).

2.2 Other Generators

Because of the lattice structure and because even the choice $M = 2^{32}$ or $M = 2^{64}$ do not have a long enough period, one considers also other types of generators. We list a few of those which have been discussed in the literature.

Nonlinear Congruential Generators An obvious variant is given by $x_i = g(x_{i-1})$ where $g : \{0, 1, \ldots M-1\} \to \{0, 1, \ldots M-1\}$ is nonlinear, e.g. $g(x) = (ax^2 + bx + c) \mod M$ oder the division modulo M if M is prime. These types of generators avoid the lattice structure of the d-tupels, but the computational effort is large. Moreover, the restriction that only M out of the M^d possible values for d-tupels can occur remains the same.

Shift Register Generators These generators are based on a binary sequence $b_n \in \{0, 1\}$ which follows the recursion

$$b_n = b_{n-p} + b_{n-q} \mod 2.$$

Then consecutive L-tupels of (b_n) with distance t are used to represent u_n in a binary expansion

$$u_n = \sum_{j=1}^{L} b_{nt+j} 2^{-j}.$$

The value of t controls the overlap of L-tupels. The maximal period of (b_n) is $2^p - 1$, and there are choices of p and q such that this value of the period is attained.

Lagged Fibonacci This generator uses a recursion of the form

$$x_i = F(x_{i-p}, x_{i-q}),$$

which is analogous to the shift register generator, but the x_i need not be binary and F is arbitrary. Possible choices of F are

$$F(X_{i-p}, X_{i-q}) = (X_{i-p} + X_{i-q}) \mod M$$

or in case the x_i's are binary vectors, componentwise addition modulo 2 (the logical operation "exclusive or").

Multiplication with Carry-over This generator uses $z_i = (x_i, c_i)$ where

$$x_i = (ax_{i-1} + c_{i-1}) \mod M, \quad c_i = \left\lfloor \frac{ax_{i-1} + c_{i-1}}{M} \right\rfloor$$

In words, one uses not only the remainder, but also the result of integer division. This is very easy to implement for $M = 2^k$, $a < M$ and $c_0 \leq a$ because then $ax_0 + c_0 \leq aM \leq M^2$. Thus by induction $c_n \leq a$ and $ax_{n-1} + c_{n-1} \leq M^2$ for all n. Hence if we write x_i in the binary system, $ax_{n-1} + c_{n-1}$ has at most $2k$ digits, the first k digits of $ax_{n-1} + c_{n-1}$ are equal to x_n and the last k digits are c_n.

One can show that the period is $(aM - 2)/2$ if $M = 2^k$ and if both $aM - 1$ and $(aM - 2)/2$ are prime.

Mersenne-Twister of Matsumoto and Nishimura (1998) This generator uses a recursion of the following form

$$x_k = x_{k-227} + x_{k-623} \begin{pmatrix} 0 & 0 \\ 0 & I_{31} \end{pmatrix} A + x_{k-624} \begin{pmatrix} I_1 & 0 \\ 0 & 0 \end{pmatrix} A$$

where x_k is a row vector in $\{0, 1\}^{32}$, A is a 32×32 matrix with binary elements and all operations are modulo 2. This means that we have a linear recursion in a space with $2^{32 \cdot 623 + 1} = 2^{19937}$ elements (all components of the 623 preceeding vectors and one component of the 624-th preceeding vector). The two inventors have constructed a matrix A such that the recursion is easy to implement and the period is $2^{19937} - 1$ for any starting value different from the one with all zeroes. This period is larger than the number of atoms in the universe, and thus it is sufficient for all simulations that will ever be run.

The pseudorandom numbers are then obtained by multiplying x_n by another binary matrix T and using the 32 binary elements as digits in a binary expansion of u_n. The inventors also show that the d-tupels are nicely equidistributed for all $d \leq 623$. R uses this generator as its default.

2.3 Combination of Generators

Most generators which are used nowadays combine several basic generators. This not only increases the period, but usually also makes the d-tuples more evenly distributed. There are at least two possibilities to combine generators. The first one combines the the current values x'_n and x''_n of the two generators by a function of two arguments $x_n = F(x'_n, x''_n)$ for a given F, e.g. addition modulo M if x'_n und x''_n take values in $\{0, 1, 2, \ldots, M - 1\}$ or bitwise addition modulo 2 if x'_i und x''_i are written in the binary system. The period of the combined generator is then less or equal to the least common multiplier of the two periods.

Let us look at the distribution of d-tupels of the combined generator. We denote by L and L' the set of possible values of d-tuples for the two individual generators which are subsets of $W = \{0, 1, 2, \ldots, M - 1\}^d$. We assume that the seeds are chosen uniformly and independently for the two generators. This induces distributions p', p'' on W for the individual and thus also for the combined generator:

$$p(w) = \sum_{F(w', w'') = w} p'(w') p''(w''). \tag{2.2}$$

Here, by abuse of notation, F is the componentwise application of the function which combines the two generators. The following lemma shows that the distribution p of the combined generator is at least as uniform as p' and p'' of the two individual generators.

Lemma 2.2. *Let p' and p'' be two distributions on a finite set W and let F be a function from $W \times W$ to W such that $F(.,w)$ and $F(w,.)$ are both bijections for any $w \in W$. Moreover, let p be the distribution on W defined by (2.2). Then it holds that*

$$\sum_w \left| p(w) - \frac{1}{|W|} \right| \leq \min \left(\sum_w \left| p'(w) - \frac{1}{|W|} \right|, \sum_w \left| p''(w) - \frac{1}{|W|} \right| \right).$$

Proof. Let us define a transition matrix

$$Q(w', w) = \sum_{w'', F(w', w'') = w} p''(w'').$$

We have $\sum_w Q(w', w) = 1$ for any fixed $w' \in W$: Each $w'' \in W$ appears exactly once as solution of $F(w', w'') = w$ if w runs through W. Similarly, $\sum_{w'} Q(w', w) = 1$. Therefore

$$\sum_{w'} p'(w) Q(w', w) = p(w), \quad \sum_{w'} \frac{1}{|W|} Q(w', w) == \frac{1}{|W|}.$$

From this, we obtain

$$\sum_w \left| p(w) - \frac{1}{|W|} \right| = \sum_w \left| \sum_{w'} \left(p'(w') - \frac{1}{|W|} \right) Q(w', w) \right| \leq \sum_w \sum_{w'} \left| p'(w') - \frac{1}{|W|} \right| Q(w', w).$$

Exchanging the role of p' and p'' concludes the proof. □

This lemma can be seen as a special case of a coupling inequality which we will meet again in the last chapter.

The second combination possibility is by shuffling: The second sequence determines the order in which the members of the first sequence are used. At the beginning we set $t = (x'_1, x'_2, \ldots, x'_k)$. In the n-th call we generate with the help of x''_n a random index $i_n \in \{1, \ldots k\}$, returns $x_n = t_{i_n}$ and replaces t_{i_n} with x'_{n+k}. Although this is appealing it is very difficult to analyze this method theoretically.

2.4 Testing of Random Numbers

Any test for uniformity on $[0,1]^d$ can be used to test the quality of generators. Usually, the case $d = 1$ is not interesting because practically all generators pass such a test. The picture is quite different for $d > 1$.

In particular we can use the chisquare test based on partitioning $[0,1]^d$ into K mutually disjoint classes and counting how many d-tuples of consecutive values lies in each class. The question is which partition one should use. Partitioning in subcubes with faces parallel to the coordinate axes quickly leads to too many classes. Moerover, if one uses overlapping d-tuples, the multinomial distribution does not hold because of the dependence between overlapping d-tuples and this has to be taken into account when calculating p-values.

One way to deal with these problems, is to count not how often each class occurs, but simply that number W of classes which never occur. This is easy to store even if the number of classes is large, and computing the expectation and variance of W reduces to a combinatorial problem which can be solved in many cases. For instance for $d = 2$, $k = 1024$ and a sequence of $n = 2^{21}$ random numbers one obtains $\mathbf{E}[W] = 141'909$ and $\operatorname{Var} W = 290^2$. The critical value is then obtained by assuming W to be approximately normal. Such tests are called "monkey tests" because one counts how many words of length d a monkey who hits a key board with k keys n times never writes (Remember that by the lemma of Borel-Cantelli a monkey who hits a key board randomly for ever writes the collected works of Shakespeare not only once, but infinitely often).

Chapter 3

Direct Generation of Random Variables

We consider the following problem: Given a distribution π on a space $\mathbb{X}$ with a σ-field $\mathcal{F}$ and a sequence of uniform random variable U_1, $U_2, \ldots$ i.i.d. $\sim$ Uniform$(0,1)$, find an algorithm which produces an i.i.d. sequence $X_1, X_2, \ldots$ with distribution π. We therefore ignore that in practice our uniform random variables are not really random.

3.1 Quantile Transform

Let F be a cumulative distribution function on $\mathbb{R}$.

Definition 3.1. *The quantile function F^{-1} is defined on $(0,1)$ by*

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

I assume that the following result is known.

Theorem 3.1. *If $U \sim$ Uniform$(0,1)$, then $X = F^{-1}(U) \sim F$.*

Example 3.1 (Exponential distribution)**.** *If $F(x) = 1 - e^{-\lambda x}$, then $F^{-1}(u) = -\frac{1}{\lambda}\log(1-u)$.*

Example 3.2 (Cauchy distribution)**.** *The densitiy and the distribution function are*

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \quad F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x).$$

Hence $F^{-1}(u) = \tan(\pi(u - 0.5))$.

Example 3.3 (Discrete distributions)**.** *If X takes the values $x_1 < x_2 < \ldots$ with probabilities $p_1, p_2, \ldots$ an. Then both F and F^{-1} are step functions:*

$$F(x) = \sum_{x_k \leq x} p_k, \quad F^{-1}(u) = x_k \text{ for } p_1 + p_2 + \cdots + p_{k-1} < u \leq p_1 + p_2 + \cdots + p_k.$$

Clearly, this can also be used to generate variables with values in any countable set. Note that the number of terms we have to add is $k-1$ if $F^{-1}(U) = x_k$, hence the expected number of additions is equal to $\sum_k kp_k - 1$. We should therefore list the values in descending order of their probabilities. If we need more than one realization, it is advantageous to compute and store the cumulative sums $p_1 + p_2 + \cdots + p_k$. Then we only need comparisons, and there are clever algorithms which minimize the number of comparisons to find the interval to which U belongs.

3.2 Rejection Sampling

The key idea is to simulate with a different distribution τ (called the proposal) and then to correct to obtain a sample from the target π.

Theorem 3.2. *Let π and τ be distributions on an arbitrary space $(\mathbb{X}, \mathcal{F})$ with densities f and g (with respect to some reference measure μ.) Assume that there is a constant $M < \infty$ such that $f(x) \leq Mg(x)$ for all x and thus*

$$a(x) := \frac{f(x)}{Mg(x)} \leq 1.$$

If X and U are independent random variables with $X \sim \tau$ und $U \sim$ Uniform$(0,1)$, then the conditional distribution of X given $U \leq a(X)$ is π:

$$\mathbb{P}(X \in A \mid U \leq a(X)) \;=\; \pi(A) \quad \forall A \in \mathcal{F}.$$

Proof. By definition of the conditional probability

$$\mathbb{P}(X \in A \mid U \leq a(X)) = \frac{\mathbb{P}(\{X \in A\} \cap \{U \leq a(X)\})}{\mathbb{P}(U \leq a(X))}$$

By the assumption on the distribution of (X, U)

$$\begin{aligned} \mathbb{P}(\{X \in A\} \cap \{U \leq a(X)\}) &= \int_{\mathbb{X}\times[0,1]} \mathbf{1}_A(x)\mathbf{1}_{[0,a(x)]}(u)\tau(dx)du = \int_A a(x)\tau(dx) \\ &= \frac{1}{M}\int_A \frac{f(x)}{g(x)} g(x)\mu(dx) = \frac{1}{M}\int_A f(x)\mu(dx) = \frac{1}{M}\int_A \pi(dx). \end{aligned}$$

The same argument shows that the denominator is equal to $1/M$. □

This leads to the following algorithm

Algorithm 3.1.

1. *Generate (X, U) independent with $X \sim \tau$ and $U \sim$ Uniform$(0,1)$.*
2. *If $U \leq a(X)$, output X, otherwise go back to 1. .*

For obvious reasons, a is called the acceptance function. Note that because f and g integrate to one, $M \geq 1$. The following Lemma shows that M controls the number of rejected values.

Lemma 3.1. *Let T denote the number of pairs (X, U) that have to be generated until $U \leq a(X)$ for the first time. Then T is geometrically distributed with parameter $1/M$, in particular $\mathbb{E}(T) = M$.*

Proof. It is clear that T has a geometric distribution since all generated pairs (U, X) are i.i.d. In the proof of Theorem 3.2, we have already shown that $\mathbb{P}(U \leq a(X)) = 1/M$. □

For this algorithm, it is sufficient to know f up to a normalizing constant. If $f(x) = \frac{1}{Z} f_*(x)$ and $f_*(x) \leq Mg(x)$, then the acceptance function is $a(x) = \frac{f_*(x)}{Mg(x)}$, and thus we do not need to know Z.

Example 3.4 (Uniform distribution on a bounded subset of $\mathbb{R}^p$)**.** *Let* $\mathbb{X} = \mathbb{R}^p$ *and* $A \subset \mathbb{R}^p$ *be a bounded and open subset. The uniform distribution on* A *has the density*

$$f(x) = \begin{cases} \textit{const.} & x \in A \\ 0 & x \notin A \end{cases}$$

As proposal, we choose the uniform distribution on a rectangle R *with* $A \subset R$*. The acceptance function is then nothing else than the indicator of* A*.*

Example 3.5 (Beta-distribution)**.** *The density is*

$$f(x) \ \propto \ f_*(x) = x^{\alpha-1}(1-x)^{\beta-1} \quad (0 < x < 1)$$

For $\alpha \geq 1$ *and* $\beta \geq 1$ f_* *is bounded and we can therefore choose* $g(x) = 1$*. One obtains*

$$\sup_x f_*(x) = \frac{(\alpha-1)^{\alpha-1}(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}}$$

For $\alpha < 1$ *and* $\beta \geq 1$ *we can choose* $g(x) = \alpha x^{\alpha-1}$*. Then*

$$\frac{f_*(x)}{g(x)} = \frac{(1-x)^{\beta-1}}{\alpha} \ \leq \ \frac{1}{\alpha}.$$

In many cases, we can find a proposal density by partitioning the space $\mathbb{X}$. Consider

$$\mathbb{X} = \bigcup B_i, \quad B_i \cap B_j = \emptyset \ (i \neq j),$$

and suppose that for each B_i we have a density g_i with support B_i and $f_*(x) \leq M_i g_i(x)$ $(x \in B_i)$. Then we can take

$$g(x) = \sum_{i=1}^{k} \frac{M_i}{M_1 + \cdots + M_k} g_i(x) I_{B_i}(x). \tag{3.1}$$

In order to compute the acceptance function, note that by construction for $x \in B_i$

$$\frac{f_*(x)}{g(x)} \ = \ \frac{f_*(x)}{\frac{M_i}{M_1+\cdots+M_k} g_i(x)} \leq M_1 + \cdots + M_k,$$

and therefore

$$a(x) \ = \ \frac{f_*(x)}{M_i g_i(x)} \ (x \in B_i).$$

Example 3.6 (Beta-distribution with $\alpha < 1$, $\beta < 1$)**.** *We choose the partition with* $B_1 = (0, 0.5)$ *and* $B_2 = [0.5, 1)$ *and the densities*

$$g_1(x) = 2^\alpha \alpha x^{\alpha-1} \ (x \in B_1), \quad g_2(x) = 2^\beta \beta (1-x)^{\beta-1} \ (x \in B_2).$$

Example 3.7 (Gamma distribution with $\gamma < 1$)**.** *If we set the scale parameter equal to one, the density is*

$$f_*(x) = x^{\gamma-1}\exp(-x) \quad (x \geq 0).$$

For $\gamma < 1$ *we can use* $B_1 = [0, 1)$*,* $B_2 = [1, \infty)$ *and the densities*

$$g_1(x) = \gamma x^{\gamma-1}, \quad g_2(x) = \exp(-(x-1)).$$

Then $M_1 = 1/\gamma$*,* $M_2 = 1/e$*.*

Example 3.8 (log-concave densities)**.** *If $B_i = (c_i, c_{i+1}] \subset \mathbb{R}$ we can choose $g_i(x) = Z_i^{-1}\exp(b_i x)$ with arbitrary b_i because we can simulate from such a density with the quantile transform. The constant M_i must be such that for $x \in B_i$ $\log f(x) \leq \log M_i - \log Z_i + b_i x$. If $\log f$ is concave, we can therefore compute the tangent $a_i + b_i x$ to $\log f$ at some point in B_i and take $M_i = a_i Z_i$. If the intervals are small, this will give a good approximation of f and thus a high acceptance probability. The idea can be used in an adaptive way: Start with 2 intervals, and whenever a proposal is rejected, compute a new tangent at the proposed value and a corresponding finer partition.*

3.3 Ratio of Uniform Random Variables

This method is limited to the one-dimensional case. We need example 3.4, which is why we first had to discuss rejection sampling.

We have seen above that one can generate a Cauchy variable as the tangent of a uniformly distributed variable. If one wants to avoid the calculation of the tangent, one can take the quotient V/U of two variables (U,V) which are uniformly distributed on the semi-circle $\{(u,v); u^2+v^2 \leq 1, u \geq 0\}$: In polar coordinates, $V/U = \tan(\varphi)$, and φ is uniform on $[0,\pi]$, see Lemma 3.3 below. We show here that many distributions can be generated as quotients of random variables which are uniform in an appropriate set G.

Theorem 3.3. *Let $f \propto f_*$ be any density on $\mathbb{R}$ (there is no need to know the normalizing constant). If (U,V) is uniform on $G = \left\{(u,v); 0 \leq u \leq \sqrt{f_*(v/u)}\right\}$, then V/U has the density f.*

The easiest way to understand the set G is to consider the intersection of G with the line $v = ux$ for fixed slope x: This intersection is the segment between the points $(0,0)$ and $(\sqrt{f_*(x)}, x\sqrt{f_*(x)})$. In particular, G is bounded if the endpoints of these segments are bounded. Hence it follows

Lemma 3.2. *If $f_*(x)$ and $|x|\sqrt{f_*(x)}$ are bounded, then G is contained in the rectangle*

$$R = [0, \sup_x \sqrt{f_*(x)}] \times [\inf_x x\sqrt{f_*(x)}, \sup_x x\sqrt{f_*(x)}].$$

We can therefore sample from the uniform distribution on G by rejection sampling.

Moreover, we obtain a heuristic proof of Theorem 3.3. The density of V/U at x is equal to

$$\frac{\mathbb{P}(x \leq \frac{V}{U} \leq x+dx)}{dx} = \frac{\mathbb{P}(U \cdot x \leq V \leq U \cdot (x+dx))}{dx}.$$

Because (U,V) is uniform on G, the numerator is proportional to the area of G intersected with the cone $\{(u,v); ux \leq v \leq u(x+dx)\}$. Up to terms of higher order, this intersection is equal to the triangle with vertices $(0,0)$, $(\sqrt{f_*(x)}, x\sqrt{f_*(x)})$ and $(\sqrt{f_*(x)}, (x+dx)\sqrt{f_*(x)})$. Hence its area is to first order equal to $\frac{1}{2}f_*(x)dx$.

For a rigorous proof, we need a theorem about the transformation of multivariate densities under invertible differentiable mappings.

Theorem 3.4. *Let $g: G \subseteq \mathbb{R}^p \to G' \subseteq \mathbb{R}^p$ be a continuously differentiable, invertible mapping whose Jacobian determinant $D(\mathbf{u}) = \det(\frac{\partial \mathbf{g_i}}{\partial \mathbf{u_j}}(\mathbf{u}))$ vanishes nowhere in G. Moreover,* *let U be a random vector with values in G and probability density $f_*(u)$. Then $X = g(U)$ has density, namely*

$$f_X(x) = \frac{f_*(g^{-1}(x))}{|D(g^{-1}(x))|}.$$

Proof. Let $h : \mathbb{R}^p \to \mathbb{R}$ be continuous and bounded. Then with the substitution $x = g(u)$ we obtain

$$\mathbb{E}(h(X)) = \mathbb{E}(h(g(U))) = \int_G h(g(u))f_*(u)du = \int_{G'} \frac{h(x)f_*(g^{-1}(x))}{|D(g^{-1}(x))|}dx$$

Such expectations determine the distribution of X, so the claim follows. □

Proof of Theorem 3.3: $g : (u, v) \to (x, y) = (u, \frac{v}{u})$ is a bijection from $\mathbb{R}^+ \times \mathbb{R}$ into itself, with Jacobian

$$D(u, v) = \det(\frac{\partial g}{\partial(u, v)}) = \det \begin{pmatrix} 1 & 0 \\ -\frac{v}{u^2} & \frac{1}{u} \end{pmatrix} = \frac{1}{u}$$

Thus, $f_{X,Y}(x, y) \propto x \cdot 1_{[0<x<\sqrt{f_*(y)}]}$, so the marginal density of Y is

$$f_Y(y) = \int f_{X,Y}(x, y)dx \propto \int_0^{\sqrt{f_*(y)}} xdx = \frac{f_*(y)}{2}.$$

□

Example 3.9 (Standard Normal Distribution). *Since $f_*(x) = \exp(-x^2/2)$, we obtain $\sup_x \sqrt{f_*(x)} = 1$ and $\sup_x x\sqrt{f_*(x)} = \sqrt{2/e}$. This defines the rectangle R. The inequality $u < \sqrt{f_*(v/u)}$ is equivalent to $\log u \leq -v^2/(4u^2)$, or $v^2 \leq -4u^2 \log u$*

Similarly, we can use this procedure for the Gamma $(\gamma, 1)$-distribution with $\gamma > 1$. For the Cauchy distribution, we find that G is a semi-circle in accordance with previous results.

3.4 Relations between Distributions

Relations between distributions can be exploited for simulation. For example, if X and Y are independent, X has a standard normally distribution and kY^2 is chi-square distributed with k degrees of freedom, then X/Y has a t-distribution with k degrees of freedom. Furthermore, the chi-square distribution with k degrees of freedom is the distribution of the sum $\sum Z_i^2$ of k independent squared standard normal variables, or for k even, the distribution of the sum of $k/2$ independent exp(1/2)-distributed variables.

3.4.1 The Normal Distribution

Let (X, Y) be a two-dimensional random variable. Consider the polar coordinates:

$$R = \sqrt{X^2 + Y^2}, \Phi = \arctan(Y/X).$$

Lemma 3.3. 1. *Let X, Y be i.i.d. $N(0, 1)$ distributed. Then R and Φ are independent, Φ is uniform on $(0, 2\pi)$and R has the distribution function $1 - e^{-\frac{1}{2}r^2}$.*

2. *Let (X, Y) be uniformly distributed on $(x^2 + y^2 \leq 1)$. Then R and Φ are independent with Φ uniform $(0, 2\pi)$ and R^2 uniform $(0, 1)$.*

Proof. Let $A \subset R^2$ be the set $\{(x,y); \sqrt{x^2+y^2} \leq r, \arctan(\frac{y}{x}) \leq \phi\}$. The first assertion follows from

$$\begin{aligned}
\mathbb{P}(R \leq r, \Phi \leq \phi) &= \frac{1}{2\pi}\int_A \exp(-\frac{1}{2}(x^2+y^2))dxdy \\
&= \frac{1}{2\pi}\int_0^\phi \int_0^r s\exp(-\frac{s^2}{2})dsd\psi \\
&= -\frac{\phi}{2\pi}\exp(-\frac{s^2}{2}) \mid_0^r = \frac{\phi}{2\pi}(1-\exp(-\frac{r^2}{2})).
\end{aligned}$$

The second assertion follows similarly:

$$\mathbb{P}(R^2 \leq r^2, \Phi \leq \phi) = \frac{\text{Total area of A}}{\pi} = r^2 \cdot \frac{\phi}{2\pi}$$

□

This leads to two methods to generate a pair of independently standard normal distributed random variables. Both are based on U, V i.i.d. $\sim$ Uniform $(0,1)$. For the first method, one uses that $2\pi V$ and $\sqrt{-2\log(U)}$ are independent and have the same distribution as Φ and R respectively (this follows from the quantile transformation). So the random variables

$$(X,Y) = \sqrt{-2\log(U)}(\cos(2\pi V), \sin(2\pi V))$$

are iid $\sim N(0,1)$ distributed. This is the Box-Muller algorithm.

For the second method, we first generate with the rejection method (U,V) uniformly on $(x^2+y^2 \leq 1)$ and then form

$$(X,Y) = \sqrt{-2\log(R^2)}(U/R, V/R) = \sqrt{\frac{-2\log(U^2+V^2)}{U^2+V^2}}(U,V)$$

This method does not need trigonometric functions.

Once we have univariate normally distributed random variables, then we can also simulate multivariate normal distributions. The multivariate normal distribution $\mathcal{N}_p(\mu,\Sigma)$ has the density

$$(2\pi)^{-p/2}(\det\Sigma)^{-1/2}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)),$$

where $\mu = \mathbb{E}(X)$ is the vector of expected values of X, and Σ is the matrix of variances and covariances of X. For the simulation we use that X can be represented in the form

$$X = \mu + AY, \text{ with } Y_1,...,Y_p \;\; i.i.d. \sim \mathcal{N}(0,1)$$

We do not give a full derivation of this result, but indicate how to choose the matrix A. From the rules for the expectation of random vectors, we obtain

$$\Sigma = \mathbb{E}((X-\mu)(X-\mu)^T) = A\mathbb{E}(YY^T)A^T = AA^T$$

This equation has many solutions. The easiest way is to require that in addition A is a lower triangular matrix. Then we have a fast and numerically stable algorithm to compute A, the Cholesky decomposition. Of course, if we can compute the eigenvalues and eigenvectors of Σ easily, then we can also use a symmetric A.

If Σ^{-1} is given instead of Σ, we decompose Σ^{-1} as BB^T, where B is a lower triangular matrix. It then follows that $\Sigma = (BB^T)^{-1} = B^{-T}B^{-1}$ i.e. $A = B^{-T}$. For the calculation of X from Y, we solve $B^T X = Y$ by backward elimination, i.e. there is no need to invert matrices.

This approach is feasible for dimension $p \leq 1000$. Larger p's occur in the simulation of Gaussian stochastic processes. However, there Σ usually has the special structure $\Sigma_{ij} = R(i-j)$, i.e. Σ is a so-called Toeplitz matrix. In this case, there are special algorithms which are based on the Fourier transform.

3.4.2 The Poisson Distribution

Lemma 3.4. *Let (X_i) be i.i.d. $\sim$ Exp(1) and $S_n = \sum_{i=1}^n X_i$ with $S_0 = 0$. Then S_n is Gamma$(n,1)$ distributed and*

$$\mathbb{P}(S_n \leq t \leq S_{n+1}) = e^{-t}\frac{t^n}{n!}.$$

Proof. The first claim follows by induction on n, using the convolution formula for the density of the sum of two independent random variables. For the second claim, we observe that:

$$\begin{aligned}\mathbb{P}(S_n \leq t < S_{n+1}) &= \mathbb{P}(S_n \leq t) - \mathbb{P}(S_{n+1} \leq t)\\ &= \int_0^t e^{-x}(\frac{x^{n-1}}{(n-1)!} - \frac{x^n}{n!})dx = e^{-x}\frac{x^n}{n!}\,|_0^t\end{aligned}$$

□

Interpretation: We consider X_i as the time between arrivals of customers in a queueing system. Then S_n is the arrival time of the n-th customer, and $S_n \leq t < S_{n+1}$ means that the number of customers that have arrived up to time t is equal to n. This number has therefore a Poisson distribution with paramer t.

Application for simulation: If U_i is uniform, then $X_i = -\log(U_i)$ has an Exp(1)-distribution, and according to the above result:

$$Y = \min\{n \mid \sum_{i=1}^n (-\log(U_i)) > t\} - 1 = \min\{n \mid U_1 \cdot U_2 \cdots U_n < e^{-t}\} - 1$$

is Poisson(t) distributed.

3.5 Summary: Simulation of Selected Distributions

Normal Distribution. Use the ratio of uniform random variables $X = \frac{V}{U}$ with (U,V) uniformly on

$$\{v^2 \leq -4u^2 \log(u)\} \subset [0,1] \times [-2\sqrt{2/e}, \sqrt{2/e}],$$

or the representation of independently normal distributed pairs in polar coordinates:

$$(X,Y) = \sqrt{-2\log(U)}(\sin(2\pi V), cos(2\pi V)),$$

with U and V independent and uniform on $(0,1)$, or

$$(X,Y) = \sqrt{\frac{-2\log(U^2+V^2)}{U^2+V^2}}(U,V),$$

with (U, V) uniform on $\{u^2 + v^2 \leq 1\}$. In R, the quantile transform with a numerical approximation of the quantile function is used as default.

Binomial Distribution. For small n use the representation of the sum of independent binary variables. For large n use the quantile transformation with an enumeration of possible values that start at $[np]$ (see Example 3.3).

Poisson Distribution. Use the quantile transformation (for large λ starting with $[\lambda]$) or the connection with i.i.d. exponentially distributed arrival times:

$$X = \min\{n \geq 1; U_1 U_2 \cdots U_n < \exp(-\lambda)\} - 1$$

with (U_i) i.i.d. Uniform $(0, 1)$.

Cauchy Distribution. Use the quantile transformation $X = \tan(\pi(U - 0.5))$ with $U \sim$ Uniform $(0, 1)$, or the quotient of uniform random variables $X = \frac{V}{U}$ with (U, V) uniform on $\{u^2 + v^2 \leq 1\}$.

Gamma $(\gamma, 1)$ **Distribution.** For $\gamma < 1$ use rejection sampling with proposal density

$$g(x) = \frac{e}{e+\gamma}\gamma x^{\gamma-1}\mathrm{I}_{[0,1]}(x) + \frac{\gamma}{e+\gamma}\exp(-(x-1))\mathrm{I}_{(1,\infty)}(x),$$

For $\gamma = 1$ (the exponential distribution), use the quantile transformation $X = -\log(U)$. For $\gamma > 1$ use the ratio of uniform random variables $X = \frac{V}{U}$ with (U, V) uniform on

$$\{2\log(u) < (\gamma - 1)\log(v/u) - v/u\} \subset [0, a] \times [0, b]$$

with $a^2 = ((\gamma - 1)/e)^{\gamma-1}$ and $b^2 = ((\gamma + 1)/e)^{\gamma+1}$. For large γ, write $\gamma = k + \gamma_1$ with k an integer and $1 < \gamma_1 < 2$ and use the representation as a sum of k exponentially-distributed and a $\Gamma(\gamma_1, 1)$-distributed independent random variables.

Beta(α, β) **Distribution.** Use either the representation of

$$X = \frac{X_1}{X_1 + X_2},$$

where X_1 and X_2 are independent and Gamma$(\alpha, 1)$ and Gamma$(\beta, 1)$ distributed respectively; or the rejection method with a proposal density

$$\begin{aligned}
g(x) &= 1 \quad (\alpha > 1, \beta > 1), \\
g(x) &= \alpha x^{\alpha-1} \quad (\alpha < 1, \beta > 1), \\
g(x) &= \beta(1-x)^{\beta-1} \quad (\alpha > 1, \beta < 1), \\
g(x) &= \frac{\beta}{\alpha+\beta}2^{\alpha}\alpha x^{\alpha-1}1_{[0,0.5]}(x) + \frac{\alpha}{\alpha+\beta}2^{\beta}\beta(1-x)^{\beta-1}1_{[0.5,1]}(x) \quad (\alpha < 1, \beta < 1)
\end{aligned}$$

t-distribution. with ν degrees of freedom . One uses the representation

$$X = \frac{X_1}{\sqrt{2X_2/\nu}}$$

where X_1 and X_2 are independent and normal- and Gamma$(\nu/2, 1)$- distributed, respectively.

3.6 Random Sampling and Random Permutations

We assume that we want to draw a random sample $S = (i_1, i_2, ..., i_n)$ without replacement from the population $\{1, 2, ..., N\}$. If the sample is ordered, each such sample has probability

$$\frac{1}{N(N-1)\cdots(N-n+1)}.$$

For $n = N$ we obtain a random permutation. If the order within the sample does not matter, each sample has probability $\binom{N}{n}^{-1}$.

The following algorithms are available

Algorithm 3.2. *Sampling without consideration of the order.*

1. *Set* $S = (1, 2, ..., n)$ *and* $k = n$.
2. *If* $k = N$, *then* S *is the desired sample; otherwise set* $k = k + 1$.
3. *Choose* $U \sim$ *Uniform* $(0, 1)$. *If* $U < \frac{n}{k}$, *let* I *be uniformly distributed on* $(1, 2, ..., n)$, *and replace the* I*-th element of* S *by* k*; otherwise* S *does not change.*
4. *Go back to step 2.*

For this algorithm we do not need to know N in advance; for each k, S is a random subset of size n from k elements. The proof that it is correct, proceeds by induction: Assume that for some k, the algorithm gives a set S with uniform distribution among all subsets of size n from k elements. By removing one element of S at random, we obtain a subset of size $n - 1$ from k elements which is again uniform. Hence for $k + 1$, any set of size n which contains $k + 1$ has probability

$$\frac{n}{(k+1)\binom{k}{n-1}} = \frac{n \cdot (n-1)! \cdot (k-n+1)!}{(k+1) \cdot k!} = \frac{1}{\binom{k+1}{n}}.$$

Similarly a set which does not contain $k + 1$ has probability

$$\frac{k+1-n}{(k+1)\binom{k}{n}} = \frac{(k+1-n) \cdot n! \cdot (k-n)!}{(k+1) \cdot k!} = \frac{1}{\binom{k+1}{n}}.$$

Algorithm 3.3. *Sampling with consideration of the order.*

1. *Generate* $U_1, ..., U_N$ *i.i.d.* $\sim$ *Uniform* $(0, 1)$.
2. *Sort the* U_i*'s and set* $R_i = rank(U_i)$.
3. $S = \{rank(U_1), ..., rank(U_n)\}$.

This algorithm is easy to program, but sorting requires $N \log N$ comparisons, hence it is slow for large N.

Algorithm 3.4. *Sampling with consideration of the order.*

1. *Set* $M = (1, 2, ..., N)$ *and* $k = 1$.
2. *Choose* I *uniformly distributed on* $(k, k + 1, ..., N)$, *and swap* M_k *with* M_I.
3. *If* $k = n$, $S = (M_1, ..., M_n)$ *is the result, otherwise set* $k = k + 1$ *and go back to step 2.*

The proof that these two algorithms are correct is easy.

3.7 Importance Sampling

Recall that the rejection algorithm simulates a distribution π by first simulating a variable X according to an incorrect proposal distribution τ and then accepting X with probability $a(X) = f(X)/(Mg(X))$. Here f and g are the densities of π and τ respectively, and M is an upper bound for the ratio $w = f/g$.

Importance sampling is based on a similar idea, but the correction does not occur in the generation of variables, but by weighing the average. It is based on the identity

$$\mathbb{E}_\pi(h(X)) = \int h(x)\pi(dx) = \int h(x)\frac{f(x)}{g(x)}g(x)\mu(dx) = \int h(x)w(x)\tau(dx) = \mathbb{E}_\tau(h(X)w(X)).$$

If we have variables X_i, which are i.i.d. and τ-distributed, the estimator

$$\tilde{\theta} = \frac{1}{N}\sum_{i=1}^{N} h(X_i)w(X_i)$$

is therefore unbiased.

In contrast to the rejection algorithm, importance sampling does not need an upper bound for the ratio w. It is obvious that

$$\mathrm{Var}(\tilde{\theta}) = \frac{1}{N}\mathrm{Var}(h(X_i)w(X_i)) \leq \frac{1}{N}\int h(x)^2 w(x^2\tau(dx) = \frac{1}{N}\int h(x)^2\frac{f(x)^2}{g(x)}\mu(dx).$$

Thus in order to have a finite variance for all bounded functions h, $\int f(x)^2/g(x)\mu(dx)$ must be finite. This is a weaker condition than f/g is bounded. The density g, however, should have heavier tails than f. In order to avoid a huge variance, g must also be similar to f. It is more difficult to achieve this in high dimensions as most distributions in high dimensions tend to differ greatly, and there are very few candidates as proposal distribution g. Therefore, the use of importance sampling in high dimensions is limited.

Importance sampling is not linear: If a constant is added to h, then the importance sampling does not add the same constant to the estimate:

$$\frac{1}{N}\sum_{i=1}^{N}(h(Y_i) + c)w(Y_i) = \frac{1}{N}\sum_{i=1}^{N} h(Y_i)w(Y_i) + c\frac{1}{N}\sum_{i=1}^{N} w(Y_i)$$

and $N^{-1}\sum_i^N w(Y_i) \neq 1$ although it converges to one as $N \to \infty$. Alternatively, we can use the estimate

$$\frac{\frac{1}{N}\sum_{i=1}^{N} h(Y_i)w(Y_i)}{\frac{1}{N}\sum_{i=1}^{N} w(Y_i)}$$

which is linear, but no longer unbiased. Note that the second version can be used also if f is known only up to a normalization constant, whereas for the first version we need the normalization constant.

3.8 Markov Chains and Markov Processes

In principle, one can simulate recursively from a p-dimensional distribution: Let π_1 be the marginal distribution of X_1 and $\pi_{j|j-1,\ldots,1}$ be the conditional distribution of X_j given

$X_1 = x_1, \ldots, X_{j-1} = x_{j-1}$. Then one can first generate X_1 from π_1 and then iteratively X_j from $\pi_{j|j-1,\ldots,1}$ for $j = 2, \ldots, p$. However, this only works if π_1 and $\pi_{j|j-1,\ldots,1}$ can be calculated explicitly. In general, one has to calculate a number of integrals for $\pi_{j|j-1,\ldots,1}$, which is usually not possible. Remember that we simulate because we want to avoid numerical integration.

One exception are multivariate normal distributions. Their conditional distributions are again normal, and we obtain the conditional expected values and variances from the Choleski decomposition.

Another important example where this approach is feasible are Markov chains. These are stochastic processes in discrete time, where the conditional distribution of X_j given $X_1 = x_1, \ldots, X_{j-1} = x_{j-1}$ only depends on x_{j-1}, and is explicitly given by a so-called transition kernel, see Section 4.1 in the next chapter.

Markov processes in continuous time with a continuous state space are more difficult to simulate. Such processes are generated by stochastic differential equations, which we briefly discuss next.

3.8.1 Simulation of Stochastic Differential Equations

A stochastic differential equation arises from an ordinary differential equation by adding a stochastic noise N_t:

$$\frac{dX_t}{dt} = f(X_t) + \sigma(X_t)N_t.$$

The noise is assumed to be white, that is N_t and N_s are stochastically independent for $t \neq s$. White noise is however a pathological object, because independence in different intervals implies that

$$\mathrm{Var}(\int_0^t N_s ds) = \text{const } t$$

However if the constant is positive and finite, then $\int_0^t N_s ds$ is of the order $\sqrt{t}$ and thus not differentiable, i.e. (N_t) does not exist.

The solution is to give an interpretation of the above equation which does not involve N_t, only the integral $B_t = \int_0^t N_s ds$ which is called Brownian motion. It is defined by the following two properties:

1. $B_0 = 0$ almost surely.
2. For all $t_0 = 0 < t_1 < t_2 < \cdots < t_n$, the increments $B_{t_i} - B_{t_{i-1}}$ $(i = 1, \ldots n)$ are independent and $\mathcal{N}(0, t_i - t_{i-1})$-distributed.

Wiener has shown that there is such a process and it can be chosen so that the paths are almost surely continuous. Therefore, a Brownian motion is often also called a Wiener process. The following lemma gives a proof by constructing a sequence of piecewise linear approximations which converges almost surely to a Brownian motion and which can be used for simulation.

Lemma 3.5. *Let $(B_1, X_{n,j}, n = 1, 2, \ldots, j = 1, 2, \ldots, 2^{n-1})$ be i.i.d. standard normal random variables and for each n let $(Y_t^{(n)})$ be the continuous process on $[0, 1]$ which is* *equal to $X_{n,j}$ for $t = (2j-1)2^{-n}$, equal to zero for $t = (2j)2^{-n}$ and interpolates linearly in between. Then the sequence $B_t^{(0)} = tB_1$,*

$$B_t^{(n)} = B_t^{(n-1)} + Y_t^{(n)} 2^{-(n+1)/2}.$$

converges with probability one uniformly in t to a Brownian motion on $[0,1]$ with continuous sample paths.

Proof. For uniform convergence, we use that

$$\mathbb{P}(\max_j |X_{n,j}| > c_n) \le 2^{n-1} 2(1-\Phi(c_n)) \le 2^n \int_{c_n}^{\infty} \frac{x}{c_n}\phi(x)dx = 2^n \frac{\phi(c_n)}{c_n}.$$

Hence if we choose $c_n = \theta\sqrt{2n\log 2}$ with $\theta > 1$, then

$$\sum_n \mathbb{P}(\max_j |X_{n,j}| > c_n) \le \sum_n \frac{\text{const.}}{\sqrt{n}} 2^{(1-\theta^2)n} < \infty.$$

By the Borel-Cantelli Lemma, it follows that with probability one $\sup_t |Y_t^{(n)}| 2^{-(n+1)/2} \le c_n 2^{-(n+1)/2}$ for all but finitely many n's. Hence uniform convergence follows because $c_n 2^{-(n+1)/2}$ is summable, and by a standard result from analysis the limit is continuous.

In order to check that the limit process is a Brownian motion, we show by induction that the increments $(B^{(n)}_{j2^{-n}} - B^{(n)}_{(j-1)2^{-n}})$ are uncorrelated and normally distributed with mean zero and variance 2^{-n}. This follows easily because two such consecutive increments are equal to

$$\frac{1}{2}\left(B^{(n-1)}_{j2^{-(n-1)}} - B^{(n-1)}_{(j-1)2^{-(n-1)}}\right) \pm X_{n,j} 2^{-(n+1)/2}.$$

□

Formally, N_t is the derivative of B_t, but the paths of B_t are nowhere differentiable. We therefore write the above stochastic differential equation in integral form

$$X_t = X_0 + \int_0^t f(X_s)ds + \int_0^t \sigma(X_s)dB_s,$$

or in shorthand version

$$dX_t = f(X_t)dt + \sigma(X_t)dB_t.$$

The crucial point is that we can define the "stochastic integral" $\int Z_s dB_s$ in a mathematically rigorous way for processes (Z_t), which have finite second moments and where Z_t depends only on B_s for $s \le t$. This does not work as a Lebesgue or Stieltjes integral because (B_t) not only is not differentiable but also has infinite total variation. Instead we use a Riemann approximation of the form

$$\sum_{j=1}^{n} Z_{t_{j-1}}(B_{t_j} - B_{t_{j-1}})$$

and show that this converges in L_2 as the partition gets finer. It is essential that we take the left boundary point of the subinterval and not an arbitrary point.

In this way, one can define rigorously what is meant by a solution. Next, one must show that solutions exist. This is done as for ordinary differential equations by the iterative approximation

$$X_t^{(m)} = X_0 + \int_0^t f(X_s^{(m-1)})ds + \int_0^t \sigma(X_s^{(m-1)})dB_s.$$

The simulation of solutions of these stochastic differential equations has received much interest in the past 20 years. The simplest method generates an approximate solution at time points $k\Delta$ according to the Euler scheme

$$X_{(k+1)\Delta} = X_{k\Delta} + f(X_{k\Delta})\Delta + \sigma(X_{k\Delta})(B_{(k+1)\Delta} - B_{k\Delta}).$$

Because the increments of B are normally distributed, the implementation is trivial. One can show that it converges to the solution for $\Delta \to 0$, though slowly. In analogy to the numerical solution of ordinary differential equations, one immediately thinks of higher order approximation schemes (Runge-Kutta). It turns out that one cannot improve the convergence rate in this way. There are procedures that have a faster convergence rate, but these are complicated. Recent work by Beskos, Papaspiliopoulos and Roberts shows that one can simulate exactly by the rejection method under certain assumptions on f and σ.

3.9 Variance Reduction

We have seen in Section 1.7 that if we estimate

$$\theta = \mathbb{E}(h(\mathbf{X})) = \int h(\mathbf{x})\pi(d\mathbf{x}).$$

by

$$\hat{\theta}_N = \frac{1}{N}\sum_{i=1}^N h(\mathbf{X}_i), \quad \mathbf{X}_i \sim \pi,$$

then the precision is given by the standard deviation

$$\frac{\sigma(h)}{\sqrt{N}} = \sqrt{\frac{\int(h(\mathbf{x}) - \theta)^2\pi(d\mathbf{x})}{N}}.$$

We discuss now several methods to reduce the variance, i.e to increase the precision.

3.9.1 Antithetic Variates

The variance of the arithmetic mean of dependent random variables depends on the covariances. One immediately sees that the variance decreases (vs. the independent case) if all the correlations between the variables are negative. Antithetic variables are a way to introduce such negative correlations.

We consider the following situation:

$$\theta = \int_0^1 h(x)dx = \mathbb{E}(h(U)),\ U \sim \text{Uniform (0,1)}.$$

Instead of $\hat{\theta}_N = \frac{1}{N}\sum_{i=1}^N h(U_i)$, we consider here

$$\tilde{\theta}_N = \frac{1}{2N}\sum_{i=1}^N (h(U_i) + h(1-U_i)).$$

We obtain

$$\begin{aligned} \operatorname{Var}(\tilde{\theta}_N) &= \frac{N}{4N^2}\operatorname{Var}(h(U_i) + h(1-U_i)) \\ &= \frac{1}{2N}\operatorname{Var}(h(U_i)) + \operatorname{Cov}(h(U_i), h(1-U_i)). \end{aligned}$$

If $\operatorname{Cov}(h(U_i), h(1-U_i)) < 0$, $\tilde{\theta}_N$ has a smaller variance than $\hat{\theta}_{2N}$. The following lemma gives conditions for this to hold.

Lemma 3.6. *If the function h is monotonic, then* $\operatorname{Cov}(h(U), h(1-U)) < 0$*, unless h is constant on* $(0,1)$*.*

Proof. Let U_1 and U_2 be independent and Uniform(0,1) distributed. Then we have

$$\operatorname{Cov}(h(U), h(1-U)) = \frac{1}{2}E[(h(U_1) - h(U_2)) \cdot (h(1-U_1) - h(1-U_2))].$$

We assume that h is for example monotonically increasing. If $U_1 < U_2$, then the first factor is negative and the second positive, and vice versa for $U_1 > U_2$. Thus, the integrand is always non-positive.

To verify that the covariance is strictly negative, we investigate when the integrand is zero. One factor must be 0, that is almost surely either $h(U_1) = h(U_2)$ or $h(1-U_1) = h(1-U_2)$. Because h is monotone, this is only possible if h is constant. □

This can be applied in particular for the approximation of $\int h(x)F(dx)$, if h is monotone and we simulate F with the quantile transformation (because F^{-1} is monotone).

3.9.2 Control Variates

We assume that there exists a function r such that $\mathbb{E}(r(X_i))$ is known. W.l.o.G., let $\mathbb{E}(r(X_i)) = 0$. Then for any c

$$\tilde{\theta}_{N,c} = \frac{1}{N}\sum_{i=1}^N (h(X_i) - cr(X_i)).$$

is an unbiased estimator of θ. Its variance is

$$\begin{aligned} \operatorname{Var}(\tilde{\theta}_{N,c}) &= \frac{1}{N}\operatorname{Var}(h(X_i) - cr(X_i)) \\ &= \frac{1}{N}[\operatorname{Var}(h(X_i)) - 2cCov(h(X_i), r(X_i)) + c^2\operatorname{Var}(r(X_i))] \end{aligned}$$

The optimal c_{opt} that minimizes this variance is therefore

$$c_{opt} = \frac{\operatorname{Cov}(h(X_i), r(X_i))}{\operatorname{Var}(r(X_i))}$$

and the minimal variance is

$$\mathrm{Var}(\tilde{\theta}_{N,c_{opt}}) = \frac{1}{N}\mathrm{Var}(h(X_i))(1-\mathrm{Corr}(h(X_i),r(X_i))^2) \leq \frac{1}{N}\mathrm{Var}(h(X_i)).$$

In general, c_{opt} is unknown, but it can be estimated by

$$\hat{c}_{opt} = \frac{\sum_{i=1}^N (h(X_i)-\hat{\theta}_N)r(X_i)}{\sum_{i=1}^N r(X_i)^2}.$$

This is consistent, and we obtain asymptotically the same variance as if c_{opt} is known.

For $r(X_i) > 0$, one usually scales r such that $\mathbb{E}(r(X_i)) = 1$ and uses a multiplicative correction:

$$\tilde{\theta}_N = \frac{\frac{1}{N}\sum_{i=1}^N h(X_i)}{\frac{1}{N}\sum_{i=1}^N r(X_i)}$$

Here one cannot calculate the expected value and variance of $\tilde{\theta}_N$ exactly. However,

$$\tilde{\theta}_N - \theta = \frac{\frac{1}{N}\sum_{i=1}^N [h(X_i)-\theta r(X_i)]}{\frac{1}{N}\sum_{i=1}^N r(X_i)}.$$

Hence for $N \to \infty$, $\tilde{\theta}_N$ converges almost surely to θ, and $\sqrt{N}(\tilde{\theta}_N - \theta)$ is asymptotically normal with mean zero and variance $\mathrm{Var}(h) - 2\theta\,\mathrm{Cov}(h,r) + \theta^2\,\mathrm{Var}(r))$. (This is another application of Slutsky's theorem). Multiplicative correction thus provides an improvement, if $h(X_i)$ and $r(X_i)$ are strongly correlated.

Example 3.10 (Variance of Trimmed Means)**.** *We consider the estimation of the variance of the trimmed means of n standard normal random variables, see Example 1.5.1. As a control variable we can use n times the square of the (untrimmed) mean.*

3.9.3 Importance Sampling and Variance Reduction

We have introduced importance sampling in Section 3.7 as an alternative to the rejection algorithm. The idea is to estimate $\theta = \int h(x)\pi(dx)$ with the help of random variables with the "wrong" distribution of τ. To correct it, one then computes at the weighted average

$$\tilde{\theta}_N = \frac{1}{N}\sum_{i=1}^N h(Y_i)w(Y_i), \quad Y_i \text{ i.i.d. } \sim \tau,$$

where

$$w(x) := \frac{f(x)}{g(x)}$$

and f and g are the density of π and τ respectively (with respect to some reference measure μ).

This method can be applied not only in situations where it is impossible or difficult to generate random variables with distribution π. There are also cases where a suitable choice of τ leads to a more accurate estimate than the one based on the direct approximation

$$\hat{\theta}_N = \frac{1}{N}\sum_{i=1}^N h(X_i), \quad X_i \text{ i.i.d. } \sim \pi.$$

It easily follows that

$$\begin{aligned}
\mathbb{E}(\hat{\theta}_N) &= \mathbb{E}(\tilde{\theta}_N) = \theta \\
N\,\mathrm{Var}(\hat{\theta}_N) &= \int h(x)^2 \pi(dx) - \theta^2 \\
N\,\mathrm{Var}(\tilde{\theta}_N) &= \int h(x)^2 w(x)^2 \tau(dx) - \theta^2 \\
&= \int h(x)^2 w(x) \pi(dx) - \theta^2.
\end{aligned}$$

The following lemma shows how one must choose τ so that $(\tilde{\theta}_N)$ has minimal variance.

Lemma 3.7. *For any choice of* g*, we have*

$$N\,\mathrm{Var}(\hat{\theta}_N) \geq (\int |h(x)|\pi(dx))^2 - \theta^2,$$

and we have equality if and only if $g(x) = const.|h(x)|f(x)$*.*

Proof. By the Cauchy-Schwarz inequality

$$\begin{aligned}
(\int |h(x)|\pi(dx))^2 &= (\int |h(x)|w(x)\tau(dx))^2 \\
&\leq \int h(x)^2 w(x)^2 \tau(dx) \cdot 1
\end{aligned}$$

□

The optimal g can practically never be used because for the weioghts we need the normalized density $g(x) = |h(x)|f(x)/\int |h(x)|f(x)\mu(dx)$ and we cannot compute the normalization if we cannot compute $\int h(x)f(x)\mu(dx)$. In particular, if $h \geq 0$ the optimal choice of g would give $\mathrm{Var}(\hat{\theta}_N) = 0$. Still, the Lemma is useful, because it implies that $\mathrm{Var}(\hat{\theta}_N)$ will be small if $g(x)$ is approximately proportional to $|h(x)|f(x)$.

Example 3.11. *Let* $h(x) = \mathbf{1}_A(x)$ *where* A *is a rare event under the distribution* π*. Then* θ_N *needs many replicates, see section 1.7. If we use importance sampling, by the previous Lemma the optimal* τ *is simply the conditional distribution of* X *under* π *given that* $X \in A$*. Clearly, this cannot be implemented. It is however sufficient to choose a density* g *which is small outside of* A *and approximately proportional to* f *in* A*.*

"Did Mendels Facts Fit His Model?" is the title of a section in Freedman, Pisani, Purves and Adhikari (1991). The answer is that the agreements between theory and Mendel's data are too good to still be credible. In other words, it is almost certain that Mendel's data were systematically "massaged". The basis for this verdict is as follows: If one calculates the chi-square test statistics for each of Mendel's experiment, and adds these values, we obtain a value of 42, and this sum is a random deviation from a chi-square distribution with 84 degrees of freedom. How big is the probability of observing this distribution with a value less or equal to 42? It cannot be found in any table.

The density of χ^2_{84}*, i.e. Gamma*$(42, \frac{1}{2})$ *is*

$$f(x) = \frac{(\frac{1}{2})^{42}}{\Gamma(42)} x^{41} e^{-\frac{x}{2}}.$$

We choose τ as a Gamma$(42, 1)$ distribution, which has an expected value of 42. Then

$$w(x) = \frac{f(x)}{g(x)} = \left(\frac{1}{2}\right)^{42} e^{21} e^{\frac{(x-42)}{2}}$$

where

$$\left(\frac{1}{2}\right)^{42} e^{21} = 3 \cdot 10^{-4}$$

therefore,

$$\mathbb{P}(X \leq 42) \approx 3 \cdot 10^{-4}] \frac{1}{N} \sum_{i=1}^{N} e^{\frac{1}{2}(Y_i - 42)} \mathbf{1}_{[Y_i \leq 42]}.$$

A simulation with N = 1000 yields an approximation of $3.6 \cdot 10^{-5}$. *The exact value generated by a (more complicated) numerical approximation is* $3.54 \cdot 10^{-5}$.

We can also look at the second version of importance sampling

$$\frac{\frac{1}{N}\sum_{i=1}^{N} h(Y_i) w(Y_i)}{\frac{1}{N}\sum_{i=1}^{N} w(Y_i)}$$

which can be used also in cases where f is known only up to a normalization constant. This is nothing else than an example of a multiplicative control variate. Which of the two versions is more precise, depends on h. For estimation of the probability of rare events, this version is typically less precise.

3.9.4 Quasi-Monte Carlo

So far we have concentrated on reducing the variance. We discuss now briefly Quasi-Monte Carlo, a method which reduces also the rate of convergence. It assumes that we want to simulate from the uniform distribution on $[0, 1]^d$ (in principle this can be achieved by a transformation of variables). The points $(u_i; 1 \leq i \leq N)$ constructed by Quasi-Monte Carlo are more regular than random points, but less regular than points from a regular grid.

The simplest way to construct Quasi-random points, is due to Halton. In the one-dimensional case, one chooses a natural number $b \geq 2$ and represents natural numbers k in the basis b:

$$k = \sum_{i=1}^{\infty} a_i(k) b^{i-1} \quad (a_i(k) \in \{0, 1, \ldots, b-1\}).$$

The k-th element of the Halton sequence is then

$$u_k = H(k, b) = \sum_{i=1}^{\infty} a_i(k) b^{-i} \in (0, 1).$$

This means one writes k in the basis b, reverses the order of the digits and adds 0. in front of it. In the d-dimensional case, one uses for the j-th component the Halton sequence with basis b_j where b_j is the j-the prime number:

$$u_k = (H(k, 2), H(k, 3), H(k, 5), \ldots, H(k, b_p))^T.$$

(Sometimes, the first component is taken as the equidistant sequence $(k - 0.5)/N$. This gives some improvement, but one has to generate all points again if one wants to increase the sample size N).

More advanced Quasi-Monte Carlo methods use so-called (t, m)-nets in base $b \geq 2$. These are sets of points $\{u_0, u_1, \ldots, u_{b^m-1}\} \subset [0, 1)^d$ such that any "elementary cube"

$$\prod_{i=1}^{d} \left[\frac{a_i}{b^{c_i}}, \frac{a_i + 1}{b^{c_i}} \right)$$

with $c_1 + c_2 + \ldots + c_d = m - t$ and arbitrary $0 \leq a_i < b^{c_i}$ contains exactly b^t points (note that such a cube has the volume b^{t-m}; therefore such cubes contain exactly the expected number of points). The construction of such nets is based on algebraic results that we cannot discuss here.

One can show that Quasi-Monte Carlo methods allow to approximate an integral $\int_{[0,1]^d} h(x)dx$ where h has bounded variation with an error of the order $N^{-1}(\log N)^{d-1}$.

Chapter 4

Markov Chain Monte Carlo

In many cases, especially in high dimensions, there are no good methods to simulate from a general target distribution π. The rejection algorithm fails because it almost always rejects (the bound for the ratio of the densities is too large). Importance sampling fails, because the variance of the weights is too large.

In this chapter, we discuss the current standard method for the simulation of distributions in high dimensions. The basic idea is to generate a sequence $(X_0, X_1, \ldots)$ recursively such that X_t for large t has approximately the desired distribution π and then to use the approximation

$$\mathbb{E}_\pi(h(X)) \approx \frac{1}{N-r+1} \sum_{t=r}^{N} h(X_t). \tag{4.1}$$

Here r is a so-called "burn-in" time, the time required until we reach the target π.

Recursive generation means that X_{t+1} depends on X_t and new (uniform) random variables, but not on previous values X_s with $s < t$, i.e. the generated variables form a Markov chain. We have to specify the transition rule of the Markov chain, that is how to get X_{t+1} if we have X_t, in such a way that the approximation (4.1) is valid. The minimum requirement is that if the starting value X_0 has already the desired distribution π, then all the following variables $X_1, X_2, \ldots$ also have the distribution π. We call such a π an invariant or stationary distribution of the Markov chain. So the first question we will address is: How to construct for a given π a transition rule, so that π is invariant ? We will see that there are many possible transition rules and there are some general methods to find some of them.

The second question is: Does the distribution of X_r converge to π for any arbitrary initial distribution of X_0? This also has a relatively simple answer. However, in order to use the approximation (4.1), we should also answer two more questions, namely: How big should r be, i.e. how quickly does the distribution of X_r converge to π ?, and, How accurate is this approximation ? These two questions are difficult to answer explicitely. We will see that they are connected and give a few answers in simple situations. First, however, we present some basic concepts and results about Markov chains.

4.1 Basic Concepts about Markov Chains

Let $\mathbb{X}$ be any space with a σ-algebra $\mathcal{F}$. A Markov chain describes a discrete time evolution on $\mathbb{X}$ with a simple dependence structure: The next state depends on the present, but not on past states. Hence for a Markov chain we have to choose the probabilities

$$\mathbb{P}(X_{t+1} \in A | X_t = x) = P(x, A)$$

for all $x \in \mathbb{X}$ and all $A \in \mathcal{F}$ Mathematically, we require the following properties of the so-called *transition kernel* P.

Definition 4.1. *A transition kernel* P *of* $(\mathbb{X}, \mathcal{F})$ *onto itself is a mapping from* $\mathbb{X} \times \mathcal{F}$ *to* $[0, 1]$ *such that*

- $P(x, .)$ *is a probability on* $(\mathbb{X}, \mathcal{F})$ *for every* $x \in X$.
- $P(., A)$ *is a measurable function for every* $A \in \mathcal{F}$.

If $\mathbb{X}$ is discrete, we usually write $i, j, \ldots$ instead of $x, y, \ldots$. In the discrete case, all we need are the transition probabilities

$$\mathbb{P}(X_{t+1} = j | X_t = i) = P(i, j).$$

which must be non-negative and $\sum_j P(i, j) = 1$ for all i. The transition kernel is then $P(i, A) = \sum_{j \in A} P(i, j)$. We denote the matrix with elements $P(i, j)$ also by P.

A Markov chain requires in addition to the transition kernel also an initial distribution.

Definition 4.2. *A (time-homogeneous) Markov chain on* $(\mathbb{X}, \mathcal{F})$ *with initial distribution* ν_0 *and transition kernel* P *is a sequence* $(X_0, X_1, X_2, ...)$ *of random variables with values in* $\mathbb{X}$ *such that*

$$\mathbb{P}(X_0 \in A) = \nu_0(A),$$

and

$$\mathbb{P}(X_{t+1} \in A | X_t = x_t, ..., X_0 = x_0) = \mathbb{P}(X_{t+1} \in A | X_t = x_t) = P(x_t, A).$$

The joint distribution of $(X_0, X_1, ..., X_t)$ is then

$$\nu_0(dx_0) \prod_{s=1}^{t} P(x_{s-1}, dx_s).$$

In order to simplify the notation for the rest of the chapter, we introduce now some operations with transition kernels. A kernel defines a mapping of the set of measurable and bounded functions on $(\mathbb{X}, \mathcal{F})$ into itself by

$$Pf(x) = \int P(x, dy) f(y).$$

A kernel also defines a mapping of set of probabilities on $(\mathbb{X}, \mathcal{F})$ by

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

Further, one can compose two kernels to form a new kernel by

$$PQ(x, A) = \int P(x, dy) Q(y, A).$$

The verification of these claims is an exercise in measure theory. By P^k we mean the kernel P composed k times with itself.

In the discrete case, Pf corresponds to the multiplication of P with a column vector from the right, νP corresponds to the multiplication of P with a row vector from the left, and the composition corresponds to matrix multiplication:

$$\begin{aligned} Pf(i) &= \sum_{j=1}^{n} P(i,j)f(j) = (Pf)(i), \\ \nu P[i] &= \sum_{k=1}^{n} \nu(k)P(k,i) = (\nu^T P)(i), \\ PQ(i,j) &= \sum_{k=1}^{n} P(i,k)Q(k,j) = (PQ)(i,j). \end{aligned}$$

Using these definitions, one can easily show that for any $k > 0$

$$\begin{aligned} \mathbb{P}(X_k \in A) &= \nu_0 P^k(A), \\ \mathbb{P}(X_{t+k} \in A | X_t = x_t, X_{t-1} = x_{t-1}, ..., X_0 = x_0) &= P^k(x_t, A), \\ \mathbb{E}(f(X_{t+k})|X_t = x_t) &= P^k f(x_t). \end{aligned}$$

Definition 4.3. *A probability distribution π on $(\mathbb{X}, \mathcal{F})$ is called invariant or stationary for a transition kernel P, if*

$$\pi P = \pi.$$

The meaning should be clear: if one chooses π as the initial distribution, then all the X_t have the distribution π.

Definition 4.4. *A probability distribution π on $(\mathbb{X}, \mathcal{F})$ is called reversible for a transition kernel P if*

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

This means that with the initial distribution π, (X_0, X_1) and (X_1, X_0) have the same distribution. One can easily conclude that then $(X_0, X_1, ...X_t)$ and $(X_t, X_{t-1}, ...X_0)$ have the same distribution for every t, i.e. the direction of time does not matter. By integrating over $\mathbb{X} \times A$, it immediately follows that a reversible probability distribution is always invariant. The converse is not always true.

Definition 4.5. *A transition kernel is called irreducible if a probability distribution ψ on $(\mathbb{X}, \mathcal{F})$ exists such that $\sum_{k=1}^{\infty} P^k(x, A) > 0$ for all $A \in \mathcal{F}$ with $\psi(A) > 0$ and for all $x \in \mathbb{X}$.*

Irreducibility means intuitively that the chain can reach with positive probability all states for any initial distribution . Often one can combine reducible kernels $P_i (i = 1, ..., k)$ so that the resulting kernel is irreducible. The combination may be either the sequential composition in the order $(i(1), i(2), ..., i(k))$

$$P = P_{i(1)} P_{i(2)} \cdots P_{i(k)}$$

or random selection amongst the k possible transitions

$$P = \frac{1}{k}(P_1 + P_2 + \cdots + P_k).$$

If π is invariant for all P_i's, then π is invariant for P in both versions. On the other hand, reversibility is preserved only in the second version.

Irreducibility implies that an invariant distribution is unique and that the law of large numbers applies.

Theorem 4.1. *Let P be an irreducible transition kernel with a stationary distribution π. Then π is the only stationary distribution, and the following statements are true*

- *For all $x \in \mathbb{X}$ and all $A \in \mathcal{F}$ with $\pi(A) > 0$*

$$\mathbb{P}(X_t \in A \text{ infinitely often } |X_0 = x) > 0.$$

- *For π-almost all $x \in \mathbb{X}$ and all $A \in \mathcal{F}$ with $\pi(A) > 0$*

$$\mathbb{P}(X_t \in A \text{ infinitely often } |X_0 = x) = 1.$$

- *For π-almost all $x \in \mathbb{X}$ and all f with $\int |f(x)|\pi(dx) < \infty$*

$$\mathbb{P}\left(\frac{1}{n+1}\sum_{t=0}^{n} f(x_t) \to \int f(x)\pi(dx)|X_0 = x\right) = 1.$$

Proof. See for instance Meyn and Tweedie (1993). □

The exceptional set in the last two statements is not satisfactory. There are sufficient conditions under which the last two statements are valid even for all x. Such a condition is, for example, that there exists a k, so that $P^k(x,.)$ for all x has a component which is absolutely continuous with respect to π. For the proof, I refer again to Meyn and Tweedie (1993).

Remember that our goal is to approximate $\int h(x)\pi(dx)$ according to (4.1). For this purpose we have to choose a transition kernel satisfying the following three conditions:

1. P is irreducible.
2. π is stationary or reversible for P.
3. Simulation from $P(x,.)$ should be easy for all x.

In the following sections, we will construct kernels with these three properties in progressively more complex situations.

4.2 The Metropolis-Hastings Algorithm

We first consider the discrete case. The condition for reversibility is then

$$\pi(i)P(i,j) = \pi(j)P(j,i). \tag{4.2}$$

For each pair of $i < j$, one can therefore choose either $P(i,j)$ or $P(j,i)$; the other value is then determined by (4.2). However, with this construction the condition $\sum_j P(i,j) = 1$ is not always satisfied. If the sum is less than one, we can correct this by modifying $P(i,i)$, but if the sum is bigger than one, we have a problem.

Before solving this problem, let us discuss in some detail what (4.2) means if some probabilities are zero. If both $\pi_i = 0$ and $\pi_j = 0$, the condition is satisfied automatically. If $\pi_i = 0$ and $\pi_j \neq 0$, then $P(j,i)$ must be zero: We must not go to a state which has probability zero from a state with positive probability. If both $\pi_i \neq 0$ and $\pi_j \neq 0$, then we must have $P(i,j) > 0 \Leftrightarrow P(j,i) > 0$: If a transition between two states which both have positive probability is possible, then also the reverse transition must be possible.

The following construction leads to the goal of satisfying both (4.2) and $\sum_j P(i,j) = 1$. We start with an arbitrary transition matrix $Q(i,j)$ such that

$$\pi(i)Q(i,j) > 0 \Leftrightarrow \pi(j)Q(j,i) > 0 \tag{4.3}$$

For each pair $i < j$, we set either $P(i,j) = Q(i,j)$ or $P(j,i) = Q(j,i)$ and determine the other value from (4.2). We do this in such a way that both $P(i,j) \leq Q(i,j)$ and $P(j,i) \leq Q(j,i)$ are satisfied. Then obviously

$$\sum_{j;j\neq i} P(i,j) \leq \sum_{j;j\neq i} Q(i,j) \leq 1,$$

and we obtain an transition matrix, if we set

$$P(i,i) = 1 - \sum_{j;j\neq i} P(i,j)$$

If we set $P(i,j) = Q(i,j)$, then we must have $P(j,i) = \pi(i)Q(i,j)/\pi(j)$ and this is less than or equal to $Q(j,i)$ if and only if $\pi(i)Q(i,j) \leq \pi(j)Q(j,i)$. If this is not satisfied, then setting $P(j,i) = Q(j,i)$ and $P(i,j) = \pi(j)Q(j,i,)/\pi(i)$ has the required properties. This definition can be written in compact form as follows for any $i \neq j$

$$P(i,j) = \min\left(Q(i,j), \frac{\pi(j)}{\pi(i)}Q(j,i)\right) = Q(i,j)a(i,j)$$

where

$$a(i,j) = \min\left(1, \frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)}\right) \leq 1.$$

Simulation according to the transition matrix $P(i,.)$ is not difficult. We have the following algorithm:

Algorithm 4.1. *1. Choose $Y \sim Q(i,.)$ and $U \sim Uniform(0,1)$.*

2. If $U \leq a(i,Y)$, then set $X = Y$, otherwise $X = i$.

This is similar to rejection sampling, but when we reject the new proposed value Y, we keep the current value, in accordance with the definition $P(i,i) = 1 - \sum_{j\neq i} P(i,j)$. A new value is proposed only in the next iteration.

Q is called the *proposal distribution* and a is called the *acceptance probability*. Note that always one of the two acceptance probabilities $a(i,j)$ or $a(j,i)$ is equal to one, i.e. we accept with the largest possible probability.

In the continuous case, the procedure is similar, with an appropriate definition of the acceptance probability. The condition (4.3) becomes crucial. If the state space is large, we usually allow only certain kinds of transitions, that is most $Q(i,j) = 0$, and we must make sure that the reverse of a possible transitions is always possible. Mathematically, this is expressed by absolute continuity.

The general result is as follows:

Theorem 4.2 (Metropolis-Hastings)**.** *Let π be a probability on $(\mathbb{X}, \mathcal{F})$ and Q be a kernel in the same space such that the two probabilities $\pi(dx)Q(x, dy)$ and $\pi(dy)Q(y, dx)$ on $(\mathbb{X}, \mathcal{F}) \times (\mathbb{X}, \mathcal{F})$ are equivalent in the sense of measure theory, i.e. the two probabilities should have the same null sets. Then the Radon-Nikodym density of $\pi(dy)Q(y, dx)$ with respect to $\pi(dx)Q(x, dy)$ exists, which we denote $r(y, x)$. Furthermore, let $a(x, y) = \min(1, r(y, x))$. Then the following kernel is reversible regarding π:*

$$P(x, A) = \int_A a(x, y)Q(x, dy) + \mathbf{1}_A(x) \cdot \left(1 - \int_X a(x, y)Q(x, dy)\right) \tag{4.4}$$

The first term in (4.4)is the probability that the kernel $Q(x, .)$ proposes a value in A and that this value is accepted. The second term is the probability that the process remains at the current value x because the proposed value is not accepted.

Proof. We must show that for any bounded function h that

$$\int\int h(x, y)\pi(dx)P(x, dy) = \int\int h(x, y)\pi(dy)P(y, dx).$$

The left side according to the definition of P is

$$\int\int h(x, y)a(x, y)\pi(dx)Q(x, dy) + \int h(x, x)\left(1 - \int_{\mathbb{X}} a(x, y)Q(x, dy)\right)\pi(dx),$$

and the right side is

$$\int\int h(x, y)a(y, x)\pi(dy)Q(y, dx) + \int h(y, y)\left(1 - \int_{\mathbb{X}} a(y, x)Q(y, dx)\right)\pi(dy).$$

Both second terms are obviously equal (how the variables are denoted does not matter).

In order to see that both first terms are equal, we first show that

$$a(x, y) = r(y, x)a(y, x).$$

By the definition of r, we have $r(y, x)r(x, y) = 1$. If $r(y, x) \leq 1$, then $a(x, y) = r(y, x)$ and $r(x, y) \geq 1$. Therefore $a(y, x) = 1$ and both sides above are equal to $r(y, x)$. If $r(y, x) \geq 1$, then $a(x, y) = 1$ and $r(x, y) \leq 1$. Therefore $a(y, x) = r(x, y)$ and both sides above are equal to 1.

Using this equality, it follows from the definition of $r(y, x)$ that

$$\begin{aligned}\int\int h(x, y)a(y, x)\pi(dy)Q(y, dx) &= \int\int h(x, y)a(y, x)r(y, x)\pi(dx)Q(x, dy)\\ &= \int\int h(x, y)a(x, y)\pi(dx)Q(x, dy).\end{aligned}$$

□

How can we verify the condition of the above proposition, and how do we calculate the Radon-Nikodym density $r(y, x)$? The following Lemma, which is a simple exercise in measure theory, covers the important cases:

Lemma 4.1. *Let P_1 and P_2 be two probabilities on $(\mathbb{X}, \mathcal{F})$.*

1. *If P_1 and P_2 have densities p_1 and p_2 w.r. to a σ-finite measure μ, then P_1 is absolutely continuous with respect to P_2 iff $\{x|p_2(x) = 0, p_1(x) > 0\}$ is a null set with respect to μ. The Radon-Nikodym density of P_1 with respect to P_2 is then $p_1(x)/p_2(x)$, independent of the choice of μ.*

2. *Let ϕ be a measurable injective mapping from $(\mathbb{X}, \mathcal{F})$ to $(\mathbb{Y}, \mathcal{G})$ and let P_i' denote the distribution $P_i \circ \phi^{-1}$ (i.e. $\phi(X) \sim P_i'$ if $X \sim P_i$). If P_1 has the density r with respect to P_2, then P_1' has the density $r(\phi^{-1}(y))$ with respect to P_2'.*

We apply this lemma with $P_1 = \pi(dx)Q(x, dy)$ and $P_2 = \pi(dy)Q(y, dx)$. In the simplest examples both $\pi(dx)$ and the proposal distributions $Q(x, dy)$ have densities $\pi(x)$ and $q(x, y)$ respectively with respect to the Lebesgue measure in the continuous case or counting measure in the discrete case. Then $\pi(dx)Q(x, dy)$ and $\pi(dy)Q(y, dx)$ are equivalent if for all pairs (x, y)

$$\pi(x)q(x, y) > 0 \Leftrightarrow \pi(y)q(y, x) > 0.$$

This implies $q(x, y) = 0$ if $\pi(x) > 0$ and $\pi(y) = 0$, as well as $q(x, y) > 0 \Leftrightarrow q(y, x) > 0$. In this case

$$r(y, x) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \Rightarrow a(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right).$$

Because only the ratio $\pi(y)/\pi(x)$ appears in this formula, it is sufficient to know π up to a normalizing constant.

Whether the Metropolis-Hastings transition is irreducible or not depends on Q. Irreducibility of Q is transferred to P. In particular, $q(x, y) > 0$ for all x, y is sufficient (but not necessary).

Two simple examples are

Example 4.1 (Independence sampler)**.** *Let $q(x, y) = q(y)$ for all x, i.e., the proposed value y is independent of the current state x. This vlaue is then accepted with probability*

$$a(x, y) = \min\left(1, \frac{\pi(Y)q(x)}{q(Y)\pi(x)}\right).$$

This is similar to the rejection algorithm and to importance sampling. If we also choose X_0 with distribution Q, then we can write

$$\frac{1}{N+1}\sum_{t=0}^{N} h(X_t) = \sum_{i=1}^{n} w_i h(Y_i)$$

where n is the number of accepted values, every Y_i is generated according to Q and w_i denotes the relative frequency of Y_i in $(X_0, X_1, ...X_N)$. In contrast to importance sampling, the Y_i are dependent here.

Example 4.2 (Random walk Metropolis)**.** *Let $\mathbb{X} = \mathbb{R}^p$ and $q(x, y) = q(y - x)$ with $q(x) = q(-x)$. This means, if $X_t = x$, then $Y = x + \varepsilon$ with $\epsilon \sim q(z)dz$, independent of x. In other words, the proposal is a random walk in $\mathbb{R}^p$. The acceptance probability becomes*

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

i.e. if the probability of the proposed value is greater than the probability of the current value, one always accepts the proposed value, otherwise only with some probability < 1.

4.2.1 Componentwise Modification

In many cases choosing a proposal distribution $Q(x, dy)$ with a density does not lead to an efficient algorithm. With such a choice, the proposed value can be anywhere in the space $\mathbb{X}$. If the current value x is plausible for π and $\mathbb{X}$ is high-dimensional, then the proposed value will almost always be less plausible than the current one, hence rejection is practically certain. In such cases it is much better if the proposed value differs from the current one in only a few components.

We formulate the algorithm in the case of a product space with two components (the second component may again be a product space). Hence, let $\mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2$, i.e., $x \in \mathbb{X}$ has the form $(x_1 x_2)$ with $x_k \in \mathbb{X}_k$. We further assume that π is absolutely continuous with respect to the product measure $\mu_1(dx_1)\mu_2(dx_2)$ and we denote the density also by π. We will consider a proposal distribution Q, which modifies only the first component with an absolutely continuous distribution while the second component remains the same:

$$Q(x, A_1 \times A_2) = \int_{A_1} q(x, y_1)\mu_1(dy_1) \cdot \mathbf{1}_{A_2}(x_2).$$

Then $\pi(dx)Q(x, dy)$ and $\pi(dy)Q(y, dx)$ are concentrated on the set of pairs with $y_2 = x_2$ whereas the three components (x_1, x_2, y_1) have densities $\pi((x_1 x_2))q((x_1 x_2), y_1)$ and $\pi((y_1 x_2))q((y_1 x_2), x_1)$ respectively. If we set $\phi(x_1, x_2, y_1) = (x_1, x_2, y_1, x_2)$, then the second statement of Lemma 4.1 shows that the conditions of Theorem 4.2 are satisfied and

$$r(x, y) = \frac{\pi((x_1 x_2))q((x_1 x_2), y_1)}{\pi((y_1 x_2))q((y_1 x_2), x_1)} \mathbf{1}_{y_2 = x_2}$$

if the numerator and denominator are both non-zero. Note that for pairs with $x_2 \neq y_2$, it does not matter how $r(x, y)$ is defined, since we only propose values with $x_2 = y_2$. This expression can also be writen with the help of the conditional density $\pi_{1|2}(x_1|x_2)$ of the first component given the second:

$$r(x, y) = \frac{\pi_{1|2}(x_1|x_2)q((x_1 x_2), y_1)}{\pi_{1|2}(y_1|x_2)q((y_1 x_2), x_1)}.$$

If one modifies only one component, one can of course never get an irreducible kernel. But one can analogously consider a second kernel, which modifies only the second component, and then apply both kernels in an alternating sequence. Likewise, instead of 2 components, one can also consider k components and for each component a kernel which keeps all other components fixed. The combination can be made according to a fixed or random order. For a fixed order, the reversibility is usually lost, but the stationary distribution does not change.

As a proposal density, we can in particular use

$$q(x, y_1) = q(x_2, y_1) = \pi_{1|2}(y_1|x_2) \tag{4.5}$$

Then the Radon-Nikodym density r above is always one, and so the proposal is always accepted. The combination of these kernels is nothing else than the Gibbs sampler already mentioned in the first chapter.

We can also generalize the idea of random walk Metropolis algorithm and use a proposal density of the form

$$q(x, y_1) = q(x_1, y_1) = q(y_1 - x_1).$$

If the random walk is symmetric, the acceptance probabilities are equal to

$$\min\left(1, \frac{\pi_{1|2}(y_1|x_2)}{\pi_{1|2}(x_1|x_2)}\right).$$

4.2.2 Metropolis-Hastings on the Space of Step Functions

The most complicated case that we discuss is the one where $\mathbb{X}$ is the union of the subspaces of different dimensions and where the Markov chain jumps between these subspaces. We explain the problem and the idea in this section in the example where we want to simulate random piecewise constant functions on $[0, 1]$. That is, we consider the space

$$\mathbb{X} = \bigcup_{k=0}^{\infty} \mathbb{X}_k$$

where $\mathbb{X}_k$ describes the space of piecewise constant functions with exactly k jumps. We parametrize the elements of $\mathbb{X}_k$ by the jump points $(t_i; i = 1, ..., k)$ and the function values $(g_i; i = 1, ..., k + 1)$, i.e.

$$x(t) = \sum_{i=1}^{k+1} g_i \mathbf{1}_{(t_{i-1},t_i]}(t)$$

with $t_0 = 0$ and $t_{k+1} = 1$. The space $\mathbb{X}_k$ is a subset of $\mathbb{R}^{2k+1}$, and we will identify $xin\mathbb{X}_k$ with $((t_i), (g_i)) \in \mathbb{R}^{2k+1}$. We denote the distribution on $\mathbb{X}$ from which we want to simulate by π, and we assume that for all k π restricted to $\mathbb{X}_k$ has a density π_k with regard to the Lebesgue measure on $\mathbb{R}^{2k+1}$.

This situation occurs for example in Bayesian nonparametric regression which uses the model

$$Y_i = x(s_i) + \varepsilon_i \quad (i = 1, ..., n)$$

where the observation points s_i are fixed (e.g., equidistant, i.e., $s_i = (i - 0.5)/n$), and the error terms ε_i are i.i.d. $\sim \mathcal{N}(0, 1)$. The mean function of x is unknown. For simplicity, the variance of the error ϵ_i is assumed to be known, but there would be no problem treating it as additional unknown parameter. As a prior distribution for x, we choose a distribution on our space $\mathbb{X}$ of jump functions: We choose a Poisson(λ)-distribution for the number of jumps, and given the number of jumps, we choose the jump times t_i as i.i.d. uniform and the function values g_i as i.i.d. $\mathcal{N}(0, \tau^2)$-distributed. The density of π on $\mathbb{X}_k$ is then

$$\pi_k(x) = \exp(-\lambda)\frac{\lambda^k}{k!}k!\ \mathbf{1}_{[t_1<t_2<\ldots<t_k]}\frac{1}{(\sqrt{2\pi}\tau)^{k+1}} \exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{k+1} g_i^2\right).$$

By Bayes formula, the posterior distribution of x given the observations $y_1, \ldots, y_n$ has on $\mathbb{X}_k$ the density

$$\pi_k(x|(y_j)) \propto \pi_k(x)\frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{k+1}\sum_{j=1}^{n} \mathbf{1}_{(t_{i-1},t_i]}(s_j)(y_j - g_j)^2\right).$$

We want to use the Metropolis-Hastings recipe. This means that we propose transitions according to a distribution Q and then by an appropriate acceptance probability ensure that the target distribution π (the posterior) is reversible. In order to access the whole space $\mathbb{X}$, we need to have transitions from $\mathbb{X}_k$ to $\mathbb{X}_j$ with $j \neq k$. The simplest transitions go from $\mathbb{X}_k$ to $\mathbb{X}_{k-1}$ and $\mathbb{X}_{k+1}$ and they do not modify the step-function x everywhere: We only add or delete a jump. Because the two functions are partially identical, the transitions are then certainly not absolutely continuous. Specifically, we propose for a current value $x = ((t_i), (g_i)) \in \mathbb{X}_k$ a value $z = ((r_i), (h_i)) \in \mathbb{X}_{k+1}$ according to the following algorithm:

1. Choose the j-th interval $I_j = (t_{j-1}, t_j]$ for subdivision with probability $t_j - t_{j-1}$ (long intervals have a greater probability of being divided). Set $r_i = t_i$ and $h_i = g_i$ for $i < j$, $r_i = t_{i-1}$ for $i > j$ and $h_i = g_{i-1}$ for $i > j+1$ (i.e. x is unchanged outside of I_j).
2. Choose the new jump point r_j uniformly on I_j.
3. Select two new function values h_j and h_{j+1} according to a density f, independently of r_j.

This defines a transition kernel $Q_k^+(x, dz)$ from $\mathbb{X}_k$ to $\mathbb{X}_{k+1}$. The distribution $\pi_k(dx)Q_k^+(x, dz)$ on $\mathbb{X}_k \times \mathbb{X}_{k+1}$ is concentrated on the union of the sets

$$A_{j,k} = \{(x, z) | x(t) = z(t) \ \forall t \notin (t_{j-1}, t_j]\} \quad (j = 1, 2, ..., k+1).$$

Each pair (x, z) in A_{jk} has $2k + 4$ free components, and we parametrize (x, z) with the components of x, the new jump point and the two new heights on both sides of the new jump. Then $\pi_k(dx)Q_k^+(x, dz)$ has for these parameters the density

$$\pi_k(t_1, ..t_k, g_1, ...g_{k+1})(t_j - t_{j-1})\mathbf{1}_{(t_{j-1}, t_j]}(r_j)\frac{1}{t_j - t_{j-1}}f(h_j)f(h_{j+1}). \tag{4.6}$$

In order to apply Theorem 4.2, we have to allow also transitions from $\mathbb{X}_{k+1}$ to $\mathbb{X}_k$ according to some kernel Q_{k+1}^-. Moreover, $\pi_{k+1}(dz)Q_{k+1}^-(z, dx)$ must be concentrated on the same sets A_{jk} and it must also have a density. Hence $Q_{k+1}^-(z, dx)$ must remove exactly one of the $k+1$ jumps of z, each jump must have positive probability to be removed, and the new jump height must be selected according to a density.

Specifically, the following definition of the transition kernel $Q_{k+1}^-(z, dx)$ from $\mathbb{X}_{k+1}$ to $\mathbb{X}_k$ satisfies all the requirements

1. Choose the jump point r_j to be eliminated at random, i.e. let j be uniform on $(1, ..., k+1)$. Set $t_i = r_i$ and $g_i = h_i$ for $i < j$, $t_i = r_{i+1}$ for $i \geq j$ and $g_i = h_{i+1}$ for $i > j$.
2. Choose the new function value g_j according to the density f.

(We take the same parametrization of A_{jk} as above, that is z has jump times $t_1, ..., t_{j-1}, r_j, t_j, ..., t_k$ and function values $g_1, ...g_{j-1}, h_j, h_{j+1}, g_{j+1}, ...g_{k+1}$) The density of $\pi_{k+1}(dz)Q_{k+1}^-(z, dx)$ on A_{jk} is then equal to

$$\pi_{k+1}(z)\frac{1}{k+1}f(g_j). \tag{4.7}$$

Suppose the current state is $x \in \mathbb{X}_k$. In order to completely specify the proposal, we have to decide whether we add or eliminate a jump. We do this by tossing a coin with parameter β_k (where $\beta_0 = 1$). Then our proposal distribution Q can be written as

$$Q(x, dz) = \beta_k Q_k^+(x, dz)\mathbf{1}_{\mathbb{X}_{k+1}}(z) + (1 - \beta_k)Q_k^-(x, dz)\mathbf{1}_{\mathbb{X}_{k-1}}(z) \quad (x \in \mathbb{X}_k).$$

The condition of Theorem 4.2 is satisfied by this choice: $\pi(dx)Q(x,dz)$ and $\pi(dz)Q(x,dz)$ are concentrated on the sets A_{jk}, and according to Lemma 4.1 (ii) the Radon-Nikodym density on an A_{jk} is equal to the quotient of β_k times the density (4.6) and $(1-\beta_{k+1})$ times the density (4.7). Thus the acceptance probability are

$$a(x,z) = \min\left(1, \frac{\pi_{k+1}(z)(1-\beta_{k+1})f(g_j)}{\pi_k(x)\beta_k f(h_j)f(h_{j+1})(k+1)}\right) \quad ((x,z)\in A_{jk})$$

and

$$a(z,x) = \min\left(1, \frac{\pi_k(x)\beta_k f(h_j)f(h_{j+1})(k+1)}{\pi_{k+1}(z)(1-\beta_{k+1})f(g_j)}\right) \quad ((x,z)\in A_{jk}).$$

This allows simulation both from the prior and the posterior: For the latter, we simply take $\pi_k(x|y)$ and $\pi_{k+1}(z|y)$ instead of $\pi_k(x)$ and $\pi_{k+1}(z)$, respectively.

The transition mechanism described above is obviously not the only possible one. Sometimes it is advantageous to propose only modifications where the mean $\int_0^1 x(t)dt$ is constant. This is achieved by the following algorithm for adding a jump:

1. Choose the j-th interval $I_j = (t_{j-1}, t_j]$ for subdivision with probability $t_j - t_{j-1}$. Then set $r_i = t_i$ and $h_i = g_i$ for $i < j$, $r_i = t_{i-1}$ for $i > j$ and $h_i = g_{i-1}$ for $i > j+1$.

2. Choose the new jump point r_j uniformly on I_j.

3. Choose as new values of the function

$$h_j = g_j + \frac{u}{r_j - t_{j-1}}, \; h_{j+1} = g_j - \frac{u}{t_j - r_j}$$

   where u has the density f and is independent of r_j.

This defines another transition kernel $Q_k^+(x,dz)$ from $\mathbb{X}_k$ to $\mathbb{X}_{k+1}$. The distribution $\pi_k(dx)Q_k^+(x,dz)$ on $\mathbb{X}_k \times \mathbb{X}_{k+1}$ is now concentrated on the union of the sets

$$B_{j,k} = \{(x,z)|x(t) = z(t) \; \forall t \notin (t_{j-1}, t_j], \int_0^1 x(t)dt = \int_0^1 z(t)dt\}$$

Every pair (x,z) in B_{jk} has $2k+3$ free components. We can for instance freely choose the components of x, the new jump point r and the above-defined variable u. For these variables $\pi_k(dx)Q_k^+(x,dz)$ has the density

$$\pi_k(t_1,...t_k,g_1,...g_{k+1})(t_j - t_{j-1})\mathbf{I}_{(t_{j-1},t_j]}(r_j)\frac{1}{t_j - t_{j-1}}f(u). \tag{4.8}$$

In order to apply Theorem 4.2, $\pi_{k+1}(dz)Q_{k+1}^-(z,dx)$ must again be concentrated on the same sets B_{jk} and also have a density. This means that $Q_{k+1}^-(z,dx)$ must remove exactly one of the $k+1$ jumps, each jump must have positive probability to be removed and the new jump height must be equal to the weighted average of the two old jump heights. Except for choosing which jump is removed, the transition is therefore deterministic, and the density of $\pi_{k+1}(dz)Q_{k+1}^-(z,dx)$ on B_{jk} is thus essentially the density $\pi_{k+1}(z)$. We must however express $z = (r_1,\ldots,r_{k+1},h_1,\ldots,h_{k+2})$ with the variables $(t_1,\ldots,t_k,g_1,\ldots,g_{k+1},r_j,u)$, and we must also take this change of variables into account in the density by multiplying with the Jacobian determinant. For most components the change of variables is trivial: $r_i = t_i$ and $h_i = g_i$ for $i < j$, $r_i = t_{i-1}$ for $i > j$ and $h_i = g_{i-1}$ for $i > j+1$. Finally, the relation between (h_j, h_{j+1}) and (g_j, u) is given in step 3. above. This relation has the Jacobian determinant

$$\frac{t_j - t_{j-1}}{(t_j - r_j)(r_j - t_{j-1})}.$$

Thus if each jump has the same probability to be removed, the density of $\pi_{k+1}(dz)Q^-_{k+1}(z, dx)$ is equal to

$$\pi_{k+1}(z)\frac{1}{k+1}\frac{t_j - t_{j-1}}{(t_j - r_j)(r_j - t_{j-1})}. \tag{4.9}$$

Using this, we can compute the acceptance probability.

4.2.3 General Transitions between Spaces of Different Dimensions

Now we generalize the approach that we have seen in the previous section. Let

$$\mathbb{X} = \cup_{k=0}^{\infty}\mathbb{X}_k,$$

where $\mathbb{X}^k$ is an open subset of $\mathbb{R}^k$ and assume π has a strictly positive density π_k (with respect to Lebesgue measure) on each $\mathbb{X}_k$. In order to simulate from π, we consider transitions from $\mathbb{X}_k$ to $\mathbb{X}_m$ of the type

$$x \to z = z(x, U_{km}),$$

where U_{km} is a d_{km} dimensional random variable with strictly positive density f_{km}. The relationship $z = z(x, u_{km})$ is assumed to be deterministic. Let $Q_{km}(x, dz)$ be the corresponding transition kernel. Then the distribution $\pi_k(dx)Q_{km}(x, dz)$ is concentrated on a $k + d_{km}$-dimensional surface in $\mathbb{R}_{k+m}$; namely, the surface consisting of all pairs of the form $(x, z(x, u_{km}))$. Moreover, the density of (x, u_{km}) is equal to $\pi_k(x)f_{km}(u_{km})$.

To satisfy the conditions of Theorem 4.2, we must therefore allow also a transition $Q_{mk}(z, dx)$ from $\mathbb{X}_m$ to $\mathbb{X}_k$ with the property that $\pi_m(dz)Q_{mk}(z, dx)$ is concentrated on the same surface. If $Q_{mk}(z, dx)$ has the same structure as $Q_{km}(x, dz)$, i.e. if a d_{mk}-dimensional random variable u_{mk} is drawn and x is obtained deterministically as $x(z, u_{mk})$, then $\pi_m(dz)Q_{mk}(z, dx)$ is concentrated on the surface $(z, x(z, u_{mk}))$. So we have two parametrizations of the same surface. In particular the dimensions must match:

$$k + d_{km} = m + d_{mk},$$

and there must be a bijection

$$(x, u_{km}) \leftrightarrow (z, u_{mk})$$

Finally, for the Radon-Nikodym density, we need to convert the density $\pi_m(z)f_{mk}(u_{mk})$ for (z, u_{mk}) into the density for (x, u_{km}), i.e., we need to multiply by the Jacobian

$$\left|\frac{\partial(z, u_{mk})}{\partial(x, u_{km})}\right|.$$

To complete the proposal distribution, we also have to choose the dimension m of the proposal. This can be done with a stochastic matrix (β_{km}), which leads to the following transition kernel from $\mathbb{X}$ into itself:

$$Q(x, dz) = \sum_{j=0}^{\infty}\beta_{kj}Q_{kj}(x, dz)\mathbf{1}_{\mathbb{X}_j}(z) \quad (x \in \mathbb{R}^k).$$

In summary, we give the formula for the acceptance probabilities as a result of the above considerations:

$$a(x,z) = \min\left(1, \frac{\pi_m(z)\beta_{mk}f_{mk}(u_{mk})}{\pi_k(x)\beta_{km}f_{km}(u_{km})}\left|\frac{\partial(z,u_{mk})}{\partial(x,u_{km})}\right|\right) \quad (x \in \mathbb{X}_k, z = z(x,u_{km}) \in \mathbb{X}_m).$$

4.3 The Accuracy of MCMC Approximations

If we use Markov chain Monte Carlo, then the random variables X_1, X_2, X_3,...are dependent and not identically distributed. In particular, X_t has the distribution π only asymptotically for $t \to \infty$. If we use the estimator

$$\hat{\theta}_N = \frac{1}{N}\sum_{t=1}^{N} h(X_t)$$

for $\theta = \int h(x)\pi(dx)$, then we make an systematic error:

$$\mathbb{E}(\hat{\theta}_N) = \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}(h(X_t)) \neq \theta.$$

This bias is of the order $O(1/N)$, provided

$$\sum_{t=1}^{\infty}\left|\mathbb{E}(h(X_t)) - \int h(x)\pi(dx)\right| < \infty.$$

Furthermore, the dependence of X_t's changes also the distribution of the random error $\hat{\theta}_N - \mathbf{E}\left[\hat{\theta}_N\right]$. In particular,

$$\mathrm{Var}(\hat{\theta}_N) = \frac{1}{N^2}\sum_{s=1}^{N}\sum_{t=1}^{N} \mathrm{Cov}(h(X_s), h(X_t)),$$

and the covariances are not zero in general.

The mean square error (MSE) takes into account both the bias and the random error:

$$\mathbb{E}\left((\hat{\theta}_N - \theta)^2\right) = \left(\mathbb{E}(\hat{\theta}_N) - \theta\right)^2 + \mathrm{Var}\left(\hat{\theta}_N\right).$$

For an error bound, we need to estimate both the bias and the variance. These are unfortunately rather difficult problems. We will see that typically the variance of $\hat{\theta}_N$ is still of order $O(1/N)$. As mentioned above, the bias is typically of $O(1/N)$, and its contribution to the mean square error is asymptotically negligible (because the bias is then squared). It is not clear if this is also true in the case of a finite N.

For the bias we are often satisfied with graphical tools, e.g. a plot of $h(X_t)$ versus t. Based on this, we try to find a time t_0 after which systematic deviations no longer occur. Then we only use the values after iteration t_0 and ignore the bias.

In the following, we first make a few theoretical considerations about bias and variance of arithmetic means in the case of Markov chains, and then discuss the treatment of dependence in the stationary case, where all X_t have the target distribution π, but are not independent.

4.3.1 Convergence Results on Markov Chains

We discuss here the bias of a Markov Chain Monte Carlo method. Let (X_t) be a Markov chain with initial distribution ν, transition P and invariant distribution π. We want to estimate how quickly

$$\mathbb{E}(h(X_t)) - \int h(x)\pi(dx) = \int P^t h(x)\nu_0(dx) - \int P^t h(x)\pi(dx)$$

converges to zero.

We restrict ourselves to the simplest case, where π is a probability on the discrete space $\{1, 2, ...n\}$. Then we have

$$|\mathbb{E}(h(X_t)) - \int h(x)\pi(dx)| = |\sum_j (\nu_0 P^t(j) - \pi P^t(j))h(j)| \leq \max_i |h(i)| \sum_j |\nu_0 P^t(j) - \pi P^t(j)|.$$

In particular, it is sufficient to bound the L_1-distance

$$||\nu_0 P^t - \pi P^t||_1 = \sum_j |\nu_0 P^t(j) - \pi P^t(j)|.$$

An algebraic approach uses a result on the eigenvalues of the transition matrix P: The Frobenius theorem states that the eigenvalue with the largest absolute value of a stochastic irreducible and aperiodic matrix equals to 1 and its multiplicity is 1. The convergence speed is then determined by the eigenvalue with the second largest absolute value.

We use here a stochastic method, the *coupling* of Markov chains. This means that one constructs a Markov process $(X_t^{(1)}, X_t^{(2)})$ on the state space $(1, 2, ...n)^2$ with the following characteristics: Marginally, $(X_t^{(1)})$ and $(X_t^{(2)})$ are both Markov chains with transition matrix P and initial distributions μ and ν, respectively. These two chains are *dependent* because they stay together after they have met for the first time, that is if $X_t^{(1)} = X_t^{(2)}$ for some t, then $X_s^{(1)} = X_s^{(2)}$ for all $s > t$. How we make the transition as long as $X_{t-1}^{(1)} \neq X_{t-1}^{(2)}$ is left open.

If we introduce the transition matrix for the coupled process

$$Q(i, j; k.l) = \mathbb{P}(X_t^{(1)} = k, X_t^{(2)} = l | X_{t-1}^{(1)} = i, X_{t-1}^{(2)} = j)$$

then the above requirements translate into

$$\begin{aligned}
\sum_l Q(i, j; k, l) &= P(i, k) \text{ for all } i \neq j, k \\
\sum_k Q(i, j; k, l) &= P(j, l) \text{ for all } i \neq j, l \\
Q(i, i; k, k) &= P(i, k), \\
Q(i, i; k, l) &= 0 \text{ if } k \neq l
\end{aligned}$$

There are still many choices for Q, the easiest one being $Q(i, j; k, l) = P(i, k)P(j, l)$ if $i \neq j$ which means that transitions occur independently of each other as long as the chains have not met.

Lemma 4.2. *For any coupling satisfying the above properties,*

$$||\nu P^t - \mu P^t||_1 \leq 2\mathbb{P}(X_t^{(1)} \neq X_t^{(2)}).$$

Moreover, there is a coupling such that

$$\mathbb{P}(X_t^{(1)} \neq X_t^{(2)}) \leq \alpha^t \mathbb{P}(X_0^{(1)} \neq X_0^{(2)})$$

where

$$\alpha = \frac{1}{2} \max_{i,j} ||P(i,.) - P(j,.)||_1.$$

The proof of Lemma 4.2 is based on the following Lemma

Lemma 4.3. *For two probabilities $P^{(1)}$ and $P^{(2)}$ on a discrete space, we have*

$$\begin{aligned}\frac{1}{2}||P^{(1)} - P^{(2)}|| &= \sum_j (p^{(1)}(j) - p^{(2)}(j))_+ = 1 - \sum_j \min(p^{(1)}(j), p^{(2)}(j)) = \sup_A |P^{(1)}(A) - P^{(2)}(A)| \\ &= \min\left\{\sum_{i \neq j} r(i,j); r \geq 0, \sum_j r(i,j) = p^{(1)}(i), \sum_j r(j,i) = p^{(2)}(i)\right\}.\end{aligned}$$

(x_+ is the positive part of x, i.e $x_+ = \max(x, 0)$.)

The last expression is nothing else than the minimum of $\mathbb{P}(X \neq X')$ over all joint distributions of (X, X') such that $X \sim P^{(1)}$ and $X' \sim P^{(2)}$. The distribution R which realizes the minimum in the last expression, is called the optimal coupling of $P^{(1)}$ and $P^{(2)}$. If we use the optimal coupling of $P(i,.)$ and $P(j,.)$ in order to define the joint transition matrix Q, then the probability that $X_t^{(1)} = X_t^{(2)}$ given $X_{t-1}^{(1)} \neq X_{t-1}^{(2)}$ is at least α for any t, and so the second claim in Lemma 4.2 follows. The first claim follows because any coupling of the two Markov chains also induces a coupling of μP^t and νP^t.

Proof. (of Lemma 4.3). We set $x_- = \max(-x, 0)$. Then $x = x_= - x_-$ and $|x| = x_+ + x_-$. Hence the first equality follows from $\sum_j (p^{(1)}(j) - p^{(2)}(j)) = 0$. For the second equality, we observe that

$$(p^{(2)}(j) - p^{(1)}(j))_+ = p^{(2)}(j) - \min(p^{(1)}(j), p^{(2)}(j)).$$

For the third equality, we use that for any A

$$P^{(1)}(A) - P^{(2)}(A) = \sum_{j \in A} (p^{(1)}(j) - p^{(2)}(j)) \leq \sum_{j; p^{(1)}(j) > p^{(2)}(j)} (p^{(1)}(j) - p^{(2)}(j)) = \sum_j (p^{(1)}(j) - p^{(2)}(j))_+,$$

and for $A = \{j; p^{(1)}(j) > p^{(2)}(j)\}$ we have equality. Exchanging the role of $P^{(1)}$ and $P^{(2)}$ thus proves the third equality.

For the last equality, we again prove two inequalities. For every r satisfying the specified conditions, and for every A we have

$$|P^{(1)}(A) - P^{(2)}(A)| = |\sum_{i,j} r(i,j)(\mathbf{1}_A(i) - \mathbf{1}_A(j))| \leq \sum_{i,j} r(i,j)|\mathbf{1}_A(i) - \mathbf{1}_A(j)| \leq \sum_{i \neq j} r(i,j).$$

For the other inequality, we choose r as follows

$$r(i,j) = \begin{cases} \min(p^{(1)}(i), p^{(2)}(i)) & \text{if} \quad i = j \\ \frac{2}{||P^{(1)} - P^{(2)}||_1}(p^{(1)}(i) - p^{(2)}(i))_+ (p^{(2)}(j) - p^{(1)}(j))_+ & \text{if} \quad i \neq j \end{cases}.$$

One can easily check that this r satisfies the specified conditions. Obviously

$$\sum_{i \neq j} r(i,j) = 1 - \sum_i r(i,i) = 1 - \sum_i \min(p^{(1)}(i), p^{(2)}(i)).$$

□

The proof also shows how we can sample from the optimal coupling r. We set $\gamma = \sum_i \min(p^{(1)}(i), p^{(2)}(i))$ and then with probability γ we generate $X = X'$ according to the distribution $(\min(p^{(1)}(i), p^{(2)}(i))/\gamma)$; and with probability $1-\gamma$ X and X' are independent with the distributions $((p^{(1)}(i) - p^{(2)}(i))_+/(1-\gamma))$ and $((p^{(2)}(i) - p^{(1)}(i)))_+/(1-\gamma))$ respectively.

If $\alpha = 1$, there must be two states i and j such that for all k either $P(i,k) = 0$ or $P(j,k) = 0$, that is there is no state which can be reached both from i and j. This happens for instance in the following example:

Example 4.3. *Consider the random walk on* $(1,2,3,4,5)$ *with reflection at the boundary:*

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

The uniform distribution $\pi = (\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$ *is reversible. If we make the transition for all chains with the same step direction* U_t, *then the chains couple the first time four steps in the same direction are chosen. In other words, for any* i, j

$$\mathbb{P}(X_t^{(1)} = X_t^{(2)} | X_{t-4}^{(1)} = i, X_{t-4}^{(2)} = j) \geq 2(\frac{1}{2})^4 = \frac{1}{8} > 0.$$

This again gives exponential convergence. The idea to consider several steps at once is helpful in general.

For more complex transitions, however, the problem remains difficult to give sharp a priori estimates for the bias. Alternatively, one can try after the simulation has been carried out to infer from the plot of $h(X_t)$ against t "when the chain has converged".

4.3.2 Estimation of the Variance in the Stationary Case

Next, we consider the stochastic error under the assumption that $(X_1, X_2, ..., X_k)$ and $(X_{i+1}, X_{i+2}, ..., X_{i+k})$ have the same distribution for all i and for all k, i.e. (X_t) is stationary.

If (X_i) is a Markov chain with a transition kernel that does not depend on time, then we have stationarity if and only if $X_1 \sim \pi$ (π is the stationary distribution). Stationarity is therefore often reasonable after an initial "burn in" period has been deleted.

Lemma 4.4. *Let* (X_i) *be stationary,* $Y_i = h(X_i)$, *and* $R(k) = \mathrm{Cov}(Y_i, Y_{i+k})$. *Then*

1.

$$\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) = \frac{1}{N}\sum_{k=-N+1}^{N-1}(1 - \frac{|k|}{N})R(k).$$

2. If $\sum_{k=1}^{\infty} |R(k)| < \infty$, *then as* $N \to \infty$

$$N\,\mathrm{Var}(\hat{\theta}_N) \to \sigma_\infty^2 = \sum_{k=-\infty}^{\infty} R(k).$$

3. If $\sum_{k=1}^{\infty}|R(k)|<\infty$, then

$$\mathit{Corr}\left(\frac{1}{N}\sum_{i=1}^{N}Y_i,\frac{1}{N}\sum_{i=N+1}^{2N}Y_i\right)\to 0.$$

Proof. Expression 1. follows from

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^{N}Y_i\right) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\underbrace{\operatorname{Cov}(Y_i,Y_j)}_{=R(i-j)}\\ &= \sum_{k=-N+1}^{N-1}R(k)\cdot\underbrace{(\text{number of pairs with } i-j=k)}_{(=N-|k|)}\end{aligned}$$

For 2. we write

$$\begin{aligned}N\operatorname{Var}\left(\hat{\theta}_N\right) &= \sum_{k=-N+1}^{N-1}(1-\frac{|k|}{N})R(k)\\ &= \sum_{k=-\infty}^{\infty}\underbrace{\max(0,1-\frac{|k|}{N})R(k)}_{\to R(k)\text{ as }N\to\infty}\end{aligned}$$

The claim therefore follows from the convergence theorem of Lebesgue.

For 3. we start with

$$\operatorname{Cov}\left(\sum_{i=1}^{N}Y_i,\sum_{i=N+1}^{2N}Y_i\right)=\sum_{k=1}^{2N-1}\min(k,2N-k)\cdot R(k).$$

Because of 2., it is sufficient to show that the expression on the right is growing less rapidly than N. This follows from the following estimate:

$$\begin{aligned}|\sum_{k=1}^{2N-1}\min(k,2N-k)\cdot R(k)| &\leq \sqrt{N}\sum_{k=1}^{\sqrt{N}}|R(k)|+N\sum_{k=\sqrt{N}}^{\infty}|R(k)|\\ &\leq \sqrt{N}\sum_{k=1}^{\infty}|R(k)|+N\sum_{k=\sqrt{N}}^{\infty}|R(k)|=o(N)\end{aligned}$$

□

Under the assumption $\sum|R(k)|<\infty$ we therefore have (using the Chebyshev inequality)

$$\mathbb{P}\left(|\hat{\theta}_N-\theta|>\epsilon\right)\leq\frac{\operatorname{Var}\left(\hat{\theta}_N\right)}{\epsilon^2}\sim\frac{\sigma_\infty^2}{N\epsilon^2}$$

Hence we need to estimate σ_∞. Moreover, since the Chebyshev inequality is usually not sharp, we would like to use the normal approximation instead.

This raises the following questions:

1. When is $\sum|R(k)|<\infty$?

2. How can we estimate σ_∞?
3. Does a central limit theorem hold?

About 1.: Let (X_i) be a stationary Markov chain with transition kernel P. Then we have

$$\begin{aligned}\mathrm{Cov}(h(X_0), h(X_t)) &= \mathbb{E}(h(X_0) - \theta)(h(X_t) - \theta)) \\ &= \mathbb{E}\left((h(X_0) - \theta)\mathbb{E}((h(X_t) - \theta)|X_0)\right) \\ &= \int (h(x) - \theta)(P^t h(x) - \theta)\pi(dx).\end{aligned}$$

So what matters is how quickly $P^t h(x) - \theta$ goes to zero, as in the analysis of the bias. In particular, the following condition is sufficient:

$$\sup_x \sum_t |P^t h(x) - \theta| < \infty.$$

About 2.: A natural estimate for $R(k)$ is:

$$\hat{R}(k) = \frac{1}{N} \sum_{i=1}^{N-|k|} (Y_i - \hat{\theta}_N)(Y_{i+|k|} - \hat{\theta}_N)$$

(The reason for the denominator N instead of $N - |k|$, is discussed in the course on time series analysis). However, $\sum_{k=-N+1}^{N-1} \hat{R}(k)$ is not suitable as an estimator of σ_∞^2 because

$$\sum_{k=-N+1}^{N-1} \hat{R}(k) = \left(\sum_{i=1}^{N}(Y_i - \hat{\theta}_N)\right)^2 = 0.$$

A better estimator is

$$\hat{\sigma}_\infty^2 = \sum_{k=-m}^{m} w_k \hat{R}(k), \tag{4.10}$$

where w_k are symmetric weights with $w_0 = 1 \geq w_1 \geq \ldots \geq w_{m+1} = 0$. We therefore downweight the covariances with increasing distance. The choice of the point m, from where on the estimated covariances have weight zero, is then crucial. In theory, $m \to \infty$ and $m = o(N)$, i.e. m grows, but slower than N, is sufficient. Empirically, $m \approx N^{1/3}$ is often a sensible choice.

About 3.: There is a large literature on the problem of the validity of the central limit theorem for stationary random variables. One of the simplest and most important result for Markov chain Monte Carlo is the following: If (X_i) is a Markov chain with P and π is reversible, then $\frac{1}{N}\sum h(X_i)$ is asymptotically normal if $\sum_k |R(k)| < \infty$.

To conclude, we consider the construction of a confidence interval for θ: Based on what has been said before, the following interval is obvious:

$$\hat{\theta}_N \pm \Phi^{-1}(1 - \frac{\alpha}{2})\frac{1}{\sqrt{N}}\,\hat{\sigma}_\infty.$$

Another possibility is the so-called "batch means" method. There one computes the means of b consecutive Y_i's:

$$\hat{\theta}_{i,b} = \frac{1}{b} \sum_{j=(i-1)b+1}^{ib} Y_j$$

These means $\hat{\theta}_{i,b}$, $i = 1, 2, ..., k = N/b$ are considered as independent and normally distributed, see the third statement of Lemma 4.4. The usual t-confidence interval is then

$$\hat{\theta}_N \pm \frac{1}{\sqrt{k}} t_{k-1,1-\frac{\alpha}{2}} \sqrt{\frac{1}{k-1} \sum_{i=1}^{k} (\hat{\theta}_{i,b} - \hat{\theta}_N)^2}$$

The advantage is that you one does not need to estimate σ_∞. The choice of b is however as difficult as the choice of m in (4.10).

4.3.3 Coupling from the Past

This is a rather recent idea to avoid the problem of bias in Markov Chain Monte Carlo. It would be nice if we could reach the stationary distribution in a finite number of steps. For this purpose we take up the concept of coupling that we have already introduced in the proof of convergence to the stationary distribution. If the paths from all possible starting values have coupled, then one also knows in particular the state of the stationary Markov chain that starts with π. However, one can not conclude that after the coupling the joint state distribution is equal to π. The time of the coupling is random, and at a random point the distribution is different from π even in the stationary chain.

This is evident in the example of the random walk with reflection at the boundary. The chains meet almost certain, but at the time T of the coupling all $X_T^{(i)} \in \{1, 5\}$, i.e. the distribution is definitely not equal to π. So we cannot use forward coupling to generate random variables which have exactly the distribution π.

The way out is to look at a fixed time $t = 0$, and introduce the coupling backwards (from the past). This means the following: First we go back to $t = -2$ and consider the chains with all starting values. If not all of these chains couple until time 0, we start at time $t = -4$. If this still is not sufficient to achieve coupling of all chains with arbitrary starting values, we go back to $t = -16$ etc. One can show that the value at time $t = 0$ generated in this way has exactly the distribution π. It is essential that as we go back further, we always use the same random numbers for the transitions at a fixed time.