MPS-LN 2

ISSN 1399-2538

Stochastic Simulation
with a view towards stochastic processes[1]

Søren Asmussen[2]

Concentrated Advanced Course
MaPhySto, University of Aarhus, Denmark, February 22–26, 1999
Version date March 18, 1999

[2]Department of Mathematical Statistics, University of Lund, Box 118, S–22100 Lund, Sweden; e–mail asmus@maths.lth.se

Contents

Preface

The present notes are based upon a Concentated Advanced Course given at MaPhySto, Aarhus, in February 1999. An earlier and incomplete version was written up in connection with a PhD course given at the Department of Mathematical Statistics, University of Lund, in the spring of 1996.

The aim has been to present some main aspects of simulation methodology at a reasonably advanced mathematical level. Some topics like random variate generation, Markov chains Monte Carlo metods and stochastic optimization are left out or treated briefly, since they form areas in themselves and extensive treatments are available elsewhere. However, in a later version of the notes I expect to complement the treatment in these areas; sections or chapters marked with a $*$ indicate that more will be filled in later.

Some general standard textbooks in stochastic simulation are Banks & Carson [22], Bratley, Fox & Schrage [29], Fishman [52], Law & Kelton [94], and Morgan [105]. They contain much practically oriented discussion not at all covered by these notes. References at a somewhat higher mathematical level are Ripley [126], Ross [131], Rubinstein [132] and Rubinstein & Melamed [133] (parts of Hammersley & Handscombe [75] are also still worth to read). However, much of the material in these notes can at present only be found in research papers or more specialized monographs; references are given at the appropriate places. In addition to standard journals in statistics and applied probability, the reader interested in pursuing the literature should be aware of journals like *ACM TOMACS* (*ACM Transactions of Modeling and Computer Simulation*), *Management Science*, and the *IEEE* journals.

The course was held from Monday February 22 to Friday February 26. Each day had 2–3 hours of lectures, covering main parts of Chapters **II**, **III**, **IV**, **VI** and **VIII**, and in addition there were computer labs on the afternoons of Tuesday to Friday, based upon some of the assignments given at the end of the notes. The labs used MatLab; the choice of this was just for convenience since MatLab is the standard at my home institution at Lund University, and I had the opportunity to have Sofia Andersson from Lund to guide the labs. The advantage of a package like MatLab is the availability of many subroutines for random variate generation, graphics etc. The drawback is that as a general programming language, MatLab is much slower that say Pascal or C++.

Aarhus and Lund, March 1999
Søren Asmussen

Chapter I

Basics

1 Uniform r.v.'s

The basic vehicle in the area of (stochastic) simulation is a stream of pseudorandom numbers produced by a computer, which is treated a sequence $U_1, U_2, \ldots$ of i.i.d. r.v.'s with a uniform distribution on $(0,1)$. In practice, $U_1, U_2, \ldots$ are typically produced by deterministic recursive algorithms. The most popular ones today are linear congruential generators,

$$U_n = \frac{X_n}{M} \quad \text{where} \quad X_{n+1} = (AX_n + C) \bmod M.$$

Here a particular popular choice, implemented in many computers and in many software packages, is $M = 2^{31} - 1$, $A = 16807$, $C = 0$.

Being deterministic, no stream of pseudorandom numbers is truly random and will fail to pass a sufficiently elaborate statistical goodness–of–fit test to the i.i.d. uniform setting. The generators used in practice typically have the property that the marginal empirical distribution of the U_n is uniform (up to rounding errors) and that observations look independent within a narrow time range. The essence is that *the generators work well in practice*. The novice is tempted to blame apparently erroneous simulation output to deficiencies in the generator. However, almost always the problem is an error of his own.

References: L'Ecuyer [96], Dodge [46] (suggests the decimals of π).

2 Non–uniform r.v.'s

Accepting the U_n as i.i.d. uniform, the next step is to use them produce a r.v. X with a prescribed distribution B. A simple case is a Bernoulli r.v.,

$\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = p$, where one can just let $X = I(U \leq p)$. This construction generalizes in a straightforward way to distributions with a finite support and is a special case of *inversion*: if B^{-1} is the inverse of the c.d.f. of B, $B^{-1}(u) = \min\{x : F(x) \geq u\}$, then $X = B^{-1}(U)$ has distribution B since

$$\mathbb{P}(X \leq x) = \mathbb{P}(B^{-1}(U) \leq x) = \mathbb{P}(U \leq B(x)) = B(x) \ .$$

For example, if B is exponential with rate δ, then $B^{-1}(x) = -\log(1-x)/\delta$ (in practice, one would use $X = -\log U/\delta$ rather than $X = -\log(1-U)/\delta$!).

From exponential r.v.'s, one can build Erlang(p)'s by simply adding p independent copies. Scaling an exponential r.v. by 2, one obtains a χ^2 with 2 degrees of freedom, which is the distribution of the squared radial part $R^2 = Y_1^2 + Y_2^2$ of independent normal(0,1) variates Y_1, Y_2. Since the conditional distribution of Y_1, Y_2 given $R = r$ is uniform on the circle with radius r, we obtain the *Box–Muller method* for generating normal random variates (in pairs):

$$Y_1 = \sqrt{-2\log U_1}\sin 2\pi U_2, \quad Y_2 = \sqrt{-2\log U_1}\cos 2\pi U_2.$$

In a similar vein, a number of standard distributions can be build. However, there are at least two barriers: efficiency concerns, since the evaluation of functions like the logarithm, square root, sine or cosine is typically much more time–consuming than the generation of uniform r.v.'s; and also the fact that B^{-1} is not explicitly available in quite a few cases.

A commonly used method is therefore instead *acceptance–rejection*. The idea is to start from a r.v. Y with a density $g(x)$ which is easily simulated and has the property $b(x) \leq Cg(x)$ where $b(x)$ is the density of X and $C < \infty$ is a constant. Given $Y = x$, one accepts Y and let $X = Y$ w.p. $b(x)/Cg(x)$. Otherwise, a new Y is generated and one continues until eventual acceptance. Algorithmically:

1. Generate Y from the density $g(x)$

2. Generate U as uniform(0,1)

3. If $U \leq b(Y)/Cg(Y)$, let $X \leftarrow Y$. Otherwise return to 1.

That this produces a r.v. with the desired density $b(x)$ follows from

$$\begin{aligned}\mathbb{P}(X \in dx) &= \mathbb{P}(Y \in dx | A) = \frac{\mathbb{P}(Y \in dx; A)}{\mathbb{P}A} \\ &= \frac{g(x) \cdot b(x)/Cg(x)}{\int_{-\infty}^{\infty} g(y) \cdot b(y)/Cg(y)\, dy} = \frac{b(x)}{\int_{-\infty}^{\infty} b(y)\, dy} \\ &= b(x),\end{aligned}$$

where $A = \{U \le b(Y)/Cg(Y)\}$ is the event of acceptance. For example, this applies to the case where $b(x)$ is a bounded density on $(0,1)$, $C = \sup_{0<x<1} b(x)$ and Y is uniform on $(0,1)$.

Here are some more elaborate examples or variants of the acceptance–rejection idea:

Example 2.1 (Ratio of uniforms*) See Ripley [126]. □

Example 2.2 (Inhomogeneous Poisson processes) The epochs σ_n of a standard Poisson process $\{N_t\}$ with rate β are easily generated by noting that the interarrival times $T_n = \sigma_n - \sigma_{n-1}$ can be generated as i.i.d. exponential. However, in many situations it would be reasonable to assume that the Poisson process $\{N_t^*\}$ has a rate $\beta(t)$ depending on time. Assuming $\beta(t) \le \beta$ for some constant $\beta < \infty$, $\{N_t^*\}$ can then be constructed by thinning $\{N_t\}$ with retention probability $\beta(t)/\beta$ at time t. That is, an epoch σ_n of $\{N_t\}$ is accepted (retained) as an epoch σ_m^* of $\{N_t^*\}$ w.p. $\beta(\sigma_n)/\beta$. Algorithmically:

1. Let $n \leftarrow 0$, $n^* \leftarrow 0$, $\sigma \leftarrow 0$, $\sigma^* \leftarrow 0$
2. Generate T as exponential with rate β;
   let $\sigma \leftarrow \sigma + T$, $n \leftarrow n + 1$
3. Generate U as uniform(0,1);
   if $U \le \beta(\sigma)/\beta$, let $n^* \leftarrow n^* + 1$, $\sigma^* \leftarrow \sigma$; item Return to 2

That this produces the correct intensity $\beta^*(t)$ for $\{N_t^*\}$ follows from

$$\begin{aligned}\beta^*(t)dt &= \mathbb{P}(\sigma_m^* \in [t, t+dt] \text{ for some } m = 0, 1 \ldots) \\ &= \mathbb{P}(\sigma_n \in [t, t+dt] \text{ for some } n = 0, 1 \ldots) \cdot \frac{\beta(t)}{\beta} \\ &= \beta dt \cdot \frac{\beta(t)}{\beta} = \beta(t)dt.\end{aligned}$$

□

Example 2.3 (Uniformization of Markov processes) Let $\{J_t\}_{t\geq 0}$ be a Markov process with a finite state space E and intensity matrix $\boldsymbol{\Lambda} = (\lambda_{ij})_{i,j\in E}$. One can simulate $\{J_t\}$ at transition epochs by noting that the holding time of state i is exponential with rate $\lambda_i = -\lambda_{ii}$, and that the next state j is chosen w.p. λ_{ij}/λ_i:

1. Let $t \leftarrow 0$, $J \leftarrow i_0$
2. Let $i \leftarrow J$;
   Generate T as exponential with rate λ_i and K with $\mathbb{P}(K = j) = \lambda_{ij}/\lambda_i$, $j \neq i$;
3. Let $t \leftarrow t + T$, $J \leftarrow K$ and return to 2

In this way, the rate λ_i of an event being created depends on the current state $i = J_t$. The uniformization algorithm creates instead events at a uniform rate η. A transition from i to $j \neq i$ then occurs w.p. λ_{ij}/η (thus, η should satisfy $\eta \geq \max_{i\in E} \lambda_i$) when the current state is $i = J_t$; if $\eta > \lambda_i$, this leaves the possibility of a dummy transition $i \to i$ (t is rejected as transition epoch). Algorithmically:

1. Let $t \leftarrow 0$, $J \leftarrow i_0$
2. Let $i \leftarrow J$;
   Generate T as exponential with rate η and K with $\mathbb{P}(K = j) = \lambda_{ij}/\eta$, $j \neq i$, and $\mathbb{P}(K = i) = \lambda_i/\eta$;
3. Let $t \leftarrow t + T$, $J \leftarrow K$ and return to 2

The algorithm makes the event times a homogeneous Poisson process with rate η, and the corresponding values of $\{J_t\}$ a Markov chain with transition matrix $\boldsymbol{I} + \boldsymbol{\Lambda}/\eta$.

The method applies also to the countable case provided $\sup_{i\in E} \lambda_i < \infty$. □

Example 2.4 (Markov–modulated Poisson processes)
Consider a Markov–modulated Poisson process with arrival rate β_i when $J_t = i$ (here $\{J_t\}$ is Markov with transition rates λ_{ij} as in Example 2.3). The intensity of an event (a transition $i \to j$ or a Poisson arrival) is $\lambda_i + \beta_i$ when $J_t = i$. Thus, choosing $\eta \geq \max_{i\in E}(\lambda_i + \beta_i)$ and letting Δ be some point $\notin E$ (marking an arrival), we may generate the arrival epochs σ as follows:

1. Let $t \leftarrow 0$, $J \leftarrow i_0$, $\sigma \leftarrow 0$
2. Let $i \leftarrow J$;
   Generate T as exponential with rate η and K with
   $$\mathbb{P}(K = j) = \begin{cases} \frac{\beta_i}{\eta} & j = \Delta \\ \frac{\lambda_{ij}}{\eta} & j \in E, j \neq i \\ \frac{\eta - \lambda_i - \beta_i}{\eta} & j = i \end{cases}$$
   Let $t \leftarrow t + T$
3. If $K = \Delta$, let $\sigma \leftarrow t$;
   otherwise, let $J \leftarrow K$ and return to 2

□

Generation of bivariate normal r.v.'s

The goal is to generate X_1, X_2 with a joint normal distribution and means μ_1, μ_2, variances σ_1^2, σ_2^2 and covariance $\rho\sigma_1\sigma_2$ ($\rho = \mathbf{Corr}(X_1, X_2)$).

Assume first $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$. Take Y_1, Y_2, Y_3 independent $\mathcal{N}(0,1)$ and

$$X_1 \;=\; \sqrt{1-|\rho|}Y_1 + \sqrt{|\rho|}Y_3, \quad X_2 \;=\; \sqrt{1-|\rho|}Y_2 \pm \sqrt{|\rho|}Y_3$$

where $+$ is for $\rho \geq 0$, $-$ for $\rho \leq 0$. In the general case, let

$$X_1 \;\longleftarrow\; \mu_1 + \sigma_1 X_1, \quad X_2 \;\longleftarrow\; \mu_2 + \sigma_2 X_2\,.$$

We discuss multivariate normals in Chapter **VIII** in connection with Gaussian processes.

How to efficiently generate non–uiform random numbers is an area in itself; see for example Devroye [44]. It should be stressed that for the average user, optimal efficiency is not necessarily a major concern, and that one may want to use a naive but easily programmed method rather than invoking more sophisticated methods or various library packages. The same remark applies to special simulation programming languages like Simula, Simscript, GPSP etc.

3 Discrete events systems and GSMP's*

Glynn [64].

Chapter II

Output analysis

1 The crude Monte Carlo method

Let Z be some random variable and assume that we want to evaluate $z = \mathbb{E}Z$, in a situation where z is not available analytically but Z can be simulated. The crude Monte Carlo (CMC) method then amounts to simulating i.i.d. replicates $Z_1, \ldots, Z_n$, estimating z by the empirical mean

$$\hat{z} \;=\; \frac{1}{n}(Z_1 + \cdots + Z_n)$$

and the variance $\sigma^2 = \sigma_Z^2 = \mathbf{Var}\, Z$ of Z by the empirical variance

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \hat{z})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} Z_i^2 - n\hat{z}^2\right) \tag{1.1}$$

[the occurence of $n-1$ rather than n follows statistical tradition; in most Monte Carlo experiments, n is very large and the difference is minor]. According to standard central limit theory,

$$\sqrt{n}(\hat{z} - z) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(0, \sigma^2\right). \tag{1.2}$$

Hence

$$\hat{z} \pm \frac{1.96\,\hat{\sigma}}{\sqrt{n}} \;=\; \left[\hat{z} - \frac{1.96\,\hat{\sigma}}{\sqrt{n}}, \hat{z} + \frac{1.96\,\hat{\sigma}}{\sqrt{n}}\right] \tag{1.3}$$

is an asymptotic 95% confidence interval, and this is the form in which the result of the simulation experiment is commonly reported.

Remark 1.1 The choice of 95% is common, but other values are, of course, possible. Say 99%, corresponding to $\hat{z} \pm 2.58\,\hat{\sigma}/\sqrt{n}$. Also one–sided confidence intervals may sometimes be relevant. Assume for example that Z

is an indicator function telling whether a certain system failure occurs or not and z the corresponding failure probability. Then $\hat{z} + 1.64\hat{\sigma}/\sqrt{n}$ is an upper 95% confidence limit for z. $\square$

Remark 1.2 In some examples, Z has a simple form and can be generated very quickly. A classical toy example is estimation of $z = \pi$ by simulating U_1, U_2 independent and uniform on $(-1, 1)$ and letting $Z = 4I(U_1^2 + U_2^2 < 1)$. In many situations, the representation of the number z of interest as $\mathbb{E}Z$ may, however, involve considerable sophistication, and the generation of a single Z may be time–consuming.

For an example, consider rare events problems, cf. Chapter **IV**, where A is a rare event (of small probability). A common terminology is then that CMC refers to the case $Z = I(A)$. However, we will consider more sophisticated choices like likelihood ratio estimators. In these notes, we adapt therefore the terminology that 'CMC' means that output analysis is performed as above and that the term does not involve the way in which Z is chosen. $\square$

Remark 1.3 (Run length considerations)

An often asked question is how large n should be to achieve a given precision. Here 'precision' can be understood for example as the half–width $1.96\,\sigma/\sqrt{n}$ of the confidence interval, and the answer is then straightforward: aiming for a precision of ϵ (say $\epsilon = 0.01$), one should take $n = n_\epsilon = 1.96^2\sigma^2/\epsilon^2$. In particular, it is seen that n_ϵ is proportional to σ^2.

In practice, σ^2 is of course unknown. A common procedure is then to start the simulation with a short pilot series with n_p replications, give an estimate σ_p^2 based upon this, and then choose n as $1.96^2\sigma_p^2/\epsilon^2$..

It may appear that the variance is the universal measure of efficiency of a CMC estimator. In particular, given the choice between two CMC schemes based upon r.v.'s Z, Z' with variances $\sigma_Z^2, \sigma_{Z'}^2$, one should choose the one with smallest variance. However, this argument cheats because it does not take into account that the expected CPU times T, T' required to generate one replicate may be very different. Instead, one can formulate the problem in terms of a *simulation budget* t: given we are prepared to spend t units CPU time for the simulation, will Z or Z' give the lower variance? The answer is that by renewal theory, the number of replications obtained within time t will be $n \sim t/T$, resp. $n' \sim t/T'$, and so the variances on the estimates are $\sigma_Z^2 T/t$, resp. $\sigma_{Z'}^2 T'/t$ (that the appropriate CLT holds

also for such random sample sizes follows from Anscombe's theorem, cf. the Appendix). Thus, we should prefer Z if $\sigma_Z^2 T < \sigma_{Z'}^2 T'$ and Z' otherwise.
□

Remark 1.4 (Bias and mean square error)
Let more generally z be an unknown number to be estimated by simulation and consider an asymptotical scheme ($n \to \infty$) where $\hat{z}_n$ is an estimator obeying a CLT of the form (1.2), i.e. $\sqrt{n}(\hat{z}_n - \mathbb{E}\hat{z}_n) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right)$, and having bias $\delta_n = \mathbb{E}\hat{z}_n - z$; such bias occurs widely since it is often hard to produce an estimator $\hat{z}_n$ such that the relation $\mathbb{E}\hat{z}_n = z$ holds exactly (for an example, consider steady–state simulation, cf. Chapter **III**). Assume further that a natural variance estimator $\hat{\sigma}_n^2$ with the property $\hat{\sigma}_n^2 \to \sigma^2$ is available.

Without explicit knowledge of the bias, it is obviously impossible to produce an asymptotic confidence interval for z. The only reasonable candidate is $\hat{z}_n \pm 1.96\hat{\sigma}_n/\sqrt{n}$, and this is in fact asymptotically valid provided $\sqrt{n}\delta_n \to 0$. Thus, one has the general principle: *the standard deviation should dominate the bias* in order that confidence intervals are meaningful.

Quite often, one can show that $n\delta_n \to \delta$ for some δ (most often not available analytically!), which is sufficient for $\sqrt{n}\delta_n \to 0$.

In the presence of non–negliglible bias, it is common to measure the efficiency of a simulation estimator in terms of the mean square error $\mathbb{E}(\hat{z}_n - z)^2$ rather than the variance. Note that the mean square error can be written as the sum of the variance and the squared bias,

$$\mathbb{E}(\hat{z}_n - z)^2 \;=\; \mathbf{Var}(\hat{z}_n) \;+\; (\mathbb{E}\hat{z}_n - z)^2. \tag{1.4}$$

□

Remark 1.5 (Confidence intervals based upon the t distribution*)
□

2 Some applications of the Delta method

We consider some further aspects and extensions of the CMC method in the notation of Section 1.

Estimating a function

In some cases, one is interested in estimating not z, but some function $f(z)$.

Example 2.1 Let $Z_1, Z_2, \ldots$ be i.i.d. failure times of some system. Then $f(z) = 1/z = 1/\mathbb{E}Z$ is the long–run failure intensity. □

Consider thus an estimator $\hat{z}_n$ obeying a CLT of the form (1.2) (not necessarily based upon the CMC method and with a possible non–zero bias δ_n). We can then write

$$f(\hat{z}_n) - f(z) \;=\; f'(z)(\hat{z}_n - z) + \frac{f''(z)}{2}(\hat{z}_n - z)^2 + o((\hat{z}_n - z)^2). \tag{2.1}$$

Multiplying by $\sqrt{n}$, the two last terms on the r.h.s. are negligible from the point of view of distributions, and so we conclude that

$$\sqrt{n}(f(\hat{z}_n) - f(z)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \omega^2\right)$$

where $\omega^2 = f'(z)^2\sigma^2$; this is the essence of the Delta method in its simplest form. We have the obvious estimator $\hat{\omega}^2 = f'(\hat{z})^2\hat{\sigma}^2$ for ω^2, and hence $\hat{z}_n \pm 1.96\,\hat{\omega}/\sqrt{n}$ is an asymptotic 95% confidence interval.

If the bias δ_n is of order δ/n, taking expectations in (2.1) yields the following asymptotics for the bias of $\hat{z}_n$[1]

$$\mathbb{E}[f(\hat{z}_n) - f(z)] \;\sim\; \frac{1}{n}\left(\delta f'(z) + \frac{\sigma^2 f''(z)}{2}\right). \tag{2.2}$$

One implication is that even if $\hat{z}_n$ is unbiased, $f(\hat{z}_n)$ is not: taking a non–linear function typically introduces a bias of order $1/n$.

Multivariate output

Let next $\boldsymbol{Z} = \left(Z^{(1)}, \ldots, Z^{(k)}\right)$ be a random vector with possibly dependent component so that $\sigma_{ij} = \mathbf{Cov}(Z^{(i)}, Z^{(j)})$ may be non–zero for $i \neq j$, and assume that we want to estimate $f = f\left(z^{(1)}, \ldots, z^{(k)}\right)$ where $z^{(i)} = \mathbb{E}Z^{(i)}$ (we will meet the example $\mathbb{E}Z^{(2)}/\mathbb{E}Z^{(1)}$ in connection with regenerative simulation). We then simulate n i.i.d. replications

$$\boldsymbol{Z}_1 = \left(Z_1^{(1)}, \ldots, Z_1^{(k)}\right), \ldots, \boldsymbol{Z}_n = \left(Z_n^{(1)}, \ldots, Z_n^{(k)}\right)$$

[1] Whereas the derivation of the CLT is rigorous, this argument is not; one needs various types of uniform integrability assumptions. The point is just to give an idea of what type of result can typically be expected.

of $\boldsymbol{Z}$ and let $\hat{f} = f(\hat{z}^{(1)}, \ldots, \hat{z}^{(k)})$ where

$$\hat{z}^{(i)} = \frac{Z_1^{(i)} + \cdots + Z_n^{(i)}}{n}, \quad \hat{\sigma}_{ij} = \frac{1}{n-1} \sum_{m=1}^{n} (Z_m^{(i)} - \hat{z}^{(i)})(Z_m^{(j)} - \hat{z}^{(j)}) .$$

Then $\sqrt{n}(\hat{f} - f) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \omega^2\right)$ where

$$\omega^2 = \sum_{i,j=1}^{k} \frac{\partial f}{\partial z^{(i)}} \left(z^{(1)}, \ldots, z^{(k)}\right) \frac{\partial f}{\partial z^{(j)}} \left(z^{(1)}, \ldots, z^{(k)}\right) \sigma_{ij}$$

and our confidence interval is $\hat{f} \pm 1.96\, \hat{\omega}/\sqrt{n}$ where

$$\hat{\omega}^2 = \sum_{i,j=1}^{k} \frac{\partial f}{\partial z^{(i)}} \left(\hat{z}^{(1)}, \ldots, \hat{z}^{(k)}\right) \frac{\partial f}{\partial z^{(j)}} \left(\hat{z}^{(1)}, \ldots, \hat{z}^{(k)}\right) \hat{\sigma}_{ij} .$$

The variance of the variance estimator $\hat{\sigma}^2$

Given the choice between two CMC schemes based upon r.v.'s Z, Z' with variances $\sigma_Z^2, \sigma_{Z'}^2$, the first concern would be to obtain minimal variance for a given CPU time (simulation budget) t as discussed in Section 1.

If the variances $\sigma_Z^2 T$, $\sigma_{Z'}^2 T'$ per unit CPU time are roughly the same (in particular, if σ_Z^2 and $\sigma_{Z'}^2$ are roughly the same and T, T' are roughly the same), the next concern might be to choose the method with the most reliable variance estimate. We go next through the computations relevant for this comparison. Write $m_k = \mathbb{E} Z^k$ (then $z = m_1$, $\sigma^2 = m_2 - m_1^2$).

Proposition 2.2 $\sqrt{n} \begin{pmatrix} \hat{z} - z \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \xrightarrow{\mathcal{D}}$

$$\mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 = m_2 - m_1^2 & 2m_1^3 + m_3 - 3m_1 m_2 \\ 2m_1^3 + m_3 - 3m_1 m_2 & -4m_1^4 + 8m_1^2 m_2 + m_4 - m_2^2 - 4m_1 m_3 \end{pmatrix} \right)$$

Proof Let $\overline{Z} = \sum_1^n Z_i/n = \hat{z}$, $\overline{Z^2} = \sum_1^n Z_i^2/n.$. Obviously,

$$\sqrt{n} \begin{pmatrix} \overline{Z} - m_1 \\ \overline{Z^2} - m_2 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} \right)$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{Var}(Z) & \mathbf{Cov}(Z, Z^2) \\ \mathbf{Cov}(Z, Z^2) & \mathbf{Var}(Z^2) \end{pmatrix} = \begin{pmatrix} \sigma^2 = m_2 - m_1^2 & m_3 - m_1 m_2 \\ m_3 - m_1 m_2 & m_4 - m_2^2 \end{pmatrix}$$

Letting $\boldsymbol{f}(x,y) = (x, y-x^2)$, $(\hat{z}, \hat{\sigma}^2)$ has the same asymptotics as $f(\overline{Z}, \overline{Z^2})$ so the result follows by the Delta method, with the asymptotic covariance matrix

$$\boldsymbol{Df}(m_1,m_2)\boldsymbol{\Sigma}\ \boldsymbol{Df}(m_1,m_2)' \ = \ \begin{pmatrix} 1 & 0 \\ -2m_1 & 1 \end{pmatrix} \boldsymbol{\Sigma} \begin{pmatrix} 1 & -2m_1 \\ 0 & 1 \end{pmatrix}$$

which is the same as asserted. □

3 Simulations driven by empirical distributions

Most often, the r.v. Z in the CMC method is produced from other r.v.'s belonging to parametric distributions. Say in a queueing simulation one uses a Gamma service time distribution. In this example, the exact form is, however, seldom known but some statistical observations are available, and it is then appealing not to fit a parametric distribution but to simulate directly from the empirical distribution of the observed service times.

To illustrate how to perform output analysis in such situations, we consider a simple case where $u_1, \ldots, u_m$ are i.i.d. observations from an unknown distribution F. The aim is to estimate $\psi(F) = \mathbb{E}_F g(U_1, U_2, \ldots)$ where $U_1, U_2, \ldots$ are i.i.d. with distribution F, by simulation based upon drawings from the empirical distribution F_m of $u_1, \ldots, u_m$ (F_m is the distribution putting mass $1/m$ at each u_k). Say g is the length of the busy period in a D/G/1 queue with service time distribution F and constant interarrival times, or $\max_{1\le r\le R} W_r$, where R is fixed and W_j the waiting time of customer j.

The naive procedure is to use CMC with $Z_i = g(u_{K(i,1)}, u_{K(i,2)}, \ldots)$ where the $K(i,j)$ are i.i.d and uniform on $\{1, \ldots, m\}$. The problem is that $z = \mathbb{E}Z = \psi(F_m)$ so that the confidence interval produced in this way is a confidence interval for $\psi(F_m)$, not $\psi(F)$ as desired: the stochastic variation in $\psi(F_m)$ is ignored.

To resolve this problem, note that typically $\psi(F_m)$ has an asymptotic normal distribution with mean $\psi(F)$ and variance of the form ω^2/m for some ω^2 as $m \to \infty$. More precisely, the program for rigorously verifying this is to find a Banach space $(V, \|\cdot\|)$ of (signed) measures such that one is willing to assume that $F \in V$, that $F_n \in V$ and that ψ is defined on all of V and Frechet– or Hadamard differentiable at F. See, e.g., Gill [56] for a good survey.

It follows that in the CMC setting, we can write

$$\hat{z} \;=\; \psi(F_m) + \frac{\sigma_{F_m}}{\sqrt{n}} X_1 \;=\; \psi(F) + \frac{\sigma_{F_m}}{\sqrt{n}} X_1 + \frac{\omega}{\sqrt{m}} X_2$$

where X_1, X_2 are independent and asymptotically standard normal when both n and m are large. To produce a confidence interval, wee need therefore an additional estimate of ω^2. In some cases, ω^2 has been evaluated analytically (see, e.g., Grübel [73]), but typically, an estimate needs to be produced by simulation and to this end a variant of the CMC method is required.

The idea is to divide the m observations into groups, say k groups of size ℓ (assuming for convenience that m can be written as $m = k\ell$) and to perform p (say) simulations within each group, using the empirical distribution $F_{\ell,i}$ in group i. For each group, we thus in a straightforward way obtain an estimate $\hat{\psi}(F_{\ell,i})$ of $\psi(F_{\ell,i})$ and an associated estimate $\hat{\sigma}^2_{F_{\ell,i}}$ of $\sigma^2_{F_{\ell,i}}$. The estimator of $\psi(F)$ is

$$\hat{\psi} \;=\; \frac{1}{k}\left(\hat{\psi}(F_{\ell,1}) + \cdots + \hat{\psi}(F_{\ell,k})\right) .$$

We can write

$$\hat{\psi}(F_{\ell,i}) \;=\; \psi(F_{\ell,i}) + \frac{\sigma_{F_{\ell,i}}}{\sqrt{p}} X_{1,i} \;=\; \psi(F) + \frac{\sigma_{F_{\ell,i}}}{\sqrt{p}} X_{1,i} + \frac{\omega}{\sqrt{\ell}} X_{2,i}$$

where the X_{ij} are independent and asymptotically standard normal when both p and ℓ are large. When ℓ is large, we can replace $\sigma^2_{F_{\ell,i}}$ by $\sigma^2 = \sigma^2_F$, and so the asymptotic variance of $\hat{\psi}$ becomes

$$\frac{\sigma^2}{kp} + \frac{\omega^2}{k\ell} \tag{3.1}$$

The natural estimates of σ^2, ω^2 are

$$\hat{\sigma}^2 = \frac{1}{k}\left(\hat{\sigma}^2_{F_{\ell,1}} + \cdots + \hat{\sigma}^2_{F_{\ell,k}}\right), \quad \hat{\omega}^2 = \frac{\ell}{k-1}\sum_{i=1}^{k}\left(\hat{\psi}(F_{\ell,i}) - \hat{\psi}\right)^2$$

and so the confidence interval is

$$\hat{\psi} \pm 1.96\sqrt{\frac{\hat{\sigma}^2}{kp} + \frac{\hat{\omega}^2}{k\ell}} \tag{3.2}$$

An obvious question is how k and p should be chosen based upon a budget of t drawings from an empirical distribution. Clearly, $t = kp$ so since both t and $m = k\ell$ are fixed, (3.1) shows that in terms of minimizing the variance, the answer is that the choice is unimportant. However, consider next the variance of the variance estimator. Since $\mathbf{Var}(\hat{\sigma}^2) \sim c_1/kp = c_1/t$, $\mathbf{Var}(\hat{\omega}^2) \sim c_2\ell/k$,

$$\mathbf{Var}\left(\frac{\hat{\sigma}^2}{kp} + \frac{\hat{\omega}^2}{k\ell}\right) \sim \frac{c_1}{t^3} + \frac{c_2\ell^2}{m^3}\ .$$

This indicates that choosing ℓ small or, equivalently, the number of groups k large, is preferable. But note that the largest possible choice $k = m$ is not feasible because the asymptotics used in the arguments requires that also $\ell = m/k$ is sufficiently large for the CLT for $\psi(F_\ell)$ to be in force.

4 Variance reduction methods

setcounterequation0

The aim of variance reduction is to produce an alternative estimator $\hat{z}_{\mathrm{VR}}$ of a number z having hopefully much smaller variance than the CMC estimator $\hat{z}_{\mathrm{CMC}}$. The study of such methods is a classical area in simulation and the literature is considerable.

It should be noted that variance reduction is typically most readily available in well structured problems. Also, variance reduction typically involves a fair amount of both theoretical study of the problem in question and added programming effort. For this reason, variance reduction is most often only worthwhile if it is substantial. Assume for example that a sophisticated method reduces the variance with 25%, i.e. $\sigma^2_{\mathrm{VR}} = 0.75\sigma^2_{\mathrm{CMC}}$ and consider the numbers n_{CMC}, n_{VR} to obtain a given precision (say in terms of halfwidth of the confidence interval). Then

$$\epsilon = \frac{1.96\sigma_{\mathrm{CMC}}}{\sqrt{n_{\mathrm{CMC}}}} = \frac{1.96\sigma_{\mathrm{VR}}}{\sqrt{n_{\mathrm{VR}}}},\quad n_{\mathrm{VR}} = \frac{\sigma^2_{\mathrm{VR}}}{\sigma^2_{\mathrm{CMC}}} n_{\mathrm{CMC}} = 0.75\, n_{\mathrm{CMC}}$$

so that at best (assuming that the expected CPU times $T_{\mathrm{CMC}}, T_{\mathrm{VR}}$ for one replication are about equal) one can only reduce the computer time by 25% which is most cases is of no relevance compared to the additional effort to develop and implement the variance reduction method. If $T_{\mathrm{VR}} > T_{\mathrm{CMC}}/0.75$, as may easily be the case, there is no gain at all.

We concentrate here upon two methods, importance sampling and control variates, for which there is a large number of examples where the variance reduction has turned out to be considerable, but briefly mention also some further classical methods. The examples we give are at a toy level, but we will meet more substantial examples later in the text.

Importance sampling

The idea is to compute $z = \mathbb{E}Z$ by simulating from a probability measure $\tilde{\mathbb{P}}$ different from the given probability measure $\mathbb{P}$ and having the property that there exists a r.v. L such that

$$z = \mathbb{E}Z = \tilde{\mathbb{E}}[LZ]. \tag{4.3}$$

Thus, using the CMC method one generates $(Z_1, L_1), \ldots, (Z_n, L_n)$ from $\tilde{\mathbb{P}}$ and uses the estimator

$$\hat{z}_{IS} = \frac{1}{n}\sum_{i=1}^{n} L_i Z_i$$

and the confidence interval

$$\hat{z}_{IS} \pm \frac{1.96\, s_{IS}}{\sqrt{n}} \quad \text{where} \quad s_{IS}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(L_i Z_i - \hat{z}_{IS})^2 .$$

In order to achieve (4.3), the obvious possibility is to take $\mathbb{P}$ and $\tilde{\mathbb{P}}$ mutually equivalent in the Radon–Nikodym sense and $L = d\mathbb{P}/d\tilde{\mathbb{P}}$ as the likelihood ratio.

Variance reduction may or may not be obtained: it depends on the choice of the alternative measure $\tilde{\mathbb{P}}$, and the problem is to make an efficient choice.

To this end, a crucial observation is that there is an optimal choice of $\tilde{\mathbb{P}}$: define $\tilde{\mathbb{P}}$ by $d\tilde{\mathbb{P}}/d\mathbb{P} = Z/\mathbb{E}Z = Z/z$, i.e. $L = z/Z$ (the event $\{Z = 0\}$ is not a concern because $\tilde{\mathbb{P}}(Z = 0) = 0$). Then

$$\begin{aligned}
\mathbf{Var}(LZ) &= \tilde{\mathbb{E}}(LZ)^2 - \Big[\tilde{\mathbb{E}}(LZ)\Big]^2 = \mathbb{E}\left(\frac{z^2}{Z^2}Z^2\right) - \Big[\mathbb{E}\left(\frac{z}{Z}Z\right)\Big]^2 \\
&= z^2 - z^2 = 0.
\end{aligned}$$

Thus, it appears that we have produced an estimator with variance zero. However, the argument cheats because we are simulating since z is not avaliable analytically. Thus we cannot compute $L = Z/z$.

Nevertheless, even if the optimal change of measure is not practical, it gives a guidance: choose $\tilde{\mathbb{P}}$ such that $d\tilde{\mathbb{P}}/d\mathbb{P}$ is as proportional to Z as possible. This may also be difficult to assess, but tentatively, one would try to choose $\tilde{\mathbb{P}}$ to make large values of Z more likely; from this, the term importance sampling.

Example 4.1 (Monte Carlo integration) Assume that $z = \int_0^1 g(x)\,dx$ where the integral is not available analytically. We can then use the CMC method by taking $Z = g(U)$ with U uniform$(0,1)$.

For importance sampling, the idea of choosing $d\tilde{\mathbb{P}}/d\mathbb{P}$ close to Z/z leads to taking $d\tilde{\mathbb{P}}/d\mathbb{P} = \tilde{g}(U)/\tilde{z}$ where $\tilde{g}$ is close to g and $\tilde{z} = \mathbb{E}\tilde{g}(U)$ is analytically available. This means that

$$Z_{\mathrm{IS}} \;=\; g(\tilde{U})\frac{\tilde{z}}{\tilde{g}(\tilde{U})}$$

where $\tilde{U}$ is simulated from the density $\tilde{g}/\tilde{z}$.

In simple examples like this, Monte Carlo integration is inferior to numerical integration but the method plays a role above all in problems of high dimensionality. □

Example 4.2 (Likelihood ratio calculations) If $\tilde{\mathbb{P}}$ is obtained by changing the density of a single r.v. U from $f(x)$ to $\tilde{f}(x)$, we have $L = f(U)/\tilde{f}(U)$. If more generally Z has the form $g(U_0, U_1, \ldots, U_\tau)$ where τ is a constant or a stopping time and $U_0, U_1, \ldots$ are i.i.d. with density f, then

$$L \;=\; \prod_{n=0}^{\tau} \frac{f(U_n)}{\tilde{f}(U_n)}.$$

If instead the U_n form a countable Markov chain with transition probabilities p_{ij} and $\tilde{\mathbb{P}}$ corresponds to by changing the p_{ij} to a different set $\tilde{p}_{ij}$ of transition probabilities, then

$$L \;=\; \prod_{n=0}^{\tau-1} \frac{p_{U_n U_{n+1}}}{\tilde{p}_{U_n U_{n+1}}}.$$

□

Control variates

The idea is to look for a r.v. W which has a strong correlation (positive or negative) with Z and a *known* mean w, generate $(Z_1, W_1), \ldots, (Z_n, W_n)$

rather than $Z_1, \ldots, Z_n$ and combine the empirical means $\hat{z}, \hat{w}$ to an estimator with lower variance than $\hat{z}$.

The naive method is to choose some arbitrary constant α and consider the estimator $\hat{z} + \alpha(\hat{w} - w)$. The point is that since w is known, we are in position to just add a term with mean zero so that the mean of the new estimator still is z. The variance is

$$\sigma_Z^2 + \alpha^2 \sigma_W^2 + 2\alpha\sigma_{ZW}^2, \tag{4.4}$$

where

$$\sigma_Z^2 = \mathbf{Var}\, Z, \quad \sigma_W^2 = \mathbf{Var}\, W, \quad \sigma_{ZW}^2 = \mathbf{Cov}(Z, W).$$

In general, nothing can be said about how (4.4) compares to the variance σ_Z^2 of the CMC estimator $\hat{z}$ (though sometimes a naive choice like $\alpha = 1$ works to produce a lower variance). However, it is easily seen that (4.4) is minimized for $\alpha = -\sigma_{ZW}^2/\sigma_W^2$, and that the minimum value is

$$\sigma_Z^2(1 - \rho^2) \quad \text{where} \quad \rho = \mathbf{Corr}(Z, W) = \frac{\sigma_{ZW}^2}{\sqrt{\sigma_Z^2 \sigma_W^2}} \tag{4.5}$$

One then simply estimates the optimal α via the empirical covariance matrix,

$$\hat{\alpha} = -\frac{s_{ZW}^2}{s_W^2},$$

where

$$s_Z^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^n (Z_i - \hat{z})^2, \quad s_W^2 = \frac{1}{n-1}\sum_{i=1}^n (W_i - \hat{w})^2,$$

$$s_{ZW}^2 = \frac{1}{n-1}\sum_{i=1}^n (Z_i - \hat{z})(W_i - \hat{w}),$$

and uses the estimator $\hat{z}_{\mathrm{CV}} = \hat{z} + \hat{\alpha}(\hat{w} - w)$ which has the same asymptotic properties as $\hat{z} + \alpha(\hat{w} - w)$; in the particular, the asymptotic variance is $\sigma_Z^2(1 - \rho^2)/n$, and a confidence interval is constructed by replacing σ_Z^2, ρ^2 by their empirical values s_Z^2, $s_{ZW}^4/s_Z^2 s_W^2$.

The procedure reduces the variance by a factor $1 - \rho^2$. Thus, one needs to look for a control variate W with $|\rho|$ as close to 1 as possible. The exact value of ρ will be difficult to asses apriori, so that in practice one would just try to make W and Z as dependent as possible (in some vague sense).

It is, however, an appealing feature that even if one is not very succesful, the resulting variance is never increased.

There is also an interesting relation to standard regression analysis. In fact, the calculation of $\hat{z}_{\mathrm{CV}}$ amounts to use a regression of Z upon W, fit a regression line by least squares and calculate the level of the line at the known value w of $\mathbb{E}W$; see Fig. II.4.1. This is seen as follows: the assumption underlying the regression (viewing the W_i as constants and not r.v.'s) is

$$\mathbb{E}Z_i \;=\; m' + \beta W_i \;=\; m + \beta(W_i - \hat{w}) \tag{4.6}$$

$(m = m' + \beta\hat{w})$, with least squares estimates

$$\hat{m} = \hat{z}, \quad \hat{\beta} \;=\; \frac{\sum_1^n (Z_i - \hat{z})(W_i - \hat{w})}{\sum_1^n (W_i - \hat{w})^2} \;=\; -\hat{\alpha}\,,$$

so that the level of the fitted regression line at w is

$$\hat{m} + \hat{\beta}(w - \hat{w}) \;=\; \hat{z}_{\mathrm{CV}}\,.$$

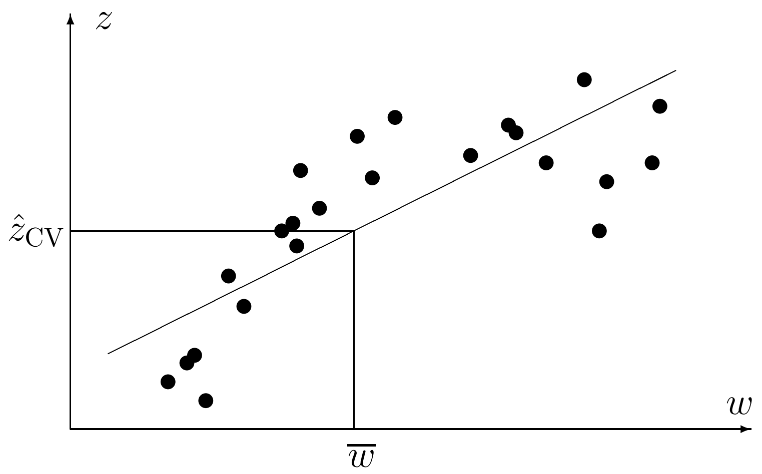

Figure II.4.1

For this reason, often the term *regression–adjusted control variates* is used. The similarity is, however, formal: regression analysis via least squares is based upon the assumption of linear dependence (and preferably normal errors) whereas nothing like this is needed for regression–adjusted control variates (one may, however view the method as inference in the limiting bivariate normal distribution of $(\hat{z}, \hat{w})$). The literature pays quite a lot of attention to control variates without regression–adjustment (i.e., α is assigned some arbitrary value), but to the present author's mind, it is difficult to imagine situations where one would prefer this to regression–adjustment.

Example 4.3 Consider again the Monte Carlo integration problem in Example 4.1. A suitable control for $Z = g(U)$ is then $W = f(U)$ with where f is close to g (to get $|\rho|$ close to 1) and $w = \mathbb{E}f(U) = \int_0^1 f(x)\,dx$ is analytically available. □

Antithetic sampling

Here one generates $Z_1, \ldots, Z_n$ not as i.i.d. but as pairwise dependent and as negatively correlated as possible. That is, one takes $n = 2m$ and generates m i.i.d. random pairs $(Z_1, Z_2), (Z_3, Z_4), \ldots, (Z_{n-1}, Z_n)$ such that the marginal distribution of Z_i is the same (as for the CMC method) for all i (even and uneven) but Z_{2j-1} and Z_{2j} may be dependent. The estimator is $\hat{z}_{\mathrm{Anth}} = (Z_1 + \cdots + Z_n)/n$ with variance

$$\begin{aligned}\frac{1}{n}\sigma^2_{\mathrm{Anth}} &= \frac{1}{m}\mathbf{Var}\left(\frac{Z_1+Z_2}{2}\right) = \frac{1}{4m}(\sigma^2_{\mathrm{CMC}} + \sigma^2_{\mathrm{CMC}} + 2\sigma^2_{\mathrm{CMC}}\rho) \\ &= \frac{1}{n}\sigma^2_{\mathrm{CMC}}(1+\rho)\end{aligned}$$

where $\rho = \mathbf{Corr}(Z_1, Z_2)$. Thus, ρ should be negative for obtaining variance reduction, and preferably as close to -1 as possible for the method to be efficient.

For example, in Monte Carlo integration (Example 4.1) one could take $Z_1 = g(U)$, $Z_2 = g(-U)$. If g is monotone, Chebycheff's covariance inequality[2] then yields $\rho \le 0$.

We know of no example where the variance reduction obtained by antithetic sampling is dramatic.

Conditional Monte Carlo

Here Z_{CMC} is replaced by $Z_{\mathrm{Cond}} = \mathbb{E}[Z_{\mathrm{CMC}}|W]$ for some r.v. W (more generally, one could consider $\mathbb{E}[Z_{\mathrm{CMC}}|\mathcal{G}]$ for some σ–field $\mathcal{G}$). Clearly, $\mathbb{E}Z_{\mathrm{Cond}} = \mathbb{E}Z_{\mathrm{CMC}} = z$. Since

$$\begin{aligned}\sigma^2_{CMC} &= \mathbf{Var}(Z_{\mathrm{CMC}}) = \mathbf{Var}(\mathbb{E}[Z_{\mathrm{CMC}}|W]) + \mathbb{E}(\mathbf{Var}[Z_{\mathrm{CMC}}|W]) \\ &= \sigma^2_{\mathrm{Cond}} + \mathbb{E}(\mathbf{Var}[Z_{\mathrm{CMC}}|W]) \ \ge\ \sigma^2_{\mathrm{Cond}},\end{aligned}$$

[2]stating that $\mathbf{Corr}(f_1(X), f_2(X)) \ge 0$ if X is a r.v. and f_1, f_2 non–decreasing functions

conditional Monte Carlo always provides variance reduction which is appealing. The difficulty is to find W such that the conditional expectation is computable.

Example 4.4 Consider the estimation of $z = \pi$ as in Remark 1.2 via $Z_{\mathrm{CMC}} = 4I(U_1(2+U_2^2 < 1)$ with U_1, U_2 independent and uniform on $(-1, 1)$. We can then take

$$Z_{\mathrm{Cond}} \;=\; \mathbb{E}[Z_{\mathrm{CMC}}|U_1] \;=\; 4\mathbb{P}(U_2^2 < 1 - U_1^2 \,|\, U_2) \;=\; 4\sqrt{1 - U_1^2}.$$

□

Example 4.5 Let $z = \mathbb{P}(X_1+X_2 > x)$ where X_1, X_2 are independent with distribution F (F is known, one can simulate from F but the convolution F^{*2} is not available). Then $Z_{\mathrm{CMC}} = I(X_1 + X_2 > x)$ and taking $W = X_1$, we get $Z_{\mathrm{Cond}} = \overline{F}(x - X_1)$.

For a related algorithm, see Section IV.2c. □

Common random numbers*

Stratification*

Isolating known components

In many cases, some parts of the expectation z of Z can be evaluated analytically. One may then attempt to organize the output analysis so that these known parts need not be simulated.

Example 4.6 Let $T_1, T_2, \ldots$ be i.i.d. and non–negative, and let $Z = \sup\{n : S_n \le t\}$ be the number of renewals up to time t where $S_n = T_1 + \cdots + T_n$ ($z = \mathbb{E}Z$ is then the renewal function at t). Letting $\tau = \inf\{n : S_n > t\}$, we then have $Z = \tau - 1$. By Wald's identity,

$$\mathbb{E}S_\tau \;=\; \mu\mathbb{E}\tau \;=\; \mu(z+1).$$

But we can write $S_\tau = t + \xi$ where $\xi = S_\tau - t$ is the overshoot. This yields

$$z \;=\; \frac{t + \mathbb{E}\xi}{\mu} - 1$$

and an alternative estimator is

$$\tilde{Z} \;=\; \frac{t + \xi}{\mu} - 1.$$

For example, if the T_i are standard exponential and $t = 50$, then Z is Poisson(50) so that $\mathbf{Var} Z = 50$. In contrast, since ξ is again standard exponential, $\mathbf{Var} \tilde{Z} = 1$. $\square$

For a further example, see the Minh–Sorli algorithm in **IX**.1.

Chapter III

Steady–state simulation

Let $\{X_t\}$ be a stochastic process in discrete or continuous time, and assume that X_t converges in distribution as $t \to \infty$, say with limit distribution π. The problem we study is to obtain information on π from a simulated (in general non–stationary) version of $\{X_t\}$.

The ideal is of course to generate a r.v. Z with distribution π. However, there are no obvious general methods for doing this. In Section 1, we study to which extent this is possible for Markov chains with a discrete state space E. The answer is that algorithms exist for E finite ($|E| < \infty$) but not in general for E countably infinite.

The case $|E| < \infty$ is, however, quite special and even there the algorithms are often prohibitively inefficient in terms of computer time. Thus, the prominent methods in the area of steady–state simulation are based on alternative representations of functionals z like the mean $\int x\pi(dx)$. In particular, we look at estimators based upon sample averages in various variants (Section 2) and regenerative simulation (Section 3).

1 Exact simulation

Let $\{X_n\}$ be a Markov chain with state space E (finite or countable), transition probabilities p_{ij} and stationary distribution π (assuming ergodicity, i.e. irreducibility, positive recurrence and aperiodicity):

$$\pi_j \;=\; \sum_{i \in E} \pi_i p_{ij} \,.$$

We are interested in the problem of whether it is possible to generate a r.v. Z with distribution π. If so, we speak of *exact simulation*, sometimes also called *perfect simulation*.

E **finite**

It is not apriori obvious whether exact simulation is possible. However, the answer is positive in the case of a finite Markov chain. The first algorithm for generating a r.v. Z with exactly the stationary distribution seems to be that of Asmussen, Glynn & Thorisson [14] but was prohibitively inefficient in terms of computer time. We describe here some algorithms developed by Propp & Wilson [121]. Write $E = \{1, \ldots, p\}$.

It will be convenient to represent the Markov chain simulation in terms of an *updating rule*. By this we understand a random vector

$$\boldsymbol{Y} \;=\; (Y(1), \ldots, Y(p))$$

such that $Y(i)$ has distribution $p_{i\cdot}$, $\mathbb{P}(Y(i) = j) = p_{ij}$ (note that the p components of $\boldsymbol{Y}$ are not necessarily independent; we return to this point later). From $\boldsymbol{Y}$, we construct a doubly infinite sequence $(\boldsymbol{Y}_n)_{n=0,\pm1,\pm2,\ldots}$ of i.i.d. random vectors distributed as $\boldsymbol{Y}$. We can then construct $\{X_n\}_{n=0,1,\ldots}$ recursively by $X_{n+1} = Y_n(X_n)$ and $X_0 = i$ where i is the initial state. More generally, we can for each $N \in \mathbf{Z}$ and each $i \in E$ define a version $\left\{X_n^N(i)\right\}_{n=N,N+1,\ldots}$ of $\{X_n\}$ starting at i at time N by

$$\begin{aligned} X_N^N(i) &= i \\ X_{N+1}^N(i) &= Y_N(i) = Y_N\left(X_N^N(i)\right), \ \ldots, X_{n+1}^N(i) = Y_n\left(X_n^N(i)\right) . \end{aligned}$$

Note the important point that if $N, N' \le n$, then the updatings of $X^N(i)$ and $X^{N'}(i')$ from n to $n+1$ use the same $\boldsymbol{Y}_n$.

The *forwards coupling time* is defined as

$$\tau_+ \;=\; \inf\left\{n = 1, 2, \ldots : X_n^0(1) = \ldots = X_n^0(p)\right\} .$$

i.e., as the first time where the Markov chains

$$\left\{X_n^0(1)\right\}_{n=0,1,\ldots}, \ \ldots \ , \ \left\{X_n^0(p)\right\}_{n=0,1,\ldots}$$

started at time 0 in the p different states coalesce. See Fig. III.1.1.

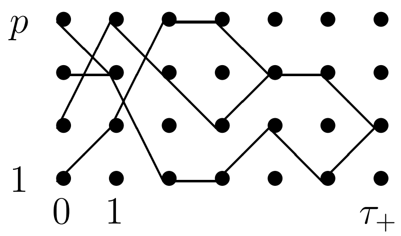

Figure III.1.1: The forwards coupling

Whether this forward coupling time is a.s. finite or not depends on the updating rule, i.e. the specific dependence between the p components of $\boldsymbol{Y}_n$. Call the updating rule *independent* if these p components are independent.

Proposition 1.1 *In the case of independent updating,* $\mathbb{P}(\tau_+ < \infty) = 1$.

Proof Let $p_{ij}^{(n)} = \mathbb{P}(X_n = j \mid X_0 = i)$ be the n–step transition probabilities. Since $p_{ij}^{(n)} \to \pi_j > 0$, $n \to \infty$, we can choose first some arbitrary state j and next N such that $p_{ij}^{(N)} > \epsilon > 0$ for all $i \in E$. Since the probability of coalescence before N is at least the probability that p independent Markov chains starting at time 0 in the p different states will all be in state j at time N, we get $\mathbb{P}(\tau_+ \leq N) \geq \epsilon^k$. Similarly, $\mathbb{P}(\tau_+ \leq 2N \,|\tau_+ > N) \geq \epsilon^k$ so ('geometric trial argument')

$$\mathbb{P}(\tau_+ > N) \leq 1 - \epsilon^k,\ \mathbb{P}(\tau_+ > 2N) \leq (1 - \epsilon^k)^2,\ \ldots$$

which implies that $\tau_+ < \infty$ a.s. □

Rather than forwards coupling, the Propp–Wilson algorithm uses *coupling from the past.* The *backwards coupling time* is defined as

$$\tau \;=\; \inf\left\{n = 1, 2, \ldots :\ X_0^{-n}(1) = \ldots = X_0^{-n}(p)\right\}\ ,$$

i.e., as the first time where the Markov chains $X^{-n}(1), X^{-n}(p)$ started at time $-n$ in the p different states coalesce. Equivalently, coalescence means that the value set

$$\left\{X_0^{-n}(1), \ldots, X_0^{-n}(p)\right\}$$

contains only one point (note that the cardinality of this set is a non–increasing function of n). See Fig. III.1.2.

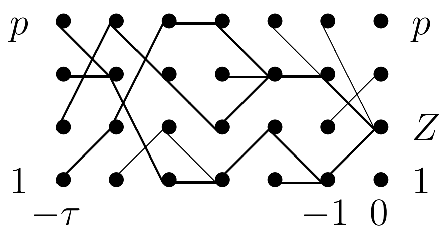

Figure III.1.2: The backwards coupling

Theorem 1.2 *Assume that the updating rule is such that* $\tau_+ < \infty$ *a.s. Then* $\tau < \infty$ *a.s. as well,* $Z = X_0^{-\tau}(i)$ *does not depend on* i *and* Z *has distribution* π.

Proof The first statement follows since $\mathbb{P}(\tau \le k) = \mathbb{P}(\tau_+ \le k)$ goes to 1 as $k \to \infty$. That Z does not depend on i is immediate from the definition of τ.

Now consider $X_0^{-n}(i)$ for some fixed i. On $\tau \le n$, we have $X_0^{-n}(i) = Z$ and hence $\mathbb{P}(X_0^{-n}(i) = j) \to \mathbb{P}(Z = j)$ for all j. On the other hand, $\mathbb{P}(X_0^{-n}(i) = j) = p_{ij}^n \to \pi_j$. Hence $\mathbb{P}(Z = j) = \pi_j$ as desired. □

Remark 1.3 *The forwards coupling time only enters as a tool to show that the backwards coupling time is finite. It is definitely not correct that $X_{\tau_+}^0(i)$ has the stationary distribution!* □

Corollary 1.4 (GENERAL MARKOV CHAIN, INDEPENDENT UPDATING) *In the case of independent updating, $\tau < \infty$ a.s., $Z = X_0^{-\tau}(i)$ does not depend on i and Z has distribution π.*

Now assume that there is defined some partial order $\preceq$ on $\{1, \dots, p\}$, such that 1 is minimal element and p maximal, $1 \preceq i \preceq p$ for all $i = 1, \dots, p$.

Recall that $X = \{X_n\}$ is called *stochastically monotone* if $i \preceq j$ implies that $X_1^0(i) \preceq X_1^0(j)$ in stochastic order. Im terms of transition probabilities: for any ℓ,

$$\sum_{k:\,\ell \preceq k} p_{ik} \;\le\; \sum_{k:\,\ell \preceq k} p_{jk} \quad \text{if } i \preceq j$$

Example 1.5 An example of a monotonous Markov chain is a random walk reflected at the barriers 0 and p,

$$X_{n+1} \;=\; \min(p, \max(0, X_n + U_n))$$

where $U_1, U_2, \dots$ are i.i.d. on $\mathbf{Z}$. Such chains show up in many finite buffer queuing problems. A particular case is *Moran's model for the dam*, where X_n is the content of a water reservoir at time n and $U_n = V_n - m$ where V_n is the amount of water flowing into the reservoir at time n and m the maximal amount of water which can be released. □

Example 1.6 In many applications in mathematical physics and image analysis, the state space E is set of all $0, 1$ configurations on a finite lattice, say $\{0, \dots, N\}^2$. See Fig. 1.3 where the filled circles correspond to spin 1 at a site and the unfilled ones to spin 0.

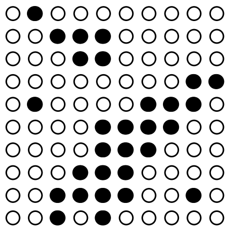

Figure III.1.3

Thus, the number of states is 2^{N^2}. The order is defined componentwise so that if we identify the configuration of all 1's with state 1 and the configuration of all 0's with state $p = 2^{N^2}$, we have $1 \preceq i \preceq p$ for all configurations i. □

Under such monotonicity assumptions, a variant of the Propp–Wilson algorithm is often more efficient. It is defined by *monotone updating*, requiring

$$Y(i) \preceq Y(j) \ \text{ if } i \preceq j.$$

This implies $X_n^N(i) \preceq X_n^N(j)$ for all N and all $n \geq N$, in particular

$$X_n^N(1) \preceq X_n^N(i) \preceq X_n^N(p) \tag{1.1}$$

for all i and all $n \geq N$. For example, in Example 1.5 the natural monotone updating rule is $Y(i) = \min(p, \max(0, i+U))$ with the same U for all i (in the case of independent updating, one would need to take the $U's$ to be independent for different i). We define

$$\tau^m = \inf\left\{n = 1, 2, \ldots : X_0^{-n}(1) = X_0^{-n}(p)\right\}$$

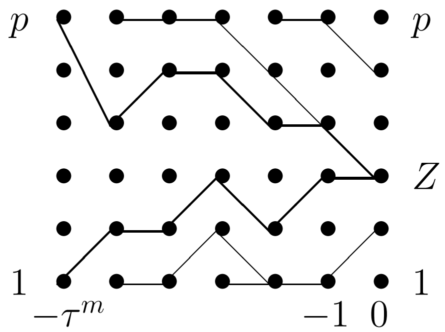

Figure III.1.3

Corollary 1.7 (STOCHASTICALLY MONOTONE MARKOV CHAIN, MONOTONE UPDATING) *In the case of monotone updating, $\tau^m = \tau < \infty$ a.s., $Z = X_0^{-\tau^m}(i)$ does not depend on i and Z has distribution π.*

Proof That $\tau_+ < \infty$ a.s. follows since $\tau(p,1) = \inf\{n : X_n^0(p) = 1\}$ is finite by recurrence and $X_{\tau(p,1)}^0(i) = 1$ for all i by (1.1).

Clearly, by definition $\tau^m \le \tau$. On the other hand, $X_0^{-\tau^m}(i) = X_0^{-\tau^m}(1) = X_0^{-\tau^m}(p)$ for all i by (1.1). □

Remark 1.8 In the monotone case, Propp & Wilson suggest to take alternatively

$$\tilde{\tau}^m \;=\; \inf\left\{n = 1, 2, 4, 8, \ldots : X_0^{-n}(1) = X_0^{-n}(p)\right\},$$

$\tilde{Z} = X_0^{-\tilde{\tau}^m}(1) = X_0^{-\tilde{\tau}^m}(1)$. That $\tilde{\tau}^m < \infty$ follows since $\tilde{\tau}^m \le 2^k$ when $\tau^m = \tau \le 2^k$, and $\mathbb{P}(\tilde{Z} = j) = \pi_j$ then follows exactly as above. For the advantages of using monotone updating and $\tilde{Z}$, see the original paper [121]. □

Some further interesting papers related to the Propp–Wilson algorithm are Fill [51], Foss & Tweedie [53], Møller [106] and Propp & Wilson [122].

E countably infinite

We will say that the stationary distribution for a class $\mathcal{P}$ of ergodic transition matrices on E (i.e., a class of ergodic Markov chains) is *simulatable* if there exists a (randomized) stopping time σ for $\{X_n\}$ and a r.v. Z, measurable w.r.t. $\mathcal{F}_\sigma$ where $\mathcal{F}_n = \sigma(X_0, \ldots, X_n, U_0, \ldots, U_n)$ with $U_0, U_1, \ldots$ uniform$(0,1)$ and independent of $\{X_n\}$, such that

$$\mathbb{P}_P(Z \in A) \;=\; \pi_P(A) \;\; \text{for all } A \subseteq E \text{ and all } P \in \mathcal{P} \tag{1.2}$$

where $\mathbb{P}_P$ indicates that $\{X_n\}$ is simulated according to P and π_P is the stationary distribution for P.

Remark 1.9 The 'rules of the game' are thus to use nothing more than a simulated version of $\{X_t\}$ and some possible additional randomization. In particular, the algorithm is not allowed to use analytic information on the p_{ij}. The purpose of this restriction is two–fold: first, if the p_{ij} are analytically available, one can argue that there exist deterministic algorithms for computing π by solving linear equations. Next, the natural description of a Markov chain is most often in terms of an updating rule rather than the p_{ij}, and one would simulate directly from the updating rule rather than use it to compute the p_{ij}. For an example, consider the Kiefer–Wolfowitz vector $\boldsymbol{W}_n = \left(W_n^{(1)}, \ldots, W_n^{(c)}\right)$ in a GI/G/c queue (the components give the workload at the c servers in non–descending order at the nth arrival).. Here the updating rule is

$$\boldsymbol{W}_{n+1} = \mathcal{R}\left(\left[W_n^{(1)} + U_n - T_n\right]^+, \left[W_n^{(2)} - T_n\right]^+, \ldots, \left[W_n^{(c)} - T_n\right]^+\right)$$

where U_n is the service time of customer n, T_n the nth interarrival time, and $\mathcal{R} : [0,\infty)^c \to [0,\infty)^c$ the operator rearranging the components in non–descending order. □

It follows from Section 2 that:

Theorem 1.10 *If E is finite, then the stationary distribution for the class $\mathcal{P}_E$ of all ergodic transition matrices on E is simulatable.*

However (Asmussen, Glynn & Thorisson [14]):

Theorem 1.11 *If E is countably infinite, then the stationary distribution for the class $\mathcal{P}_E$ of all ergodic transition matrices on E is not simulatable.*

Proof We argue by contradiction by assuming that (1.2) holds for $\mathcal{P} = \mathcal{P}_E$. Assume w.l.o.g. that $E = \{0, 1, 2, \ldots\}$.

Let $P^{(0)} = (p_{ij}^{(0)})_{i,j\in E}$ be arbitrary, write $\mathbb{P}_0 = \mathbb{P}_{P^{(0)}}$, $\pi^{(0)} = \pi_{P^{(0)}}$ and choose $K < \infty$ such that $\mathbb{P}_0(Z \le K, M \le K) > 1 - \epsilon$ where $M = \max_{n\le\sigma} X_n$ and $0 < \epsilon < 1/2$. For $\alpha \in (0,1)$, define

$$p_{ij}^{(\alpha)} = \begin{cases} p_{ij}^{(0)} & i \le K \\ \alpha + (1-\alpha)p_{ij}^{(0)} & i = j > K \\ (1-\alpha)p_{ij}^{(0)} & i > K, i \ne j \end{cases}$$

That is, $P^{(\alpha)}$ is obtained from $P^{(0)}$ by adding a geometric number (with parameter α) of 'self–loops' in states $i > K$; on states $i \le K$, the transitions are just the same and hence

$$\mathbb{P}_\alpha(Z \le K, M \le K) = \mathbb{P}_0(Z \le K, M \le K) > 1 - \epsilon\,.$$

Write $\mathbb{P}_\alpha = \mathbb{P}_{P^{(\alpha)}}$, $\pi^{(\alpha)} = \pi_{P^{(\alpha)}}$.

Let $\tau = \inf\{n > 0 : X_n = 0 \mid X_0 = 0\}$ and recall the formula

$$\pi_i \;=\; \frac{1}{\mathbb{E}\tau}\mathbb{E}\sum_{i=0}^{\tau-1} I(X_n = i)$$

for the stationary distribution of a Markov chain. For $i \le K$, this yields

$$\begin{aligned}\pi_i^{(\alpha)} \;&=\; \frac{1}{\mathbb{E}_\alpha\tau}\mathbb{E}_\alpha\sum_{i=0}^{\tau-1} I(X_n = i) \;=\; \frac{1}{\mathbb{E}_\alpha\tau}\mathbb{E}_0\sum_{i=0}^{\tau-1} I(X_n = i)\\ &=\; \frac{\mathbb{E}_0\tau}{\mathbb{E}_\alpha\tau}\pi_i^{(0)}\,. \end{aligned} \tag{1.3}$$

From the self–loop property it follows that

$$\mathbb{E}_\alpha\tau \;\ge \frac{1}{1-\alpha}\mathbb{P}_\alpha\left(\max_{0\le n<\tau} X_n > K\right) \;=\; \frac{1}{1-\alpha}\mathbb{P}_0\left(\max_{0\le n<\tau} X_n > K\right) \tag{1.4}$$

As $\alpha \uparrow 1$, the r.h.s. of (1.4) goes to ∞, and hence (1.3) goes to 0. Hence with $A = \{1, \ldots, K\}$, we have $\pi^{(\alpha)}(A) < \epsilon$ for all α close enough to 1, and get

$$\begin{aligned}\mathbb{P}_\alpha(Z \in A) - \pi^{(\alpha)}(A) \;&\ge\; \mathbb{P}_\alpha(Z \le K, M \le K) - \pi^{(\alpha)}(A)\\ &\ge\; 1 - \epsilon - \epsilon \;> 0\,,\end{aligned}$$

contradicting (1.2). $\square$

The implication of Theorem (1.11) is not necessarily that one should consider exact simulation impossible when faced with a particular non–finite Markov chain $\{X_n\}$. Rather, Theorem (1.11) says that exact simulation cannot be based upon simulated values of $\{X_n\}$ alone but one needs to combine with some specific properties of $\{X_n\}$, i.e. to involve knowledge of the form $P \in \mathcal{P}_0$ where $\mathcal{P}_0 \subset \mathcal{P}_E$. Examples are in Foss & Tweedie [53] in the framework of Harris chains, in [14] in a regenerative setting, and in Ensor & Glynn [49] for the GI/G/1 waiting time (see further IX.1 for the algorithm of [49]).

Example 1.12 In the notation of the proof of Theorem (1.11), the cycle length τ is obviously simulatable. Assume also that the stationary excess variable τ^* with distribution $\mathbb{P}(\tau^* = n) = \mathbb{P}(\tau > n)/\mathbb{E}\tau$ is simulatable. Then it is shown in [14] that exact simulation from π is possible.

Theorem 1.11 shows that τ^* is not in general simulatable even if τ is so (consider X_n = the residual cycle at time n). However, let $\mathcal{P}_0$ be the class of Markov chains such that $\mathbb{P}(\tau > n) \le Cg(n)$ for some variable but explicit constant C and some fixed function g, w.l.o.g. satisfying $g(0) + g(1) + \cdots = 1$. By a result of Keane & O'Brien [87], it is then possible to generate τ^* from an i.i.d. sequence of r.v.'s distributed as τ. I.e., τ^* and hence π is simulatable. □

2 Sample averages

For simplicity, we consider the discrete time case (usually, the continuous time is only notationally different). Let $\{X_n\}_{n=0,1,2,\ldots}$ be a stochastic process in discrete time and with state space $[0,\infty)$ with limit distribution π as $n \to \infty$. We consider the problem of estimating the mean z of π, using a budget of t simulated values of $\{X_n\}$.

The process $\{Y_n\}$ to be simulated could be more complicated and X_n a function of Y_n.

The most obvious estimator is the Cesaro average

$$\hat{z}_t \;=\; \frac{1}{t}\sum_{n=0}^{t-1} X_n\,. \tag{2.1}$$

The reason for this is that $\hat{z}_t \xrightarrow{\text{a.s.}} z$, $t \to \infty$, (consistency) under very general conditions. In fact, it is sometimes argued that it is this property which makes z a relevant performance measure, not the interpretation in terms of stationarity.

Further typical asymptotic properties of (2.1) are a CLT with variance constant σ^2/t,

$$\sqrt{t}(\hat{z}_t - z) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(0,\sigma^2\right) \tag{2.2}$$

where

$$\sigma^2 \;=\; \lim_{t\to\infty} t\,\mathbf{Var}(\hat{z}_t)\,, \tag{2.3}$$

and a bias of order $1/t$,

$$\mathbb{E}\hat{z}_t \;=\; z + \frac{u}{t} + o\left(\frac{1}{t}\right) \tag{2.4}$$

for some constant u.

2a Asymptotics

We start by the bias relation (2.4) which is elementary to obtain:

Proposition 2.1 *Define* $u_n = \mathbb{E}X_n - z$ *and assume* $\sum_1^\infty |u_n| < \infty$. *Then* (2.4) *holds with* $u = \sum_1^\infty u_n$.

Proof Since $\sum_t^\infty u_n = o(1)$, we obtain

$$\begin{aligned} \mathbb{E}\hat{z}_t &= \frac{1}{t}\sum_{n=0}^{t-1}\mathbb{E}X_n \;=\; z + \frac{1}{t}\sum_{n=0}^{t-1}u_n \\ &= z + \frac{u}{t} - \frac{1}{t}\sum_{n=t}^{\infty}u_n \;=\; z + \frac{u}{t} + o\left(\frac{1}{t}\right) . \end{aligned}$$

□

It should be noted that (not unexpectedly) the constant u is typically extremely hard to get a hand on except for the following classical case:

Example 2.2 Let $\{Y_n\}$ be a finite Markov chain with transition matrix $\boldsymbol{P}$ and $X_n = f(Y_n)$. Representing π as a row vector $\boldsymbol{\pi}$ and f as a column vector $\boldsymbol{f}$, we then have $z = \boldsymbol{\pi f}$. The *fundamental matrix* (e.g. Kemeny, Snell & Knapp [89]) is defined as

$$\boldsymbol{F} \;=\; (\boldsymbol{I} - \boldsymbol{P}_1)^{-1} \;=\; \sum_{n=0}^{\infty}\boldsymbol{P}_1^n$$

where $\boldsymbol{P}_1 = \boldsymbol{P} - \boldsymbol{e\pi}$ and $\boldsymbol{e}$ is the column vector of 1's. One can check that the inverse always exists and that $\boldsymbol{P}_1^n = \boldsymbol{P}^n - \boldsymbol{e\pi}$. Thus, if $\boldsymbol{\nu}$ is the initial vector for X_0, $\boldsymbol{\nu e} = 1$ yields

$$u_n \;=\; \boldsymbol{\nu P}^n\boldsymbol{f} - \boldsymbol{\pi f} \;=\; \boldsymbol{\nu P}^n\boldsymbol{f} - \boldsymbol{\nu e \pi f} \;=\; \boldsymbol{\nu}(\boldsymbol{P}^n - \boldsymbol{e\pi})\boldsymbol{f} \;=\; \boldsymbol{\nu P}_1^n\boldsymbol{f}$$

so that

$$u \;=\; \boldsymbol{\nu F f}\,. \tag{2.5}$$

The fundamental matrix also determines the variance constant, see the Appendix where also some generalizations beyond the finite case are discussed in the framework of Poisson's equation. □

It follows in particular from (2.2), (2.4) that we have the desirable property discussed in II.1 that the standard deviation ($= O(1/\sqrt{t})$) should dominate the bias ($= O(1/t)$). Nevertheless, much attention is given in the literature to controlling the bias.

One approach is to perform exact simulation, when feasible, to generate X_0 acccording to π and then go on simulating $X_1, \ldots, X_{t-1}$ according to the update rule; then the bias is 0. The reason that one would not just simulate i.i.d. replications of the stationary r.v. and take the average is that exact simulation is usually time–consuming. See further Assignment 9.

Another approach is to neglect the observations $X_0, \ldots, X_{t_0-1}$ in a suitable 'warm–up' period of length t_0 and use the estimator $\sum_{t_0}^{t-1} X_n/(t-t_0)$. In this way (assuming that $t \to \infty$ with t_0 fixed), the bias is asymptotically reduced from u/t to $\overline{u}_{t_0}/t$ where $\overline{u}_{t_0} = \sum_{t_0}^{\infty} u_n$ whereas the variance remains σ^2/t. The difficulty is that it is usually very hard to asses for a given process how large t_0 must be for $\{X_{t_0}, X_{t_0+1}, \ldots\}$ to be 'almost stationary' in the sense that $\overline{u}_{t_0}$ is substantially smaller than $u = \overline{u}_0$. Nevertheless, the method is widely used in practice.

Remark 2.3 For say the GI/G/1 waiting time process, one can show that $\big\{(1-\rho)W_{t/(1-\rho)^2}\big\}$ can be approximated by a reflected Brownian motion $\big\{\overline{B}(t)\big\}$ with negative drift $-\mu$ and variance constant σ^2, say, when ρ is close to 1 (heavy traffic). The same holds for many other queueing processes, and Whitt [154] then suggests to choose t_0 such that $\big\{\overline{B}(t)\big\}$ has become 'almost stationary at time $t_0(1-\rho)^2$. For example when $\mu = 1$, $\sigma^2 = 1$, the stationary mean of $\big\{\overline{B}(t)\big\}$ is 2, whereas $\mathbb{E}\overline{B}(t)$ is increasing in t and reaches 95% of its steady–state value 1/2 at time when 2.15. Thus, a possible choice is $t_0 = 2.15$.

Whitt [154] and Asmussen [9] contain a number of further applications of heavy traffic analysis in simulation. $\square$

Now turn to the CLT. Various approaches are available for a rigorous proof of (2.2) (under suitable conditions). For example:

Proposition 2.4 *Assume that $\{X_n\}$ is regenerative w.r.t. $\{\tau_n\}$. Then (2.2) holds, with σ^2 related to the variance constant ω^2 in (3.2) via $\sigma^2 = \omega^2\mathbb{E}\tau$.*

The proof is given in connection with regenerative variance estimation below.

2b Variance estimation methods

The purpose is to obtain an estimate $\hat{\sigma}_t^2$ (based upon $X_0, \ldots, X_{t-1}$) of the variance constant σ^2 in (2.2) which is consistent so that

$$\hat{z}_t \pm \frac{1.96\hat{\sigma}_t}{\sqrt{t}}$$

is an asymptotic 95% confidence interval.

In the i.i.d. case, the obvious estimator is

$$\hat{\sigma}_t^2 \;=\; \frac{1}{t-1}\sum_{n=0}^{t-1}(X_n - \hat{z}_t)^2\,. \tag{2.6}$$

However, in a stochastic process context this estimator is not even consistent: its a.s. limit is the variance of π which is general does not equal σ^2. We shall survey a number of methods valid for dependent data.

The time series method

This is based upon stationary process theory so we will assume for simplicity that $\{X_n\}$ is strictly stationary.

Proposition 2.5 *Define* $\rho_k = \mathbf{Cov}(X_n, X_{n+k})$. *Then the limit in* (2.3) *exists and is given by*

$$\sigma^2 \;=\; \rho_0 + 2\sum_{k=1}^{\infty}\rho_k\,,$$

provided the sum converges absolutely.

[the verification of the CLT (2.2) is in the general stationary setting usually performed by involving some mixing condition].

Proof

$$\begin{aligned}
t\mathbf{Var}(\hat{z}_t) &= \frac{1}{t}\sum_{n,m=0}^{t-1}\mathbf{Cov}(X_n, X_m)\\
&= \rho_0 + \frac{2}{t}\sum_{n=0}^{t-1}\sum_{m=n+1}^{t-1}\mathbf{Cov}(X_n, X_m) \;=\; \rho_0 + \sum_{k=1}^{t-1}\frac{t-k}{t}\rho_k\\
&\to \rho_0 + 2\sum_{k=1}^{\infty}\rho_k,
\end{aligned}$$

using dominated convergence in the last step. □

To use Proposition 2.5, we estimate ρ_k by

$$\hat{\rho}_k \;=\; \frac{1}{t-k}\sum_{n=1}^{t-k}\left(X_n-\hat{z}_t\right)\left(X_{n+k}-\hat{z}_t\right)$$

and σ^2 by $\sum_{-N}^{N}\hat{\rho}_k$ where $N<t$ goes to ∞ with t. A difficulty is that the $\hat{\rho}_k$ with k close to t are extremely unprecise; this is reflected in that the rate of convergence of the variance estimator obtained in this way is not equally good as the one $O(t^{-1/2})$ obtained say by the regenerative method.

Batch means

The method of batch means is probably the most common practical choice of variance estimation method. The idea is to divide $X_0,\ldots,X_{t-1}$ into k groups ('batches') of ℓ each ($k\ell=t$)

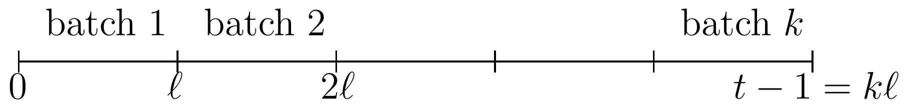

Figure III.2.1

and treat batches as if they were i.i.d. The averages within the k batches are

$$V_1=\frac{1}{\ell}\sum_{n=0}^{\ell-1}X_n,\;V_2=\frac{1}{\ell}\sum_{n=\ell}^{2\ell-1}X_n,\;\ldots\;,\;V_k=\frac{1}{\ell}\sum_{n=(k-1)\ell}^{t-1}X_n\,,$$

and the estimator for z is the grand batch mean

$$\overline{V}\;=\;\frac{1}{k}(V_1+\cdots+V_k)\;=\;\frac{1}{k\ell}\sum_{n=0}^{t-1}X_n$$

which is simply the same as the sample average $\hat{z}_t$.

For estimating the variance on $\overline{V}$, $V_1,\ldots,V_k$ are treated as if they were i.i.d. say with variance ω^2. This leads to $\mathbf{Var}\overline{V}=\omega^2/k$ and, recalling that $\mathbf{Var}\hat{z}_t\approx\sigma^2/t$, $k\ell=t$, that $\sigma^2=\ell\omega^2$. The estimators for ω^2,σ^2 are

$$\hat{\omega}^2=\frac{1}{k-1}\sum_{i=1}^{k}(V_i-\overline{V})^2,\quad\text{resp.}\quad\hat{\sigma}^2=\ell\hat{\omega}^2\,.$$

Under quite general conditions, $\hat{\omega}^2$ is consistent when both k and ℓ go to ∞; see e.g. Damerdji [41], who use strong approximation techniques, and references there. In Carlstein [34] and Goldsman & Meketon [71], it is shown that taking $\ell = O(t^{1/3})$ is optimal in the mean square sense. The choice of k, ℓ subject to the constraint $k\ell = t$ is a trade–off: taking k too small makes $\hat{\omega}^2$ an unprecise estimate, taking ℓ too small makes the assumption of independence between the V_i bad. The method of batch means leaves the bias (of order u/t) unaffected since the estimator remains $\hat{z}_t$.

Multiple replications

Again, we perform the simulation in k groups ('replications') of ℓ each ($k\ell = t$) but now groups are truly i.i.d., not just approximately as for the batch means method. The difference is illustrated on Fig. III.2.2.

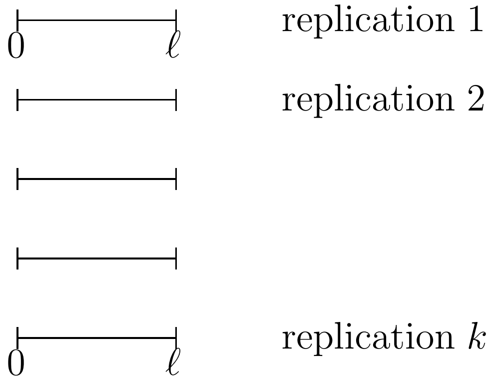

Figure III.2.2

Thus the k replications use independent versions $\left\{X_n^{(1)}\right\}, \ldots, \left\{X_n^{(k)}\right\}$ of $\{X_n\}$. The average from the ith replication is

$$V_i \;=\; \frac{1}{\ell}\sum_{n=1}^{\ell} X_n^{(i)} ,$$

and the estimator for z is $\overline{V} = (V_1 + \cdots + V_k)/k$. The variance is ω^2/k, with ω^2 estimated by $\sum_1^k (V_i - \overline{V})^2/(k-1)$. Since

$$\mathbb{E}V_i \;=\; \mathbb{E}\hat{z}_\ell \;\sim\; z + \frac{u}{\ell} ,$$

also $\mathbb{E}\overline{V} \sim z + u/\ell$ so that we have an asymptotic bias of order $1/\ell$. Thus, the bias is increased compared to the batch mean method but on the other hand, the V_i are truly independent.

For asymptotic purposes, we want both k and ℓ to go to infinity with t, and in such a way that the standard deviation $\omega/t^{1/2}$ dominates the bias u/ℓ. Since $\omega^2 \sim \sigma^2/\ell$, this is achieved if $1/\ell$ is much smaller than $1/\sqrt{k\ell}$, i.e. if ℓ is much larger than k. Since $t = k\ell$, the maximal value of ℓ is t. However, we cannot choose ℓ too close to t because k must be sufficiently large for the appropriate CLT to be in force.

Further asymptotic discussion of the method of replications can be found in Glynn [63].

3 Regenerative simulation

Let $\{X_n\}_{n=0,1,2,\ldots}$ be a regenerative process (see the Appendix) in discrete time (usually, the continuous time is only notationally different), with regeneration points $\{\tau_n\}$, state space $[0,\infty)$ and limit distribution π as $n \to \infty$. We want to estimate the mean z of π and to give a confidence interval.

Note that in practice, the process $\{Y_n\}$ to be simulated would be more complicated and X_n a function of Y_n. Say Y_n is the state of a queueing network as seen from the view of the nth customer and X_n is his sojourn time. This is no problem in a regenerative setting: if $\{Y_n\}$ is regenerative w.r.t. $\{\tau_n\}$, then so is $\{\varphi(Y_n)\}$ for any φ.

The reason that the regenerative structure is particularly convenient from the point of view of producing confidence intervals is the independence of cycles, which allows everything to be reduced to standard i.i.d. theory. To this end, we recall the formula

$$z \;=\; \frac{1}{\mathbb{E}\tau}\mathbb{E}\sum_{k=0}^{\tau-1} X_k\,. \tag{3.1}$$

The idea is now to simulate the process until n cycles have been completed, estimate $\mathbb{E}\tau$ by the empirical mean of $\tau_1,\ldots,\tau_n$, $\mathbb{E}\sum_0^{\tau-1} X_k$ by the empirical mean of the corresponding quantities over the n cycles, and z by the ratio. A confidence interval can then be build from the i.i.d. structure of the cycles. The details follow.

To conform with the notation of II.1, we let

$$Z^{(1)} \;=\; \tau, \quad Z^{(2)} \;=\; \sum_{k=0}^{\tau-1} X_k, \quad \boldsymbol{Z} \;=\; \left(Z^{(1)}, Z^{(2)}\right),$$

$$z^{(i)} \;=\; \mathbb{E}Z^{(i)}, \quad \sigma_{ij} \;=\; \mathbf{Cov}\left(Z^{(i)}, Z^{(j)}\right), \quad f(x,y) \;=\; \frac{y}{x}$$

Then $z = f\left(z^{(1)}, z^{(2)}\right)$, and our estimator is

$$\hat{z} \;=\; f\left(\hat{z}^{(1)}, \hat{z}^{(2)}\right) \;=\; \frac{\hat{z}^{(2)}}{\hat{z}^{(1)}} \quad \text{where} \quad \hat{z}^{(i)} \;=\; \frac{Z_1^{(i)} + \cdots + Z_n^{(i)}}{n}$$

with

$$Z_n^{(1)} \;=\; \tau_n - \tau_{n-1}, \quad Z_n^{(2)} \;=\; \sum_{k=\tau_1+\cdots+\tau_{n-1}}^{\tau_1+\cdots+\tau_n-1} X_k\,.$$

Then $\sqrt{n}(\hat{z} - z) \xrightarrow{\mathcal{D}} N(0, \omega^2)$ where

$$\begin{aligned} \omega^2 \;&=\; \sum_{i,j=1}^{2} \frac{\partial f}{\partial z^{(i)}} \frac{\partial f}{\partial z^{(j)}} \sigma_{ij} \\ &=\; \frac{z^{(2)^2}}{z^{(1)^4}}\sigma_{11} + \frac{1}{z^{(1)^2}}\sigma_{22} - 2\frac{z^{(2)}}{z^{(1)^3}}\sigma_{12}\,, \end{aligned} \tag{3.2}$$

using

$$\frac{\partial f}{\partial x}(x,y) = -\frac{y}{x^2}, \quad \frac{\partial f}{\partial y}(x,y) = \frac{1}{x}\,.$$

Our confidence interval is

$$\hat{z} \pm \frac{1.96\,\hat{\omega}}{\sqrt{n}} \tag{3.3}$$

where

$$\begin{aligned} \hat{\omega}^2 \;&=\; \frac{\hat{z}^{(2)^2}}{\hat{z}^{(1)^4}}\hat{\sigma}_{11} + \frac{1}{\hat{z}^{(1)^2}}\hat{\sigma}_{22} - 2\frac{\hat{z}^{(2)}}{\hat{z}^{(1)^3}}\hat{\sigma}_{12}\,, \\ \hat{\sigma}_{ij} \;&=\; \frac{1}{n-1}\sum_{m=1}^{n}(Z_m^{(i)} - \hat{z}^{(i)})(Z_m^{(j)} - \hat{z}^{(j)})\,. \end{aligned}$$

Note that since

$$Z_1^{(1)} + \cdots + Z_n^{(1)} \;=\; \tau_n\,, \quad Z_1^{(2)} + \cdots + Z_n^{(2)} \;=\; \sum_{k=0}^{\tau_n - 1} X_k$$

the regenerative estimator $\hat{z}$ can also be interpreted as the sample average $\hat{z}_{\tau_n}$ with random sample size $t = \tau_n$. The crux of the regenerative method is the natural way in which the variance can be estimated.

Remark 3.1 A sometimes useful twist of regenerative simulation is that is fact the strict independence of cycles is not all that essential. The relevant generalization is that of Palm theory ([21] or [150]). Here the assumption is that the cycles

$$\left\{X_{k+\tau_1+\cdots+\tau_{n-1}}\right\}_{k=0,\dots,\tau_n-\tau_{n-1}-1}$$

are not i.i.d. but form a strictly stationary sequence (the process $\{X_k\}$ is *event–stationary* w.r.t. $\{\tau_n\}$). We can then form a regenerative process $\left\{\tilde{X}_k\right\}$ by replacing the dependent cycles by i.i.d. cycles with the same distribution and estimate the variance as above; the practical feasability of this program of course assumes that the event–stationary distribution of X_{τ_n} can be generated. A particular important case is Harris chains, where cycles are one–dependent ([6] Ch. VI.3); with one–dependence, one can in principle write up a natural variance estimator but the approach in terms of i.i.d. cycles is more straightforward to implement. □

3a Regenerative variance estimation

We assume here that $\{X_n\}$ is regenerative. We do not use the regenerative estimator for z studied above, but just the sample average $\hat{z}_t$. However, the variance is estimated regeneratively from cycles $1, 2, \dots, N_t - 1$ where

$$N_t \;=\; \inf\{n:\, \tau_1+\cdots+\tau_n > t\}$$

(thus, $N_t - 1$ is the number of cycles copleted by time t).

The first important remark is that $\hat{z}_t$ obeys the same CLT as the two estimators

$$\hat{z}_t^{(-)} \;=\; \frac{Z_1^{(2)}+\cdots+Z_{N_t-1}^{(2)}}{Z_1^{(1)}+\cdots+Z_{N_t-1}^{(1)}}, \quad \hat{z}_t^{(+)} \;=\; \frac{Z_1^{(2)}+\cdots+Z_{N_t}^{(2)}}{Z_1^{(1)}+\cdots+Z_{N_t}^{(1)}}\,.$$

Note that $\hat{z}_t^{(-)}$, $\hat{z}_t^{(+)}$ are regenerative estimators with a random number ($N_t - 1$, resp. N_t) of cycles, and that they can also be interpreted as sample means with random sample sizes ($t^{(-)} = \tau_1 + \cdots + \tau_{N_t-1}$, resp. $t^{(+)} = \tau_1 + \cdots + \tau_{N_t}$) in view of

$$Z_1^{(1)}+\cdots+Z_{N_t}^{(1)} \;=\; t^{(+)}\,, \quad Z_1^{(2)}+\cdots+Z_{N_t}^{(2)} \;=\; \sum_{n=0}^{t^{(+)}-1} X_n$$

(and similarly for $\hat{z}_t^{(-)}$).

Proposition 3.2 *The estimators $\hat{z}_t$, $\hat{z}_t^{(-)}$, $\hat{z}_t^{(+)}$ are asymptotically equivalent in the sense of Definition A.2.*

In the proof, we need the following obvious lemma:

Lemma 3.3 *The processes*

$$\{X_t + \cdots + X_{t^{(+)}-1}\}_{t=0,1,2,\ldots}, \quad \left\{t^{(+)} - t\right\}_{t=0,1,2,\ldots}$$

are regenerative w.r.t. $\{\tau_n\}$, hence convergent in distribution.

Proof of Proposition 3.2. We show that $\hat{z}_t$ and $\hat{z}_t^{(+)}$ are asymptotically equivalent (the case of $\hat{z}_t^{(-)}$ is similar). Let

$$\hat{z}_t^{*(+)} = \frac{Z_1^{(2)} + \cdots + Z_{N_t}^{(2)}}{t} = \frac{t^{(+)}}{t}\hat{z}_t^{(+)}.$$

Then

$$\sqrt{t}\left|\hat{z}_t^{(+)} - \hat{z}_t^{*(+)}\right| = \frac{t^{(+)} - t}{\sqrt{t}}\hat{z}_t^{(+)} \xrightarrow{\mathcal{D}} 0 \cdot z,$$

using Lemma 3.3 in the last step, so that it suffices to show asymptotic equivalence of $\hat{z}_t$ and $\hat{z}_t^{*(+)}$. This follows by another application of Lemma 3.3:

$$\sqrt{t}\left|\hat{z}_t^{*(+)} - \hat{z}_t\right| = \left|\frac{X_t + \cdots + X_{t^{(+)}-1}}{\sqrt{t}}\right| \xrightarrow{\mathcal{D}} 0.$$

□

We can now complete the
Proof of Proposition 2.4. Since $N_t/t \overset{\text{a.s.}}{\to} 1/\mathbb{E}\tau$, a variant of Anscombe's theorm yields $\sqrt{N_t}(\hat{z}_t^{(+)} - z) \xrightarrow{\mathcal{D}} N(0, \omega^2)$ or, equivalently $\sqrt{t}(\hat{z}_t^{(+)} - z) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$ where $\sigma^2 = \omega^2\mathbb{E}\tau$. Now just appeal to the asymptotic equivalence of $\hat{z}_t$ and $\hat{z}_t^{(+)}$. □

Remark 3.4 In many examples, there are several possible choices of regeneration points. Say $\{\tau_n\}$, $\{\tau_n'\}$ are two possible choices. Should we then use $\{\tau_n\}$ or $\{\tau_n'\}$?

One may be tempted to think that $\{\tau_n\}$ is preferable if $\mathbb{E}\tau < \mathbb{E}\tau'$ and vice versa (more cycles will then be obtained within a fixed simulation budget t). However, since σ^2 (as a long-run variance constant) is independent of the choice of cycle structure, *the choice between $\{\tau_n\}$ and $\{\tau_n'\}$ does not affect the variance σ^2/t on the regenerative estimator.*

The possible difference is therefore in small–sample properties or in the variance on the variance estimator $\hat{\sigma}^2$. However, even here there are no simple rules. The asymptotic variance of $\hat{\sigma}^2$ can be computed by a similar but more tedious application of the Delta method as in Proposition II.2.2, cf. Glynn & Iglehart [66]. The resulting expression is complicated but can be used to show that taking the smaller expected cycle length does not necessarily lead to the smaller variance on $\hat{\sigma}^2$. □

4 Duality representations

Our starting point is

Example 4.1 In the GI/G/1 queue, we have the representation $W \stackrel{\mathcal{D}}{=} M$ where W is the steady–state waiting time, $M = \sup_{n=0,1,\ldots} S_n$ and $\{S_n\}$ is a random walk with increments X_k distributed as the difference between a service time and an independent interarrival time, cf. the Appendix. Assume that we want to estimate $z = \mathbb{P}(W > x)$. We can then write $z = \mathbb{P}(M > x)$ $= \mathbb{P}(\tau(x) < \infty)$ where $\tau(x) = \inf\{n > 0 : S_n > x\}$. In this way the problem of simulating a stationary characteristics is converted to the problem of simulating a first passage probability.

The difficulty is of course that CMC is not feasible since $\mathbb{P}(\tau(x) < \infty) < 1$ under the stability condition $\rho < 1$ (equivalent to $\mathbb{E}X < 0$) and so $I(\tau(x) < \infty)$ cannot be generated by simulating $\{S_n\}$ up to a stopping time. However, in this example a classical importance sampling technique based upon exponential change of measure exists and does not only resolve the infinite horizon problem but does in fact give extremely accurate estimates also for very large x. We return to this in Chapter IV. □

Here are two further less standard examples:

Example 4.2 The GI/G/1 queue waiting time process can be viewed as a random walk reflected at zero in view of the two equivalent representation $W_{n+1} = (W_n + X_n)^+$ (the Lindley recursion) or $W_n = S_n - \inf_{0\le k\le n} S_k$. In many problems involving a finite buffer $b < \infty$, one has instead a random walk reflected both at 0 and b,

$$V_{n+1} \;=\; \min(b, \max(0, V_n + X_n))$$

Also in this case, there is a first passage representation of the stationary r.v. V (Siegmund [148]):

$$\mathbb{P}(V \ge z) \;=\; \mathbb{P}(S_{\tau(z-b,z)} \ge b), \quad 0 \le z \le b,$$

where $\tau(z-b,z) = \inf\{n > 0 : S_n \notin (z-b,z)\}$. In this example, $\tau(z-b,z) < \infty$ a.s., and hence CMC simulation with $Z = I(\tau(z-b,z) < \infty)$ is feasible. □

Example 4.3 Let $\{V_t\}_{t\geq 0}$ be a storage process with release rule $r(x)$ and compound Poisson input $\{A_t\}$, $A_t = \sum_1^{N_t} U_i$ where $\{N_t\}$ is a Poisson process with intensity β and the U_i are independent of $\{N_t\}$ with distribution B concentrated on $(0,\infty)$,

$$V_t \;=\; A_t \;-\; \int_0^t r(A_s)\,ds.$$

Then, similarly to the two preceeding examples, the stationary distribution can be represented as a first passage probability

$$\mathbb{P}(V \geq x) \;=\; \mathbb{P}(\tau(x) < \infty) \tag{4.1}$$

where $\tau(x) = \inf\{t > 0 : R_t \leq 0 \,|\, R_0 = x\}$ and $\{R_t\}_{t\geq 0}$ is given by

$$R_t \;=\; \int_0^t r(A_s)\,ds \;-\; A_t.$$

Again, there is an infinite horizon problem, and we briefly mention the relevant importance sampling scheme in Chapter IX. □

The examples above all fit into a common framework where one has two processes $\{V_t\}$, $\{R_t\}$ with state space $[0,\infty]$, or subintervals, connected via the formula (4.1). The first such general construction is due to Siegmund [148] in a Markov process context. Starting from $\{V_t\}$, Siegmund constructed $\{R_t\}$ in terms of its transition probabilities by

$$\mathbb{P}(R_t \leq y \,|\, R_0 = x) \;=\; \mathbb{P}(V_t \geq x \,|\, V_0 = y). \tag{4.2}$$

For this to define a transition semi–group, it is necessary and sufficient that $\{V_t\}$ is stochastically monotone and that $\mathbb{P}(V_t \geq x \,|\, V_0 = y)$ is right–continuous in y for any fixed x. Note that taking $x = y = 0$ shows that state 0 is absorbing for $\{R_t\}$. If (4.2) holds, one then obtains (4.1) by taking $y = 0$ and letting $t \to \infty$. The processes $\{V_t\}$ and $\{R_t\}$ are said to be in *Siegmund duality*.

In Example 4.1 with $V = W$, one can define $R_n = x - S_n$ as long as $x - S_n > 0$; when $(-\infty, 0)$ is hit, R_n is reset to 0 and remains there for ever. Example 4.2 is the same except for a further resetting to ∞ when $(b, \infty]$ is hit. Example 4.3 is also a particular case; a direct verification of (4.2) was performed in Asmussen & Schock Petersen [19] by sample path time reversion.

Asmussen & Sigman [20] gave a generalization beyond the Markov case by considering a recursive discrete time setting where $\{V_n\}$ is given recursively by $V_0 = y$, $V_{n+1} = f(V_n, U_n)$ where the driving sequence $\{U_n\}$ is not i.i.d. as for the Markov case, but strictly stationary, w.l.o.g. with doubly infinite time ($n \in \mathbf{Z}$). Assuming $f(v, u)$ to be non–decreasing in v for fixed u, one then defines g by letting $g(\cdot, u)$ be a generalized inverse of $f(\cdot, u)$ and lets $R_0 = x$, $R_{n+1} = g(R_n, U_{-n})$. Then again (4.2) holds and hence so does (4.1).

Steady–state simulation via (4.1) is typically elegant and (when combined with say importance sampling) efficient when it is feasible. The main limitation is that monotonicity and existence of the generalized inverse of $f(\cdot, u)$ (which most often limits the state space to be one–dimensional) are required; some progress to get beyond this was recently obtained by Blaszczyszyn & Sigman [27] but the practical usefulness for simulation seems questionable so far. Also a general non–Markovian theory in continuous time is missing. Some further relevant references for duality are Asmussen & Rubinstein [17] (simulation aspects) and Asmussen [11] (Markov–modulated continuous time models)

Chapter IV

Rare events simulation

1 Introduction

The problem is to estimate $z = \mathbb{P}A$ when z is small, say of the order 10^{-3} or less. I.e., $Z = I(A)$ and A is a *rare event*. Examples occur in telecommunications ($z =$ bit loss rate, probability of buffer overflow), reliability ($z =$ the probability of failure before t), insurance risk ($z =$ the ruin probability) etc.

The CMC method leads to a variance of $\sigma_Z^2 = z(1-z)$ which tends to zero as $z \downarrow 0$. However, the issue is not so much that the absolute error is small as that the relative error is high:

$$\frac{\sigma_Z}{z} \;=\; \frac{\sqrt{z(1-z)}}{z} \;\sim\; \frac{1}{\sqrt{z}} \;\to\; \infty, \quad z \downarrow 0.$$

In other words, a confidence interval of width 10^{-4} may look small, but if the point estimate $\hat{z}$ is of the order 10^{-5}, it does not help telling whether z is of the magnitude 10^{-4}, 10^{-5} or even much smaller. Another way to illustrate the problem is in terms of the sample size n needed to acquire a given relative precision, say 10%, in terms of the half–width of the confidence interval. This leads to the equation $1.96\sigma_Z/(z\sqrt{n}) = 0.1$, i.e.

$$n \;=\; \frac{100 \cdot 1.96^2 z(1-z)}{z^2} \;\sim\; \frac{100 \cdot 1.96^2}{z} \tag{1.1}$$

increases like z^{-1} as $z \downarrow 0$. Thus, if z is small, large sample sizes are required, and when we get to probabilities of the order $z \sim 10^{-9}$, which occurs in many telecomunications applications, CMC simulation is not only inefficient but in fact impossible.

For a formal set–up allowing to discuss such efficiency concepts, let $\{A(x)\}$ be a family of rare events where $x \in (0,\infty)$ or $x \in \mathbb{N}$, assume that $z(x) = \mathbb{P}A(x) \to 0$ as $x \to \infty$ and for each x let $Z(x)$ be as estimator of $z(x)$, i.e. $\mathbb{E}Z(x) = z(x)$. An algorithm is defined as a family $\{Z(x)\}$ of such r.v.'s.

The best performance which has been observed in realistic rare events setting is *bounded relative error* as $x \to \infty$,

$$\limsup_{x\to\infty} \frac{\mathbf{Var}Z(x)}{z(x)^2} \ < \ \infty\,. \tag{1.2}$$

In particular, such an algorithm will have the feature that n as computed in (1.1) (with $z(1-z)$ replaced by $\mathbf{Var}Z(x)$) remains bounded as $x \to \infty$.

An efficiency concept slightly weaker than (1.2) is *logarithmic efficiency*: $\mathbf{Var}(Z(x)) \to 0$ so quickly that

$$\limsup_{x\to\infty} \frac{\mathbf{Var}(Z(x))}{z(x)^{2-\epsilon}} \ = \ 0 \tag{1.3}$$

for all ϵ, or, equivalently, that

$$\liminf_{x\to\infty} \frac{|\log \mathbf{Var}(Z(x))|}{|\log z(x)^2|} \ \geq \ 1\,. \tag{1.4}$$

Note that this is slightly weaker than bounded relative error. For example, if $z(x) \sim Ce^{-\gamma x}$, it allows $\mathbf{Var}(Z(x))$ to decrease like $x^p e^{-2\gamma x}$ or even $e^{-2\gamma x+\beta\sqrt{n}}$. The reason for working with logaritmic efficiency rather than bounded relative error is that the difference is minor from a practical point of view and that logarithmic efficiency often is much easier to verify.

In accordance with discussions of run lengths in Chapter II, it would have been more logical to replace $\mathbf{Var}(Z(x))$ by $T(x)\mathbf{Var}(Z(x))$ in (1.2), (1.3). One can check that in the examples we discuss, $T(x)$ grows so slowly with x that this makes no difference.

Much of the work on rare events simulation is focused on importance sampling as a potential (though not the only) way to design efficient algorithms; in fact, two of the three algorithms in Section 2 which are our main examples employ this method. The optimal change of measure (as discussed generally for importance sampling in II.4) is given by

$$\tilde{\mathbb{P}}(B) = \mathbb{E}\left[\frac{Z}{z};B\right] = \frac{1}{z}\mathbb{P}(AB) = \mathbb{P}(B|A).$$

I.e., *the optimal* $\tilde{\mathbb{P}}$ *is the conditional distribution given* A. However, just the same problem as for importance sampling in general comes up: it is usually not practicable to simulate from $\mathbb{P}(\cdot|A)$, and we cannot compute the likelihood ratio since z is unknown. Again, we may try to make $\tilde{\mathbb{P}}$ look as much like $\mathbb{P}(\cdot|A)$ as possible; in fact, the two importance sampling algorithms in Section 2 may be seen in this light.

Some general references on rare events simulation are Heidelberger [77] and Asmussen & Rubinstein [17].

2 Three efficient algorithms

In this section, we give three examples of algorithms meeting the efficiency criteria discussed in Section 1. One will note that they all deal with extremely simple problems. In more complex situations, one should not expect to be able to find rare events estimators which are say logarithmically efficient. Rather, the ideas behind algorithms like the ones we study will then provide guidelines on how to proceed for getting some substantial variance reduction without necessarily meeting the efficiency criteria in full.

2a Siegmund's algorithm

Let $X_1, X_2, \ldots$ be i.i.d. with common distribution F not concentrated on $(-\infty, 0]$ or $[0, \infty)$, assume that $\mathbb{E}X < 0$, and define

$$S_n = X_1 + \cdots + X_n, \quad \tau(x) = \inf\{n : S_n > x\}.$$

The problem is to estimate $z(x) = \mathbb{P}(\tau(x) < \infty)$ when x is large and hence $z(x)$ small. This problem has many applications: GI/G/1 waiting times, ruin probabilities, sequential tests (see the Appendix).

Exponential change of measure

Define the exponential family $\{F_\theta\}$ generated by F as in the Appendix. Applying Wald's fundamental identity (Proposition A.1) with $A = \{\tau(x) < \infty\}$, we get

$$\mathbb{P}(\tau(x) < \infty) \;=\; \mathbb{E}_\theta\left[L_{\tau(x),\theta}; \tau(x) < \infty\right] \tag{2.1}$$

where

$$L_{n,\theta} \;=\; \prod_{k=1}^{n} \frac{\hat{F}[\theta]}{e^{\theta X_k}} \;=\; e^{-\theta S_n}\hat{F}[\theta]^n.$$

Now consider the choice of θ. The first step is to choose θ such that $\mathbb{P}_\theta(\tau(x) < \infty) = 1$, i.e. $\mathbb{E}_\theta X \geq 0$. Noting that

$$\mathbb{E}_\theta X \;=\; \mathbb{E}\frac{Xe^{\theta X}}{\hat{F}[\theta]} \;=\; \frac{\hat{F}'[\theta]}{\hat{F}[\theta]},$$

this means that $\theta \geq \gamma_0$ (the solution of $\hat{F}'[\theta] = 0$ or, equivalently, the minimizer of $\hat{F}[\theta]$; cf. Fig. 2.1). For such a θ, (2.1) becomes

$$z(x) \;=\; \mathbb{P}(\tau(x) < \infty) \;=\; \mathbb{E}_\theta L_{\tau(x),\theta}\,. \tag{2.2}$$

Thus, we may perform the simulation by the CMC method with $Z(x) = L_{\tau(x),\theta}$.

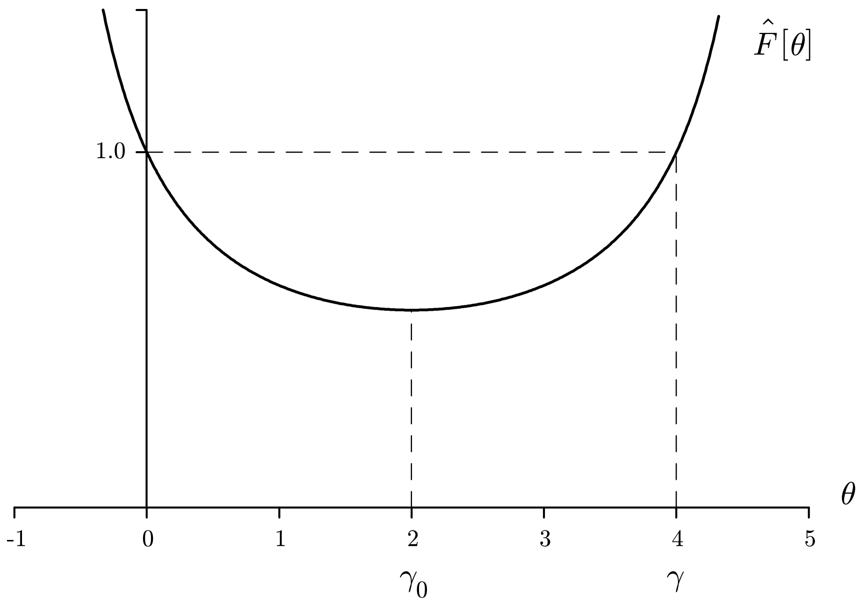

Figure IV.2.1

The crucial fact is now that typically a certain value γ of θ is superior. We will assume that there exist a $\gamma > 0$ such that $\hat{F}[\gamma] = 1$, $\hat{F}'[\gamma] < \infty$; in view of $\mathbb{E}X < 0$ and convexity, this basically only says that enough exponential moments exist, cf. Fig. 2.1. For this special case, (2.2) becomes

$$z(x) \;=\; \mathbb{P}(\tau(x) < \infty) \;=\; \mathbb{E}_\gamma e^{-\gamma S_{\tau(x)}} \;=\; e^{-\gamma x}\mathbb{E}_\gamma e^{-\gamma \xi(x)}\,, \tag{2.3}$$

where $\xi(x) = S_{\tau(x)} - x$ is the overshoot.. Indeed we shall show:

Theorem 2.1 *The algorithm given by* $Z(x) = e^{-\gamma x}e^{-\gamma\xi(x)}$ [simulated from $\mathbb{P}_\gamma$] *has bounded relative error.*

Example 2.2 Assume that F is $\mathcal{N}(-\mu, 1)$ where $\mu > 0$.
Then $\hat{F}[s] = \exp\left\{-\mu s + s^2/2\right\}$, so that γ solves $0 = -\mu\gamma + \gamma^2/2$ which in view of $\gamma > 0$ implies $\gamma = 2\mu$. We then get

$$\hat{F}_\gamma[s] \;=\; \hat{F}[s+\gamma] \;=\; \exp\left\{\mu s + s^2/2\right\}$$

which shows that F_γ is $\mathcal{N}(\mu, 1)$. □

Example 2.3 Assume that $X = U - T$ is the independent difference between two exponential r.v.'s with rates δ, resp. β ($\beta < \delta$). This corresponds to the M/M/1 queue with arrival rate β and service rate δ. Then $\hat{F}[\gamma] = 1$ means

$$1 \;=\; \mathbb{E}e^{\gamma U}\mathbb{E}e^{-\gamma T} \;=\; \frac{\delta}{\delta-\gamma}\frac{\beta}{\beta+\gamma}$$

which has the positive solution $\gamma = \delta - \beta$. We then get

$$\hat{F}_\gamma[s] \;=\; \hat{F}[s+\gamma] \;=\; \frac{\beta}{\beta-s}\frac{\delta}{\delta+s}$$

which shows that F_γ is the distribution of the independent difference between two exponential r.v.'s with rates β, resp. δ. I.e., the changed measure corresponds to the M/M/1 queue with arrival rate δ and service rate β (the rates are switched). □

Asymptotics for $z(x)$

The process $\{\xi(x)\}_{x\ge 0}$ is regenerative (regenerates at each partial maximum of $\{S_n\}$). Thus $\xi(x) \xrightarrow{\mathcal{D}} \xi$ and

$$\mathbb{E}_\gamma e^{-\gamma\xi(x)} \;\to\; C \;=\; \mathbb{E}_\gamma e^{-\gamma\xi}\,.$$

Thus

$$z(x) \;=\; \mathbb{P}(\tau(x) < \infty) \;\sim\; Ce^{-\gamma x}\,, \tag{2.4}$$

a celebrated result going back to Cramér (1930).

Asymptotics for $\mathbf{Var}_\gamma Z(x)$

The calculations are almost the same as for $x(x)$. Recalling that $Z = e^{-\gamma x}e^{-\gamma\xi(x)}$, we get

$$\mathbb{E}_\gamma Z^2 \;=\; e^{-2\gamma x}\mathbb{E}_\gamma e^{-2\gamma\xi(x)} \;\approx\; C_1 e^{-2\gamma x}$$

where $C_1 = \mathbb{E}_\gamma e^{-2\gamma\xi}$. By Jensen's inequality, $C_1 > C^2$, and hence

$$\mathbf{Var}_\gamma Z(x) \;\sim\; C_1 e^{-2\gamma x} - \left(Ce^{-\gamma x}\right)^2 \;\sim\; C_2 e^{-2\gamma x}\,, \tag{2.5}$$

where $C_2 = C_1 - C^2 > 0$. The relative error is thus

$$\frac{\sqrt{\mathbf{Var}_\gamma Z}}{z(x)} \;\sim\; \frac{C_2^{1/2}e^{-\gamma x}}{Ce^{-\gamma x}} \;=\; C_3$$

($C_3 = C_2^{1/2}/C$) which does not increase with x, completing the proof of Theorem 2.1. □

Note that in this example we have $T(x) \sim \mathbb{E}_\gamma \tau(x) = O(x)$.

Uniqueness of the change of measure in Siegmund's algorithm

Consider as above an importance sampling algorithm for estimating $z(x) = \mathbb{P}(\tau(x) < \infty)$ for a random walk with negative drift $\mu = \mu_F$, with the extension that we allow an arbitrary candidate G for the changed distribution of the X_k. That is, we simulate $X_1, X_2, \ldots$ from G and use the estimator

$$Z(x) \;=\; W_{\tau(x)}(F|G) \;=\; \frac{dF}{dG}(X_1)\ldots\frac{dF}{dG}(X_{\tau(x)})\,, \tag{2.6}$$

where dF/dG means Radon–Nikodym derivative (e.g., if F and G have both densities f, g w.r..t. Lebesgue measure, then $(dF/dG)(x) = f(x)/g(x)$). Note that we must impose two conditions on G: that dF/dG exists and that G has positive mean μ_G (otherwise, the simulation does not terminate in finite time).

Theorem 2.4 *The importance sampling algorithm* (2.6) *is logarithmically efficient if and only if* $G = F_\gamma$.

Proof (Asmussen & Rubinstein [17]) Sufficiency [even with the stronger conclusion of bounded relative error] is contained in Theorem 2.1, so we assume the IS distribution is $G \neq F_\gamma$.

By the chain rule for Radon–Nikodym derivatives,

$$\begin{aligned}\mathbb{E}_G Z(x)^2 &= \mathbb{E}_G W^2_{\tau(x)}(F|G) = \mathbb{E}_G \left[W^2_{\tau(x)}(F|F_\gamma) W^2_{\tau(x)}(F_\gamma|G)\right] \\ &= \mathbb{E}_\gamma \left[W^2_{\tau(x)}(F|F_\gamma) W_{\tau(x)}(F_\gamma|G)\right] = \mathbb{E}_\gamma \exp\{K_1 + \cdots + K_{\tau(x)}\},\end{aligned}$$

where

$$K_i = \log\left(\frac{dF_\gamma}{dG}(X_i)\left(\frac{dF}{dF_\gamma}(X_i)\right)^2\right) = -\log\frac{dG}{dF_\gamma}(X_i) - 2\gamma X_i.$$

Here

$$\mathbb{E}_\gamma K_i = \epsilon' - 2\gamma \mathbb{E}_{F_\gamma} X_i = \epsilon' - 2\gamma\mu_\gamma,$$

where $\mu_\gamma = \mu_{F_\gamma} > 0$ and

$$\epsilon' = -\mathbb{E}_\gamma \log\frac{dG}{dF_\gamma}(X_i) > 0,$$

by the information inequality (see the Appendix). Since $K_1, K_2, \ldots$ are i.i.d., Jensen's inequality and Wald's identity yield

$$\mathbb{E}_G Z(x)^2 \geq \exp\left\{\mathbb{E}_\gamma(K_1 + \cdots + K_{\tau(x)})\right\} = \exp\left\{\mathbb{E}_\gamma \tau(x)(\epsilon' - 2\gamma\mu_\gamma)\right\}.$$

Since $\mathbb{E}_{F_\gamma}\tau(x)/x \to 1/\mu_\gamma$, it thus follows (using (2.4)) that for $0 < \epsilon'' < \epsilon'$, $0 < \epsilon < \epsilon''/\gamma\mu_\gamma$,

$$\begin{aligned}\liminf_{x\to\infty} \frac{\mathbb{E}_G Z(x)^2}{z(x)^{2-\epsilon}} &= \liminf_{x\to\infty} \frac{\mathbb{E}_G Z(x)^2}{C^{2-\epsilon} e^{-2\gamma x + \epsilon\gamma x}} \\ &\geq \liminf_{x\to\infty} \frac{e^{x(\epsilon''/\mu_\gamma - 2\gamma)}}{C^{2-\epsilon} e^{-2\gamma x + \epsilon\gamma x}} = \infty,\end{aligned}$$

which completes the proof. □

2b Efficient simulation of $\mathbb{P}(S_n > n(\mu + \epsilon))$

Consider again a random walk $S_n = X_1 + \cdots + X_n$ where $X_1, X_2, \ldots$ are i.i.d. with common distribution F with mean μ. The rare event in question is now $A(n) = \{S_n > n(\mu + \epsilon)\}$ where $\epsilon > 0$ [thus the index n is discrete in this example]. That $z(n) = \mathbb{P}A(n) \to 0$ as $n \to \infty$, and hence that the event is rare indeed, is immediate from the LLN.

We shall again employ exponential change of measure. Writing $\hat{F}[s] = e^{\kappa(s)}$, i.e. $\kappa(s) = \log \hat{F}[s]$, we have

$$\frac{dF_\theta}{dF}(x) = e^{\theta x - \kappa(\theta)}, \quad L_{n;\theta} = e^{-\theta S_n + n\kappa(\theta)}. \tag{2.7}$$

Thus $Z(n) = e^{-\theta S_n + n\kappa(\theta)} I(S_n > n(\mu + \epsilon))$.

The relevant choice of θ turns out to be as for the saddlepoint method:

$$\mathbb{E}_\theta X = \kappa'(\theta) = \mu + \epsilon \tag{2.8}$$

which in particular implies $\theta > 0$ (since κ' is strictly increasing according to the strict convexity of κ) and $I > 0$ where $I = \theta(\mu + \epsilon) - \kappa(\theta)$. Cf. Fig. IV.2.2.

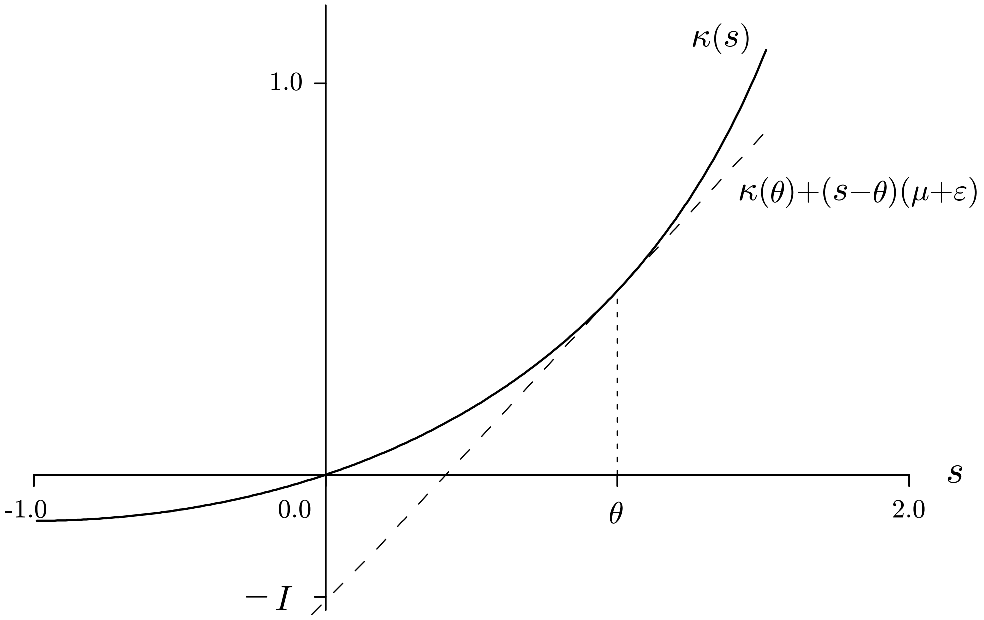

Figure IV.2.2

Theorem 2.5 *The exponential change of measure* (2.7), (2.8) *is logaritmically efficient and the only importance sampling distribution G with this property.*

Lemma 2.6 (Chernoff bound) $z(n) \le e^{-nI}$.

Proof By Wald's fundamental identity,

$$
\begin{aligned}
z(n) &= \mathbb{P}(A(n)) \;=\; \mathbb{E}_\theta\left[L_{n;\theta};\, A(n)\right] \\
&= \mathbb{E}_\theta\left[e^{-\theta S_n + n\kappa(\theta)};\, S_n > n(\mu+\epsilon)\right] \\
&= e^{-nI}\mathbb{E}_\theta\left[e^{-\theta(S_n - n(\mu+\epsilon))};\, S_n > n(\mu+\epsilon)\right] \\
&\le e^{-nI}\,.
\end{aligned} \tag{2.9}
$$

□

Lemma 2.7 $\mathbf{Var}_\theta\, Z(n) \;\le\; e^{-2nI}$.

Proof As in (2.9),

$$
\begin{aligned}
\mathbb{E}_\theta Z(n)^2 &= \mathbb{E}_\theta\left[e^{-2\theta S_n + 2n\kappa(\theta)};\, S_n > n(\mu+\epsilon)\right] \\
&= e^{-2nI}\mathbb{E}_\theta\left[e^{-2\theta(S_n - n(\mu+\epsilon))};\, S_n > n(\mu+\epsilon)\right] \\
&\le e^{-2nI}\,.
\end{aligned}
$$

□

Lemma 2.8 $\liminf_{n\to\infty}\left(e^{nI+\theta\sqrt{n}}\right) z(n) > 0$.

Proof Since

$$
\frac{S_n - n(\mu+\epsilon)}{\sqrt{n}} \;\to\; \mathcal{N}\left(0, \sigma_\theta^2\right)
$$

in $\mathbb{P}_\theta$–distribution ($\sigma_\theta^2 = \kappa''(\theta) > 0$), we have

$$
\liminf_{n\to\infty} \mathbb{P}_\theta\left(\frac{S_n - n(\mu+\epsilon)}{\sqrt{n}} \in (0,1)\right) \;=\; \Phi\left(\frac{1}{\sigma_\theta}\right) - \Phi(0) \;:= c \;>\; 0\,.
$$

Hence

$$
\begin{aligned}
&\liminf_{n\to\infty}\left(e^{nI+\theta\sqrt{n}}\right) z(n) \\
&\ge \liminf_{n\to\infty} e^{nI+\theta\sqrt{n}} e^{-\eta n}\mathbb{E}_\theta\left[e^{-\theta(S_n - n(\mu+\epsilon))};\, \frac{S_n - n(\mu+\epsilon)}{\sqrt{n}} \in (0,1)\right] \\
&\ge \liminf_{n\to\infty} e^{\theta\sqrt{n}} e^{-\theta\sqrt{n}}\mathbb{P}_\theta\left(\frac{S_n - n(\mu+\epsilon)}{\sqrt{n}} \in (0,1)\right) \\
&= c \;>\; 0\,.
\end{aligned}
$$

□

The first part of Theorem 2.5 now follows by combining Lemmas 2.7, 2.8. The second (uniquenes of the importance sampling distribution) can be proved by similar arguments as for Siegmund's algorithm. See also Bucklew, Ney & Sadowsky [33]. □

Remark 2.9 As sharpening of Lemmas 2.6, 2.7, one can prove as one of the basic results in the theory of saddlepoint approximations that subject to some smoothness assumptions,

$$z(n) \sim \frac{e^{-nI}}{\theta\kappa''(\theta)\sqrt{2\pi n}},$$

see e.g. Petrov [119] and Jensen [84] for this and sharper versions. □

2c An efficient algorithm for heavy–tailed distributions

The problem is to simulate $z(x) = \mathbb{P}(S_n > x)$ where $S_n = X_1 + \cdots + X_n$ with $X_1, X_2, \ldots$ i.i.d. with common distribution F with a heavy tail. Here F is concentrated on $(0, \infty)$ and satisfies $\overline{F}(x) = L(x)/x^\alpha$ where $L(x)$ is slowly varying: $L(tx)/L(x) \to 1$, $x \to \infty$ (e.g. $L(x)$ bounded with a limit in $(0, \infty)$, $L(x) = (\log x)^\beta$ or $(\log\log x)^\beta$, $-\infty < \beta < \infty$), and x is large so that $z(x)$ is small. It is well known and not hard to prove, cf. e.g. Feller [50], Bingham, Goldie & Teugels [25] or Embrechts, Klüppelberg & Mikosch [48] that

$$z(x) \sim nL(x)/x^\alpha, \quad x \to \infty, \tag{2.10}$$

(but the approximation requires x to be very large to be precise so that simulation may be required).

Note that $\hat{F}[s] = \infty$ for all $s > 0$ (otherwise, we could apply the algorithm related to the Chernoff bound, at least if in addition n is large as well). Here we think of n as fixed and consider the limit $x \to \infty$ rather than $n \to \infty$.

The CMC estimator is $Z_1(x) = I(S_n > x)$. By (2.10), its variance $z(x)(1 - z(x))$ is of the order of magnitude $\overline{F}(x)$. We shall consider two algorithms based upon a conditional Monte Carlo idea.

The first and obvious idea (cf. Example II.4.5) is to condition upon $X_1, \ldots, X_{n-1}$ which leads to

$$Z_2(x) = \mathbb{P}(S_n > x \mid X_1, \ldots, X_{n-1}) = \overline{F}(x - S_{n-1}).$$

Thus, we generate only $X_1, \ldots, X_{n-1}$. As a conditional Monte Carlo estimator, $Z_2(x)$ has a smaller variance than $Z_1(x)$. However, asymptotically it presents no improvement: the variance is of the same order of magnitude $\overline{F}(x)$. To see this, just note that

$$\mathbb{E}Z_2(x)^2 \;\geq\; \mathbb{E}[\overline{F}(x - S_{n-1}); X_1 > x] \;=\; \mathbb{P}(X_1 > x) \;=\; \overline{F}(x)$$

(here we used that by positivity of the X_i, $S_{n-1} > x$ when $X_1 > x$, and that $\overline{F}(y) = 1$, $y < 0$).

The reason that this algorithm does not work well is that the probability of one single X_i to become large is too big. We avoid this problem by discarding the largest X_i and considering only the remaining ones. For the simulation, we thus generate $X_1, \ldots, X_n$, form the order statistics

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}\,,$$

throw away the largest one $X_{(n)}$, and let

$$Z_3(x) \;=\; \mathbb{P}(S_n > x \mid X_{(1)}, X_{(2)}, \ldots, X_{(n-1)}) \;=\; \frac{\overline{F}((x - S_{(n-1)}) \vee X_{(n-1)})}{\overline{F}(X_{(n-1)})}\,,$$

where $S_{(n-1)} = X_{(1)} + X_{(2)} + \cdots + X_{(n-1)}$. To check the formula for the conditional probability, note first that

$$\mathbb{P}(X_{(n)} > x | X_{(1)}, X_{(2)}, \ldots, X_{(n-1)}) = \frac{\overline{F}(X_{(n-1)} \vee x)}{\overline{F}(X_{(n-1)})},$$

We then get

$$\begin{aligned}
&\mathbb{P}(S_n > x | X_{(1)}, X_{(2)}, \ldots, X_{(n-1)}) \\
&\quad= \mathbb{P}(X_{(n)} + S_{(n-1)} > x | X_{(1)}, X_{(2)}, \ldots, X_{(n-1)}) \\
&\quad= \mathbb{P}(X_{(n)} > x - S_{(n-1)} | X_{(1)}, X_{(2)}, \ldots, X_{(n-1)}) \\
&\quad= \frac{\overline{F}((x - S_{(n-1)}) \vee X_{(n-1)})}{\overline{F}(X_{(n-1)})}\,.
\end{aligned}$$

Theorem 2.10 *Assume that $\overline{F}(x) = L(x)/x^\alpha$ ($\alpha > 1$) with $L(x)$ slowly varying. Then the algorithm given by $\{Z_3(x)\}$ is logarithmically efficient.*

The key step in the proof of Theorem 2.10 is the following estimate:

Lemma 2.11

$$\mathbb{E}\left[Z_3(x)^2\right] \;\le\; n\,(n-1)\;\left[\frac{1}{2}\overline{F}^2\left(\frac{x}{2}\right) - \overline{F}^2(\frac{x}{n})\log\overline{F}\left(\frac{x}{2}\right)\right] \tag{2.11}$$

Proof We first recall (e.g. Gut [74] p. 106) that the density $f_{X_{(n-1)}}(y)$ of the r.v. $X_{(n-1)}$ is

$$f_{X_{(n-1)}}(y) \;=\; n\,(n-1)\;F^{n-2}(y)\;\overline{F}(y)\;f(y).$$

We then get

$$\begin{aligned}
\mathbb{E}\left[Z_3(x)^2\right] \;&=\; \mathbb{E}\left[\frac{\overline{F}(x-S_{(n-1)})}{\overline{F}(X_{(n-1)})};\; X_{(n-1)} \le \frac{x}{n}\right]^2 &(2.12)\\
&+\mathbb{E}\left[\frac{\overline{F}((x-S_{(n-1)})\vee X_{(n-1)})}{\overline{F}(X_{(n-1)})};\; \frac{x}{n} < X_{(n-1)} \le \frac{x}{2}\right]^2 &(2.13)\\
&+\mathbb{E}\left[1; X_{(n-1)} > \frac{x}{2}\right]^2. &(2.14)
\end{aligned}$$

The first summand (2.12) can be bounded as follows. If $X_{(n-1)} \le \frac{x}{n}$ then $\overline{F}(x-S_{(n-1)}) \le \overline{F}(\frac{x}{n})$, so that

$$\begin{aligned}
&\mathbb{E}\left[\frac{\overline{F}(x-S_{(n-1)})}{\overline{F}(X_{(n-1)})};\; X_{(n-1)} \le \frac{x}{n}\right]^2\\
&\quad\le\; \overline{F}^2\left(\frac{x}{n}\right)\int_0^{x/n}\frac{f_{X_{(n-1)}}(y)}{\overline{F}^2(y)}dy\\
&\quad\le\; n\,(n-1)\,\overline{F}^2\left(\frac{x}{n}\right)\int_0^{x/n}\frac{f(y)}{\overline{F}(y)}dy\\
&\quad=\; -n\,(n-1)\,\overline{F}^2\left(\frac{x}{n}\right)\,\log\overline{F}\left(\frac{x}{n}\right).
\end{aligned}$$

The second summand (2.13) can be bounded in the same way. For $\frac{x}{n} <$

$X_{(n-1)} \leq \frac{x}{2}$, $\overline{F}((x - S_{(n-1)}) \vee X_{(n-1)}) \leq \overline{F}\left(\frac{x}{n}\right)$, yielding

$$
\begin{aligned}
\mathbb{E}&\left[\frac{\overline{F}((x - S_{(n-1)}) \vee X_{(n-1)})}{\overline{F}(X_{(n-1)})};\ \frac{x}{n} < X_{(n-1)} \leq \frac{x}{2}\right]^2 \\
&\leq\ \overline{F}^2\left(\frac{x}{n}\right) \int_{x/n}^{x/2} \frac{f_{X_{(n-1)}}(y)}{\overline{F}^2(y)} dy \\
&\leq\ n\,(n-1)\,\overline{F}^2\left(\frac{x}{n}\right) \int_{x/n}^{x/2} \frac{f(y)}{\overline{F}(y)} dy \\
&=\ -n\,(n-1)\,\overline{F}^2\left(\frac{x}{n}\right)\ \left[\log \overline{F}\left(\frac{x}{2}\right) - \log \overline{F}\left(\frac{x}{n}\right)\right].
\end{aligned}
$$

To find an upper bound for (2.14) we write

$$
\begin{aligned}
\mathbb{E}\left[1; X_{(n-1)} > \frac{x}{2}\right]^2 &=\ \int_{x/2}^{\infty} f_{X_{(n-1)}}(y)dy \\
&=\ n\,(n-1) \int_{x/2}^{\infty} F^{n-2}(y)\ \overline{F}(y)\ f(y)dy \\
&\leq\ n\,(n-1) \int_{x/2}^{\infty} \overline{F}(y)\ f(y)dx \\
&=\ n\,(n-1)\,\frac{1}{2}\overline{F}^2\left(\frac{x}{2}\right).
\end{aligned}
$$

Adding the above inequalities leads to the desired result. □

Proof of Theorem 2.10. It follows from (2.10) and $x^{\epsilon}L(x) \to \infty$, $x^{-\epsilon}L(x) \to 0$, $x \to \infty$, for any $\epsilon > 0$ that $|\log z(x)| \sim \alpha \log x$. Since $L(x/d) \approx L(x)$, we have $\overline{F}(x/d) \approx d^{\alpha}\overline{F}$, and hence Lemma 2.11 yields

$$
\begin{aligned}
|\log Var\, Z_3(x)| &=\ -\log \mathbf{Var}\, Z_3(x)\ \geq\ -\log \mathbb{E}\left[Z_3(x)^2\right] \\
&\sim\ -\log \overline{F}(x)^2\ \sim\ 2\alpha \log x
\end{aligned}
$$

□

Theorem 2.10 is from Asmussen & Binswanger [12]. The whole area of rare events simulation for heavy–tailed distributions is largely open. Asmussen, Binswanger & Højgaard [13] have one more working algorithm involving importance sampling but also a number of counterexamples showing that the main ideas from the light–tailed case do not carry over.

The current interest in heavy–tailed distributions is considerable. Their relevance is argued strongly in Embrechts, Klüppelberg & Mikosch [48] in the setting of insurance– and finance problems, and by a number of authors in telecommunication problems, see e.g. Resnick [124] and references there.

3 Conditioned limit theorems

The optimal change of measure for the IS is the conditional distribution $\mathbb{P}^{(x)}(\cdot) = \mathbb{P}(\cdot|A(x))$ given $A(x)$. Therefore an obvious way to look for a good IS distribution is to try to find a simple asymptotic description of $\mathbb{P}^{(x)}(\cdot)$ and to simulate using this asymptotic description.

For the random walk setting in Section 2a where $A(x) = \{\tau(x) < \infty\}$, it turns out that an asymptotic description of $\mathbb{P}^{(x)}(\cdot)$ is available. The results state roughly that up to $\tau(x)$, the random walk behaves as if it changed increment distribution from F to F_γ, which is precisely the type of behaviour needed to infer (at least heuristically) the optimality of γ. A variety of precise statements supporting this informal description were given by Asmussen [4]. For example:

Proposition 3.1 *Let $\{B(x)\}$ be any sequence of events with $B(x) \in \mathcal{F}_{\tau(x)}$, $B(x) \subseteq \{\tau(x) < \infty\}$, $\mathbb{P}_\gamma(B(x)) \to 1$, $x \to \infty$. Then $\mathbb{P}^{(x)}(B(x)) \to 1$ as well.*

Proof From Wald's fundamental identity, we get

$$\begin{aligned}\mathbb{P}^{(x)}(B^c(x)) &= \frac{\mathbb{P}(B^c(x);\,\tau(x)<\infty)}{\mathbb{P}(\tau(x)<\infty)} = \frac{\mathbb{E}_\gamma(L_{\tau(x);\gamma};\,B^c(x))}{\mathbb{P}(\tau(x)<\infty)}\\ &\le \frac{e^{-\gamma x}\mathbb{P}_\gamma(B^c(x))}{\mathbb{P}(\tau(x)<\infty)} \sim \frac{\mathbb{P}_\gamma(B^c(x))}{C} \to 0.\end{aligned}$$

□

As a main example, consider the one–dimensional empirical distribution of the X_i. Define

$$\hat{F}^{(n)}(y) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le y).$$

Corollary 3.2 *As $x \to \infty$, $\mathbb{P}^{(x)}\left(\left\|\hat{F}^{(\tau(x))} - F_\gamma\right\| > \epsilon\right) \to 0$, where $\|\cdot\|$ denote the supremum norm.*

Proof By the Glivenko–Cantelli theorem, $\left\|\hat{F}^{(n)} - F_\gamma\right\| \to 0$ $\mathbb{P}_\gamma$–a.s. as $n \to \infty$. Hence also $\|\hat{F}^{(\tau(x))} - F_\gamma\| \to 0$, and we can take

$$B(x) = \left\{\|\hat{F}^{(\tau(x))} - F_\gamma\| > \epsilon\right\}.$$

□

The results of [4] are in fact somewhat more general by allowing inference also on the dependency structure in the conditional limit. For example, it is straightforward to show that

$$\frac{1}{\tau(x)}\sum_{i=1}^{\tau(x)} I(X_i \le y_1,\dots,X_{i+k-1}\le y_k) \;\to\; F_\gamma(y_1)\dots F_\gamma(y_k)$$

in $\mathbb{P}(\cdot\,|\,\tau(x)<\infty)$–probability. Perhaps, the most convincing indication that the X_i are asymptotically conditionally independent is the fact that variance constants coming up in conditional approximations by Brownian motion and Brownian bridge are the same as in the unconditional F_γ–random walk. See [4] for more detail.

A counterexample

It is important to point out that examples have started to show up in the literature which clearly show that the idea of simulating using an asymptotic description of $\mathbb{P}^{(x)}(\cdot)$ has its limitations. We give one of them, taken from Glasserman & Wang [61].

As in Section 2, we consider exceedances in the LLN but this time two–sided,

$$A(n) \;=\; \{S_n > n\epsilon \text{ or } S_n < -n\epsilon'\}$$

(taking $\mu=0$). We choose ϵ' such that

$$\frac{\mathbb{P}(S_n>n\epsilon)}{\mathbb{P}(S_n<-n\epsilon')} \;\to\; \infty \tag{3.1}$$

so that

$$\mathbb{P}^{(n)}(\cdot) \;\sim\; \mathbb{P}(\cdot\,|\,S_n>n\epsilon)$$

Thus, it is suggested to use the same exponential change of measure as in Section 2b and

$$Z(n) \;=\; e^{-\theta S_n+n\kappa(\theta)}I(S_n>n\epsilon \text{ or } S_n<-n\epsilon').$$

However, we will see then that the contribution to $\mathbf{Var}_\theta Z(n)$ from the event $\{S_n<-n\epsilon'\}$ blows up the variance.

Fix ϵ, and define θ,θ' as the solutions of $\kappa'(\theta)=\epsilon$, resp. $\kappa'(\theta')=-\epsilon'$ and let $I=\theta\epsilon-\kappa(\theta)$, $I'=-\theta'\epsilon'-\kappa(\theta')$. We choose ϵ' such that $I'>I$ (then (3.1) holds by Lemmas 2.6, 2.7) and

$$\delta' \;=\; \theta(\epsilon+\epsilon')+I-I' \;>\; 0 \tag{3.2}$$

which can be otained by first choosing ϵ' such that $I' = I$ and next replacing ϵ' by a slightly larger value to get $I' > I$ without violating (3.2).

Proposition 3.3 *If* $0 < \delta < \delta'$, *then* $\liminf_{n\to\infty} \frac{\mathbf{Var}_\theta Z(n)}{z(n)^2 e^{\delta n}} = \infty$.

Proof

$$
\begin{aligned}
\mathbb{E}_\theta Z(n)^2 &\geq \mathbb{E}_\theta \left[L_{n,\theta}^2;\, S_n < -n\epsilon'\right] = \mathbb{E}\left[L_{n,\theta};\, S_n < -n\epsilon'\right] \\
&= \mathbb{E}\left[e^{-\theta S_n + n\kappa(\theta)};\, S_n < -n\epsilon'\right] \geq e^{n\theta\epsilon' + n\kappa(\theta)}\mathbb{P}(S_n < -n\epsilon') \\
&\geq c_1 e^{n\theta\epsilon' + n\kappa(\theta)} e^{-n\theta' - \theta'\sqrt{n}} \\
&= c_1 \exp\left\{n\left[\theta(\epsilon + \epsilon') - I - I'\right] - \theta'\sqrt{n}\right\}
\end{aligned}
$$

using Lemma 2.7 for the last inequality. Hence by Lemma 2.6

$$
\begin{aligned}
&\liminf_{n\to\infty} \frac{\mathbf{Var}_\theta Z(n)}{z(n)^2 e^{\delta n}} \\
&\geq c_1 \liminf_{n\to\infty} \exp\left\{n\left[\theta(\epsilon + \epsilon') - I - I'\right] - \theta'\sqrt{n} + 2nI - n\delta\right\} \\
&= c_1 \liminf_{n\to\infty} \exp\left\{n\left[\delta' - \delta\right] - \theta'\sqrt{n}\right\} = \infty.
\end{aligned}
$$

□

4 Large deviations

The LD approach to OECM (optimal exponential change of measure) has several variants. We give one involving the concept of *optimal path* which may be seen as an alternative approach to conditioned limit theorems.

We will work in the i.i.d. setting of the Siegmund algorithm and write $\kappa(\theta) = \log \hat{F}[\theta]$. We first introduce the function

$$
I(y) = \sup_{\theta\in\Theta}(\theta y - \kappa(\theta)), \quad y \in \mathcal{Y} = \{\kappa'(\theta) : \theta \in \Theta\},
$$

which in the literature goes under names like the *LD rate function*, the *Legendre transform*, the *Legendre–Fenchel transform*, the *Cramér transform* etc. (sometimes the sign is reversed). For simplicity, we assume that the interval Θ is open. For $y \in \mathcal{Y}$, we define $\theta(y)$ (the saddlepoint of y) by $\kappa'(\theta(y)) = y$ so that $I(y) = \theta(y)y - \kappa(\theta(y))$. Note that we have already encountered $I(y)$ once, in Section 2b, where $I = I(\mu + \epsilon)$.

It is not to hard to show that I is non–negative, convex and attains its minimum 0 for $y = \kappa'(0) = \mu$. The crucial fact for OECM is

Lemma 4.1 $\min\limits_{0<y<\infty} \dfrac{I(y)}{y}$ *is attained for* $y^* = \kappa'(\gamma)$.

Proof Obviously, $I(y)/y \to \infty$ as $y \downarrow 0$ and (cf. (4.3) below) $I(y)/y$ is non–decreasing for large y so that the minimum is attained. By straightforward differentiation, we get

$$\begin{aligned} I'(y) &= \theta'(y)y + \theta(y) - \theta'(y)\kappa'(\theta(y)) \;=\; \theta(y), \\ \frac{d}{dy}\frac{I(y)}{y} &= \frac{yI'(y) - I(y)}{y^2} \;=\; \frac{y\theta(y) - \theta(y)y + \kappa(\theta(y))}{y^2} \\ &= \frac{\kappa(\theta(y))}{y^2}. \end{aligned} \tag{4.3}$$

Putting the last expression equal to 0 yields $\theta(y^*) = \gamma$ (since we look at minimum values for $y > 0$ only, $\theta(y^*) = 0$ is excluded) from which we immediately get $y^* = \kappa'(\gamma)$. □

One of the main themes of LD theory is to give estimates of the probability that the random walk (or some more general process) follows an atypical path. In the random walk setting, this means that $S^{(n)}(\cdot)$ follows a path different from the one $\varphi_0(t) = \mu t$ given by the LLN where $S^{(n)}(t) = S_{\lfloor nt \rfloor}/n$, $0 \le t \le 1$ (here $\lfloor \cdot \rfloor$ = integer part). The LD results state that under appropriate regularity conditions,

$$\mathbb{P}(S^{(n)}(\cdot) \in \mathcal{S}) \;\sim\; \exp\left\{-n \inf_{\varphi \in \mathcal{S}} \int_0^1 I\left(\varphi'(t)\right) dt\right\} \tag{4.4}$$

for suitable subsets $\mathcal{S}$ of continuous paths with $\varphi_0 \notin \mathcal{S}$. In many examples, there is a single path φ^* for which the minimum is attained, and this is the *optimal path.*

In order to understand how the random walk reaches the high level x, we perform the optimization not only over φ but also over n. We then write x in the form $x = ny$ and let $\mathcal{S}$ be the set of continuous functions on $[0,1]$ with $\varphi(0) = 0$, $\varphi(1) = y$. Then (4.4) takes the form

$$\mathbb{P}(S_n \sim x) \;\sim\; \exp\left\{-x\frac{1}{y} \inf_{\varphi \in \mathcal{S}} \int_0^1 I\left(\varphi'(t)\right) dt\right\} \tag{4.5}$$

By Jensen's inequality and the convexity of I,

$$\int_0^1 I\left(\varphi'(t)\right) dt \;\ge\; I\left(\int_0^1 \varphi'(t)dt\right) \;=\; I(y),$$

with equality if and only $\varphi(t) = ty$. Hence for fixed n,

$$\mathbb{P}(S_n \sim x) \;\sim\; \exp\left\{-x\frac{I(y)}{y}\right\}.$$

Viewing x as fixed, taking into the account that minimizing over n is the same as minimizing over y, we obtain by Lemma 4.1 that the minimizer is $y^* = \kappa'(\gamma)$. In conclusion, if x is large, the most likely way in which the random walk can cross level x is by crossing at time

$$\tau(x) \;=\; n \;=\; \frac{x}{y} \;=\; \frac{x}{\kappa'(\gamma)}$$

and by moving linearly at rate $y = \kappa'(\gamma)$ up to that time. But *this is precisely the same way as in which the random walk with increment distribution F_γ crosses level x*, which motivates that the conditional distribution given the rare event $\tau(x) < \infty$ is the one of the random walk with increment distribution F_γ.

Notice that the LD argument is somewhat more heuristic than the ones in the preceding subsections (an obvious gap is that it identifies the optimal IS only in terms of its mean). Also, in simple settings LD results are typically not the strongest possible, involving only logarithmic asymptotics.

However, the motivation for the LD approach is its generality and the fact that the mathematical state of the area is very advanced, providing a considerable body of theory to draw upon. The philosophy is that once it has been understood how to the paraphraze the optimality properties of OECM for simple systems in LD language, the generality of LD theory will allow to find suitable IS distributions also in more complicated settings.

At least in the present authors' opinion, the success in implementing this program has been slightly more moderate than sometimes claimed. As we see it:

1. The alternative approaches to OECM (in the sense of how to derive it and how to study its optimality properties) in simple models are both simpler, more precise and lead to stronger conclusions. Actually, one can argue from many specific cases that the use of LD's theory say for the GI/G/1 queue most often shoots over the goal.

2. When getting to more complex (and thereby also more realistic and practically challenging!) models, LD theory leads into variational problems which do not have an explicit solution. Thus, LD theory allows to identify explicitly the OECM for simple models *only*..

3. Counterexamples like the one in Section 3 are worrying and seem to indicate that care is needed when applying the LD approach beyond simple problems.

Despite of these remarks, the LD approach is often the natural one and the only feasible one.

Some relevant references on the LD point of view in simulation are Cottrell *et al.* [38], Bucklew *et al.* [33], Sadowsky & Bucklew [140], Sadowsky [136] - [138] and Lehtonen & Nyrhinen [97], [98]. For LD theory in general, we refer to Bucklew [32], Dembo & Zeitouni [43] and Shwartz & Weiss [146].

5 Multilevel splitting

Splitting is a set of ideas for estimating the probability $z(x)$ that a Markov chain $\{X_n\}$ hits a rare set $B(x)$ in a regenerative cycle. There are many variants of the method around. Some main early references are Kahn & Harris [85], Hammersley & Handscomb [75] and Bayes [23], and some more recent ones Villén–Altamirano [152], [153] (RESTART), Glasserman *et al* [58], [59], [60] and Haraszti & Townsend [76].

The method uses a decomposition of the state space E into subsets $E_0, \ldots, E_m$ with $E_m = B(x)$, see Fig. 5.1.

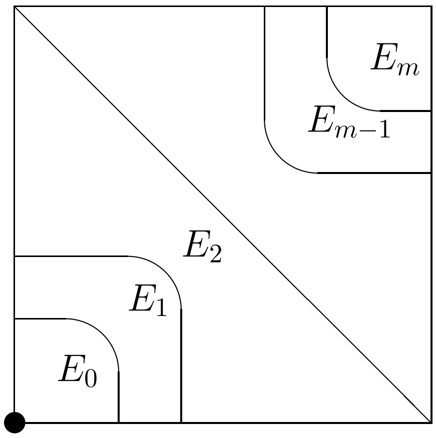

Figure IV.5.1

Think of $\mathbf{0} = (0,0)$ as the regeneration point ($\mathbf{0} \in E_0$) and that the event of hitting E_m occurs through succesive hits of the E_k. That is, define

$$\tau_k \;=\; \inf\{n:\, X_n \in E_k \cup E_{k+1} \cup \ldots \cup E_m\}\;.$$

We then assume

$$\mathbb{P}_i(X_{\tau_k} \in E_k \mid \tau_k < \infty) \;=\; 1, \quad i \in E_{k-1}, \tag{5.1}$$

which implies

$$z(x) \;=\; p_1 p_2 \dots p_m \tag{5.2}$$

where $z(x) = \mathbb{P}_{\mathbf{0}}(\tau_m < C)$ with $C = \inf\{n > 0 : \; X_n = \mathbf{0}\}$ the regenerative cycle and $p_k = \mathbb{P}_{\mathbf{0}}(\tau_k < \infty \mid \tau_{k-1} < \infty)$.

For the simulation, we generate n_1 sample paths $\left\{X_n^{(i)}\right\}$ starting from $X_0 = \mathbf{0}$; out of these, $N_1 = \sum_1^{n_1} I\left(\tau_1^{(i)} < C^{(i)}\right)$ will hit E_1 before returning to $\mathbf{0}$, and we have the obvious estimate $\hat{p}_1 = N_1/n_1$. For the N_1 successes, we let ν_1 denote the empirical distribution of the entrance state in E_1,

$$\nu_1(j) \;=\; \frac{1}{N_1}\sum_{i=1}^{n_1} I\left(X_{\tau_1^{(i)}} = j,\, \tau_1^{(i)} < C^{(i)}\right).$$

We then generate n_1 copies of $\{X_n\}$ with initial distribution ν_1, and let N_2 be the number of copies entering E_2 before returning to $\mathbf{0}$, ν_2 the corresponding empirical distribution of the successful hits of E_2, $\hat{p}_2 = N_2/n_2$, and so on. The estimator is $\hat{z} = \hat{z}(x) = \hat{p}_1 \dots \hat{p}_m$.

The generation of a copy of $\{X_n\}$ with initial distribution ν_1 (and similarly in the following steps) can be performed in more than one way. One approach is to choose the initial value X_0 by randomly sampling from the succesful hits of E_1. Another is to use each such hit, say at $x \in E_1$, to generate a new number n_1' of paths starting from x. Thus $n_2 = n_1' N_1$; from this, the term *splitting*. Irrespective of the choice:

Proposition 5.1 *$\hat{z}$ is an unbiased estimator of z.*

Proof Let $\mathcal{F} = \sigma\left(X_n^{(i)} : i = 1, \dots, n_1,\, n = 0, \dots, \tau_1^{(i)}\right)$, $q_2(y) = \mathbb{P}_y(\tau_2 < C)$, $y \in E_1$. Then for $m = 2$,

$$\begin{aligned}
\mathbb{E}[\hat{p}_2|\mathcal{F}] \;&=\; \int_{E_1} q_2(y)\,\nu_1(dy) \;=\; \frac{1}{N_1}\sum_{i=1}^{n_1} I\left(\tau_1^{(i)} < C^{(i)}\right) q_2\left(X_{\tau_1^{(i)}}\right), \\
\mathbb{E}\hat{z} \;&=\; \mathbb{E}\,\hat{p}_1\hat{p}_2 \;=\; \mathbb{E}\left[\hat{p}_1 \mathbb{E}[\hat{p}_2|\mathcal{F}]\right] \\
&=\; \frac{1}{n_1}\mathbb{E}\sum_{i=1}^{n_1} I\left(\tau_1^{(i)} < C^{(i)}\right) q_2\left(X_{\tau_1^{(i)}}\right) \;=\; p_1 p_2 \;=\; z.
\end{aligned}$$

The case $m > 2$ follows similarly by induction. $\square$

With the right choice of m and the E_k, the splitting algorithm can be highly efficient. We illustrate this via a simple example, a birth–death chain on $E = \{0, \dots, x\}$ with transition matrix

$$\begin{pmatrix} b_0 & a_0 & 0 & 0 & \dots & 0 & 0 & 0 \\ b_1 & 0 & a_1 & 0 & & 0 & 0 & 0 \\ 0 & b_2 & 0 & a_2 & & 0 & 0 & 0 \\ \vdots & & & & \ddots & & & \vdots \\ 0 & 0 & 0 & 0 & & b_{x-1} & 0 & a_{x-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & b_x & a_x \end{pmatrix}$$

$(a_i + b_i = 1)$. The rare set is $B(x) = x$ and thus

$$x \;=\; z(x) \;=\; \mathbb{P}_0(X_n = x \text{ for some } n < C).$$

The level sets are $E_k = \{x_k, x_{k+1}, \dots, x_{k+1} - 1\}$ where $0 = x_0 < x_1 < \dots < x_{m-1} < x_m = x$.

Proposition 5.2 *For a fixed imulation budget*

$$n \;=\; n_1 + \cdots + n_k, \tag{5.3}$$

Var$\hat{z}$ *is asymptotically minimized by taking*

$$p_k \sim e^{-2} \approx 0.135, \quad m \sim -\log z/2.$$

For this choice,

$$\mathbf{Var}\hat{z} \;\sim\; \frac{(ez\log z)^2}{4n}.$$

In particular, $\hat{z}$ is logarithmically efficient.

Proof We follow Garvels & Kroese [54] (see also Villén–Altamirano [153]). We consider the fixed effort variant where n_k is non–random. Note that ν_k is concentrated at x_k and thus the $\hat{p}_k$ are simply obtained by independent binomial sampling so that

$$\begin{aligned} \mathbb{E}\hat{z}^2 \;&=\; \prod_{k=1}^m \mathbb{E}\hat{p}_k^2 \;=\; \prod_{k=1}^m \left\{ \frac{p_k(1-p_k)}{n_k} + p_k^2 \right\} \\ &=\; z^2 \prod_{k=1}^m \left\{ \frac{b_k^2}{n_k} + 1 \right\} \end{aligned} \tag{5.4}$$

where $b_k^2 = (1-p_k)/p_k$.

Consider first the miniaztion of (5.4) subject to (5.3). If the n_k are large, this means that we must minimize $\sum_1^m b_k^2/n_k$. Using the principle of Lagrange multipliers[1] and treating the n_k as continuous variables, we get

$$0 \;=\; -\frac{b_k^2}{n_k^2} + K \;=\; -\frac{b_\ell^2}{n_\ell^2} + K, \quad \frac{n_k}{n_\ell} = \frac{b_k}{b_\ell}, \quad n_k = n\frac{b_k}{s_b}$$

where $s_b = b_1 + \cdots + b_m$. Then

$$\mathbb{E}\hat{z}^2 \;\sim\; z^2\sum_1^m \frac{b_k^2}{n_k} \;=\; z^2\sum_1^m b_k\frac{s_b}{n} \;=\; \frac{1}{n}z^2 s_b^2.$$

Thus, the optimal partitioning $\{x_k\}$ is obtained by minimizing

$$s_b \;=\; \sum_{k=1}^m \sqrt{\frac{1-p_k}{p_k}}$$

subject to $p_1 \ldots p_m = z$. Another use of Lagrange multipliers yields

$$0 \;=\; \frac{1}{2p_k\sqrt{(1-p_k)/p_k}} + K\frac{z}{p_k}, \quad \sqrt{(1-p_k)/p_k} = \sqrt{(1-p_\ell)/p_\ell},$$

so that the p_k must be equal, $p_k = z^{1/m}$. Then

$$\mathbb{E}\hat{z}^2 \;\sim\; \frac{1}{n}z^2\left(\sum_{k=1}^m \sqrt{\frac{1-z^{1/m}}{z^{1/m}}}\right)^2 \;=\; \frac{1}{n}z^2\frac{m^2(1-z^{1/m})}{z^{1/m}}.$$

This is minimized by taking $m = -\log z/2$, and we then get $p_k = z^{1/m} = e^{-2}$,

$$\mathbb{E}\hat{z}^2 \;\sim\; \frac{(ez\log z)^2}{4n}.$$

$\square$

In practice, one can at best hope to get close to logarithmic efficiency. First, because the p_k are not known given the x_k. Second, because the p_k are not continuous variables since the x_k are not so. Thus, one should think of Proposition 5.2 as a guideline only.

[1] stating that to minimize $f(y_1,\ldots,y_m)$ subject to $g(y_1,\ldots,y_m)=0$, we must look for a K satisfying $h_k(y_1,\ldots,y_m)=0$, $Kg(y_1,\ldots,y_m)=0$ where $h=f+Kg$ and h_k is the kth partial derivative.

6 Reliability*

Not implemented in this version. Some references are Anantharam *et al.* [3], Chang *et al.* [36], Goyal *et al.* [72], Heidelberger [77], Heidelberger, Shahabuddin & Nicola [78], Nakayama [107], [108] and Shahabuddin [144].

Chapter V

Markov chain Monte Carlo methods*

Not implemented in this version. A good reference is Gilks, Richardson & Spiegelhalter [55].

Chapter VI

Gradient estimation

We consider a stochastic system and assume (highly simplified) that its performance can be summarized as a single real number z, which typically can be expressed as $\mathbb{E}Z$ for some r.v. Z. For example, in a PERT net Z can be the (random) length of the maximal path, in a data network Z can be the steady–state delay of a packet or z the probability that a packet is transmitted without errors, in an insurance risk model z can be the probability $\mathbb{P}(C > x)$ that the claims C within one year exceeds a large value x (typically, $C = \sum_1^N U_i$ where N is the number of claims and $U_1, U_2, \ldots$ the claim sizes) and so on.

In all these examples, z will typically depend on a number of parameters (for example in the insurance risk model, N could be Poisson with rate parameter β and the claims having a distribution B_θ depending on a parameter θ). One could then be interested not only in the performance z but also in its derivatives

$$z_\beta = \frac{\partial}{\partial \beta} z, \quad , z_\theta = \frac{\partial}{\partial \theta} z, \ \ldots$$

w.r.t. the parameters $\beta, \theta, \ldots$. Here z_β is called the *sensitivity* of z w.r.t. β (and similarly for z_θ etc.), and the vector $(z_\beta, z_\theta, \ldots)$ is the *gradient*. The problem we address in this chapter is how to estimate such sensitivities by simulation.

There are numerous reasons for being interested in the sensitivities, in particular:

1. For identifying the most important system parameters;

2. To asses the effect of a small change of a parameter;

3. To produce confidence intervals for z if some parameters are not completely known but estimated. For example, if β is a Poisson parameter

estimated as the empirical rate $\overline{\beta} = N_T/T$ of Poisson events in $[0, T]$, then $\overline{\beta}$ is asymptotically normal $N(\beta, \beta/T)$ as $T \to \infty$. So, if $z(\beta)$ is analytically available

$$z(\overline{\beta}) \pm 1.96 \frac{\overline{\beta}|z_\beta|}{\sqrt{T}}$$

is an asymptotic 95% confidence interval for $z(\beta)$. More generally, if $z(\beta)$ needs to be evaluated by simulation, the confidence interval is

$$\hat{z}(\overline{\beta}) \pm 1.96\sqrt{\frac{\overline{\beta}^2 \hat{z}_\beta^2}{T} + \frac{\hat{\sigma}^2}{n}}$$

where $\hat{z}(\overline{\beta})$ of $z(\overline{\beta})$ is a CMC estimator based upon n replications and with associated variance estimator $\hat{\sigma}^2$, and $\hat{z}_\beta$ an estimator of the sensitivity. See further Heidelberger & Towsley [79] and Rubinstein & Shapiro [134] pp. 96–100.

4. In stochastic optimization, where we want to find the maximum (or minimum) of $z = z(\theta)$ w.r.t. some parameter θ, most algorithms require knowledge of z_θ (see Chapter VII).

5. Finally there are examples where the sensitivities are of intrinsic interest, e.g. in financial mathematics.

1 Finite differences

Assume that for each θ we are in a position to generate a r.v. $Z(\theta)$ with expectation $z(\theta)$. We want a simulation estimate of $z_\theta = z'(\theta)$.

The starting point for the method of finite differences is the formulas

$$f'(\theta) = \lim_{h \downarrow 0} \frac{f(\theta + h) - f(\theta)}{h} = \lim_{h \downarrow 0} \frac{f(\theta + h/2) - f(\theta - h/2)}{h} \tag{1.1}$$

for the derivative of a deterministic function $f(\theta)$. In the context of simulation, this suggests to perform a CMC experiment with either

$$\tilde{Z}_\theta = \frac{Z(\theta + h) - Z(\theta)}{h} \quad \text{or} \quad Z_\theta = \frac{Z(\theta + h/2) - Z(\theta - h/2)}{h} \tag{1.2}$$

for some small h, where we take $Z(\theta + h), Z(\theta)$ as independent for $\tilde{Z}_\theta$ and $Z(\theta + h/2), Z(\theta - h/2)$ for Z_θ.

A first important observations is that the second formula in (1.1) is preferable for numerical differentiation because

$$\begin{aligned}\frac{f(\theta+h)-f(\theta)}{h} &= f'(\theta)+\frac{h}{2}f''(\theta)+O(h^2),\\ \frac{f(\theta+h/2)-f(\theta-h/2)}{h} &= f'(\theta)+\frac{h^2}{24}f'''(\theta)+O(h^3)\end{aligned}$$

as follows by straightforward Taylor expansions. In the simulation context, just the same calculation shows that the bias of Z_θ is an order of magnitude lower than that of $\tilde{Z}_\theta$, so that obviously Z_θ should be preferred.

The choice of h is left open. If the number of replications is n and $\hat{z}_\theta$ the corresponding average of the Z_θ, it seems reasonable that $h = h_n$ should go to zero as $n \to \infty$ to reduce bias. On the other hand, taking a smaller h increases variance so there is a trade–off. The answer is that h should be of order $n^{-1/6}$:

Proposition 1.1 *The mean square error $\mathbb{E}(\hat{z}_\theta - z_\theta)^2$ is asymptotically minimized by letting*

$$h = h_n = \frac{1}{n^{1/6}}\frac{[576\mathbf{Var}(Z(\theta)]^{1/6}}{\left|\frac{d^3}{d\theta^3}\mathbb{E}Z(\theta)\right|^{1/3}}$$

Proof Clearly, $\mathbf{Var}(Z_\theta) \sim 2\mathbf{Var}(Z(\theta))/h^2$ so

$$\begin{aligned}\mathbb{E}(\hat{z}_\theta - z_\theta)^2 &= \mathbf{Var}(\hat{z}_\theta) + (\mathbb{E}\hat{z}_\theta - z_\theta)^2\\ &\sim \frac{2}{nh^2}\mathbf{Var}(Z(\theta)) + \left(\frac{h^2}{24}\frac{d^3}{d\theta^3}\mathbb{E}Z(\theta)\right)^2\end{aligned}$$

Now the function $a/x + bx^2$ of x is minimized for $x = (a/2b)^{1/3}$. Letting

$$x = h^2,\quad a = 2\mathbf{Var}(Z(\theta))/n,\quad b = \left(\frac{d^3}{d\theta^3}\mathbb{E}Z(\theta)/24\right)^2,$$

the result follows. □

In L'Ecuyer & Perron [95], it is shown that the finite differences method performs substantially better when combined with common random numbers and that the rate of convergence then is as good as the more sophisticated method of IPA to be discussed next. Of course, a main disadvantage of the method of finite differences is its bias.

2 Infinitesimal perturbation analysis

The idea of this method is *sample path derivatives*: we write $z = z(\theta)$ as $\mathbb{E}Z(\theta)$ for some r.v. depending on θ, and estimate z_θ via

$$Y = Y(\theta) = \frac{\partial}{\partial\theta} Z(\theta),$$

evaluated at $\theta = \theta_0$ where θ_0 is the parameter of the given system. Thus we simulate n i.i.d. copies $Y_1, \ldots, Y_n$ of Y and use the estimator

$$\hat{z}_\theta = \frac{Y_1 + \cdots + Y_n}{n}, \tag{2.1}$$

with the obvious confidence interval based upon the empirical variance of the Y_i. What is needed for consistency of $\hat{z}_\theta$ is

$$\frac{\partial}{\partial\theta}\mathbb{E}Z(\theta) = \mathbb{E}\frac{\partial}{\partial\theta}Z(\theta). \tag{2.2}$$

We illustrate the approach via an example (a simplified version of what is needed for a PERT net):

Example 2.1 Assume that $Z = \max(X_1, X_2)$ where θ is a scale parameter for X_2 (thus, the given system corresponds to $\theta_0 = 1$). That is,

$$Z(\theta) = \max(X_1, \theta X_2) = \begin{cases} X_1 & X_1 > \theta X_2 \\ \theta X_2 & X_1 < \theta X_2 \end{cases}.$$

It follows that

$$Y(\theta) = \begin{cases} 0 & X_1 > \theta X_2 \\ X_2 & X_1 < \theta X_2 \end{cases},$$

so that $Y = Y(\theta_0) = X_2 I(X_1 < X_2)$. The check of (2.2) goes as follows: with F_1, F_2 the c.d.f.'s of f_1, f_2, we get

$$\begin{aligned} z(\theta) &= \mathbb{E}\max(X_1, \theta X_2) = \int_0^\infty \mathbb{P}(\max(X_1, \theta X_2) > x)\,dx \\ &= \int_0^\infty (1 - F_1(x)F_2(x/\theta))\,dx \end{aligned}$$

and by differentiation under the integral sign

$$\begin{aligned} z_\theta &= \int_0^\infty F_1(x)\,\frac{x}{\theta^2}\,f_2(x/\theta)\,dx \\ &\overset{\theta=1}{=} \int_0^\infty \mathbb{P}(X_1 < x)\,x\,f_2(x)\,dx = \mathbb{E}Y. \end{aligned}$$

In general, the assumption (2.2) appears to be rather innocent, requiring differentiation and integration to be interchangeable. However, the following examples show that one has to be careful:

Example 2.2 Consider $\max(X_1, X_2)$ as in Example 2.1 but assume now that the relevant performance is $z = \mathbb{P}(X_2 > X_1)$. Then $Z(\theta) = I(\theta X_2 > X_1)$ (see Figure 2.1), which is differentiable with derivative 0 except at $\theta = X_1/X_2$.

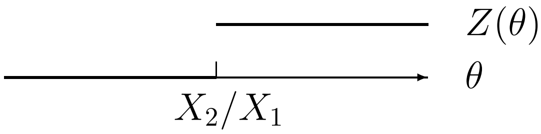

Figure VI.2.1

Since $\mathbb{P}(X_1 = X_2) = 0$, it follows that

$$Y \;=\; \frac{\partial}{\partial\theta} Z(\theta)\bigg|_{\theta=1} \;=\; 0 \quad \text{a.s.}$$

so that $0 = \mathbb{E}Y \neq z_\theta$. □

The same phenomenon occurs often when discrete r.v. s are involved (say the Poisson r.v. N in the insurance risk example). Here is a counterexample of a somewhat different type.

Example 2.3 (Shortest job first*) □

A key reference for infinitesimal perturbation analysis is Glasserman [57].

3 The likelihood ratio method

We illustrate the method via the same example as used for IPA:

Example 3.1 Assume again that $z(\theta) = \mathbb{E}\max(X_1, \theta X_2)$ and that we are interested in $\theta = \theta_0 = 1$. With f_1, f_2 the densities of X_1, X_2, we can write the density of θX_2 as

$$f(x, \theta) \;=\; \frac{1}{\theta} f_2\left(\frac{x}{\theta}\right)$$

and get

$$z(\theta) \;=\; \int\!\!\int \max(x_1,x_2)\, f_1(x) f(x_2,\theta)\, dx_1 dx_2$$

Differentiation under the integral sign yields

$$\begin{aligned} z_\theta \;&=\; \int\!\!\int \max(x_1,x_2)\, f_1(x_1) \frac{\partial}{\partial\theta} f(x_2,\theta)\, dx_1 dx_2 \\ &=\; \int\!\!\int \max(x_1,x_2)\, f_1(x_1) f_2(x,\theta) \frac{(\partial f/\partial\theta)(x_2,\theta)}{f(x_2,\theta)}\, dx_1 dx_2 \\ &=\; \mathbb{E}_\theta[ZS] \end{aligned}$$

where

$$S \;=\; S_\theta(X_2) \;=\; \frac{(\partial f/\partial\theta)(X_2,\theta)}{f(X_2,\theta)} \;=\; \frac{\partial}{\partial\theta}\log f(X_2,\theta)$$

is the *score function* familiar from statistics. □

Just the same calculation yields:

Proposition 3.2 *If $z(\theta) = \mathbb{E}_\theta[Z]$ where Z is a function of τ i.i.d. r.v.'s $X_1, X_2, \ldots$ with density $f(x,\theta)$ and τ is a constant or a stopping time, then*

$$z_\theta \;=\; \frac{d}{d\theta} z(\theta) \;=\; \mathbb{E}_\theta[ZS]$$

where

$$S \;=\; \sum_{i=1}^{\tau} \frac{d \log f(X_i,\theta)}{d\theta}.$$

Note in particular the convenient feature that the score function is additive.

Example 3.1 and Proposition 3.2 again contain an implicit condition, namely that differentation and expectation can be interchanged. I.e., we want a formula of the type

$$\frac{d}{d\theta} \int h(x) f(x,\theta)\, dx \;=\; \int h(x)\, \frac{d}{d\theta} f(x,\theta)\, dx$$

to be valid. However, this is basically a regularity condition on the density and holds in practice in much greater generality than the condition (2.2) needed for IPA to be valid.

A key reference for likelihood ratio gradient estimation is Rubinstein & Shapiro [134].

3a Rare events*

Nakayama [108], Asmussen & Rubinstein [18].

4 Examples and special methods*

Chapter VII

Stochastic optimization*

Not implemented in this version. Good references are Pflug [120] and Rubinstein & Shapiro [134].

1 The Robbins-Monro algorithm*

2 Response surfaces*

Chapter VIII

Simulation of some special processes

Let $\{X(t)\}$ be a stochastic process in discrete or continuous time. We are interested in generating a sample path $\{x(t)\}_{0\le t\le T}$ by simulation where T is a fixed number (say $T=1$) or a stopping time.

The methods that we survey in this chapter are highly dependent on the type of process in question, and also on the type of application (what is the sample path to be used for?). In some cases like Lévy– or stable processes, it may even be non–trivial or impossible to generate one –dimensional distributions (i.e., $x(T)$ for a fixed T); we are then faced with a particular problem in random variate generation. In other situations like stationary Gaussian processes, the generation of $x(T)$ for one T may be easy but the dependence structure may make it difficult to generate finite–dimensional distributions (the random vector $(x(0), x(1), \ldots, x(T))$ in discrete time, or a discrete skeleton $(x_{T/n}(kT/n))_{0\le k\le n}$ in continuous time), in particular when the order is high. A method which is suitable for a fixed T may not be suitable if T is a stopping time; say the method is based upon generating discrete skeletons by bisection, starting with generating $(x(0), x(T))$, then $(x(0), x(T/2), x(T))$, next $(x(0), x(T/4), x(T/2), x(3T/4), x(T))$ and so on. In continuous time, it may be straightforward to generate a discrete skeleton with the correct finite dimensional distributions (say in the case of Brownian motion) but using a discrete skeleton may introduce errors in the specific application, say we are interested in characteristics of the first passage time $\inf\{t>0:\ X(t)\ge x\}$.

The error criteria to be used depend on the type of application. If one is interested in just generating $x(T)$ sufficiently accurate, an appropriate error criterion for may be

$$\sup_{f\in\mathcal{H}} |\mathbb{E}f(x(T)) - \mathbb{E}f(X(T))| \tag{0.1}$$

for a suitable class $\mathcal{H}$ of smooth functions. The accuracy of the sample path approximation can be measured by criteria like

$$\mathbb{E} \sup_{0\le t\le T} |x(t) - X(t)| \tag{0.2}$$

or, for some suitably chosen p,

$$\mathbb{E} \int_0^T |x(t) - X(t)|^p \, dt, \quad \mathbb{E} \sum_{n=0}^{T} |x_n - X_n|^p \tag{0.3}$$

in continuous, resp. discrete, time, assuming there is a natural way to represent $\{x(t)\}$ and $\{X(t)\}$ on the same probability space.

For some continuous time processes such as work loads or queue lengths in queues, compound Poisson processes (with a possibly added linear drift) etc., there is an embedded sequence of points which determines the evolution of the process as a whole. For many interesting processes such as general Lévy processes or solutions to SDE's, this is not the case, and the process is then usually generated from a discrete skeleton. Sometimes it then works quite well to define $\{x(t)\}$ by linear interpolation in between grid points or by $x(t) = x(kT/n)$, $kT/n \le t < (k+1)T/n$, but in other cases such procedures may be clearly unreasonable.

1 Brownian motion

Let $\{W(t)\}_{t\ge 0}$ be standard Brownian motion (BM). The generation of the process along a discrete skeleton $0, h, 2h, \ldots$ is straightforward: just generate the increments

$$W(h) = W(h) - W(0),\ W(2h) - W(h),\ W(3h) - W(2h), \ldots$$

as i.i.d. $\mathcal{N}(0, h)$ variables.

In view of the simplicity of this procedure, there is not much literature on the simulation of Brownian motion. A notable exception is Knuth [92].

The error from linear interpolation

To generate a continuous time version of BM is intrinsically impossible because of the nature of the paths. This creates the problem that most

Brownian functionals cannot be generated exactly from neither a discrete skeleton nor from other obvious discrete events schemes.

An obvious procedure is to define $\{w_h(t)\}$ by linear interpolation in between grid points $0, h, 2h, \ldots$. Note that we can take the increments of $\{w_h(t)\}$ as the increments of $\{W(t)\}$ so assume w.l.o.g. that $w_h(kh) = W(kh)$. It is then natural to ask how far the linearly interpolated process $\{w_h(t)\}$ is from $\{W(t)\}$. We consider the criteria (0.2), (0.3).

Proposition 1.1 *Let $h = 1/n$. Then as $n \to \infty$*

$$\mathbb{E}\int_0^1 |w_{1/n}(t) - W(t)|\, dt \;=\; \frac{c}{n^{1/2}}$$

$$\liminf_{n\to\infty} \frac{\mathbb{E}\sup_{0\le t\le 1}|w_{1/n}(t) - W(t)|}{\sqrt{\log n}/n^{1/2}} \;>\; 0.$$

Proof Let $B^T(t) = W(t) - (t/T)W(t)$, $0 \le t \le T$, denote the Brownian bridge up to time T. Then $\int_0^1 |w_{1/n}(t) - W(t)|\, dt$ has the same distribution as the sum of n independent copies of $\int_0^{1/n} |B^{1/n}(t)|\, dt$. By standard scaling properties of Brownian motion, $\left\{B^T(tT)\right\}_{0\le t\le 1}$ has the same distribution as $\left\{\sqrt{T}B^1(t)\right\}_{0\le t\le 1}$, and so

$$\begin{aligned}
&\mathbb{E}\int_0^1 |w_{1/n}(t) - W(t)|\, dt\\
&= n\mathbb{E}\int_0^{1/n} |B^{1/n}(t)|\, dt \;=\; \mathbb{E}\int_0^1 |B^{1/n}(t/n)|\, dt\\
&= n^{-1/2}\mathbb{E}\int_0^1 |B^1(t)|\, dt.
\end{aligned}$$

The second assertion follows by extreme value theory. First as above, $\sup_{0\le t\le 1}|w_{1/n}(t) - W(t)|$ has the same distribution as the maximum of n independent copies of $\sup_{0\le t\le 1/n}|B^{1/n}(t)|$ which in turn has the same distribution as the maximum of n independent copies of $\sup_{0\le t\le 1}|B^1(t)|/n^{1/2}$. By extreme value theory, this maximum grows in distribution like $\sqrt{\log n}/n^{1/2}$. $\square$

Note that the rate of convergence in Proposition 1.1 is much slower than in the deterministic case: if $f(t)$ is a function of $t \in [0, 1]$ and $f_{1/n}$ the function obtained by linear interpolation with grid points $0, 1/n, \ldots, 1$, then

$$\int_0^1 |f_{1/n}(t) - f(t)|\, dt \;\sim\; \frac{1}{12n^2}\int_0^1 |f''(t)|\, dt,$$
$$\sup_{0\le t\le 1} |f_{1/n}(t) - f(t)| \;\sim\; \frac{1}{4n^2}\sup_{0\le t\le 1} |f''(t)|.$$

This is an important point in understanding for example the care which must be taken in the simulation of SDE's.

Bisection

Assume for example that we are interested in characteristics associated with the first passage time $\tau(x) = \inf\{t \ge 0 : W(t) \ge x\}$. We can then simulate a discrete skeleton $\{w_{1/n}(k/n)\}_{k=0,1,\dots}$ and use

$$\tau_n(x) \;=\; \inf\{t = 1/n, 2/n, \dots : w_{1/n}(t) \ge x\}$$

as approximation. However, this obviously overestimates $\tau(x)$ so one could consider to make the skeleton finer and finer to judge whether such a discrete approximation is sufficiently accurate, and we proceed to give the details for such an algorithm.

Consider $\{W(t)\}$ in the time interval $[0,1]$. The goal is then to generate a set of r.v.'s

$$w_k(0),\, w_k(1/2^k),\, \dots,\, w_k((2^k-1)/2^k),\, w_k(1)$$

which have the same joint distribution as

$$W(0),\, W(1/2^k),\, \dots,\, W((2^k-1)/2^k),\, W(1)$$

and the sample path consistency property

$$w_k(j/2^{k-1}) \;=\; w_{k-1}(j/2^{k-1}),\; j = 0,1,\dots,2^{k-1}. \tag{1.1}$$

First generate $w_0(0), w_0(1)$ by taking $w_0(0) = 0$, $w_0(1) \sim \mathcal{N}(0,1)$. Next use the fact that

$$W(t/2) \mid W(t) \;\sim\; \mathcal{N}(W(t)/2, t/4).$$

I.e., if the $w_{k-1}(j/2^{k-1})$ have been generated, define $w_k(i/2^k)$ by (1.1) for $i = 2j$. For $i = 2j+1$, take $w_k(i/2^k) \sim \mathcal{N}(y, 2^{-k-1})$ where

$$y \;=\; \frac{1}{2}\left(w_{k-1}(j/2^{k-1}) + w_{k-1}((j+1)/2^{k-1})\right).$$

Reflected Brownian motion with or without drift

Brownian motion $\{W_\mu(t)\}$ with drift μ is defined by $W_\mu(t) = W(t) + \mu t$ where $\{W(t)\}$ is standard Brownian motion.. Obviously, $\{W_\mu(t)\}$ is straightforward to simulate among a discrete skeleton by just adding a linear term to $\{W(t)\}$.

Reflected Brownian motion with drift μ is defined as

$$\overline{W}_\mu(t) \;=\; W_\mu(t) - \inf_{0\le s\le t} W_\mu(s).$$

If $\mu = 0$, $\{\overline{W}_0(t)\}$ has the same distribution as $\{|W(t)|\}$ and so simulating along a discrete grid is easy and there is no error at the grid points. If $\mu \neq 0$, it seems natural to simulate first a discrete skeleton $\{w_{\mu,1/n}(k/n)\}$ of $\{W_\mu(t)\}$ and let $\{\overline{w}_{\mu,1/n}(k/n)\}$ be the reflected version defined by

$$\overline{w}_{\mu,1/n}(k/n) \;=\; w_{\mu,1/n}(k/n) - \min_{\ell=0,\dots,k} w_{\mu,1/n}(\ell/n). \tag{1.2}$$

However, now there is an error also at the grid points because even if $\{w_{\mu,1/n}(k/n)\}$ fits exactly, the minimum is taken over too small a set and so $\overline{w}_{\mu,1/n}(k/n)$ is smaller than $\overline{W}_\mu(k/n)$.

A careful analysis of this situation was given by Asmussen, Glynn & Pitman [15], who showed that the error at the grid points is of order $n^{-1/2}$ and gave precise asymptotics. In [15], some algorithms which improve upon (1.2) are also discussed.

The first is as follows. For fixed $T > 0$, the joint density of

$$\Bigl(W_\mu(T), -\min_{0\le t\le T} W_\mu(t)\Bigr) \tag{1.3}$$

is known. For simulation purposes, a convenient representation of this distribution is to note that marginally, $W_\mu(T)$ is normal $(\mu T, T)$, and that by a result of Lévy,

$$F_y(x) = \mathbb{P}\left(-\min_{0\le t\le T} W_\mu(t) - y \le x \,\middle|\, W_\mu(T) = y\right) = 1 - e^{-2x(y+x)/T}.$$

By easy calculus,

$$F_y^{-1}(z) = \frac{-y + \sqrt{y^2 - 2T\log(1-z)}}{2}.$$

Thus, we may first generate $W_\mu(T)$ as normal $(\mu T, T)$ and next let

$$-\min_{0\le t\le T} W_\mu(t) = -\frac{W_\mu(T)}{2} + \frac{\sqrt{W_\mu(T)^2 - 2T\log(U)}}{2},$$

where U is uniform on $(0,1)$.

Thus, an algorithm for exact simulation of BM W, the minimum M and thereby RBM $\overline{W} = W - M$ at the epochs $t = 0, 1/n, 2/n \ldots$ is obtained as follows:

Algorithm 1.A

1. Let $t \leftarrow 0$, $W \leftarrow 0$, $\overline{W} \leftarrow 0$, $M \leftarrow 0$.
2. Generate (T_1, T_2) from the density (1.3) with $T = 1/n$;
3. Let $t \leftarrow t + 1/n$, $M \leftarrow \min(M, W - T_2)$, $W \leftarrow W + T_1$, $\overline{W} \leftarrow W - M$.
4. Return to 2.

A related algorithm is exact on a random grid associated with a Poisson process that is run independently of the RBM $\overline{W}$. It takes advantage of the following well-known lemma:

Lemma 1.2 *Let T be an exponential r.v. with rate λ which is independent of $\{W_\mu(t)\}$. Then the r.v.'s $W_\mu(T) - \min_{0\le t\le T} W_\mu(t)$ and $-\min_{0\le t\le T} W_\mu(t)$ are independent and exponentially distributed with rates η and ω respectively, where*

$$\eta = -\mu + \sqrt{\mu^2 + 2\lambda}, \qquad \omega = \mu + \sqrt{\mu^2 + 2\lambda}.$$

Thus, an algorithm for unbiased simulation of BM B, the minimum M and thereby RBM $\overline{W} \leftarrow W - M$ at the epochs t of a Poisson(λ) grid is obtained as follows:

Algorithm 1.B

1. Let $t \leftarrow 0$, $W \leftarrow 0$, $\overline{W} \leftarrow 0$, $M \leftarrow 0$.
2. Generate T, S_1, S_2 as exponential r.v.'s with rates 1, η, ω, respectively.
3. Let $t \leftarrow t + T$, $M \leftarrow \min(M, W - S_2)$, $W \leftarrow W + S_1 - S_2$, $\overline{W} \leftarrow W - M$.
4. Return to 2.

An open problem of considerable interest is too find good algorithms for simulating reflected Brownian motion in higher dimensional regions, like the models in Dai, Harrison & Williams [40].

2 Lévy jump processes

A primer on Lévy processes is given in the Appendix. Recall that such a process $\{X(t)\}_{t\geq 0}$ is defined as a continuous time process on $\mathbb{R}$ with stationary independent increments and $X(0) = 0$, and can be written as an independent sum $X(t) = ct + \sigma W(t) + J(t)$ of a linear drift ct, a Brownian component $\sigma W(t)$, and a (possibly compensated) jump process $\{J(t)\}$ given in terms of its Lévy measure $\nu(dx)$. Since we have discussed Brownian motion separately, we consider here only the case $\sigma^2 = 0$. Recall also that

$$\int_{-\infty}^{\infty} (y^2 \wedge 1)\, \nu(dy) \;<\; \infty. \tag{2.1}$$

and that the paths of $\{J(t)\}$ are of finite variation (no compensation needed) if and only if

$$\int_{-\infty}^{\infty} (|y| \wedge 1)\, \nu(dy) \;<\; \infty. \tag{2.2}$$

(say a stable process with $0 < \alpha < 1$ or a subordinator).

For simulation, the compound Poisson case is obviously straightforward from any point of view (provided at least that it is straightforward to simulate from the probability measure proportional to $\nu(dy)$) and so we concentrate in the following on the case $\int_{-\infty}^{\infty} \nu(dy) = \infty$.

Any Lévy jump process $\{J(t)\}$ can be written as the independent sum

$$J(t) \;=\; J^{(1)}(t) + J^{(2)}(t) \tag{2.3}$$

where the Lévy measures of $\{J^{(1)}(t)\}$, $\{J^{(2)}(t)\}$ are the restrictions $\nu^{(1)}$, $\nu^{(2)}$ of ν to $(-\epsilon, \epsilon)$, resp. $\{y : |y| \geq \epsilon\}$. Here $\nu^{(2)}$ is finite so $\{J^{(2)}(t)\}$ is a compound Poisson process and simulation is straightforward. As a first attempt, one would then choose $\epsilon > 0$ so small that $\{J^{(1)}(t)\}$ can be neglected and just simulate $\{J^{(2)}(t)\}$.

As a more refined procedure, it is often suggested (e.g. Bondesson [28], Rydberg [135]) to replace $\{J^{(1)}(t)\}$ by a Brownian motion with the appropriate variance $\sigma_\epsilon^2 = \int_{-\epsilon}^{\epsilon} y^2\, \nu(dy)$ and mean $\mu_\epsilon = \int_{-\epsilon}^{\epsilon} y\, \nu(dy)$ in the finite variation case (2.2), $\mu_\epsilon = 0$ in the compensated case. The justification for this is the folklore because small jumps get more and more dominant as ϵ becomes small, one should have

$$\left\{ \left(J^{(1)}(t) - \mu_\epsilon t \right) / \sigma_\epsilon \right\}_{t\geq 0} \xrightarrow{\mathcal{D}} \{W(t)\}_{t\geq 0} \tag{2.4}$$

as $\epsilon \downarrow 0$. Whereas it seems questionable whether (2.4) holds in complete generality, the following result covers most cases of practical interest:

Proposition 2.1 *Assume that ν has a density of the form $L(x)/x^{\alpha+1}$ for all small x where $L(x)$ is slowly varying and $0 < \alpha < 2$. Then (2.4) holds.*

Proof We show only that $J^{(1)}(1)$, properly normalized, has a limiting standard normal distribution. By Karamata's theorem ([25]),

$$\sigma_\epsilon^2 \;=\; \int_{-\epsilon}^{\epsilon} x^2\,\nu(dx) \;=\; \int_{-\epsilon}^{\epsilon} x^{1-\alpha} L(x)\,dx \;\sim\; \frac{L(\epsilon)+L(-\epsilon)}{2-\alpha}\epsilon^{2-\alpha}$$

Since $L(\epsilon)\epsilon^\gamma \to 0$, $\epsilon^\gamma/L(\epsilon) \to 0$ for any $\gamma > 0$ and similarly for $L(-\epsilon)$, we therefore have $\epsilon/\sigma_\epsilon \to 0$ so that

$$\begin{aligned}
&\log \mathbb{E}\,\exp\left\{s\left(J^{(1)}(1)-\mu_\epsilon t\right)/\sigma_\epsilon\right\}\\
&= \int_{-\epsilon}^{\epsilon}\left(e^{sx/\sigma_\epsilon}-1-sx/\sigma_\epsilon\right)\nu(dx) \;=\; \int_{-\epsilon}^{\epsilon}\left(\frac{s^2x^2}{2\sigma_\epsilon^2}+O\left(\frac{|s^3x^3|}{\sigma_\epsilon^3}\right)\right)\nu(dx)\\
&= \frac{s^2}{2}+o(1),
\end{aligned}$$

where the last equailty follows from

$$\int_{-\epsilon}^{\epsilon}|x^3|\,\nu(dx) \;\sim\; \frac{L(\epsilon)+L(-\epsilon)}{3-\alpha}\epsilon^{3-\alpha}$$

□

Discrete skeletons

Because of the property of stationary independent increments, the problem of simulating a discrete skeleton $\left\{j_{T/n}(kT/n)\right\}$ of a Lévy jump process is obviously equivalent to the problem of r.v. generation from a specific infinitely divisible distribution. In some cases like the Gamma, Cauchy or inverse Gaussian distributions, the density is available and standard methods may apply.

For stable distributions, there is standard algorithm due Chambers, Mallow and Stuck [35] (see also Samorodnitsky & Taqqu [142]). It has a particularly simple form for a symmetric stable distribution ($\beta = 0$): if Y_1, Y_2

are independent such that Y_1 is standard exponential and Y_2 uniform on $(-\pi/2, \pi/2)$, then

$$X = \frac{\sin(\alpha Y_2)}{(\cos Y_2)^{1/\alpha}} \left(\frac{\cos((1-\alpha)Y_2)}{Y_1} \right)^{(1-\alpha)/\alpha} \tag{2.5}$$

has a $S_\alpha(1,0,0)$ distribution. Note that if $\alpha = 2$, then (2.5) reduces to

$$X = \sqrt{Y_1}\frac{\sin(2Y_2)}{\cos Y_2} = 2\sqrt{Y_1}\sin Y_2$$

which is the Box–Muller method for generating a normal r.v. with variance 2. The algorithm is also fairly simple and well working in the asymmetric case but then of a somewhat more complicated form.

For asymmetric stable distributions and –processes, the right skewed case $\beta = 1$ can be viewed as the building block because of the fact that if Y_1, Y_2 are independent and $S_\alpha(1,0,0)$ distributed, then

$$Y = \mu + \sigma\left(\frac{1+\beta}{2}\right)^{1/\alpha} Y_1 - \sigma\left(\frac{1-\beta}{2}\right)^{1/\alpha} Y_2$$

has a $S_\alpha(\sigma, \beta, \mu)$ distribution.

For general Lévy jump processes, it is the exception rather than the rule that special methods are available as for the Gamma, Cauchy, inverse Gaussian and stable cases, and most often the r.v. generation has to be based directly upon the Lévy measure ν. It is obviously impossible to generate an infinity of jumps, and so invariably some truncation– or limiting procedure is involved.

Different algorithms were suggested by Bondesson [28] and Damien, Laud & Smith [42]. Bondesson's method is an early instance of ideas related to the series representations discussed below, and we return to it there. The starting point of Damien, Laud & Smith is the finite measure $\theta(dy) = y^2/(1+y^2)\,\nu(dy)$ in the form (A3.7) of the Lévy–Khintchine representation. Write $c = \int_{-\infty}^{\infty} \theta(dy)$.

Proposition 2.2 ([42]) *Let (U_i, V_i), $i = 1, \ldots, n$, be i.i.d. pairs such that U has distribution $\theta(dy)/c$ and the conditional distribution of V given $U = y$ is Poisson$(c(1+y^2)/ny^2)$, and let*

$$Z_n = \begin{cases} \displaystyle\sum_{i=1}^n U_i V_i & \text{in the finite variation case.} \\ \displaystyle\sum_{i=1}^n \left(U_i V_i - \frac{c}{nU_i} \right) & \text{in the compensated case.} \end{cases}$$

Then $Z_n \xrightarrow{\mathcal{D}} X_1$ *as* $n \to \infty$.

Proof Letting $\lambda_n(y) = c(1+y^2)/ny^2$, we have $\mathbb{E}[e^{sUV} \mid U = y] = e^{\lambda_n(y)(e^{sy}-1)}$. Thus in the finite variation case, we get for $\Re s = 0$ that

$$\begin{aligned}
\log \mathbb{E}e^{sZ_n} &= n \log\left(\int_0^\infty e^{\lambda_n(y)(e^{sy}-1)} \frac{y^2}{c(1+y^2)}\nu(dy)\right) \\
&= n \log\left(\int_0^\infty \left\{1 + \lambda_n(y)(e^{sy}-1) + O(1/n^2)\right\} \frac{y^2}{c(1+y^2)}\nu(dy)\right) \\
&= n \log\left(1 + \frac{1}{n}\int_0^\infty (e^{sy}-1)\nu(dy) + O(1/n^2)\right) \\
&= n \log\left(1 + \frac{1}{n}\varphi(s) + O(1/n^2)\right) \ \to \ \varphi(s)
\end{aligned}$$

as should be (using that the $O(1/n^2)$ term is uniform in y). For the compensated case, see [42]. □

A particularly appealing case is the $S_\alpha(1,1,0)$ case where U can be generated as $\sqrt{1/W-1}$ with W having a Beta$(\alpha/2, 1-\alpha/2)$ distribution.

Series representations

A common idea of the methods to be discussed is to avoid the planar point process N with intensity measure $\nu(dy) \otimes dt$ discussed in the Appendix and to work instead with a Poisson process M on $[0,\infty)$. Sometimes, the methods improve upon those discussed for discrete skeletons by better allowing to identify the location of the important jumps which is important, e.g., for reducing the uniform error (0.2) (note that even in the simple case of a Poisson process, (0.2) evaluated for a discrete skeleton does not go to zero). We do not know of practical implementations of most of the series representations presented below, and it does not seem apriori obvious either whether they represent an improvement of the simple idea of simulating the compound Poisson process obtained by truncating the Lévy measure at ϵ. Nevertheless, the representations contain potentially useful ideas.

We take the intensity of M to be λ, and denote the nth epoch by Γ_n (thus, $\{\Gamma_n - \Gamma_{n-1}\}_{n=1,2,\ldots}$ is a sequence of i.i.d. exponential r.v.'s with mean $1/\lambda$). Let further the sequences $\{U_n\}$, $\{\xi_n\}$ be independent of M and i.i.d., such that U_n is uniform on $(0,1)$ and ξ_n has some distribution varying from case to case in the following. The representations we consider typically have the form

$$\left\{\sum_{n=1}^{\infty} G(\xi_n, \Gamma_n) I(U_n \le t)\right\}_{0 \le t \le 1}$$

in the finite variation case.

We first consider the algorithm of Bondesson [28].

Proposition 2.3 ([28]) *Assume that $\{X(t)\}$ is a subordinator. Then:*
(i) *There exists a family $\{H(dy, u)\}_{u \ge 0}$ of distributions on $[0, \infty)$ and a $\lambda > 0$ such that*

$$\lambda \int_0^{\infty} H(dy, u)\, du \;=\; \nu(dy); \tag{2.6}$$

(ii) *For such a family $\{H(dy, u)\}$, let $W_1, W_2, \ldots$ be r.v.'s which are conditionally independent given M, such that W_i has distribution $H(\cdot, \Gamma_i)$ given M. Then $X = W_1 + W_2 + \cdots$ has the same distribution as X_1.*

Proof Part (i) follows from the proof of Corollaries 2.4 or 2.6 below. For (ii), we can write the Lévy exponent $\varphi(s) = \log \mathbb{E} e^{sX_1}$ as

$$\varphi(s) \;=\; \int_0^{\infty} (e^{sy} - 1)\nu(dy) \;=\; \int_{u=0}^{\infty} \int_{y=0}^{\infty} \lambda(e^{sy} - 1) H(dy, u)\, du$$

which we recognize as the c.g.f. of the total reward $X^*(\infty)$ in a time–inhomogeneous compound Poisson process $\{X^*(t)\}_{0 \le t < \infty}$ with constant arrival rate λ and jump size distribution $H(\cdot, u)$ at time u. From this the result follows. $\square$

Corollary 2.4 *Let $\lambda > 0$ and define $\overline{\nu}(x) = \int_{x+}^{\infty} \nu(dy)$,*

$$g(u) \;=\; \sup\{x : \overline{\nu}(x) \;>\; \lambda u\}$$

Then $X = g(\Gamma_1) + g(\Gamma_2) + \cdots$ has the same distribution as X_1.

Proof Let $\overline{H}(x, u) = \int_{x+}^{\infty} H(dy, u)$. Then (2.6) can be rewritten as

$$\lambda \int_0^{\infty} \overline{H}(x, u)\, du \;=\; \overline{\nu}(x), \quad x \ge 0. \tag{2.7}$$

Letting $H(\cdot, u)$ be the degenerate distribution at $g(u)$, we have $\overline{H}(x, u) = 1$, $x \le g(u)$, $\overline{H}(x, u) = 0$, $x < g(u)$. $\square$

Note the similarity of the algorithm of Corollary 2.4 to r.v. generation by inversion. In [28], several other choices of H are discussed in particular settings.

Bondesson only considers one–dimensional distributions, not processes, but in fact:

Corollary 2.5 *Under* (i) *of Proposition 2.3,*

$$\{X(t)\}_{0\le t\le 1} \stackrel{\mathcal{D}}{=} \left\{\sum_{n=1}^{\infty} W_n I(U_n \le t)\right\}_{0\le t\le 1}. \tag{2.8}$$

Proof Let $\tilde{X}(t) = \sum_{n=1}^{\infty} W_n I(U_n \le t)$. We can then think of $\tilde{X}(t)$ as the total reward in the process obtained from $\{X^*(t)\}$ from thinning with retention probability t. Hence as in the proof of Proposition 2.3, with λ replaced by λt, we get $\log \mathbb{E} e^{s\tilde{X}(t)} = t\varphi(s)$, and it only remains to show independence of increments. But if we split $\{X^*(t)\}$ into three processes $\{X^*(t;1)\}$, $\{X^*(t;2)\}$, $\{X^*(t;3)\}$ by letting a jump go to the three processes w.p.'s t, $t+s$, resp. $1-t-s$ according to the U_n, these processes are independent and hence so are the total rewards

$$\begin{aligned} X^*(\infty;1) &= \sum_{n=1}^{\infty} W_n I(U_n \le t) &= \tilde{X}(t), \\ X^*(\infty;2) &= \sum_{n=1}^{\infty} W_n I(t < U_n \le t+s) &= \tilde{X}(t+s) - \tilde{X}(t). \end{aligned}$$

□

In the following, let $\lambda = 1$. There are several series representations of Lévy processes of similar type as (2.8) around in the literature. For example, an $S_\alpha(1,\beta,0)$ process with $\alpha < 1$ can be represented as

$$C_\alpha^{1/\alpha} \sum_{n=1}^{\infty} \xi_n \Gamma_n^{-1/\alpha} I(U_n \le t) \tag{2.9}$$

where

$$C_\alpha = \left(\int_0^\infty x^{-\alpha} \sin x \, dx\right)^{-1}, \quad \mathbb{P}(\xi_n = 1) = \mathbb{P}(\xi_n = -1) = \frac{1+\beta}{2}.$$

Letting $H(\cdot, u)$ be the distribution of a r.v. which is $\pm C_\alpha^{1/\alpha} u^{-1/\alpha}$ with probabilities $(1\pm\beta)/2$, this representation is as the same form as in Corollary 2.5.

For $1 \le \alpha < 2$, there are similar expansion as (2.9) but with certain centering terms $tb_n^{(\alpha)}$ added for each term (corresponding to compensation). See [142] for details. Maybe more surprisingly, if the process is not completely skewed (i.e., $|\beta| \neq 1$) then such centering can be avoided. For example, for $\alpha \neq 1$ a possible representation is (2.9) with the distribution of ξ_n changed to $\mathbb{P}(\xi_n = a_\pm) = 1 - a_\pm/(a_+ + a_-)$ where

$$a_\pm \;=\; \pm\left[\frac{1\pm\beta}{2}\left(\frac{1+\beta}{1-\beta}\right)^{\pm 1/(\alpha-1)} + 1\right]^{1/\alpha}.$$

Cf. Janicki & Weron [83]. For further studies of series expansions without compensation, see Rosiński [129]

Here is one more example, random thinning of i.i.d. sequences (Rosiński [128]); the thinning corresponds to allowing $H(\cdot, u)$ to have an atom at 0 in Proposition 2.3.

Corollary 2.6 *Consider the finite variation case. Let the ξ_n have distribution F where F a probability distribution on $\mathbb{R}/\{0\}$ which is equivalent to ν in the Radon–Nikodym sense, and let $g = d\nu/dF$. Then the process can be represented as*

$$\sum_{n=1}^{\infty} \xi_n I(g(\xi_n) \ge \Gamma_n) I(U_n \le t).$$

Proof For $y > 0$, we get

$$H(dy, u) \;=\; \mathbb{P}(\xi_n \in dy;\, g(\xi_n) \ge u) \;=\; F(dy) I(g(y) \ge u)$$

Thus $g = d\nu/dF$ yields

$$\int_0^\infty H(dy, u)\, du \;=\; F(dy) g(y) \;=\; \nu(dy).$$

□

3 Stochastic differential equations

3a Numerical methods for ODE's

Consider the ODE $\dot{x}(t) = a(t, x(t))$ with initial condition $x(0) = x_0$ in the time interval $[0, T]$. A numerical solution is typically implemented via discrete approximations: we write $T = N\Delta$ and generate the approximation y^Δ by generating

$$y_0 = y^\Delta(0),\ y_1 = y^\Delta(\Delta),\ \ldots,\ y_n = y^\Delta(n\Delta),\ \ldots, y_N = y^\Delta(T) = y^\Delta(N\Delta)\,.$$

The error criterion is

$$e(\Delta) \;=\; \left|x(T) - y^\Delta(T)\right| \;=\; |x(T) - y_N|\ .$$

The basic method is the Euler method

$$y_0 = x_0, \quad y_{n+1} = y_n + a(n\Delta, y_n)\Delta\,. \tag{3.1}$$

Under suitable smoothness conditions (which we omit here and in the following), $e(\Delta) = O(\Delta)$.

Here are some improvements:

a). Take a Taylor expansion of order $k > 1$ rather than $k = 1$ as in (3.1). For $k = 2$, this gives

$$y_{n+1} \;=\; y_n + a(n\Delta, y_n)\Delta + \{a_t(n\Delta, y_n) + a(n\Delta, y_n)a_x(n\Delta, y_n)\}\,\frac{\Delta^2}{2}\,.$$

Here $e(\Delta) = O(\Delta^2)$.

b). In

$$x(\Delta) = x(0) + \int_0^\Delta a(t, x(t))\,dt\,,$$

approximate the integral by

$$\{a(0, x(0)) + a(\Delta, x(\Delta))\}\,\frac{\Delta}{2}$$

(the trapezoidal rule) rather than $a(0, x(0))\Delta$ as in (3.1). Here $X(\Delta)$ is unknown but can be estimated by (3.1), i.e. predicted by $x(0) + a(0, x(0))\Delta$. This gives $y_0 = x_0$,

$$\begin{aligned}\overline{y}_{n+1} &= y_n + a(n\Delta, y_n)\Delta,\\ y_n + 1 &= y_n + \{a(n\Delta, y_n) + a((n+1)\Delta, \overline{y}_{n+1})\}\,\frac{\Delta}{2}\,,\end{aligned}$$

which is an example of a *predictor–corrector* method. Again, $e(\Delta) = O(\Delta^2)$.

3b The Euler methods for SDE's

In this and the following sections, we consider the SDE $X(0) = x_0$,

$$dX(t) \;=\; a(t, X(t))dt + b(t, X(t))\, dW(t) \;\; 0 \le t \le T, \tag{3.1}$$

where $\{W(t)\}_{t\ge 0}$ is standard Brownian motion.

The numerical methods for SDE's are modelled after those for ODE's. We will start by the Euler method and in the next sections, we study SDE analogues of methods based upon higher orden Taylor expansion. We mention for completeness that also say implicit methods have been extended but shall not give the details (see Kloeden & Platen [90]).

The Euler scheme is $T = N\Delta$, $y_0 = x_0$,

$$y_{n+1} \;=\; y_n + a(n\Delta, y_n)\Delta + b(n\Delta, y_n)(\Delta W_n)\,,$$

where $(\Delta W_n) = W((n+1)\Delta) - W(n\Delta)$. For brevity, we write ΔW in the following. The (ΔW_n) are generated as i.i.d. $N(0, \Delta)$ variables.

3c Error criteria

For SDE's, one may be interested in two types of fit, strong and weak (which one depends on the type of application):

(s) $\{y_n\}$ should give a good approximation of the sample path of $\{X(t)\}$. This leads to the error criterion

$$e_{\mathrm{s}}(\Delta) \;=\; \mathbb{E}\,|X(T) - y_{N\Delta}|\;.$$

(w) $y_{N\Delta}$ should give a good approximation of the distribution of $X(T)$. That is, $\mathbb{E}g(y_{N\Delta})$ should be close to $\mathbb{E}g(X(T))$ for sufficiently many smooth functions g.

We will say that $Y^\Delta = \{y_n\}$ converges strongly to X at time T with order $\gamma > 0$ if $e_{\mathrm{s}}(\Delta) = O(\Delta^\gamma)$, and weakly if

$$|\mathbb{E}g(X(T)) - \mathbb{E}g(y_{N\Delta})| \;=\; O(\Delta^\gamma)$$

for all g such that $g', g'', \ldots, g^{(2(\gamma+1)}$ exist [in practice, the relevant values of γ are $\gamma = 1, 1.5, 2, 2.5, \ldots$ so that $2(\gamma+1)$ is integer] and have polynomial growth,

$$\left|g^{(k)}(x)\right| \;\le\; d_{g,k}x^{p_{g,k}}\,, \quad k = 0, \ldots, 2(\gamma+1)\,.$$

We state without proof the following main result:

Theorem 3.1 *The Euler scheme* (3.1) *converges strongly with order* $\gamma = 0.5$*, and weakly with order* $\gamma = 1$.

3d The Milstein scheme

The idea is that the approximation

$$\int_0^\Delta b(t, X(t))\,dW(t) \;\sim\; b(0, X(0))W_\Delta$$

is the main source of error for the Euler scheme. To improve it, we estimate the error by Ito's formula for $b(t, X(t))$:

$$\begin{aligned}
&\int_0^\Delta b(t, X(t))\,dW(t) \;-\; b(0, X(0))W_\Delta \\
&= \int_0^\Delta \{b(t, X(t)) - b(0, X(0))\}\,dW(t) \\
&= \int_0^\Delta \Big\{ \int_0^t \Big[b_t(s, X(s)) + a(s, X(s))b_x(s, X(s)) + \frac{1}{2}b^2(s, X(s))b_{xx}(s, X(s)) \Big]\,ds \\
&\qquad + \int_0^t b(s, X(s))b_x(s, X(s))\,dW(s) \Big\}\,dW(t) \\
&\sim\; O(\Delta^2) \;+\; b(0, x_0)b_x(0, x_0) \int_0^\Delta \int_0^t dW(s)\,dW(t) \\
&\sim\; b(0, x_0)b_x(0, x_0) \int_0^\Delta W(t)\,dW(t) \\
&= \; b(0, x_0)b_x(0, x_0) \left\{ \frac{1}{2}W_\Delta^2 - \frac{1}{2}\Delta \right\}.
\end{aligned}$$

This leads to the Milstein scheme $y_0 = x_0$,

$$y_{n+1} \;=\; y_n + a\Delta + b\Delta W + \frac{1}{2}bb_x\left\{(\Delta W)^2 - \Delta\right\} \qquad (3.1)$$

where $a = a_n = a(n\Delta, y_n)$ and similarly for b, b_x.

Theorem 3.2 *The Milstein scheme* (3.1) *converges strongly with order* $\gamma = 1$.

3e Ito–Taylor expansions

We proceed by refining the estimate used for the Milstein scheme. For notational convenience, let, e.g., b_x denote $b_x(0, x_0)$ when occurring outside integrals and $b_x(s, X(s))$ when occuring in an integral w.r.t. ds or $dW(s)$. We get

$$\begin{aligned}
&\int_0^\Delta b(t, X(t))\, dW(t) \; - \; b(0, X(0))W_\Delta \\
&\quad = \int_0^\Delta b\, dW(t) \; - bW_\Delta \\
&\quad = \int_0^\Delta \left\{ \int_0^t \left[b_t + ab_x + \frac{1}{2} b^2 b_x x \right] ds \; + \; \int_0^t bb_x\, dW(s) \right\} dW(t) \quad (3.1)
\end{aligned}$$

In the last term, we expand $bb_x = b(s, X(s))b_x(s, X(s))$ one more time by Ito's formula and note that the $dW(u)$ term dominates the du term. Thus approximately (3.1) is

$$\begin{aligned}
&\left[b_t + ab_x + \frac{1}{2} b^2 b_x x \right] \int_0^\Delta t\, dW(t) \; + \; bb_x \int_0^\Delta W(t)\, dW(t) \\
&\quad + \int_0^\Delta dW(t) \int_0^t dW(s) \int_0^s b \frac{\partial}{\partial x}(bb_x)\, dW(u) \\
\sim \; &\left[b_t + ab_x + \frac{1}{2} b^2 b_{xx} \right] (\Delta \cdot W - Z) + \frac{1}{2}(W^2 - \Delta) \\
&\quad + b(bb_{xx} + b_x^2) \int_0^\Delta \left(\frac{1}{2} W(t)^2 - \frac{t}{2} \right) dW(t) \qquad (3.2)
\end{aligned}$$

where

$$Z \; = \; \int_0^\Delta W(s)\, ds \; = \; \Delta W \; - \; \int_0^\Delta s\, dW(s).$$

Similarly,

$$\begin{aligned}
&\int_0^\Delta a(t, X(t))\, dt \; - \; a(0, X(0))\Delta \\
&\quad = \int_0^\Delta a\, dt \; - a\Delta \\
&\quad = \int_0^\Delta \left\{ \int_0^t \left[a_t + aa_x + \frac{1}{2} b^2 a_{xx} \right] ds \; + \; \int_0^t ba_x\, dW(s) \right\} dt \\
&\quad \sim \frac{1}{2} \left[a_t + aa_x + \frac{1}{2} b^2 a_{xx} \right] \Delta^2 \; + \; ba_x \int_0^\Delta W(t)\, dt \\
&\quad \sim \frac{1}{2} \left[a_t + aa_x + \frac{1}{2} b^2 a_{xx} \right] \Delta^2 \; + \; ba_x Z \qquad (3.3)
\end{aligned}$$

By evaluating the differential of $d(W(t)^3/3 - tW(t))$ by Ito's formula, it is seen that the double of the integral in (3.2) is

$$\frac{1}{3}W(\Delta)^3 - \Delta W(\Delta) \;=\; \frac{1}{3}W^3 - \Delta \cdot W\,.$$

Hence, approximating $X(\Delta)$ by $x_0 + a\Delta + bW$+(3.2)+(3.3), we arrive at the Ito-Taylor formula

$$\begin{aligned} X(\Delta) \;\sim\; & x_0 + a\Delta + bW + \frac{1}{2}bb_x(W^2 - \Delta) \\ & + a_x bZ + \frac{1}{2}\left[a_t + aa_x + \frac{1}{2}b^2 a_{xx}\right]\Delta^2 \\ & + \left[b_t + ab_x + \frac{1}{2}b^2 b_{xx}\right](\Delta \cdot W - Z) \\ & + \frac{1}{2}\left[b(bb_{xx} + b_x^2\right]\left(\frac{1}{3}W^3 - \Delta \cdot W\right)\,. \end{aligned}$$

For the following, we note that the covariance matrix of $(W, Z) = (W(\Delta), Z(\Delta))$ is

$$\begin{pmatrix} \Delta & \frac{1}{2}\Delta^2 \\ \frac{1}{2}\Delta^2 & \frac{1}{3}\Delta^3 \end{pmatrix}\,. \tag{3.4}$$

This follows, e.g., by writing

$$W \;=\; \int_0^\Delta dW(s), \quad Z \;=\; \int_0^\Delta dt \int_0^t dW(s) \;=\; \int_0^\Delta (\Delta - s)dW(s)$$

which yields the covariance function as

$$\begin{pmatrix} \int_0^\Delta ds & \int_0^\Delta (\Delta - s)\,ds \\ \int_0^\Delta (\Delta - s)\,ds & \int_0^\Delta (\Delta - s)^2\,ds \end{pmatrix}\,,$$

cf. (A8.4).

3f Higher order schemes

A scheme of strong order 1.5 is obtained directly from the Ito–Taylor expansion: $y_0 = x_0$,

$$
\begin{aligned}
y_{n+1} \;=\;& y_n + a\Delta + b(\Delta W) + \frac{1}{2}bb_x\left\{(\Delta W)^2 - \Delta\right\} \\
&+ a_x b(\Delta Z) + \frac{1}{2}\left[a_t + aa_x + \frac{1}{2}b^2 a_{xx}\right]\Delta^2 \\
&+ \left[b_t + ab_x + \frac{1}{2}b^2 b_{xx}\right]\left((\Delta W)\cdot\Delta - (\Delta Z)\right) \\
&+ \frac{b}{2}\left[bb_{xx} + b_x^2\right]\left(\frac{1}{3}(\Delta W)^3 - (\Delta W)\cdot\Delta\right)
\end{aligned}
$$

where $a = a_n = a(n\Delta, y_n)$ etc. and the $((\Delta W)_n, (\Delta Z)_n)$ are generated as i.i.d. bivariate normals with mean 0 and covariance matrix (3.4) (cf. also the extended Box–Muller method in Chapter I for generating dependent bivariate normals).

A scheme of weak order 2 of a slightly simpler form can be obtained by deleting the last term.

4 Gaussian processes

Let $\{X(t)\}$ be a real–valued Gaussian process in discrete or continuous time as specified with its covariance function $\gamma(s,t) = \mathbf{Cov}(X(s), X(t))$; it will not be a restriction for the following to assume that the mean is zero. We will only consider simulation of discrete skeletons so we adapt a discrete time notation $X_0, X_1, \ldots$.

In some cases, a simple description of the dynamics is available which makes it possible to simulate the process directly. A simple example is a discrete time ARMA(p,q) process with representation

$$
X_{n+1} \;=\; \beta_1 X_n + \beta_2 X_{n-1} + \cdots + \beta_p X_{n-p+1} + \alpha_1\epsilon_n + \cdots + \alpha_q\epsilon_{n-q+1} \quad (4.1)
$$

where the ϵ_n are i.i.d. standard normal variables. However, in many examples one has to work directly with the covariance function or the spectral density (see below).

Cholesky factorization. Prediction

We consider recursive algorithms based upon the covariance function, studying how to generate X_{n+1} given $X_0, \ldots, X_n$ have been generated. Thus we need to specify the conditional distribution of X_{n+1} given $X_0, \ldots, X_n$ which is a standard problem in the multivariate normal distribution. Write $\boldsymbol{\Gamma}(n)$ for the covariance matrix of $X_0, \ldots, X_n$ and let $\boldsymbol{\gamma}(n)$ be the $(n+1)$–column vector with $\gamma_k(n) = \gamma(n+1, k)$, $k = 0, \ldots, n$. Then

$$\boldsymbol{\Gamma}(n+1) \;=\; \begin{pmatrix} \boldsymbol{\Gamma}(n) & \boldsymbol{\gamma}(n) \\ \boldsymbol{\gamma}(n)' & \gamma(n+1, n+1) \end{pmatrix}.$$

Therefore by general results on the multivariate normal distribution, the conditional distribution of X_{n+1} given $X_0, \ldots, X_n$ is $\mathcal{N}\left(\hat{X}_{n+1}, \sigma_n^2\right)$ where

$$\hat{X}_{n+1} \;=\; \boldsymbol{\gamma}(n)'\boldsymbol{\Gamma}(n)^{-1}\begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_n \end{pmatrix}, \quad \sigma_n^2 \;=\; \gamma(n+1, n+1) - \boldsymbol{\gamma}(n)'\boldsymbol{\Gamma}(n)^{-1}\boldsymbol{\gamma}(n)\,,$$

and we can just generate X_{n+1} according to $\mathcal{N}\left(\hat{X}_{n+1}, \sigma_n^2\right)$.

Note that in the terminology of time series (e.g. Brockwell & Davis [30]), $\hat{X}_{n+1}$ is the best linear predictor for X_{n+1} (in terms of minimizing the mean square error) and σ_n^2 the corresponding prediction error.

Form the point of simulation, the difficulty is to organize the calculations economically, say by recursive computation of $\boldsymbol{\gamma}(n)'\boldsymbol{\Gamma}(n)^{-1}$ to avoid matrix inversion in each step, or by some other method.

An established device is *Cholesky factorization*. This is an algorithm for writing a given symmetric $(n+1) \times (n+1)$ matrix $\boldsymbol{\Gamma} = (\gamma_{ij})_{i,j=0,\ldots,n}$ as $\boldsymbol{\Gamma} = \boldsymbol{C}\boldsymbol{C}'$ where $\boldsymbol{C} = (c_{ij})_{i,j=0,\ldots,n}$ is (square) lower triangular ($c_{ij} = 0$ for $j > i$), and works as follows. By symmetry of Γ and $\boldsymbol{C}\boldsymbol{C}'$, it suffices that the ijth elements of Γ and $\boldsymbol{C}\boldsymbol{C}'$ are equal for $j \le i$ which means

$$\gamma(i, j) \;=\; \sum_{k=0}^{n} c_{ik}c_{jk} \;=\; \sum_{k=0}^{i\wedge j} c_{ik}c_{jk} \;=\; \sum_{k=0}^{j} c_{ik}c_{jk}, \quad j \le i. \tag{4.2}$$

This determines first c_{00} by $\gamma(0,0) = c_{00}^2$ ($i = j = 0$, whereas for $i = 1$ we get two equations

$$\gamma(1,0) = c_{10}c_{00}, \quad \gamma(1,1) = c_{10}^2 + c_{11}^2$$

determining c_{10}, c_{11}. In general, if $c_{i'j}$ has been computed for $i' < i$, we get

$$c_{ij} \;=\; \frac{1}{c_{jj}}\left(\gamma(i,j) - \sum_{k=0}^{j-1} c_{ik}c_{jk}\right), \quad j<i, \quad c_{ii}^2 \;=\; \left(\gamma(i,i) - \sum_{k=0}^{i-1} c_{ik}^2\right). \tag{4.3}$$

For simulation of $X_0, \ldots, X_n$, the implication is that we can take $Y_0, \ldots, Y_n$ to be i.i.d. standard normal, write $\boldsymbol{Y}(n) = (Y_0 \ldots Y_n)'$, $\boldsymbol{X}(n) = (X_0 \ldots X_n)'$, and define $\boldsymbol{C}(n)$ to be the Cholesky factorization of $\boldsymbol{\Gamma}(n)$. Then the X_i are generated by $\boldsymbol{X}(n) = \boldsymbol{C}(n)\boldsymbol{Y}(n)$. Component by component,

$$X_i \;=\; \sum_{k=0}^{i} c_{ik}Y_k, \quad i = 0, \ldots, n. \tag{4.4}$$

Note that we did not write $c_{ik}(n)$ because (4.3) shows that $c_{ik}(n)$ does not depend on n as long as $i, k \le n$. This means in particular that to get $\boldsymbol{C}(n+1)$ from $\boldsymbol{C}(n)$, one only needs to compute the last row ($i = n+1$). That $\boldsymbol{X}(n)$ has the correct distribution follows from

$$\mathbf{Cov}(\boldsymbol{X}(n)) \;=\; \mathbf{Cov}(\boldsymbol{C}(n)\boldsymbol{Y}(n)) \;=\; \boldsymbol{C}(n)\boldsymbol{I}\boldsymbol{C}(n)' \;=\; \boldsymbol{\Gamma}(n).$$

Remark 4.1 The representation (4.4) shows that $Y_0, \ldots, Y_n$ form a Gram–Schmidt orthogonalization of $X_0, \ldots, X_n$. That is, (in the L_2 sense) $Y_0, \ldots, Y_n$ are orthonormal and $\mathrm{span}(Y_0, \ldots, Y_k) = \mathrm{span}(X_0, \ldots, X_k)$. $\square$

In conclusion, simulation via Cholesky factorization is exact (no approximation is involved) and one does not need to set the time horizon in advance. Note also that no matrix inversion at all is involved. The drawback of the method is that is becomes slow and demanding in terms of storage (one needs to store all c_{ij}) as n becomes large.

In general, Cholesky factorization is just a mathematical device for matrix manipulation. However, in the case of Gaussian processes the procedure can be given an interesting interpretation in terms of the standard problem of time series analysis of *prediction* or *forecasting*: given we have observed $X_0, \ldots, X_n$, we want a predictor of X_{n+1}. Now the best linear predictor (in terms of minimizing the mean square error) of X_{n+1} is

$$\hat{X}_{n+1} \;=\; \mathbb{E}[X_{n+1} \,|\, X_0, \ldots, X_n] \;=\; \boldsymbol{\gamma}(n)'\boldsymbol{\Gamma}(n)^{-1}\begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_n \end{pmatrix}$$

and thus algorithms for recursive prediction are potentially useful for simulation as well.

Two such algorithms are given in [30] Ch. 5, the *Durbin–Levinson algorithm* and the *innovations algorithm.* We consider only the latter. The key is to represent $\hat{X}_i$ as a linear combination of the $X_k - \hat{X}_k$ with $k < i$ rather than the X_k,

$$\hat{X}_i \;=\; \sum_{k=0}^{i-1} \theta_{i,i-k}(X_k - \hat{X}_k) \tag{4.5}$$

in the notation of [30]. Obviously $\hat{X}_0 = 0$, $\sigma_n^2 = \mathbf{Var}(X_{n+1} - \hat{X}_{n+1})$, $\sigma_{-1}^2 = r(0,0)$. Define $Y_k = (X_k - \hat{X}_k)/\sigma_{k-1}$.

Proposition 4.2 *The Y_k are i.i.d. standard normal.*

Proof All that needs to be shown is independence. Let $\mathcal{H}_k$ denote the subspace of L_2 spanned by $X_0, \ldots, X_k$ and let $< X, Y >= \mathbb{E}(XY)$ denote the usual inner product in L_2. For $i < j$ we have $X_i - \hat{X}_i \in \mathcal{H}_{j-1}$ and $X_j - \hat{X}_j \perp \mathcal{H}_{j-1}$ by definition of $\hat{X}_j$. Thus the r.v.'s $Y_j = X_j - \hat{X}_j$ are orthogonal, i.e. uncorrelated and independence follows from properties of the multivariate normal distribution. □

If we let $c_{ik} = \theta_{i,i-k}\sigma_{k-1}$, $k < i$, $c_{ii} = \sigma_{i-1}$, and write $X_i = (X_i - \hat{X}_i) + \hat{X}_i$, (4.5) takes the form (4.4). That is, *determining the $\theta_{i,j}$ needed for the innovation algorithm involves just the same equations as Cholesky factorization, and* (with the right choice of sign) *the Y_k in* (4.4) *can be interpreted as the* $(X_k - \hat{X}_k)/\sigma_{k-1}$ *which in turn form a Gram–Schmidt orthonormalization of the X_k* (cf. Remark 4.1).

For a further variant of Cholesky factorization, see Hosking [80].

Spectral simulation. FFT

We will assume that $\{X(t)\}$ is strictly stationary. Consider first the case of a discrete time process $X_0, X_1, X_2, \ldots$ and write $\gamma_k = \gamma(n, n+k)$ (by stationary, this does not depend on n). Then the sequence $\{\gamma_k\}$ is positive definite and so by Herglotz's theorem, it can be represented as

$$\gamma_k = \int_0^{2\pi} e^{ik\lambda}\, \nu(d\lambda) \tag{4.6}$$

for some finite real measure ν on $[0, 2\pi)$, the *spectral measure*; the condition that the process is real–valued is equivalent to

$$\int_A \nu(d\lambda) \;=\; \int_{2\pi - A} \nu(d\lambda)\,, \quad A \subset (0, \pi) \tag{4.7}$$

(if the spectral density $s = d\nu/d\lambda$ exists, this simply means that s is symmetric around π, $s(\lambda) = s(2\pi - \lambda)$, $0 < \lambda < \pi$). The *spectral representation* of the process is

$$X_n \;=\; \int_0^{2\pi} e^{in\lambda}\, Z(d\lambda) \tag{4.8}$$

where $\{Z(\lambda)\}_{\lambda\in[0,2\pi)}$ is a complex Gaussian process which is traditionally described by having increments satisfying

$$\begin{aligned} \mathbb{E}\left[(Z(\lambda_2) - Z(\lambda_1))\overline{(Z(\lambda_4) - Z(\lambda_3))}\right] \;&=\; 0, \\ \mathbb{E}|Z(\lambda_2) - Z(\lambda_1)|^2 \;&=\; \nu(\lambda_1, \lambda_2] \end{aligned} \tag{4.9}$$

for $\lambda_1 \le \lambda_2 \le \lambda_3 \le \lambda_4$; the integral should be understood as the L_2 limit of approximating step functions (of course, the imaginary part in (4.8) has to vanish since X is real–valued). See e.g. Cramér & Leadbetter [39].

For simulation, it is then appealing to simulate Z and construct X via (4.8). However, Z is not completely specified by (4.9). But:

Proposition 4.3 *Assume that X is real–valued and define $Z(\lambda) = Z_1(\lambda) + iZ_2(\lambda)$ by first taking $\{Z_1(\lambda)\}_{0\le\pi}$, $\{Z_2(\lambda)\}_{0\le\pi}$ to be independent real–valued Gaussian with independent increments satisfying*

$$\mathbf{Var}(Z_i(\lambda_2) - Z_i(\lambda_1)) \;=\; \frac{1}{2}\nu(\lambda_1, \lambda_2]\,,$$

and next letting

$$Z(\pi + \lambda) \;=\; Z(\pi) \,+\, \overline{Z(\pi) - Z(\pi - \lambda -)}, \quad 0 < \lambda < \pi.$$

Then (4.8) *holds, i.e.*

$$X_n \;=\; 2\int_0^{\pi} \cos(n\lambda) Z_1(d\lambda) \;+\; 2\int_0^{\pi} \sin(n\lambda) Z_2(d\lambda). \tag{4.10}$$

Note that in the presence of a spectral density s we may rewrite the definition of the Z_i as

$$dZ_i(\lambda) \;=\; \sqrt{\frac{1}{2}s(Z_i(\lambda))}\, dW_i(\lambda)$$

where W_1, W_2 are independent Brownian motions.
Proof Given Z, we define X by (4.8) By definition, $dZ(\pi+\lambda) = \overline{dZ(\pi-\lambda)}$ so that $\int_\pi^{2\pi} e^{in\lambda} = \int_0^\pi e^{-in\lambda}$ which implies that (4.8) can be written in the alternative form (4.10). In particular, (4.10) shows that X is real–valued (and obviously Gaussian) so it suffices to show that the covariance function is correct. But by (A8.4)

$$\begin{aligned}
\mathbf{Cov}(X_{n+k}, X_n) &= \mathbb{E}[X_{n+k}X_n] \;=\; \mathbb{E}\left[X_{n+k}\overline{X_n}\right] \\
&= \mathbb{E}\left[\int_0^{2\pi} e^{i(n+k)\lambda}\, Z(d\lambda) \cdot \int_0^{2\pi} e^{-in\lambda}\, \overline{Z(d\lambda)}\right] \\
&= \int_0^{2\pi} e^{i(n+k)\lambda} e^{-in\lambda}\, \mathbb{E}|Z(d\lambda)|^2 \\
&= \int_0^{2\pi} e^{ik\lambda}\, \nu(d\lambda) \;=\; r_k.
\end{aligned}$$

□

For practical implementation, the stochastic integrals in (4.10) may either be computed by SDE schemes as in Section 3, or by discrete approximations as follows. Say that X has *discrete spectrum* if ν has a finite support, say mass σ_k^2 at λ_k, $k = 0, \ldots, N-1$. If X is real–valued, this means by (4.7) that N is of the form $2M$ and that we can choose $\lambda_k \le \pi$, $\lambda_{M+k} = 2\pi - \lambda_k$, $k = 0, \ldots, M-1$. Then we can write

$$Z_i(\lambda) \;=\; \sum_{k:\,\lambda_k \le \lambda} \sigma_k Z_{k,i}, \quad 0 \le \lambda \le \pi,$$

with the $Z_{k,i}$ i.i.d. real normal with variance $1/2$, and (4.10) becomes

$$X_n \;=\; 2\sum_{k=0}^{M-1} \sigma_k \left\{\cos(n\lambda_k)Z_{k,1} + \sin(n\lambda_k)Z_{k,2}\right\}. \tag{4.11}$$

A process of the form (4.11) is of course straightforward to simulate. In general, spectral simulation is then performed by approximating the spectral measure by a measure with finite support, which is always possible. But note that there is no canonical way to perform this discrete approximation, and that the method is only approximative.

The great advantage of the method is, however, the speed when (4.11) is implemented via the FFT (fast Fourier transform). This algorithm is an extremely fast algorithm for transforming a real or complex sequence $\{a_n\}_{n=0,1,\dots,N-1}$ into its inverse Fourier transform $\{\hat{a}_n\}_{n=0,1,\dots,N-1}$ when N is a power of 2, $N = 2^m$. Here

$$\hat{a}_n = \sum_{k=0}^{N-1} a_k \exp\{i2\pi kn/N\}$$

(note that often alternative ways to write the Fourier sum are encountered in the literature). One then needs to choose the λ_k of the form $2\pi k/N$ which again is always possible, let $a_k = \sigma_k Z_k$ and take X_n as $\Re\hat{a}_n$.

In continuous time, the same method of course applies to simulate any discrete skeleton $X_0, X_h, X_{2h}, \dots$. One needs then to compute the spectral measure ν_h of the skeleton, which is most often performed from formulas like

$$\nu_h(d\lambda) = \sum_{k=-\infty}^{\infty} \nu(d(\lambda + 2k\pi/h))$$

where ν is the spectral measure in the Bochner representation

$$\gamma(t) = \int_{-\infty}^{\infty} e^{it\lambda}\, \nu(d\lambda) \qquad (4.12)$$

$(\gamma(t) = \gamma(s, s+t))$. For details and implementation issues, see for example Appendix D in Lindgren [100].

An idea somewhat related to spectral simulation is to use wavelets. See e.g. Abry & Sellan [1] for a special case.

ARMA approximations

A stationary Gaussian discrete time process has the ARMA form (4.1) if and only if the spectral measure is absolutely continuous with density of the form $p(e^{i\lambda})/q(e^{i\lambda})$ where p, q are polynomials, cf. Brockwell & Davis [30] Ch. 4. For a general $\{X_n\}$, one can find polynomials p_n, q_n such that

$$\frac{p_n(e^{i\lambda})}{q_n(e^{i\lambda})}\, d\lambda \xrightarrow{w} \nu(d\lambda),$$

see again [30]. This suggests to choose p, q such that the measure with density $p(e^{i\lambda})/q(e^{i\lambda})$ is close to ν, and simulate the corresponding ARMA process as an approximation to $\{X_n\}$.

The precise form of p, q is of course subject to choice and arbitrary. See Krenk & Clausen [93] and Gluver & Krenk [62] for some relevant discussion.

Some further references relevant for the general problem of simulating Gaussian processes are Wood & Chan [155] and Dietrich & Newsom [45].

5 Fractional Brownian motion

Mandelbrot & Van Ness [102] defined fractional Brownian motion as a mean zero Gaussian process $\{B_H(t)\}_{t\geq 0}$ with covariance function of the form

$$r(t,s) \;=\; \frac{\sigma^2}{2}\left(|t|^{2H}+|s|^{2H}-|t-s|^{2H}\right) \tag{5.1}$$

for some $H \in (0,]$ (the *Hurst parameter* or self–similarity parameter). For $H = 1/2$, we are back to standard Brownian motion. This class of processes can be characterized as the only self–similar[1] Gaussian processes with stationary increments. There has been a boom of interest in fractional Brownian motion during the last years, mainly because of statistical studies (e.g. Leland *et al.* [99] and Paxson & Floyd [118]) of phenomena like internet traffic which shows long–range dependence (a covariance function decaying slower than exponentially) and self–similarity. Fractional Brownian motion is then a natural candidate as a model both because of its intrinsic properties and because it appears as limit of many more detailed descriptors of network traffic. Some selected references on performance of systems with fractional Brownian input are Norros [114], Duffield & O'Connell [47] and Narayan [109]. However, there are virtually no explicit results and so simulation becomes an important tool.

For simulation, one can use one of the methods discussed in Section 4 for general stationary Gaussian processes (in this setting, the increments of $\{B_H(t)\}$). In particular, the Cholesky factorization method has been implemented by Michna [103]. One should note that the faster methods of ARMA approximations or FFT are potentially dangerous because they destroy the long–range dependence.

Special algorithms for fractional Brownian motion could potentially be based upon some of the many stochastic integral representations of $\{B_H(t)\}$

[1]A stochastic process $\{X(t)\}$ is self–similar if for some H $\{X(at)\} \stackrel{\mathcal{D}}{=} \{a^H X(t)\}$ for all $a \geq 0$.

which are around. The classical one is

$$\begin{aligned} B_H(t) &= C_H \left[\int_{-\infty}^{0} \left\{ (t-y)^{H-1/2} - |y|^{H-1/2} \right\} dW(y) \right. \\ &\qquad \left. + \int_0^t (t-y)^{H-1/2} \, dW(y) \right] \end{aligned} \tag{5.2}$$

where

$$C_H = \sqrt{\frac{2H}{(H-1/2)\mathrm{Beta}(H-1/2, 2-2H)}}$$

and $\{W(y)\}$ is a two–sided Brownian motion. However, simulating via (5.2) faces the same difficulty as above because one has to truncate the integral to a finite range, thereby destroying long–range dependence. A representation without this difficulty is

$$B_H(t) = \int_0^t K(t,y) \, dW(y) \tag{5.3}$$

where

$$\begin{aligned} K(t,s) &= C'_H \, s^{1/2-H} \int_s^t u^{H-1/2}(u-s)^{H-3/2} \, du, \\ C'_H &= \sqrt{\frac{H(2H-1)}{\mathrm{Beta}(H-1/2, 2-2H)}} \end{aligned}$$

(cf. Norros, Valkeila & Virtamo [116]). However, we do not know of practical implementations of simulating via (5.3).

Some further references on aspects of simulating fractional Brownian motion are Mandelbrot [101] (a fast ad hoc method), Abry & Sellan [1] (wavelets), Paxson [117] (FFT), Norros, Mannersalo & Wang [115] (bisection with truncated memory), and Michna [103] and Huang *et al.* [81], [82] (importance sampling methods).

Chapter IX

Selected topics

1 Special algorithms for the GI/G/1 queue

1a Exact simulation of W

We consider a random walk maximum M in the notation of the Siegmund algorithm in IV.2a. Recall in particular that F_γ denotes the probability measure with density $e^{\gamma x}$ w.r.t. F where $\gamma > 0$ is the solution of $\hat{F}[\gamma] = 1$, that $\tau(x) = \inf\{n : S_n > x\}$, $\xi(x) = S_{\tau(x)} - x$, and that we have the representation $\mathbb{P}(M > x) = \mathbb{E}_\gamma e^{-\gamma S_{\tau(x)}}$ where $\mathbb{P}_\gamma$ is the probability measure such that the X_n are i.i.d. with distribution F_γ w.r.t. $\mathbb{P}_\gamma$.

An algorithm for exact simulation of M (which has the same distribution as the GI/G/1 waiting time W subject to a suitable choice of F) was suggested by Ensor & Glynn [49]. It uses an exponential r.v. V with rate γ (independent of $\{S_n\}$) and the ladder heights $S_{\tau_+(n)}$ where

$$\tau_+(0) = 0, \quad \tau_+(n+1) = \inf\left\{k > \tau_+(n) : S_k > S_{\tau_+(n)}\right\}.$$

Cf. Fig. IX.1.1 where the ladder heights are marked with a $\bullet$.
The r.v. generated by the simulation is the last ladder height

$$Z \;=\; \sup\left\{S_{\tau_+(n)} : S_{\tau_+(n)} \le V\right\}$$

not exceeding V, see again Fig. IX.1.1.

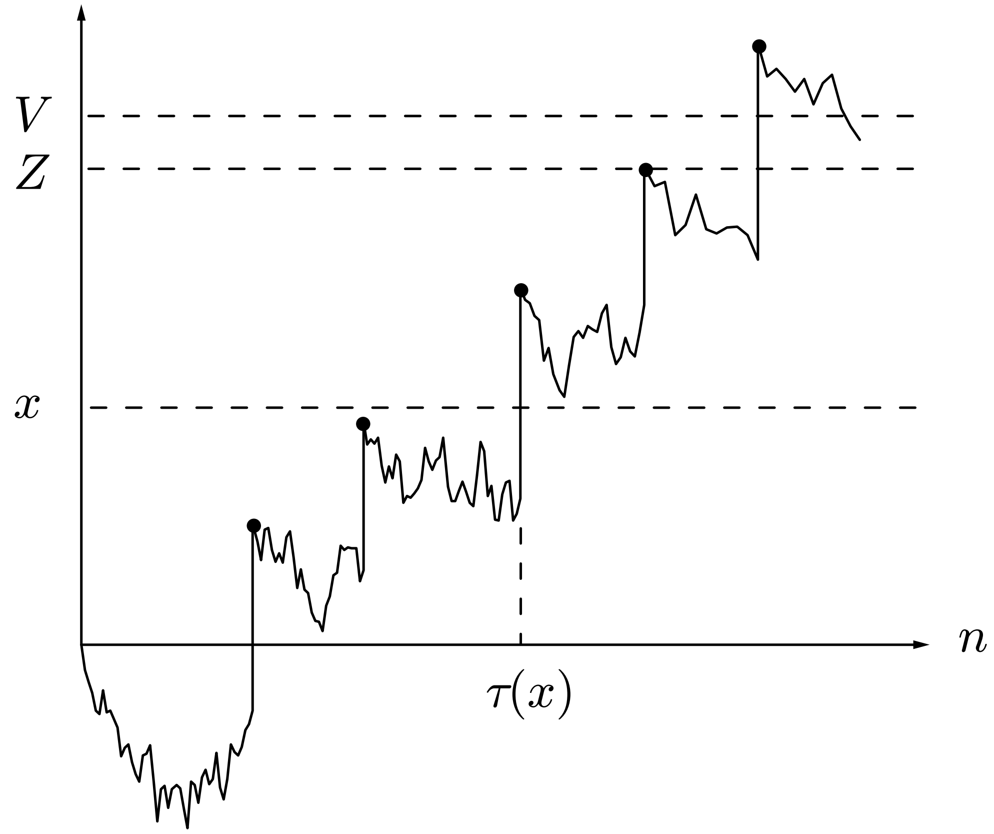

Figure IX.1.1

Proposition 1.1 *The* $\mathbb{P}_\gamma$*–distribution of* Z *is the same as the* $\mathbb{P}$*–distribution of* M*.*

Proof First note that $S_{\tau(x)}$ is necessarily a ladder height. By sample path inspection, $Z > x$ if and only $S_{\tau(x)} \le V$ so that

$$\begin{aligned}\mathbb{P}_\gamma(Z > x) &= \mathbb{P}_\gamma(S_{\tau(x)} \le V) = \mathbb{E}_\gamma\left[\mathbb{P}_\gamma\left(S_{\tau(x)} \le V \,\middle|\, S_{\tau(x)}\right)\right] \\ &= \mathbb{E}_\gamma e^{-\gamma S_{\tau(x)}} = \mathbb{P}(M > x).\end{aligned}$$

□

From exact simulation of the waiting time $W \stackrel{\mathcal{D}}{=} M$, one can easily perform exact simulation also of the steady–state queue length Q. This follows from the *distributional Little's law* stating that Q has the same distribution as $N(W+U)$ where W, U, $\{N(t)\}$ are independent, U has the service time distribution and $\{N(t)\}$ is a version of the renewal arrival process. Thus, if

Z is generated as above and $T_1, T_2, \ldots$ are independent interarrival times, the r.v.

$$\sup n = 0, 1, 2, \ldots : T_1 + \cdots + T_n > Z + U$$

has the same distribution as Q.

For the M/G/1 queue, an obvious alternative exact simulation estimator for W comes from the Pollaczek–Khintchine formula: W has the same distribution as $U_1^* + \cdots + U_N^*$ where N, U_1^*, U_2^* are independent, N is geometric with $\mathbb{P}(N = n) = (1 - \rho)\rho^n$, $n = 0, 1, 2, \ldots$, and the U_k^* have the equilibrium service time distribution. In contrast to the Ensor–Glynn estimator, this estimator can also be used in the case of heavy tails.

1b The Minh–Sorli algorithm

A classical formula for the mean delay due to Marshall is

$$\mathbb{E}W \;=\; \frac{\mathbb{E}U^2 + \mathbb{E}T^2 - 2\mathbb{E}U\mathbb{E}T}{2(\mathbb{E}T - \mathbb{E}U)} - \frac{\mathbb{E}I^2}{2} \tag{1.1}$$

where U, T are generic service– and interarrival times and I the idle period. This is obtained from the recursion $W_{n+1} = (W_n + U_n - T_n)^+$ by squaring: since

$$\begin{aligned}
\mathbb{E}(W_n + U_n - T_n)^2 &= \mathbb{E}\left[(W_n + U_n - T_n)^+ - (W_n + U_n - T_n)^-\right]^2 \\
&= \mathbb{E}(W_n + U_n - T_n)^{+^2} + \mathbb{E}(W_n + U_n - T_n)^{-^2}
\end{aligned}$$

and $(W_n + U_n - T_n)^-$ can be identified with I, we get (assuming $\{W_n\}$ to be stationary)

$$\begin{aligned}
\mathbb{E}W^2 &= \mathbb{E}(W + U - T)^2 - \mathbb{E}I^2 \\
&= \mathbb{E}W^2 + \mathbb{E}U^2 + \mathbb{E}T^2 - 2\mathbb{E}U\mathbb{E}T + 2\mathbb{E}W(\mathbb{E}U - \mathbb{E}T) - \mathbb{E}I^2,
\end{aligned}$$

and (1.1) follows.

Minh & Sorli [104] suggested to use (1.1) for simulation by noting that everything is known except for $\mathbb{E}I^2$. Thus, one can simply simulate a busy cycle, let $Z = I^2$ and use the CMC method.

As ρ approaches 1, the first term in (1.1) becomes dominant, and hence one expects the possible variance reduction to be most substantial when ρ is close to 1. Indeed, in the case of the M/M/1 queue with traffic intensity $\rho = 0.9$, the variance reduction compared to regenerative simulation is reported in [8] to be about a factor of 2.000.

1c A control variate method for $\mathbb{E}W$

The problem is to estimate $z = \mathbb{E}W$ by simulation. Of course, standard methods like regenerative simulation apply in a straightforward way to this problem. For a more sophisticated algorithm, suggested in Asmussen [8], note first the formula

$$\mathbb{E}M \;=\; \int_0^\infty \mathbb{P}(M > x)dx \;=\; \int_0^\infty \mathbb{P}(\tau(x) < \infty)dx\,,$$

where $\tau(x) = \inf\{n : S_n > x\}$. Here an extremely efficient (at least for large x) estimator for $\mathbb{P}(\tau(x) < \infty)dx$ is provided by Siegmund's algorithm, viz. $e^{-\gamma x}e^{-\gamma\xi(x)}$, so that it is appealing to try the estimator

$$\int_0^\infty e^{-\gamma x}e^{-\gamma\xi(x)}\,dx$$

for $\mathbb{E}M$. The obvious difficulty is that evaluating $\xi(x)$ for all x would require an infinitely long simulation. This can be circumvented by truncating the integral and suitably compensate. I.e., let $V > 0$ be independent of $\{S_n\}$ and define

$$Z \;=\; \int_0^V \frac{1}{\mathbb{P}(V > x)}e^{-\gamma x}e^{-\gamma\xi(x)}\,dx\,.$$

Then indeed

$$\begin{aligned}
\mathbb{E}Z \;&=\; \mathbb{E}\int_0^\infty I(x < V)\frac{1}{\mathbb{P}(V > x)}e^{-\gamma x}e^{-\gamma\xi(x)}\,dx \\
&=\; \int_0^\infty \mathbb{P}(x < V)\frac{1}{\mathbb{P}(V > x)}e^{-\gamma x}\mathbb{E}e^{-\gamma\xi(x)}\,dx \\
&=\; \int_0^\infty e^{-\gamma x}\mathbb{E}e^{-\gamma\xi(x)}\,dx \;=\; \int_0^\infty \mathbb{P}(\tau(x) < \infty\,dx \;=\; \mathbb{E}M\,.
\end{aligned}$$

Simulation experiments indicate that the variance of Z is reasonably small but not extremely small. We improve this by introducing

$$C \;=\; \int_0^V \frac{1}{\mathbb{P}(V > x)}e^{-\gamma x}\,dx$$

as control variate; note that by the same calculation as for $\mathbb{E}Z$,

$$\mathbb{E}C \;=\; \int_0^\infty e^{-\gamma x}\,dx \;=\; \frac{1}{\gamma}\,.$$

The control variate estimator is

$$\overline{Z} - \overline{\alpha}(\overline{C} - \mathbb{E}C) \tag{1.2}$$

where

$$\begin{aligned}
\overline{Z} \;&=\; \frac{1}{n}(Z_1 + \cdots + Z_n), \quad \overline{C} \;=\; \frac{1}{n}(C_1 + \cdots + C_n)\,, \\
\overline{C} \;&=\; \frac{\sum(Z_i - \overline{Z})(C_i - \overline{C})}{\sum(C_i - \overline{C})^2}\,.
\end{aligned}$$

Indeed the estimator (1.2) does the job: the observed variance reduction is often a factor of 1.000–25.000!!

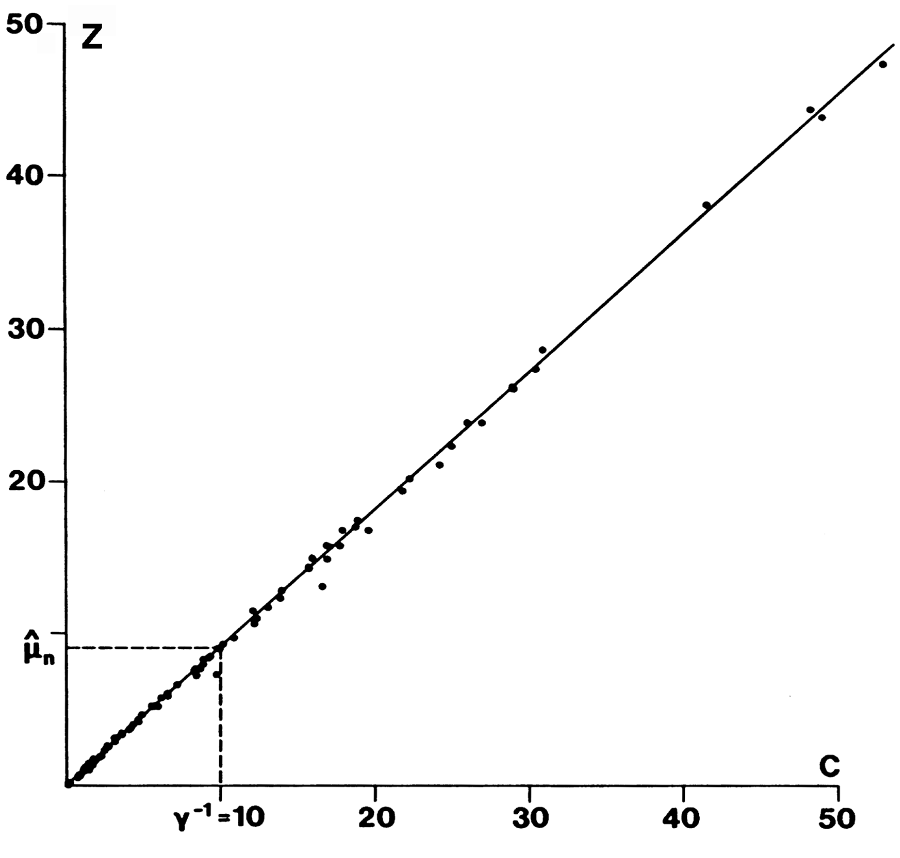

Figure IX.1.2

The relevant choice of the distribution of V turns out to be the exponential distribution with rate γ, $\mathbb{P}(V > x) = e^{-\gamma x}$, and it can be proved that this choice is asymptotically optimal in a suitable sense.

Figure IX.1.2 gives an example of the high linear dependence between Z and C: the correlation is 0.999!!

In the setting of the Ensor–Glynn algorithm in Section 1a, a related idea is to apply V as control for the r.v. Z generated by exact simulation.

2 Exponential change of measure in Markov–modulated models

Consider first the discrete time case and let $J = \{J_n\}_{n=0,1,2,\ldots}$ be an irreducible Markov chain with a finite state space E. A Markov additive process (MAP) $\{S_n\}_{n=0,1,2,\ldots}$ is an extension of random walks, defined as $S_n = Y_1 + \ldots + Y_n$ where the Y_n are conditionally independent given J such that the distribution of Y_n is $H^{(ij)}$ given $\{J_{n-1} = i, J_n = j\}$. The fundamental parameters of a MAP are thus the $H^{(ij)}$ and the transition matrix $\boldsymbol{P} = (p_{ij})_{i,j\in E}$ of J or, equivalently, the $F^{(ij)} = p_{ij}H^{(ij)}$; note that

$$F^{(ij)}(\infty) = p_{ij}, \quad F^{(ij)}(y) = \mathbb{P}_i(X_1 \le y, J_1 = j).$$

Of models where discrete time MAP's play an important role, we mention in particular Markov chain with transition matrices of GI/M/1 or M/G/1 type, see Neuts [111], [112] or [6] Ch. X.4.

The generalization of the m.g.f. is the $E \times E$ matrix $\hat{\boldsymbol{F}}[\theta]$ with ijth element $\hat{F}^{(ij)}[\theta]$, and as generalization of the cumulant g.f. one can take the Perron–Frobenius eigenvalue $\kappa(\theta)$ of $\hat{\boldsymbol{F}}[\theta]$; denote the corresponding right eigenvector by $\boldsymbol{h}^{(\theta)} = \left(h_i^{(\theta)}\right)_{i\in E}$, i.e. $\hat{\boldsymbol{F}}[\theta]\boldsymbol{h}^{(\theta)} = e^{\kappa(\theta)}\boldsymbol{h}^{(\theta)}$. The ECM corresponding to θ is then given by

$$\tilde{\boldsymbol{P}} = e^{-\kappa(\theta)}\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}^{-1}\hat{\boldsymbol{F}}[\theta]\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}, \quad \tilde{H}_{ij}(dx) = \frac{e^{\theta x}}{\hat{H}_{ij}[\theta]}H_{ij}(dx),$$

and the OECM by taking $\theta = \gamma$ where γ is the solution of $\kappa(\gamma) = 0$. Here $\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}$ is the diagonal matrix with the $h_i^{(\theta)}$ on the diagonal. The likelihood ratio is

$$W_n(\boldsymbol{P}|\tilde{\boldsymbol{P}}) \;=\; \frac{h^{(\theta)}(J_0)}{h^{(\theta)}(J_n)}e^{-\theta S_n + n\kappa(\theta)}. \tag{2.3}$$

In continuous time, a MAP with an underlying finite Markov process $\{J_t\}_{t\geq 0}$ has a simple description, cf. e.g. Neveu [113]. The clue for the understanding is the structure of a continuous time random walk (process with stationary independent increments) as the independent sum of a deterministic drift, a Brownian component and a pure jump (Levy) process, see e.g. [6] Ch. III.8. Let the intensity matrix of $\{J_t\}$ be $\boldsymbol{\Lambda} = (\lambda_{ij})_{i,j\in E}$. On an interval $[t, t+s)$ where $J_t \equiv i$, the MAP then $\{S_t\}$ evolves like a process with stationary independent increments with the drift μ_i, the variance σ_i^2 of the Brownian component and the Levy measure $\nu_i(dx)$ depending on i. In addition, a jump of $\{J_t\}$ from i to $j \neq i$ has probability q_{ij} of giving rise to a jump of $\{S_t\}$ at the same time, the distribution of which has then some distribution $B^{(ij)}$.

Let $\hat{\boldsymbol{F}}_t[\theta]$ be the matrix with ijth element $\mathbb{E}_i\left[e^{\theta S_t}; J_t = j\right]$. It is easy to see that $\hat{\boldsymbol{F}}_t[\theta] = e^{t\boldsymbol{G}[\theta]}$, where

$$G^{(ij)}[\theta] \;= \left\{ \begin{array}{cc} q_{ij}\lambda_{ij}\hat{B}^{(ij)}[\theta] + (1-q_{ij})\lambda_{ij} & i \neq j \\ \lambda_{ii} + \mu_i\theta + \frac{1}{2}\sigma_i^2\theta^2 + \int_0^\infty e^{\theta x}\nu_i(dx) & i = j \end{array} \right. .$$

We define $\kappa(\theta)$ as the dominant eigenvalue of $\boldsymbol{G}[\theta]$ and $\boldsymbol{h}^{(\theta)}$ as the corresponding right eigenvector. Equivalently, $e^{t\kappa(\theta)}$ is the Perron–Frobenius eigenvalue of $\hat{\boldsymbol{F}}_t[\theta]$ and $\boldsymbol{h}^{(\theta)}$ the right eigenvector. The ECM corresponding to θ is then given by

$$\tilde{\boldsymbol{\Lambda}} = \boldsymbol{\Delta}^{-1}_{\boldsymbol{h}^{(\theta)}}\boldsymbol{G}[\theta]\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}} - \kappa(\theta)\boldsymbol{I}, \quad \tilde{\mu}_i = \mu_i + \theta\sigma_i^2, \quad \tilde{\sigma}_i^2 = \sigma_i^2,$$

$$\nu_i(dx) = e^{\theta x}\nu_i(dx), \;\; \tilde{q}_{ij} = \frac{\lambda_{ij}q_{ij}\hat{B}_{ij}[\theta]}{\lambda_{ij} + \lambda_{ij}q_{ij}(\hat{B}_{ij}[\theta]-1)}, \;\; \tilde{B}_{ij}(dx) = \frac{e^{\theta x}}{\hat{B}_{ij}[\theta]}B_{ij}(dx).$$

In particular, the expression for $\tilde{\boldsymbol{\Lambda}}$ means

$$\tilde{\lambda}_{ij} = \frac{h_j^{(\theta)}}{h_i^{(\theta)}}\lambda_{ij}\left[1 + q_{ij}(\hat{B}_{ij}[\theta]-1)\right], \quad i \neq j$$

(the diagonal elements are determined by $\lambda_{ii} = -\sum_{j\neq i}\lambda_{ij}$), and if $\nu_i(dx)$ is compound Poisson, $\nu_i(dx) = \beta_i B_i(dx)$ with $\beta_i < \infty$ and B_i a probability measure, then also $\tilde{\nu}_i(dx)$ is compound Poisson with

$$\tilde{\beta}_i = \beta_i\hat{B}_i[\theta], \quad \tilde{B}_i(dx) = \frac{e^{\theta x}}{\hat{B}_i[\theta]}B_i(dx). \tag{2.4}$$

The likelihood ratio on $[0, T]$ is just (2.3) with n replaced by T.

Example 2.1 If all $\sigma_i^2 = 0$, $\nu_i = 0$, $q_{ij} = 0$, we have a process with piecewise linear sample paths with slope μ_i when $J_t = i$. This process (or rather its reflected version, cf. Example 2.4) is a *Markovian fluid*, a process of considerable current interest because of its relevance for ATM technology; for some recent references, see Asmussen [10] and Rogers [127] . The ECM just means to replace λ_{ij} with $\tilde{\lambda}_{ij} = h_j^{(\theta)}\lambda_{ij}/h_i^{(\theta)}$ for $i \neq j$ and is therefore another Markovian fluid. For some special structures like independent sources, the eigenvalue problem determining the OECM (i.e. γ) can be reduced quite a lot using the concept of *effective bandwidth*. □

Example 2.2 Assume that all $\sigma_i^2 = 0$, $r_i = -1$, $q_{ij} = 0$, and that $\nu_i = \beta_i B_i(dx)$ corresponds to the Poisson case with the B_i concentrated on $(0, \infty)$. This process (or rather its reflected version) corresponds to the work–load process in the Markov–modulated M/G/1 queue with arrival intensity β_i and service time distribution B_i of the customer arriving when $J_t = i$, and the ECM just means to replace these parameters by the ones given by (2.4). □

Example 2.3 Assume that $\sigma_i^2 = 0$, $r_i = 0$, $\nu_i = 0$, $q_{ij} = 0$ for all i and that $B^{(ij)}$ is concentrated on $\{0, 1\}$ for all i, j. Then the MAP is a counting process, in fact the same as the Markovian point process introduced by Neuts [110] and increasingly popular as a modeling tool. □

ECM for Markov additive processes goes back to a series of papers by Keilson & Wishart and others in the sixties, e.g. [86]. To our knowledge, the first use of the concept in simulation is Asmussen [7]. Further recent references are Bucklew [32], Bucklew *et al.* [33], Lehtonen & Nyrhinen [98] and Chang *et al.* [36]. Again, some of the most interesting applications involve combination with duality ideas, which for infinite buffer problems just means time reversal.

Example 2.4 Let S_t be the MAP described in Example 2.1. Then the fluid model of interest is

$$V_t \;=\; S_t - \min_{0 \le u \le t} S_u.$$

Define the cycle as $C = \inf\{t > 0 : S_t = 0\}$ (this definition is only interesting if $J_0 = i$ with $r_0 > 0$) and assume that the rare event $A(x)$ is the event $\left\{\sup_{0 \le t < C} V_t \ge x\right\}$ of buffer overflow within the cycle. Then (noting that $S_t = V_t$ for $t < C$) we can just perform the simulation by performing OECM for the MAP $\{S_t\}$ and running it up to it hits either x or 0. If instead $A(x) = \{V \geq x\}$ is defined in terms of the steady state, we first note the well–known representation ([11] and references there) $V \stackrel{\mathcal{D}}{=} \max_{0 \leq t < \infty} \tilde{S}_t$ where $\left\{\tilde{S}_t\right\}$ is the MAP we obtain by time–reversing $\{J_t\}$ (replacing λ_{ij} by $\tilde{\lambda}_{ij} = \pi_j \lambda_{ij} / \pi_i$ where π is the stationary distribution), leaving the r_i unchanged and letting $\tilde{J}_t$ have distribution π. Thus $\alpha(x) = \mathbb{P}(\tilde{\tau}(x) < \infty)$ where $\tilde{\tau}(x) = \inf\left\{t : \tilde{S}_t \geq x\right\}$ and the simulation is performed by running $\left\{\tilde{S}_t\right\}$ until it hits x. Similar remarks apply to Example 2.3. □

The approach can to some extent be generalized beyond finite E. An example is given in Section 8. Note however, that if E is infinite, a MAP may be quite complicated (an example is provided by the local time of a diffusion) and that the existence of dominant eigenvalues for the relevant integral operator does not always hold.

3 Further examples of change of measure*

Girsanov's formula*

Many–server queues*

Sadowsky [139], Sadowsky & Spankowsky [141].

Local exponential change of measure*

Cottrell, Fort & Malgoyres [38], Asmussen & Nielsen [16].

Queueing networks*

Appendix

A1 Some central limit theory

Anscombe's theorem

Proposition A.1 *Let $X_1, X_2, \dots$ be i.i.d. with mean μ and variance σ^2, let $S_n = X_1 + \cdots + X_n$ and let $T_n \in \mathbb{N}$ be a sequence of random times satisfying $T_n/t_n \xrightarrow{\mathbb{P}} 1$ for some deterministic sequence $\{t_n\}$ with $t_n \to \infty$. Then $(S_{T_n} - T_n\mu)/t_n^{1/2} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right)$, $n \to \infty$.*

For a proof, see e.g. Chung [37].

Asymptotic equivalence

Often two estimators $\hat{z}_t^{(1)}$, $\hat{z}_t^{(2)}$ for the same number z turn out to be close modifications of each other so that asymptotically, it is unimportant whether to use one or the other. Our formal framework in such situations is the following:

Definition A.2 *Let $\hat{z}_t^{(1)}$, $\hat{z}_t^{(2)}$ be defined on the same probability space, and assume that $\sqrt{t}(\hat{z}_t^{(1)} - z) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_1^2\right)$, $\sqrt{t}(\hat{z}_t^{(2)} - z) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_2^2\right)$ as $t \to \infty$, where $\sigma_1^2 > 0$, $\sigma_1^2 > 0$. Then $\hat{z}_t^{(1)}$, $\hat{z}_t^{(2)}$ are* asymptotically equivalent *if $\sqrt{t}\left(\hat{z}_t^{(1)} - \hat{z}_t^{(2)}\right) \xrightarrow{\mathbb{P}} 0$ as $t \to \infty$.*

It follows easily that then $\sigma_1^2 = \sigma_2^2$.

A2 Exponential change of measure: the i.i.d. case

Let $X_1, X_2, \dots$ be i.i.d. with common distribution F having m.g.f.

$$\hat{F}[s] \; = \; \mathbb{E}e^{sX} \; = \; \int_{-\infty}^{\infty} e^{sx} F(dx)\,.$$

The shape of $\hat{F}$ is as on Fig. A.1 depending on the value of $\mathbb{E}X$:

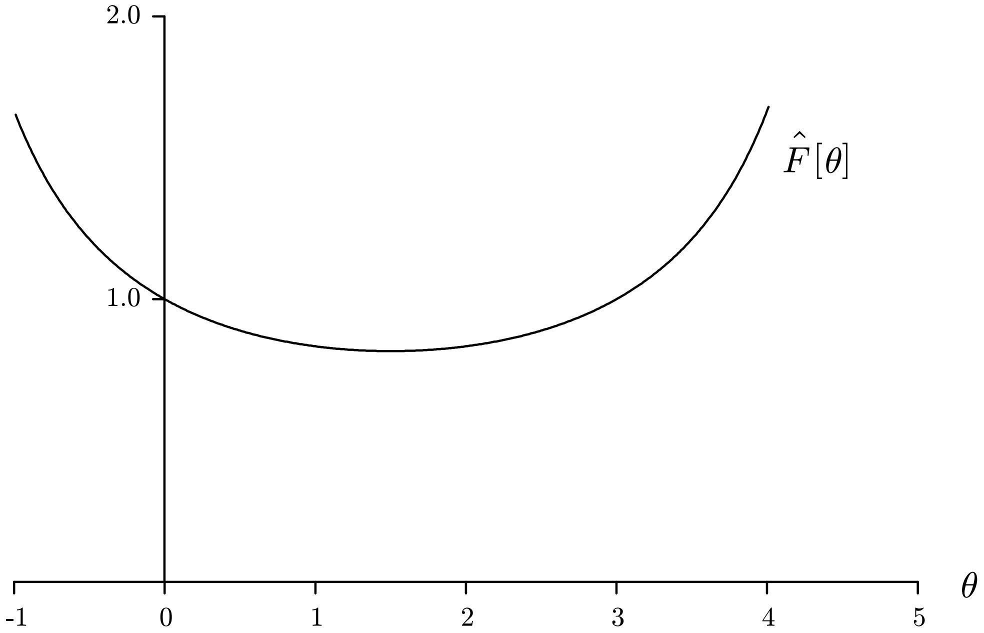

Figure A.1.a $\mu < 0$

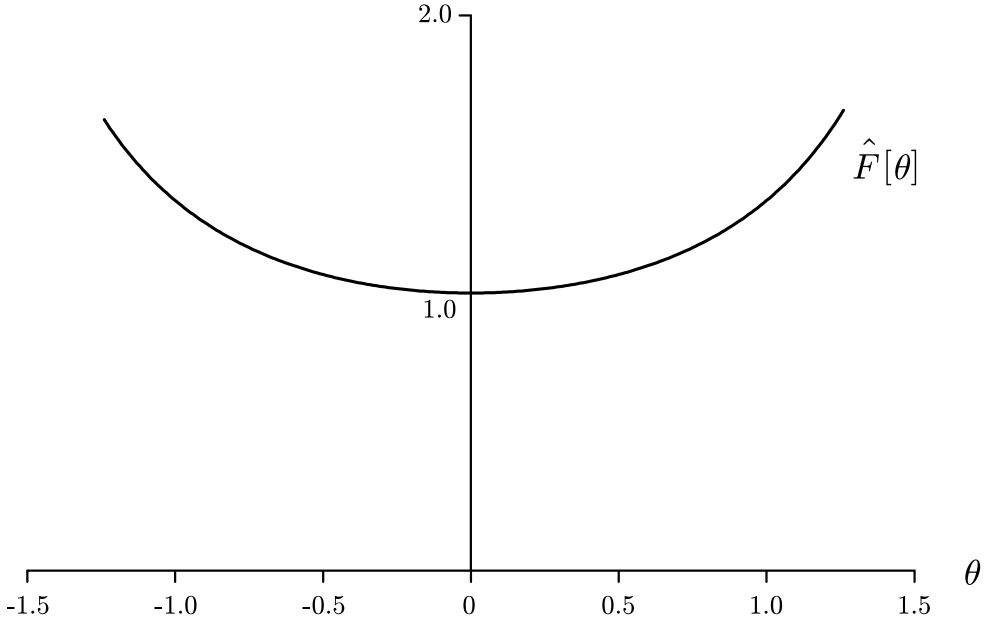

Figure A.1.b $\mu = 0$

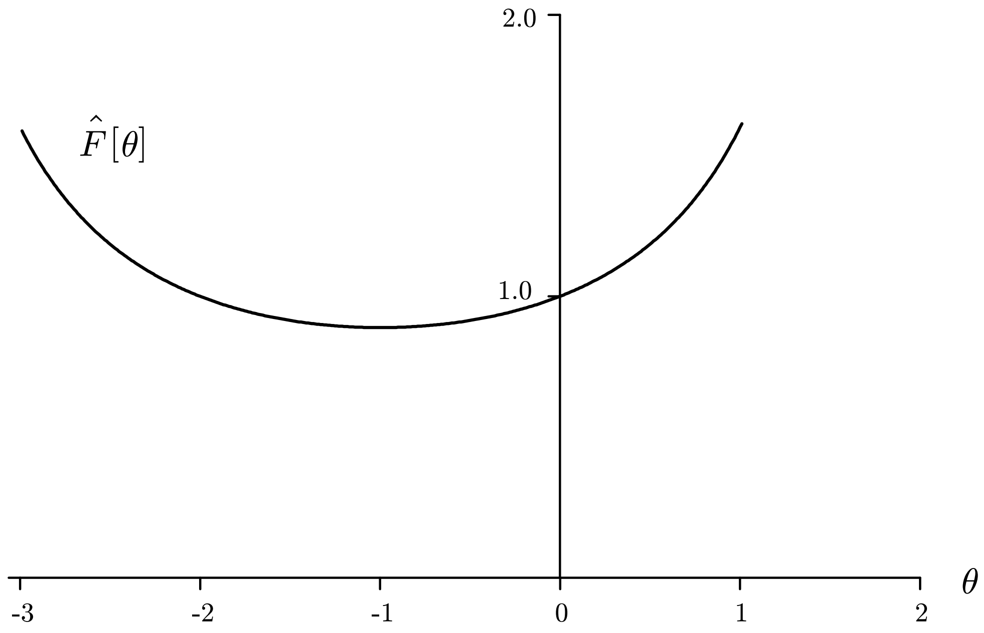

Figure A.1.c $\mu > 0$

It is well–known that $\hat{F}$ is convex and logaritmically convex.
The exponential family $\{F_\theta\}_{\theta \in \Theta}$ generated by F is defined by

$$\Theta \;=\; \Bigl\{\theta \in \mathbb{R}:\ \hat{F}[\theta] < \infty\Bigr\}, \quad \frac{dF_\theta}{dF}(x) \;=\; \frac{e^{\theta x}}{\hat{F}[\theta]}$$

(Radon–Nikodym derivative). In particular, if F has density $f(x)$, then the density of F_θ is

$$f_\theta(x) \;=\; \frac{e^{\theta x} f(x)}{\hat{F}[\theta]}\,.$$

The m.g.f. $\hat{F}_\theta[s]$ of F_θ is

$$\begin{aligned}\hat{F}_\theta[s] \;&=\; \int e^{sx}\, F_\theta(dx) \;=\; \int e^{sx}\,\frac{dF_\theta}{dF}(x)\, F(dx) \;=\; \int e^{sx+\theta x}\,\frac{1}{\hat{F}[\theta]}\, F(dx)\\ &=\; \frac{\hat{F}[s+\theta]}{\hat{F}[\theta]}.\end{aligned}$$

The likelihood ratio $L_{n,\theta}$ up to time n, as defined by the requirement

$$\mathbb{E}g(X_1,\dots,X_n) \;=\; \mathbb{E}_\theta\left[L_{n,\theta}g(X_1,\dots,X_n)\right] \tag{A2.1}$$

for all well–behaved g is

$$L_{n,\theta} \;=\; \frac{f(X_1)}{f_\theta(X_1)}\cdots\frac{f(X_n)}{f_\theta(X_n)} \;=\; e^{-\theta S_n}\hat{F}[\theta]^n\,.$$

Equation (A2.1) tells how to compute expectations involving F in terms of expectations involving F_θ over a fixed time horizon. The method carries over to stopping times. Let $\mathcal{F}_n = \sigma(X_1,\dots,X_n)$.

Proposition A.1 (Wald's fundamental identity of sequential analysis) *For any stopping time $\tau \in [0,\infty]$ and any event $A \in \mathcal{F}_\tau$ with $A \subseteq \{\tau < \infty\}$,*

$$\mathbb{P}(A) \;=\; \mathbb{E}_\theta\left[L_{\tau,\theta};\, A\right]$$

Proof Taking $g(X_1,\dots,X_n) = I(\tau = n)I(A)$, we get

$$\mathbb{P}(\tau = n;\, A) \;=\; \mathbb{E}_\theta\left[L_{n,\theta};\, \tau = n;\, A\right] \;=\; \mathbb{E}_\theta\left[L_{\tau,\theta};\,;\, \tau = n;\, A\right]\,.$$

Summing over n, the conclusion follows. □

A3 Lévy– and stable processes

A *Lévy process* $\{X(t)\}_{t\ge 0}$ is defined as a continuous time process on $\mathbb{R}$ with stationary independent increments and $X(0) = 0$. This class of processes is in one–one correspondance with the class of infinitely divisible distributions via the distribution of $X(1)$, see below.

A Lévy process can be written as an independent sum $X(t) = ct + \sigma W(t) + J(t)$ of a linear drift ct, a Brownian component $\sigma W(t)$, and a jump process $\{J(t)\}$ given in terms of its Lévy measure $\nu(dx)$ as will next be explained. Since we have discussed Brownian motion separately, we consider here only the case $\sigma^2 = 0$.

The Lévy measure ν can be any non–negative measure on $\mathbb{R}$ satisfying $\nu(\{0\}) = 0$ and

$$\int_{-\infty}^{\infty}(y^2 \wedge 1)\,\nu(dy) \;<\; \infty. \tag{A3.1}$$

Equivalently, $\int_{|y|>\epsilon}\nu(dy)$ and $\int_{-\epsilon}^{\epsilon}y^2\,\nu(dy)$ are finite for some (and then all) $\epsilon>0$. A particular important case is stable processes, where $X(1)$ has an α–stable $S_\alpha(\sigma,\beta,0)$ distribution corresponding to

$$\frac{d\nu}{dy} \;=\; \begin{cases} \frac{C_+}{y^{\alpha+1}} & y>0 \\ \frac{C_-}{|y|^{\alpha+1}} & y<0 \end{cases} \tag{A3.2}$$

with C_+,C_- depending on σ,β, see further below.

The rough description of $\{J(t)\}$ is that jumps of size x occur at intensity $\nu(dx)$. In particular, if ν has finite mass $\lambda=\int_{-\infty}^{\infty}\nu(dy)$, then $\{J(t)\}$ is a compound Poisson process with intensity λ and jump size distribution $\nu(dy)/\lambda$. In general, for any interval I not having 0 as an endpoint, the sum of the jumps of size $\in I$ in $[s,s+t)$ is a compound Poisson r.v. with intensity $\lambda_I=\int_I\nu(dy)$ and jump size distribution $\nu(dy)I(y\in I)/\lambda_I$. Jumps in disjoint intervals are independent, and so we can describe the totality of jumps by the points in a planar Poisson process $N(dy,dt)$ with intensity measure $\nu(dy)\otimes dt$. A point of N at (Y_i,T_i) then corresponds to a jump of size Y_i at time T_i for $\{J(t)\}$. If in addition to (A3.1) one has

$$\int_{-\infty}^{\infty}(|y|\wedge 1)\,\nu(dy) \;<\; \infty. \tag{A3.3}$$

(this is equivalent to the paths of $\{J(t)\}$ to be of finite variation), one can simply write

$$J(t) \;=\; \int_{\mathbb{R}\times[0,t]} yN(dy,dt). \tag{A3.4}$$

If (A3.3) fails, this Poisson integral does not converge absolutely, and $\{J(t)\}$ has to be defined by a *compensation* (centering) procedure. For example, letting

$$Y_0(t) \;=\; \int_{\{y:\,|y|>1\}\times[0,t]} yN(dy,dt),\quad Y_n(t) \;=\; \int_{|y|\in(y_{n+1},y_n]} yN(dy,dt),$$

one can let

$$J(t) \;=\; Y_0(t)+\sum_{n=1}^{\infty}\{Y_n(t)-\mathbb{E}Y_n(t)\} \tag{A3.5}$$

where $1=y_1>y_2>\ldots\downarrow 0$ (note that $\mathbb{E}Y_n(t)=t\int_{y_{n+1}}^{y_n}y\,\nu(dy)$). The series converges absolutely a.s. since

$$\sum_{n=1}^{\infty}\mathbf{Var}(Y_n(t)) \;=\; \sum_{n=1}^{\infty}t\int_{|y|\in(y_{n+1},y_n]}y^2\nu(dy) \;=\; \int_{-1}^{1}y^2\nu(dy) \;<\; \infty,$$

and the sum is easily seen to be independent of the particular partitioning $\{y_n\}$. But note that since the role of the interval $[-1, 1]$ is arbitrary, a compensated Lévy jump process is only given canonically up to a drift term.

If $J(t) \geq 0$ for all $t \geq 0$, then $\{J(t)\}$ is called a subordinator. The Lévy measure for a subordinator necessarily satisfies (A3.3), and any Lévy jump process satisfying (A3.3) can be written as the independent difference between two subordinators, defined in terms of the restriction of ν to $(0, \infty)$, resp. the reflection of the restriction of ν to $(-\infty, 0)$.

The property of stationary independent increments implies that $\log \mathbb{E} e^{sX(t)}$ has the form $t\varphi(s)$. Here $\varphi(s)$ is called the *Lévy exponent*; its domain includes the imaginary axis $\Re s = 0$ and frequently larger areas depending on properties of ν, say $\{s : \Re s \leq 0\}$ in the case of a subordinator. Thus, $\varphi(s)$ is the cumulant g.f. of an infinitely divisible distribution, having Lévy–Khintchine representation

$$\varphi(s) \;=\; cs + \frac{\sigma^2 s^2}{2} \;+\; \int_{-\infty}^{\infty} \left(e^{sy} - 1 - \frac{sy}{1+y^2} \right) \nu(dy) \qquad \text{(A3.6)}$$

Different equivalent forms which are often used are

$$\varphi(s) \;=\; cs + \frac{\sigma^2 s^2}{2} \;+\; \int_{-\infty}^{\infty} \left(e^{sy} - 1 - \frac{sy}{1+y^2} \right) \frac{1+y^2}{y^2} \theta(dy) \qquad \text{(A3.7)}$$

where $\theta(dy) = y^2/(1+y^2)\,\nu(dy)$ can be any non–negative finite measure, or

$$\varphi(s) \;=\; c_1 s + \frac{\sigma^2 s^2}{2} \;+\; \int_{-\infty}^{\infty} (e^{sy} - 1 - syI(|y| \leq 1)\,\nu(dy) \qquad \text{(A3.8)}$$

In the finite variation case (A3.3), $\sigma^2 = 0$, $c = -\int y/(1+y^2)\nu(dy)$, and (A3.6) is often written

$$\varphi(s) \;=\; \int_{-\infty}^{\infty} (e^{sy} - 1)\,\nu(dy). \qquad \text{(A3.9)}$$

Similarly, in the compensated case (A3.5), $c_1 = 0$ in (A3.8) and $\sigma^2 = 0$.

A standard reference for Lévy processes is Bertoin [24]. For stable processes, see Samorodnitsky & Taqqu [142].

Stable distributions and processes

For $1 < \alpha < 2$, $\alpha \neq 1$, the α–stable $S_\alpha(\sigma, \beta, \mu)$ distribution is defined as the distribution with c.g.f. of the form

$$\varphi(s) \;=\; -\sigma^\alpha |s|^\alpha \left(1 - \beta \mathrm{sign}(s/i)) \tan \frac{\pi\alpha}{2} \right) \;+\; s\mu, \quad \Re s = 0$$

for some $\sigma > 0$, $\beta \in [-1, 1]$ and $\mu \in \mathbb{R}$. There is a similar but somewhat different expression, which we omit, when $\alpha = 1$. The reader should note that the theory is somewhat different according to whether $0 < \alpha < 1$, $\alpha = 1$ or $1 < \alpha < 2$.

If the r.v. Y has a $S_\alpha(\sigma, \beta, \mu)$ distribution, then $Y + a$ has a $S_\alpha(\sigma, \beta, \mu + a)$ distribution and aY a $S_\alpha(\sigma|a|, \mathrm{sign(a)}\beta, \mu)$ distribution. Thus, μ is a translation parameter and σ a scale parameter. The interpretation of β is as a skewness parameter, as will be clear from the discussion of stable processes to follow.

A stable process is defined as a Lévy jump process with a Lévy measure of the form (A3.2) and X_1 having a $S_\alpha(\sigma, \beta, 0)$ distribution. If $0 < \alpha < 1$, then (A3.3) holds and the process can be defined by (A3.4). If $1 \leq \alpha < 2$, compensation is needed and care must be taken to choose the drift term to get $\mu = 0$. One can reconstruct β from the Lévy measure as $\beta = (C_+ - C_-)/(C_+ - C_-)$.

A4 Regenerative processes

A stochastic process $\{X_t\}$ (in discrete or continuous time) is called *regenerative* if it consists of i.i.d. cycles. The model example is discrete recurrent Markov chains: fixing some reference state i, a cycle starts in state i, the next at the following visit to i and so on.

For a formal definition, we first consider the *zero–delayed* case. Here $\{X_t\}$ is defined to be regenerative w.r.t. $\{\tau_n\}$, a sequence of i.i.d. r.v.'s, if the τ_n are i.i.d. and > 0, and, more generally, the segments

$$\left\{X_{\tau_1+\cdots+\tau_{n-1}+t}\right\}_{0\leq t<\tau_n} \tag{A4.1}$$

are i.i.d. stochastic processes. Here (A4.1) is called the nth cycle and τ_n its length.

The crucial property of regenerative processes is that nothing more than $\mathbb{E}\tau < \infty$ (which we will assume throughout) is sufficient to ensure ergodicity in the sense that

$$\frac{1}{t}\sum_{n=0}^{t} f(X_n) \ \stackrel{\text{a.s.}}{\to}\ \mathbb{E}f(X)\,, \tag{A4.2}$$

where X is a proper r.v. with distribution given by

$$\mathbb{E}f(X) \;=\; \frac{1}{\mathbb{E}\tau}\mathbb{E}\sum_{n=0}^{\tau-1} f(X_n)\,. \tag{A4.3}$$

If in addition the distribution of τ is aperiodic, then $X_n \to X$ in total variation (in particular, in distribution).

Example A.1 If $\{X_n\}$ is an ergodic finite Markov chain, say with stationary distribution $\boldsymbol{\pi} = (\pi_i)$, we can take the τ_n as the return times to some fixed state i. If f is the indicator of j and we write $\tau(i) = \tau$, (A4.3) then becomes

$$\pi_j \;=\; \frac{1}{\mathbb{E}_i\tau(i)}\mathbb{E}_i\sum_{n=0}^{\tau(i)} I(X_n = j).$$

In particular $(i = j)$, $\pi_i = 1/\mathbb{E}_i\tau(i)$, an expression which is used as the basic expression for the stationary distribution in many textbooks. □

A5 The GI/G/1 queue

The GI/G/1 queue is a single server queue with customers $n = 0, 1, 2, \ldots$, service time U_n of customer n and time T_n between the arrivals of customers n and $n + 1$ (usually, it is assumed that customer 0 arrived at time $t = 0$). The basic GI/G/1 assumption is that the sequences $\{U_n\}$, $\{T_n\}$ are independent and i.i.d., with distribution say B for the service time U and A for the interarrival time T.

The waiting time (delay) W_n of customer n is the time he spends from the arrival until he starts service; in contrast, the sojourn time $W_n + U_n$ is the total time in system. It is easy to see that $\{W_n\}$ satisfies the Lindley recursion

$$W_{n+1} \;=\; (W_n + U_n - T_n)^+. \tag{A5.1}$$

Letting $X_n = U_n - T_n$, $S_n = X_0 + \cdots + X_{n-1}$ is a random walk, and (A5.1) shows that $\{W_n\}$ may be seen as $\{S_n\}$ reflected at 0; equivalently, $W_n = S_n - \min_{k=0,\ldots,n} S_k$ (assuming W_0). With $M_n = \max_{k=0,\ldots,n} S_k$, this implies that $W_n \stackrel{\mathcal{D}}{=} M_n$.

The traffic intensity is defined as $\rho = \mathbb{E}U/\mathbb{E}T$. If $\rho < 1$, the queue is called *stable* and the processes of waiting time, workload, queue lenght etc. then converge in distribution. With W the limit of W_n, we get in particular that $W \stackrel{\mathcal{D}}{=} M$ where $M = \lim M_n = \max_{k=0,1,2,\ldots} S_k$.

A6 Poisson's equation. The fundamental matrix

Let $\{X_n\}$ be an ergodic regenerative Markov chain (possibly on a general state space) with generic cycle C, transition operator P ($Pf(x) = \mathbb{E}_x f(X_1)$) and stationary distribution π ($\pi P = \pi$).

Poisson's equation is $g = f + Pg$. Applying π to both sides, we see that $\pi(f) = 0$ is a necessary condition for existence of a solution. If this condition is met, one can check that a solution is

$$g(x) \;=\; \mathbb{E}_x \sum_{n=0}^{C-1} f(X_n)$$

(uniqueness holds under mild conditions).

Poisson's equation determines both bias– and variance constants in the CLT for sample averages $\hat{f}_t = \sum_0^{t-1} f(X_n)/t$:

$$\mathbb{E}_x \hat{f}_t \;\sim\; \frac{g(x)}{t}, \tag{A6.1}$$

$$\mathbf{Var}_x \hat{f}_t \;\sim\; \frac{\sigma^2}{t} \;\text{ where }\; \sigma^2 = \pi(f^2) + 2\pi(fg). \tag{A6.2}$$

For (A6.1), note that by Proposition 2.1 $t\mathbb{E}_x \hat{f}_t \to \sum_0^\infty \mathbb{E}_x f(X_n) = \tilde{g}(x)$ (say). But obviously

$$\begin{aligned} \tilde{g}(x) \;&=\; f(x) + \sum_{n=1}^{\infty} \mathbb{E}_x f(X_n) \;=\; f(x) + \mathbb{E}_x \mathbb{E}_{X_1} \sum_{n=0}^{\infty} f(X_n) \\ &=\; f(x) + \mathbb{E}_x \tilde{g}(X_1) \;=\; f(x) + P\tilde{g}(x). \end{aligned}$$

For (A6.2), we have by regenerative process theory that

$$\begin{aligned} \mathbb{E}C\,\sigma^2 \;&=\; \mathbb{E}\left(\sum_{n=0}^{C-1} f(X_n)\right)^2 \\ &=\; \mathbb{E}\sum_{n=0}^{C-1} f(X_n)^2 \;+\; 2\,\mathbb{E}\sum_{n=0}^{C-1}\sum_{m=n}^{C-1} f(X_n)f(X_m) \\ &=\; \mathbb{E}C\,\pi(f^2) \;+\; 2\,\mathbb{E}\sum_{n=0}^{C-1} f(X_n)\mathbb{E}\left[\sum_{m=n}^{C-1} f(X_m) \,\middle|\, X_0,\ldots,X_n\right] \\ &=\; \mathbb{E}C\,\pi(f^2) \;+\; 2\,\mathbb{E}\sum_{n=0}^{C-1} f(X_n)g(X_n) \;=\; \mathbb{E}C\,\pi(f^2) \;+\; 2\,\mathbb{E}C\,\pi(fg). \end{aligned}$$

Note that when applying (A6.1), (A6.2), the condition $\pi(f) = 0$ is most often not met and one has to replace f by $f_0 = f - \pi(f)$ and g by the solution to $g = f_0 + Pg$.

Example A.1 Let $\{Y_n\}$ be a finite Markov chain with transition matrix $\boldsymbol{P}$. Representing π as a row vector $\boldsymbol{\pi}$ and f as a column vector $\boldsymbol{f}$, we then have $z = \boldsymbol{\pi f}$.

For solution of Poisson's equation $\boldsymbol{g} = \boldsymbol{f} + \boldsymbol{Pg}$, we get formally that $\boldsymbol{g} = (\boldsymbol{I} - \boldsymbol{P})^{-1}\boldsymbol{f}$. However, since 1 is eigenvalue of $\boldsymbol{P}$, the inverse does not make sense and has to be replaced by a generalized inverse $\boldsymbol{F} = (\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{e\pi})^{-1}$, the *fundamental matrix* (e.g. Kemeny, Snell & Knapp [89]). To verify that $\boldsymbol{g} = \boldsymbol{Ff}$ is a solution, we have to check that

$$\boldsymbol{f} \;=\; (\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{e\pi})\boldsymbol{f} \;+\; (\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{e\pi})\boldsymbol{P}(\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{e\pi})^{-1}\boldsymbol{f}.$$

But since $\boldsymbol{P}$ and $\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{e\pi}$ commute, the last term is just $\boldsymbol{Pf}$, and the assertion follows from $\pi(f) = 0$.

Cf. also Example III.2.2. □

Beyond finite Markov chains, explicit solutions of Poisson's equation have been derived in Glynn [65] for the M/G/1 queue and Bladt [26] for the PH/PH/1 queue.

A7 Sequential tests

Let $Y_1, Y_2, \ldots$ be i.i.d. with common density $f(y)$, and consider the problem of testing

$$H_0: \; f = f_0 \quad \text{versus} \quad H_1: \; f = f_1 \,. \tag{A7.1}$$

For a given fixed n, the usual likelihood ratio test rejects if

$$L_n \;=\; \frac{f_1(Y_1)\ldots f_1(Y_n)}{f_0(Y_1)\ldots f_0(Y_n)}$$

is large, say $L_n > b'_n$. Letting

$$X_k \;=\; \log\left[f_1(Y_k)/f_0(Y_k)\right], \quad S_n \;=\; X_1 + \cdots + X_n, \quad b_n = \log b'_n\,,$$

this means $S_n > b_n$.

The sequential test is formed by fixing a, b and continue observation until time $\tau = \inf\{n : S_n \notin [a, b]\}$. One rejects if $S_\tau > b$ and accepts if

$S_\tau < a$. The level α is the probability of rejecting a true null hypothesis, i.e. $\mathbb{P}_0(S_\tau > b)$.

Note that subject to the null hypothesis, $\{S_n\}$ is a random walk with negative drift:

$$\mathbb{E}_0 X \;=\; \mathbb{E}_0 \log \frac{f_1(Y)}{f_0(Y)} \;<\; 0$$

by the information inequality. Further, the level α is typically in the range 1%–5% so that $\{S_\tau > b\}$ is (moderately) rare.

A standard reference for sequential analysis is Siegmund [149].

A8 Ito's formula

If $\{x(t)\}_{t\geq 0}$ is a solution of the ODE

$$\dot{x}(t) \;=\; a(t,x(t)), \quad \text{i.e.} \quad dx(t) \;=\; a(t,x(t))\,dt$$

and f is a smooth function of two variables (a is assumed smooth as well), then the chain rule gives

$$\begin{aligned} d\,f(t,x(t)) \;&=\; f_t(t,x(t))\,dt \;+\; f_x(t,x(t))\,dx(t) \\ &=\; f_t(t,x(t))\,dt \;+\; f_x(t,x(t))a(t,x(t))\,dt \end{aligned}$$

where $f_t(t,x) = \frac{\partial}{\partial t} f(t,x)$, $f_x(t,x) = \frac{\partial}{\partial x} f(t,x)$ denotes the partial derivatives.

Ito's formula is a similar expression for a function $f(t,X(t))$ of the solution $\{X(t)\}$ of the SDE

$$dX(t) \;=\; a(t,X(t))dt + b(t,X(t))\,dW(t) \qquad \text{(A8.2)}$$

where $\{W(t)\}_{t\geq 0}$ is standard Brownian motion, and states that

$$df(X(t)) \;=\; \left\{af_x + f_t + b^2 f_{xx}\right\} dt \;+\; bf_x\,dW(t) \qquad \text{(A8.3)}$$

where a, b, f_t, f_x and f_{xx} (the second partial derivative w.r.t. x) are evaluated at $(t, X(t))$. The precise meaning of this statement is that (A8.2), (A8.3) should be interpreted as

$$X(t) - X(0) \;=\; \int_0^t a(s,X(s))\,ds \;+\; \int_0^t b(s,X(s))\,dW(s),$$

resp.

$$
\begin{aligned}
&f(X(t)) - f(X(0)) \;= \\
&\quad \int_0^t \left\{ a(s,X(s)) f_x(s,X(s)) + f_t(s,X(s)) + b^2(s,X(s)) f_{xx}(s,X(s)) \right\} ds \\
&\quad + \int_0^t b(s,X(s)) f_x(s,X(s))\, dW(s),
\end{aligned}
$$

where $\int_0^t b(s,X(s))\,dW(s)$ etc. denotes the Ito integral.

The proof of (A8.3) can be found in any standard textbook in stochastic integration. The heuristics is the expression $(dW(t))^2 = dt$ (compare to $(dt)^2 = 0!$), which is motivated from quadratic variation properties of $\{W(t)\}$. Thus, compared to ODE's, one needs to take into account also the term containing f_{xx} in the second order Taylor expansion to correctly include all terms of order dt.

A formula that is often used is

$$
\mathbf{Cov}\left(\int_0^t f(s)\,dW(s)\,,\, \int_0^t g(s)\,dW(s) \right) \;=\; \int_0^t f(s)g(s)\,ds\,. \qquad \text{(A8.4)}
$$

A9 The information inequality

Proposition A.1 (The information inequality) *Let f, g be densities. Then*

$$
\int \log g(x) f(x)\,dx \;\le\; \int \log f(x)\, f(x)\,dx\,,
$$

where $\log 0 \cdot y = 0$, $0 \le y < \infty$. If equality holds, then f and g define the same probability measure.

Proof Let X be a r.v. with density $f(x)$ and write $\mathbb{E}_f$ for the corresponding expectation. Then by Jensen's inequality

$$
\begin{aligned}
&\int \log g(x) f(x)\,dx \;-\; \int \log f(x)\, f(x)\,dx \\
&=\; \mathbb{E}_f \log g(X) - \mathbb{E} \log f(X) \;=\; \mathbb{E}_f \log \frac{g(X)}{f(X)} \\
&\le\; \log\left(\mathbb{E}_f \frac{g(X)}{f(X)} \right) \;=\; \log\left(\int_{\{f>0\}} \frac{g(x)}{f(x)} f(x)\,dx \right) \\
&=\; \log\left(\int_{\{f>0\}} g(x)\,dx \right) \;\le\; \log 1 \;=\; 0\,,
\end{aligned}
$$

where $\{f > 0\} = \{x : f(x) > 0\}$. If equality holds, then $\int_{\{f>0\}} g(x)\,dx = 1$ and $g(X) = f(X)$ a.s. Hence for any A,

$$\begin{aligned}\int_A g(x)\,dx &= \int_{A\cap\{f>0\}} g(x)\,dx \\ &= \mathbb{E}_f\left[\frac{g(X)}{f(X)};\ X \in A\cap\{f>0\}\right] \\ &= \mathbb{P}(X \in A\cap\{f>0\}) \ = \ \int_A f(x)\,dx\,.\end{aligned}$$

$\square$

Bibliography

[1] P. Abry & D. Sellan (1996) The wawelet–based synthesis for the fractional Brownian motion proposed by F. Sellan and Y. Meyer: remarks and fast implementation. *Appl. Comp. Harmonic Anal.* **3**, 337–383.

[2] V. Anantharam (1988) How large delays build up in a $GI/GI/1$ queue. *Queueing Systems* **5**, 345–368.

[3] V. Anantharam, P. Heidelberger and P. Tsoucas (1990) Analysis of rare events in continuous time Markov chains via time reversal and fluid approximation. *Adv. Appl. Prob.* (to appear)..

[4] S. Asmussen (1982) Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the $GI/G/1$ queue. *Adv. Appl. Prob.* **14**, 143–170.

[5] S. Asmussen (1985) Conjugate processes and the simulation of ruin problems. *Stoch. Proc. Appl.* **20**, 213–229.

[6] S. Asmussen (1987) *Applied Probability and Queues,* John Wiley & Sons, New York.

[7] S. Asmussen (1989) Risk theory in a Markovian environment. *Scand. Actuarial J.,* 69–100.

[8] S. Asmussen (1990) Exponential families and regression in the Monte Carlo study of queues and random walks *Ann. Statist* **18**, 1851–1867.

[9] S. Asmussen (1992) Queueing simulation in heavy traffic. *Math. Opns. Res.* **17**, 84–111.

[10] S. Asmussen (1995) Stationary distributions for fluid flow models with or without Brownian noise. *Stochactic Models* **11**, 21-49 (1995).

[11] S. Asmussen (1995) Stationary distributions via first passage times. *Advances in Queueing: Models, Methods & Problems* (J. Dshalalow ed.), 79–102. CRC Press, Boca Raton, Florida.

[12] S. Asmussen & K.. Binswanger (1997) Simulation of ruin probabilities for subexponential claims. *ASTIN Bull.* **27**, 297–318.

[13] S. Asmussen, K. Binswanger & B. Højgaard (1999) Rare events simulation for heavy–tailed distributions. *Bernoulli* (to appear).

[14] S. Asmussen, P.W. Glynn & H. Thorisson (1992) Stationarity detection in the initial transient problem. *ACM TOMACS* **2**, 130-157.

[15] S. Asmussen, P.W. Glynn & J. Pitman (1996) Discretization error in the simulation of one–dimensional reflecting Brownian motion. *Ann. Appl. Probab.* **5**, 875–896.

[16] S. Asmussen & H.M. Nielsen (1995) Ruin probabilities via local adjustment coefficients. *J. Appl. Probab.* **32**, 736–755.

[17] S. Asmussen & R.Y. Rubinstein (1995) Steady–state rare events simulation in queueing models and its complexity properties. *Advances in Queueing: Models, Methods & Problems* (J. Dshalalow ed.), 429–466. CRC Press, Boca Raton, Florida.

[18] S. Asmussen & R.Y. Rubinstein (1999) Sensitivity analysis of insurance risk models via simulation. *Management Science* (to appear).

[19] S. Asmussen & S. Schock Petersen (1988) Ruin probabilities expressed in terms of storage processes. *Adv. Appl. Probab.* **20**, 913-916.

[20] S. Asmussen & K. Sigman (1996) Monotone stochastic recursions and their duals. *Probab. Eng. Inf. Sci.* **10**, 1–20.

[21] F. Baccelli & P. Bremaud (1994). *Elements of Queueing Theory. Palm–Martingale Calculus and Stochastic Recursions*. Springer–Verlag.

[22] J. Banks, J.S. Carson, II, & B.L. Nelson (1996) *Discrete Event–Systems Simulation* (2nd ed.). Prentice–Hall.

[23] A.J. Bayes (1970) Statistical techniques for simulation models. *Austr. Comp. J.* **2**, 180–184.

[24] J. Bertoin (1996) *Lévy Processes*. Cambridge University Press.

[25] N. H. Bingham, C. M. Goldie & J. L. Teugels. *Regular Variation*. Encyclopedia of Mathematics and its Applications **27**. Cambridge University Press, 1987.

[26] M. Bladt (1996) The variance constant for the actual waiting time of the PH/PH/1 queue. *Ann. Appl. Probab.* **6**, 766–777.

[27] B. Blaszczyszyn & K. Sigman (1999) Risk and duality in multidimensions. *Stoch. Proc. Appl.* (to appear).

[28] L. Bondesson (1982) On simulation from infinitely divisible distributions. *Adv. Appl. Probab.* **14**, 855–869.

[29] P. Bratley, B.L. Fox & L. Schrage (1987) *A Guide to Simulation*. Springer, New York.

[30] P.J. Brockwell & R.A. Davis (1991) *Time Series: Theory and Methods* (2nd ed.). Springer, New York.

[31] P.J. Brockwell & R.A. Davis (1996) *Introduction to Time Series and Forecasting*. Springer, New York.

[32] J.A. Bucklew (1990) *Large Deviation Techniques in Decision, Simulation, and Estimation.* John Wiley & Sons, New York.

[33] J.A. Bucklew, P. Ney & J.S. Sadowsky (1990) Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. *J. Appl. Prob.* **27**, 44–59.

[34] E. Carlstein (1986) The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.* **14**, 1171–1179.

[35] J.M. Chambers, C.L. Mallows & B.W. Stuck (1976) A method for simulating stable random variables. *J. Amer. Statist. Ass.* **71**, 340–344.

[36] C.S. Chang, P. Heidelberger, P. Juneja, & P. Shahabuddin (1994) Effective bandwidth and fast simulation of ATM intree networks. *Performance Evaluation* **20**, 45–65.

[37] K.L. Chung (1974) *A Course in Probability Theory* (2nd ed.). Academic Press.

[38] M. Cottrel, J.C. Fort & G. Malgouyres (1983) Large deviations and rare events in the study of stochastic algorithms. *IEEE Trans. Automatic Control* **AC–28**, 907-918.

[39] H. Cramér & M.R. Leadbetter (1967) *Stationary and Related Stochastic Processes*. Wiley.

[40] J.G. Dai, J.M. Harrison & R. Williams (1999) *Brownian Models of Stochastic Processing Networks.* Book manuscript.

[41] H. Damerdji (1994) Strong consistency of the variance estimator in steady–state simulation output analysis. *Math. Opns. Res.* **19**, 494–512.

[42] P. Damien, P.W. Laud & A.F.M. Smith (1995) Approximate random variate generation from infinitely divisible distributions with applications to Bayesian inference. *J. R. Statist. Soc.* **B57**, 547–563.

[43] A. Dembo & O. Zeitouni (1993) *Large Deviation Techniques*. Jones & Bartlett, Boston.

[44] L. Devroye (1986) *Non–Uniform Random Variate Generation*. Springer, New York.

[45] Dietrich & Newsom (1993) A fast and exact method for multidimensional Gaussian simulation. *Water Resources Res.* **29**, 2861–2869.

[46] Y. Dodge (1996) A natural random number generator. *Int. Statist. Review* **64**, 329–344.

[47] N.G. Duffield & N. O'Connell (1995) Large deviations and overflow probabilities for the general single–server queue. *Math. Proc. Camb. Philos. Soc.* **118**, 363–374.

[48] P. Embrechts, C. Klüppelberg & T. Mikosch (1997) *Modelling Extremal Events for Finance and Insurance*. Springer, Heidelberg.

[49] K.B. Ensor & P.W. Glynn (1997) Simulating the maximum of a random walk. Manuscript.

[50] W. Feller (1971) *An Introduction to Probability Theory and Its Applications* (2nd ed.) **II**. Wiley.

[51] J.A. Fill (1998) An interruptible algorithms for perfect sampling via Markov chains. *Ann. Appl. Probab.* **8**, 131–162.

[52] G. Fishman (1996) *Monte Carlo. Concepts, Algorithms and Applications*. Springer–Verlag.

[53] S.G. Foss & R.L. Tweedie (1998) Perfect simulation and backwards coupling. *Stochastic Models* **14**, 187–203.

[54] M.J.J. Garvels & D.P. Kroese (1998) A comparison of RESTART implementations. *Proceedings of the Winter Simulation Conference* **1998**, 601–609. IEEE Press.

[55] W.R. Gilks, S. Richardson & D.J. Spiegelhalter (1996) *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

[56] R.D. Gill (1989) Non– and semi–parametric maximum likelihood estimators and the von Mises method (Part I). *Scand. J. Statist.* **16**, 97–128.

[57] P. Glasserman (1991) *Gradient Estimation via Perturbation Analysis*. Kluwer, Boston, Dordrecht, London.

[58] P. Glasserman, P. Heidelberger, P. Shahabuddin & T. Zajic (1996) Multilevel splitting for estimating rare events probabilities. *Opns. Res.* (to appear).

[59] P. Glasserman, P. Heidelberger, P. Shahabuddin & T. Zajic (1996) Splitting for rare event simulation: Analysis of simple cases. *Proceedings of the Winter Simulation Conference* **1996**, 302–308. IEEE Press.

[60] P. Glasserman, P. Heidelberger, P. Shahabuddin & T. Zajic (1997) A large deviations principle on the efficiency of multilevel splitting. *IEEE Trans. Aut. Contr.* (to appear).

[61] P. Glasserman & Y. Wang (1997) Counterexamples in importance sampling for large deviations probabilities. *Ann. Appl. Probab.* **7**, 731–746.

[62] H. Gluver & S. Krenk (1990) Robust calibration and rescaling of ARMA processes for simulation. Danish Center for Applied Mathematics and Mechanics. Report no 409.

[63] P.W. Glynn (1987) Limit theorems for the method of replications. *Stochastic Models* **3**, 343–354.

[64] P.W. Glynn (1989) A GSMP formalism for discrete event systems. *Proc. IEEE* **77**, 14–23.

[65] P.W. Glynn (1989) Poisson's equation for the recurrent M/G/1 queue. *Adv. Appl. Probab.* **26**, 1044–1062.

[66] P.W. Glynn & D.L. Iglehart (1987) A joint central limit theorem for the sample mean and regenerative variance estimator. *Ann. Opns. Res.* **8** 41–55.

[67] P.W. Glynn and D.L. Iglehart (1988) Simulation methods for queues. An overview. *Queueing Systems* **3**, 221-255.

[68] P.W. Glynn & D.L. Iglehart (1989) Importance sampling for stochastic simulations. *Management Sci.* **35**, 1367–1392.

[69] P.W. Glynn & D.L. Iglehart (1990) Simulation output analysis using standardized time series. *Math Opns. Res.* **15**, 1–16.

[70] P.W. Glynn & W. Whitt (1992) The asymptotic efficiency of simulation estimators. *Opns. Res.* **40**, 505–520.

[71] D. Goldsman & M. Meketon (1986) A comparison of several variance estimators for stationary increments stochastic processes. *Technical Report* **J–85–12**, ISYE, Georgia Tech.

[72] A. Goyal, P. Shahabuddin, P. Heidelberger, V.F. Nicola & P.W. Glynn (1992) A unified framework for simulating Markovian models of highly dependable systems. *IEEE Trans. Computers* **41**, 36–51.

[73] R. Grübel (1998) Stochastic models: a functional view. Manuscript.

[74] A. Gut (1995) *An Intermediate Course in Probability.* Springer–Verlag.

[75] J.M. Hammersley & D.C. Handscomb (1964) *Monte Carlo Methods.* Methuen, London.

[76] Z. Haraszti & J.K. Townsend (1998) The theory of direct probability redistribution and its application to rare event simulation.

[77] P. Heidelberger (1995) Fast simulation of rare events in queueing and reliability models. *ACM TOMACS* **6**, 43–85.

[78] P. Heidelberger, P. Shahabuddin & V. Nicola (1994) Bounded relative error in estimating transient measures of highly dependable non–Markovian systems. *ACM TOMACS* **4**, 137–164.

[79] P. Heidelberger & D. Towsley (1989) Sensitivity analysis from sample paths using likelihood ratios. *Management Science* **35**, 1475–1488.

[80] J.R.M. Hosking (1984) Modeling persistence in hydrological time series using fractional differencing. *Water Resources Res.* **20**, 1898–1908.

[81] C. Huang, M. Devetsikiotis, I. Lambadaris & A.R. Kaye (1995) Fast simulation for self–similar traffic in ATM networks. *IEEE*, 438–444.

[82] C. Huang, M. Devetsikiotis, I. Lambadaris & A.R. Kaye (1998) Fast simulation of queues with long–range dependent traffic. *Stochastic Models* **15**.

[83] A. Janicki & A. Weron (1994) *Simulation and Chaotic Behaviour of α–Stable Stochastic Processes.* Marcel Dekker.

[84] J.L. Jensen (1995) *Saddlepoint Approximations.* Clarendon Press, Oxford.

[85] H. Kahn & T.E. Harris (1951) Estimation of particle transmission by random sampling. *National Bureau of Standard Applied Mathematics Series* **12**, 27–30.

[86] J. Keilson & D.M.G. Wishart (1964) A central limit theorem for processes defined on a finite Markov chain. *Proc. Cambridge Philos. Soc.* **60**, 547–567.

[87] M. Keane & G. O'Brien (199x) A Bernoulli factory. *ACM TOMACS.*

[88] F. Kelly (1977) *Reversibility and Stochastic Networks.* Wiley, Chichester.

[89] J.G. Kemeny, J.L. Snell & A.W. Knapp (1966) *Denumerable Markov Chains.* van Nostrand, Princeton.

[90] P.E. Kloeden & E. Platen (1992) *Numerical Solution of Stochastic Differential Equations*. Springer–Verlag, New York.

[91] P.E. Kloeden, E. Platen & H. Schurz (1994) *Numerical Solution of Stochastic Differential Equations through Computer Experiments*. Springer–Verlag, New York.

[92] D.E. Knuth (1984) An algorithm for Brownian zeroes. *Computing* **33**, 89–94.

[93] S. Krenk & J. Clausen (1987) On the calibration of ARMA processes for simulation. In *Reliability and Optimization of Structural Systems* (P. Thoft–Christensen ed.). Springer-Verlag.

[94] A.M. Law & W.D. Kelton (1991) *Simulation Modeling and Analysis* (2nd ed.). McGraw Hill.

[95] P. L'Ecuyer & G. Perron (1994) On the convergence rates of IPA and FDC derivative estimators. *Opns. Res.* **42**, 643–656.

[96] P. L'Ecuyer (1994) Uniform random number generation. *Ann. Opns. Res.* **53**, 77–120.

[97] T. Lehtonen & H. Nyrhinen (1992) Simulating level–crossing probabilities by importance sampling. *Adv. Appl. Probab.* **24**, 858–874.

[98] T. Lehtonen & H. Nyrhinen (1992) On asymptotically efficient simulation of ruin probabilities in a Markovian environment. *Scand. Actuarial J.,* 60–75.

[99] W. Leland, M. Taqqu, W. Willinger & D. Wilson (1994) On the self–similar nature of ethernet traffic. *IEEE/ACM Trans. Netw.* **2**, 1–15.

[100] G. Lindgren (1999) *Lecture Notes on Stationary and Related Stochastic Processes*. Lund University. Available from
http://www.maths.lth.se/matstat/staff/georg.

[101] B.B. Mandelbrot (1971) A fast fractional Gaussian noise generator. *Water Resources Research* **7**, 543–553.

[102] B.B. Mandelbrot & J.W. Van Ness (1968) Fractional Brownian motions, fractional noises and applications. *SIAM Review* **10**, 422–437.

[103] Z. Michna (1999) On tail probabilities and first passage times for fractional Brownian motion. *Math. Methods of Opns. Res.* (to appear).

[104] D.L. Minh & R.M. Sorli (1983) Simulating the GI/G/1 queue in heavy traffic. *Opns. Res.* **31**, 966–971.

[105] B.J.T. Morgan (1984) *Elements of Simulation*. Chapman & Hall.

[106] J. Møller (1999) Perfect simulation of conditionally specified models. *J. R. Statist. Soc.* **B61**, 251–264.

[107] M.K. Nakayama (1996) A charcaterization of the simple failure biasing method for simulations of highly reliable Markovian systems. *ACM TOMACS* **4**, 52–88.

[108] M.K. Nakayama (1996) General conditions for bounded relative error in simulations of highly reliable systems. *Adv. Appl. Probab.* **28**, 687–727.

[109] O. Narayan (1998) Exact asymptotic queue length distribution for fractional Brownian traffic. *Advances in Performance Evaluation* **1**.

[110] M.F. Neuts (1977) A versatile Markovian point process. *J. Appl. Probab.* **16**, 764–779.

[111] M.F. Neuts (1981) *Matrix-Geometric Solutions in Stochastic Models.* Johns Hopkins University Press, Baltimore. London.

[112] M.F. Neuts (1989) *Structured Stochastic Matrices of the M/G/1 Type and their Applications*. Marcel Dekker, New York.

[113] J. Neveu (1961). Une generalisation des processus a accroissementes positifs independantes. *Abh. Math. Sem. Hamburg* **23**, 36–61.

[114] I. Norros (1994) A storage model with self–similar input. *Queueing Systems* **16**, 387–396.

[115] I. Norros, P. Mannersalo & J.L. Wang (1999) Simulation of fractional Brownian motion with conditionalized random midpoint dis-pacement. Submitted.

[116] I. Norros, E. Valkeila & J. Virtamo (1999) A Girsanov type formula for fractional Brownian motion. *Bernoulli* (to appear).

[117] V. Paxson (1997) Fast, approximate synthesis of fractional Gaussian noise for generating self–similar network traffic. *Computer Communication Review 27*, 5–18

[118] V. Paxson & S. Floyd (1995) Wide area traffic: the failure of Poisson modelling. *IEEE/ACM Trans. Netw.* **3**, 226–244.

[119] V.V. Petrov (1965) On the probabilities of large deviations for sums of independent random variables. *Th. Probab. Appl.* **10**, 287–298.

[120] G. Pflug (1997) *Stochastic Optimization.* Kluwer.

[121] J.G. Propp & D.B. Wilson (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Rand. Struct. Algs.* **9**, 223–252.

[122] J.G. Propp & D.B. Wilson (1998) How to get a perfectly random sample from a generic Markov chains and generate a random spanning tree of a directed random graph. *J. Algs.* **27**, 170–217.

[123] P. Protter (1990). *Stochastic Integration and Differential Equations.* Springer-Verlag, New York.

[124] S.I. Resnick (1997) Heavy tail modeling and teletraffic data. *Ann. Statist.* **25**, 1805–1869.

[125] A. Ridder (1996) Fast simulation of Markov fluid models. *Adv. Appl. Probab.* **28**, 786–803.

[126] B. Ripley (1987) *Stochastic Simulation.* Wiley, New York.

[127] L.C.G. Rogers (1994) Fluid models in queueing theory and Wiener–Hopf factorisation of Markov chains. *Ann. Appl. Probab.* **4**, 390–413.

[128] J. Rosiński (1990) On series representations of infinitely random vectors. *Ann. Probab.* **18**, 405–430.

[129] J. Rosiński (1999) Series expansion without compenation for infinitely divisible processes. *In preparation.*

[130] S.M. Ross (1985) *Introduction to Probability Models* (3rd ed.). Academic Press.

[131] S.M. Ross (1991) *A Course in Simulation.* Macmillan.

[132] R.Y. Rubinstein (1981) *Simulation and the Monte Carlo Method.* Wiley, New York.

[133] R.Y. Rubinstein & B. Melamed (1998) *Classical and Modern Simulation.* John Wiley & Sons.

[134] R.Y. Rubinstein & A. Shapiro (1993) *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization via the Score Function Method.* John Wiley & Sons.

[135] T.H. Rydberg (1997) The normal inverse Gaussian Lévy process: simulation and approximation. *Stochastic Models* **13**, 887–910.

[136] J.S. Sadowsky (1991) Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue. *IEEE Trans. Automat. Contr.* **AC–36**, 1383–1394.

[137] J.S. Sadowsky (1993) On the optimality and stability of exponential twisting in Monte Carlo simulation. *IEEE Transaction on Information Theory* **IT-39**, 119–128.

[138] J.S. Sadowsky (1994) Monte Carlo estimation of large deviations probabilities. Manuscript, Arizona State University, Temple, AZ.

[139] J.S. Sadowsky (1995) The probability of large queue lengths and waiting times in a heterogeneous multiserver queue. Part II: Positive recurrence and logarithmic limits. *Advances in Applied Probability* **27**, 567-583.

[140] J.S. Sadowsky & J.S. Bucklew (1990) On large deviations theory and asymptotically efficient Monte Carlo estimation. *IEEE Trans. Inform. Theory* **IT-36**, 579–588.

[141] J.S. Sadowsky & W. Szpankowsky (1995) The probability of large queue lengths and waiting times in a heterogeneous multiserver queue. Part I: Tight limits. *Adv. Appl. Probab.* **27**, 532-566.

[142] G. Samorodnitsky & M.L. Taqqu (1994) *Non–Gaussian Stable Processes*. Chapman & Hall.

[143] L.W. Schruben (1986) Sequential simulation run control using standardized times series. *ASA–ACM Interface Meeting*, Colorado.

[144] P. Shahabuddin (1994) Importance sampling for the simulation of highly reliable Markovian systems. *Management Science* **40**, 333–352.

[145] G.S. Shedler (1993) *Regenerative Stochastic Simulation*. Academic Press.

[146] A. Shwartz & A. Weiss (1995) *Large Deviations for Performance Analysis: Queues, Communication and Computers*. Chapman & Hall.

[147] D. Siegmund (1976). Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.* **4**, 673–684.

[148] D. Siegmund (1976) The equivalence of absorbing and reflecting barrier problems for stochastically monotone Markov processes. *Ann. Probab.* **4**, 914–924.

[149] D. Siegmund (1985) *Sequential Analysis*. Springer–Verlag.

[150] K. Sigman (1994) *Stationary Marked Point Processes: An Intuitive Approach.* Chapman & Hall, New York.

[151] R. Suri (1989) Perturbation analysis: the state of the art and research issues explained via the GI/G/1 queue. *Proceeedings of the IEEE* **77**, 114–137.

[152] M. & J. Villén–Altamirano (1991) RESTART: A method for accelerating rare events simulations. *Queueing Performance and Control in ATM.* (J.W. Cohen & C.D. Pack, eds). *Proceedings of ITC* **13**, 71–76.

[153] M. & J. Villén–Altamirano (1994) RESTART: A straightforward method for fast simulation of rare events. *Proceedings of the Winter Simulation Conference* **1994**, 282–289. IEEE Press.

[154] W. Whitt (1989) Planning queueing simulations. *Management Science* **35**, 1341–1366.

[155] Wood & Chan (1994) Simulation of stationary Gaussian processes in $[0,1]^d$. *J. Comp. Graph. Statist.* **3**, 409–432.

Assignments

Assignment 1

Consider the GI/G/1 queue with interarrival times $T_0, T_1, \ldots$ and service times $U_0, U_1, \ldots$. The waiting time of customer n is them W_n where $W_0 = 0$,

$$W_{n+1} \;=\; (W_n + U_n - T_n)^+.$$

We consider here the specific example where $T_k \equiv 1$ (D/GI/1) and the service time distribution $B(x) = \mathbb{P}(U_k \le x)$ is only known via observations $u_1, \ldots, u_m$. Your assignment is to produce a point estimate of $\mathbb{E}M_{25}$ where

$$M_{25} \;=\; \max_{n=0,\ldots,25} W_n$$

and an associated 95% confidence interval.

Read first II.3 of the notes carefully. You should not try to verify Hadamard– or Frechet differentiability rigorously but assume this can be done. You are free to choose the simulation budget t (but note that an excessively large t does not make sense, not only because of time constraints but also because the stochastic fluctuations in the empirical distribution B_m can never be eliminated). Your choice of the number k of groups and the number p of simulations for each group should be validated through histograms showing that asymptotic normality is reasonably fulfilled.

Practical MatLab guidance: $u_1, \ldots, u_m$, $m = 10.000$, can be found on the file sophia.mat. To access, write 'load sophia'.

Assignment 2

A system develops in i.i.d. cycles. According to whether a certain catastrophic event occurs or not within a cycle, the cycle is classified as failed or non–failed. Denote by p the probability that a cycle is failed, by ℓ_1 the expected length of a cycle given it does not fail, and by ℓ_2 the expected time until failure in a cycle given it is failed. We are interested in ℓ, the expected time until a failure occurs.

1. Express ℓ in terms of p, ℓ_1, ℓ_2.

2. You are presented with statistics of 1000 cycles, of which 87 failed. The non–failed cycles had an empirical mean of 20.2 and an empirical variance of 18.6, and the average time until failure in the failed cycles was 5.4 with an empirical variance of 3.1. Give a confidence interval for ℓ.

Assignment 3

Consider a (s, S) inventory system where the number of goods stored at time t is $V(t)$. Demands arrive (one at a time) according to a Poisson process with intensity λ. When $V(t-) = s + 1$, $V(t) = s$, an order of size $S - s$ is placed and arrives after a random time Z (the lead time). Demands arriving while $V(t) = 0$ are lost. It is assumed that $S - s > s$.

Write a program for regenerative simulation of p, the long–run probability that a demand is lost. Use whichever values of s, S, λ you like and whichever distribution of Z. Save your program for Assignment 4.

Assignment 4

Consider again the model of Assignment 3, but assume that λ is small compared to Z in the sense that the event that $V(t) = 0$ in a cycle is rare.

Improve your program for Assignment 3 by combining with a variance reduction technique.

You can, e.g., use a change of λ depending on the value of Z. Be also clever and use the fact that the expected number of demands lost in a cycle only depends on the residual lead time at the time where $V(t-) = 1$, $V(t) = 0$.

Assignment 5

In the model of Assignment 3, give an estimate of the sensitivity of p w.r.t. λ and an associated confidence interval by the likelihood ratio method.

Assignment 6

In earthquake modeling, let $S(t)$ be the stress potential at time t (taken to be left–continuous) and $s = 20$ a treshold value. At time $\tau = \inf\{t : S(t) > s\}$ an earthquake occurs, and the potential is then reset to 0, $S(\tau+) = 0$. In between quakes, the stress potential builds up like a subordinator with Lévy measure

$$\nu(dx) \;=\; \frac{x^{1/2}+3}{x^{7/4}}\,dx.$$

Produce a histogram for the severity of an earthquake, as defined by the r.v. $S(\tau) - s$.

MatLab guidance: a routine for generating a r.v. U with a Pareto tail of the form

$$\mathbb{P}(U > x) \;=\; \begin{cases} 1 & x \le a \\ \left(\frac{a}{x}\right)^{\alpha} & x \ge a \end{cases}$$

is available and is called as sofia2(alpha,a) [you may or may not need this depending on which method you use].

Assignment 7

Suggest some variance reduction methods for evaluating

$$\int_0^\infty (x + 0.02x^2)\exp\left\{0.1\sqrt{1+\cos x} \;-\; x\right\}\,dx$$

by Monte Carlo integration.

Assignment 8

Perform exact simulation of the Moran dam (Example **III**.1.5) for the case where V is geometric with mean 2 and $m = 3$, and p is variable. Use both independent updating and monotone updating, and compare the two methods in terms of for how high values of p you are able to produce Z within reasonable time.

Assignment 9

Let $\{Y_n\}$ be a Markov chain and $X_n = f(Y_n)$. Assume it is possible to perform exact simulation from $\{Y_n\}$, say T_e units of computer time is needed on the average to produce one copy Y^* of the stationary r.v. For estimating $f(Y^*)$, we start $\{Y_n\}$ with $Y_0 = Y^*$, run the chain up to time $m-1$ and let $Z = Z_m = \sum_0^{m-1} Y_n/m$.

If one updating of $\{Y_n\}$ requires T units of computer time and the stationary covariance function of $\{X_n\}$ has the form $\rho_k = z^k$ with $0 < z < 1$ (cf. Proposition III.2.5), how would you choose m to minimize the variance for a given simulation budget t? Give numerical examples for different values of z, T_e, T, say $T_e = 10T, 100T, 1000T$.