Monte Carlo Methods

David Šiška
School of Mathematics, University of Edinburgh

2016/17[*]

Contents

1 Introduction **2**

1.1 Numerical integration 2

1.2 Derivative pricing 3

1.3 Some useful identities 5

2 Convergence **6**

2.1 Random variables, their distribution, density and characteristic functions 6

2.2 Convergence modes 7

2.3 Law of large numbers and Central limit theorem 8

3 Generating random samples **13**

3.1 Pseudorandom numbers, uniform distribution 13

3.2 Inversion method 13

3.3 Acceptance rejection method 15

3.4 Box–Muller method for generating normally distributed samples . . . 17

3.5 Generating correlated normally distributed samples 19

3.6 Summary 20

4 Variance reduction **21**

4.1 Antithetic variates 21

4.2 Control variates 22

4.3 Multiple control variates 26

4.4 Summary 26

4.5 Further reading 27

[*] This is the first draft of the notes. There will be mistakes that need to be corrected and material will be added and removed as will be appropriate for the lectures. Last updated 1st June 2018.

5 Some applications 28
5.1 Asian options 28
5.2 Options on several risky assets 31
6 Approximating sensitivity on parameters 34
6.1 Background 34
6.2 General Setting 35
6.3 Finite Difference Approach 35
6.4 Calculating Sensitivities Pointwise in Ω 37
6.5 The Log-likelihood Method 39
7 Solutions to some exercises 40
A Appendix 43
A.1 Multi-dimensional Itô's formula 43

1 Introduction

We are interested in Monte Carlo methods as a general simulation technique. However many (most) of our examples will come from financial mathematics.

1.1 Numerical integration

We start with examples that are not directly related to derivative pricing. This is to let us understand the main idea behind Monte Carlo methods without getting confused by general derivate pricing issues.

Example 1.1 (Numerical integration in one dimension)**.** Let $f : [a, b] \to \mathbb{R}$ be given and say that we want to approximate

$$I = \int_a^b f(x)dx.$$

Assume that $f \geq 0$ on $[a, b]$ and that f is bounded on the interval $[a, b]$ and let $M := \sup_{x\in[a,b]} f(x)$. Assume that we know how to generate samples from $U(0, 1)$ (that is the uniform distribution on the interval 0 to 1).

Let $(u_i)_{i=1}^N$ and $(v_i)_{i=1}^N$ be two collections of N samples each from $U(0, 1)$. Let $x_i := a + (b - a)u_i$ and $y_i := Mv_i$. Let $\mathbb{1}_A$ be equal to 1 if A is true and 0 otherwise. Then we can approximate I with

$$I_N := (b - a)M\frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\{f(x_i)\geq y_i\}}.$$

That is, we count the number of times when y_i is equal to or less than $f(x_i)$ and then we divide by N. Finally, we scale this by the area of the rectangle inside which we are sampling our random points. One would hope that I_N converges to I, in some sense, as $N \to \infty$.

Example 1.2 (Multidimensional numerical integration)**.** Let $\Omega \subset \mathbb{R}^d$ be bounded inside the hypercube

$$[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d].$$

Let $f : \Omega \to \mathbb{R}^+$ be measurable, integrable and bounded. We wish to approximate

$$I = \int_\Omega f(x)dx.$$

Let

$$M := \sup_{x \in \Omega} f(x).$$

For $j = 1, \ldots, d+1$ and $i = 1, \ldots, N$ sample u_{ij} independently from $U(0,1)$. Let

$$\begin{aligned} x_{i1} &:= a_1 + (b_1 - a_1)u_{i1}, \\ x_{i2} &:= a_2 + (b_2 - a_2)u_{i2}, \\ &\vdots \\ x_{id} &:= a_d + (b_d - a_d)u_{id}, \\ y_i &:= m + (M - m)u_{i(d+1)}. \end{aligned}$$

Let $x_i := (x_{i1}, x_{i2}, \ldots, x_{id})^T$. First we approximate the volume of Ω, denoted by V. Let V_N denote the approximation of the volume.

$$V_N := (b_1 - a_1)(b_2 - a_2) \cdots (b_d - a_d) \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\{x_i \in \Omega\}}.$$

We can now approximate I with

$$I_N := V_N M \frac{1}{\tilde{N}} \sum_{i=1}^N \mathbb{1}_{\{x_i \in \Omega\}} \mathbb{1}_{\{f(x_i) \geq y_i\}}, \quad \tilde{N} = \sum_{i=1}^N \mathbb{1}_{\{x_i \in \Omega\}}.$$

That is, we first calculate $\tilde{N}$, that is the number of x_i that lie inside Ω. Then we count the number of times y_i is equal to or less than $f(x_i)$ and we divide by $\tilde{N}$. Finally we scale by the volume of $\Omega \times [0, M]$. Again we hope that I_N converges to I, in some sense, as $N \to \infty$.

1.2 Derivative pricing

We now give some examples of pricing derivatives with Monte Carlo methods. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_t)_{t \in [0,T]}$ a given filtration to which the traded assets are adapted. It can shown that for any option whose payoff is given by a $\mathcal{F}_T$-measurable random variable h has the value at time $t < T$ given by

$$V_t = \mathbb{E}^{\mathbb{Q}} \left(D(t,T)h | \mathcal{F}_t \right),$$

where $D(t,T)$ is the "discounting factor" for the time period t to T, which in the simplest case can be $D(t,T) = \exp(-r(T-t))$ for some risk free rate $r \geq 0$ and where $\mathbb{E}^{\mathbb{Q}}$ denotes the expectation under the risk neutral measure $\mathbb{Q}$. This is the measure under which the discounted traded assets are martingales.

We have shown that in the particular case of European call and put options in the Black–Scholes framework we have

$$v(t,S) = \mathbb{E}^{\mathbb{Q}}(e^{-r(T-t)}g(S_T)|S_t = S),$$

where g is the function giving the option payoff at exercise time T. Of course in this case we have the well known Black–Scholes formula giving the option price.

Example 1.3 (Classical Black–Scholes)**.** Say that we have derived the Black–Scholes formula ourselves but we are not sure whether we have performed all the calculations correctly. One way for us to check would be to use Monte Carlo methods to approximate the option payoff by simulating the behaviour of the risky asset.

Recall that the model for the risky asset in the real-world measure $\mathbb{P}$ is

$$dS_t = \mu S_t dt + \sigma S_t dW_t,$$

where $(W_t)_{t\in[0,T]}$ is a $\mathbb{P}$-Wiener process with respect to $(\mathcal{F}_t)_{t\in[0,T]}$, $\mu \in \mathbb{R}$ and $\sigma > 0$. Say that $g(S) := [S-K]_+$, that is, the option is an European call option.

We have shown that in the risk-neutral measure $\mathbb{Q}$ the evolution of the risky asset is given by

$$dS_t = rS_t dt + \sigma S_t d\tilde{W}_t, \tag{1}$$

where $(\tilde{W}_t)_{t\in[0,T]}$ is a $\mathbb{Q}$-Wiener process with respect to $(\mathcal{F}_t)_{t\in[0,T]}$. We further know that

$$S_T = S_t \exp\left(\left(r - \frac{1}{2}\sigma^2\right)(T-t) + \sigma\left(\tilde{W}_T - \tilde{W}_t\right)\right).$$

The option price is thus given by

$$v(t,S) = \mathbb{E}^{\mathbb{Q}}\left(e^{-r(T-t)}\left[S\exp\left(\left(r-\frac{1}{2}\sigma^2\right)(T-t) + \sigma\left(\tilde{W}_T - \tilde{W}_t\right)\right) - K\right]_+\right).$$

By definition $\tilde{W}_T - \tilde{W}_t$ is normally distributed with mean 0 and variance $T-t$. If $Z \sim N(0,1)$ then $\sqrt{T-t}Z$ has the same distribution as $\tilde{W}_T - \tilde{W}_t$. So

$$v(t,S) = \mathbb{E}^{\mathbb{Q}}\left(e^{-r(T-t)}\left[S\exp\left(\left(r-\frac{1}{2}\sigma^2\right)(T-t) + \sigma\sqrt{T-t}Z\right) - K\right]_+\right). \tag{2}$$

Now we will use a Monte Carlo method to evaluate (2). Assume for now that we know how to draw samples from standard normal distribution. Let us take N independent samples $(z_i)_{i=1}^N$ from $N(0,1)$. The approximation is given by

$$v_N(t,S) := \frac{1}{N}\sum_{i=1}^N e^{-r(T-t)}\left[S\exp\left(\left(r-\frac{1}{2}\sigma^2\right)(T-t) + \sigma\sqrt{T-t}z_i\right) - K\right]_+.$$

We would hope that for fixed t and S we can say that $v_N(t,S)$ converges in some sense to $v(t,S)$ as $N \to \infty$.

Now we can compare $v_N(t,S)$ to the option price given by the Black–Scholes formula. If the numbers are close (and on average decreasing as N increases) then we would have every reason to believe we are using the correct Black–Scholes formula.

1.3 Some useful identities

Let X, Y be random variables and recall that

$$\begin{aligned}
\mathrm{Var}[X] &:= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right]-(\mathbb{E}[X])^2\,, \\
\mathrm{Cov}[X,Y] &:= \mathbb{E}\Big[(X-\mathbb{E}[X])\,(Y-\mathbb{E}[Y])\Big] = \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]\,.
\end{aligned}$$

If λ, μ are constants, then

$$\begin{aligned}
\mathbb{E}[\mu+X] &= \mathbb{E}[X]+\mu\,, \\
\mathrm{Var}[\mu+X] &= \mathrm{Var}[X]\,, \\
\mathbb{E}[\lambda\,X] &= \lambda\;\mathbb{E}[X]\,, \\
\mathrm{Var}[\lambda\,X] &= \lambda^2\,\mathrm{Var}[X]\,, \\
\mathbb{E}[X+Y] &= \mathbb{E}[X]+\mathbb{E}[Y]\,, \\
\mathrm{Var}[X+Y] &= \mathrm{Var}[X]+\mathrm{Var}[Y]+2\,\mathrm{Cov}[X,Y]\,.
\end{aligned}$$

2 Convergence

So far we have not discussed the convergence of Monte Carlo algorithms. It is clear that the "usual" notions of convergence are insufficient when analysing Monte Carlo methods. No matter how large sample we take, we can always be extremely "unlucky", draw a "unrepresentative" sample and get a bad estimate for the true solution of our problem. In this section we introduce the appropriate notion of convergence, law of large numbers and the central limit theorem, which provides the convergence of Monte Carlo algorithms.

2.1 Random variables, their distribution, density and characteristic functions

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If X is an $\mathbb{R}^d$-valued random variable then its *distribution function* (sometimes called the cumulative distribution function or CDF) is $F : \mathbb{R}^d \to [0,1]$ given for $\mathbb{R}^d \ni x = (x_1, \ldots, x_n)$ by

$$F(x) = F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \le x_1, \ldots, X_d \le x_d).$$

We recall that if for some $g : \mathbb{R}^d \to \mathbb{C}^{d'}$ we have $\mathbb{E}|g(X)| < \infty$ then

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^d} g(y)\, dF(y).$$

In particular taking $g(y) = \mathbb{1}_B(y)$ for some $B \in \mathcal{B}(\mathbb{R}^d)$ leads to

$$\mathbb{P}(X \in B) = \mathbb{E}[\mathbb{1}_B(X)] = \mathbb{E}[g(X)] = \int_B dF(y).$$

We say that the distribution function F has *density* f if $f : \mathbb{R}^d \to [0, \infty)$ is such that $\int_{\mathbb{R}^d} f(y)dy = 1$ and

$$F(x) = F(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(y_1, \ldots, y_d)\, dy_d \cdots dy_1.$$

If X is a random variable with a distribution that has a density then we call X *continuous*[1]. Recall that for a continuous random variable X with density f we have

$$\mathbb{E}g(X) = \int_{\mathbb{R}^d} g(y) f(y)\, dy$$

whenever $\mathbb{E}|g(X)| < \infty$.

The *characteristic function* φ of a distribution F (or a random variable with distribution F) is defined by

$$\varphi(z) := \int_{\mathbb{R}^d} e^{izx}\, dF(x)\,, \quad z \in \mathbb{R}^d.$$

Here $zx := \sum_{i=1}^d z_i x_i$ is the inner (dot) product. We see that if X has the distribution F then its characteristic function is $\varphi(z) = \mathbb{E}[e^{izX}]$.

Let $(X_k)_{k \in \mathbb{N}}$ be independent random variables and let $S_n = X_1 + \cdots + X_n$. Then

$$\varphi_{S_n}(t) = \mathbb{E}\left[\prod_{k=1}^n e^{itX_k}\right] = \prod_{k=1}^n \mathbb{E}\left[e^{itX_k}\right] = \prod_{k=1}^n \varphi_{X_k}(t)\,. \tag{3}$$

[1] This is a completely different concept to continuity of functions!

Theorem 2.1. *Let X be an $\mathbb{R}$-valued random variable with distribution F and characteristic function $\varphi(z) = \mathbb{E}[e^{izX}]$. Then φ satisfies the following:*

1. $|\varphi(t)| \leq \varphi(0) = 1$.
2. $t \mapsto \varphi(t)$ *is uniformly continuous.*
3. $\varphi(t)$ *equals the complex conjugate of* $\varphi(-t)$.
4. $\varphi(t)$ *is real-valued if and only if F is symmetric in the sense that $\mathbb{P}(B) = \mathbb{P}(-B)$, where $-B := \{x : -x \in B\}$.*
5. *If $\mathbb{E}|X|^n < \infty$ for some $n \geq 1$ then*

$$\frac{d^r}{dt^r}\varphi(t) = \varphi^{(r)}(t) = \int_{\mathbb{R}} (ix)^r e^{itx}\, dF(x)$$

exists for all $r \leq n$ and $\mathbb{E}[X^r] = i^{-r}\varphi^{(r)}(0)$. Moreover

$$\varphi(t) = \sum_{r=0}^{n} \frac{(it)^r}{r!}\mathbb{E}[X^r] + \frac{(it)^n}{n!}E_n(t), \tag{4}$$

with $|E_n(t)| \leq 3\mathbb{E}|X|^n$ and $E_n(t) \to 0$ as $t \to 0$.

Compare the expansion in (4) to the Taylor expansion:

$$\varphi(t) = 1 + t\varphi'(0) + \frac{t^2}{2!}\varphi^{ii}(0) + \cdots + \frac{t^n}{n!}\varphi^{(n)}(0) + \frac{(it)^n}{n!}E_n(t).$$

2.2 Convergence modes

We now look at various types of convergence. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables.

Definition 2.2 (Pointwise Convergence). *We say that the random variables converge pointwise to X if for all $\omega \in \Omega$ we have $X_n(\omega) \to X(\omega)$ as $n \to \infty$.*

We say that an event E occurs *almost surely* (or a.s. for short) if $\mathbb{P}(\Omega \backslash E) = \mathbb{P}(E^c) = 0$. From this follows the definition of almost sure convergence.

Definition 2.3 (Almost sure Convergence). *We say that the random variables converge* almost surely *to X if there is an event E with $\mathbb{P}(E^c) = 0$ such that for all $\omega \in E$ we have $X_n(\omega) \to X(\omega)$ as $n \to \infty$.*

We can immediately see that pointwise convergence implies almost sure convergence.

Definition 2.4 (L^p Convergence). *Let $p > 0$. We say that the random variables converge* in L^p *to X if $\mathbb{E}[|X_n - X|^p] \to 0$ as $n \to \infty$.*

Definition 2.5 (Convergence in probability). *We say that the random variables converge* in probability *to random variable X if for all $\varepsilon > 0$ we have*

$$\mathbb{P}\left[|X_n - X| \geq \varepsilon\right] \to 0 \quad \textit{as } n \to \infty.$$

Definition 2.6 (Convergence in Distribution)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be random variables with distributions $(F_n)_{n\in\mathbb{N}}$. We say that the random variables converge in distribution to a random variable X with distribution F if $F_n(x) \to F(x)$ as $n \to \infty$ for all real numbers x at which F is continuous.*

We make the following remarks.

1. We will sometimes use the notation $X_n \xrightarrow{d} X$ to denote that $(X_n)_{n\in\mathbb{N}}$ converges to X is distribution as $n \to \infty$.
2. The random variables need not be defined on the same probability space when one considers convergence in distribution. Indeed the statement only involves the distribution functions.

The following two theorems give the relations between different types of convergence.

Theorem 2.7. *We have:*

i) Almost sure convergence implies convergence in probability.

ii) L^p convergence implies convergence in probability.

iii) Convergence in probability implies convergence in distribution.

For proof see Shiryaev [7, Ch. II, 10, Theorem 2]. The following theorem says that there are (at least) three equivalent ways to see convergence in distribution.

Theorem 2.8. *Let φ_{X_n} and φ_X be the characteristic functions of X_n and X respectively. Then the following are equivalent:*

i) $X_n \to X$ as $n \to \infty$ in distribution.

ii) $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ as $n \to \infty$ for all bounded and continuous functions f.

iii) $\varphi_{X_n}(t) \to \varphi_X(t)$ as $n \to \infty$ for all $t \in \mathbb{R}$.

For proofs (of a more general result) and further reading see Shiryaev [7, Ch. III, 1, Theorem 1 and Ch. III, 3, Theorem 1].

2.3 Law of large numbers and Central limit theorem

We now have all the tools we will need to prove the Law of large numbers and the Central Limit Theorem. The proofs are those given in Shiryaev [7, Ch. 3].

Theorem 2.9 (Law of large numbers)**.** *Let $(X_k)_{k\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables such that $\mathbb{E}|X_1| = m < \infty$. Let*

$$S_n = X_1 + \cdots + X_n.$$

Then $\frac{S_n}{n} \to m$ in probability.

Proof. Let φ and $\varphi_{S_n/n}$ be the distribution functions of the random variables X_1 and S_n/n respectively. That is

$$\varphi(t) = \mathbb{E}[e^{itX_1}], \quad \varphi_{S_n/n}(t) = \mathbb{E}\left[e^{it\frac{S_n}{n}}\right].$$

Then, since X_i are independent, with same calculation as for (3), we have

$$\varphi_{S_n/n}(t) = \left(\varphi\left(\frac{t}{n}\right)\right)^n.$$

From (4) we know that we can write φ as

$$\varphi\left(\frac{t}{n}\right) = 1 + \frac{itm}{n} + \frac{it}{n}E_1\left(\frac{t}{n}\right).$$

From Theorem 2.1 we know that $E_1\left(\frac{t}{n}\right) \to 0$ as $n \to \infty$ for each fixed t. So we can write[2]

$$\varphi\left(\frac{t}{n}\right) = 1 + \frac{itm}{n} + o\left(\frac{1}{n}\right)$$

and so

$$\varphi_{S_n/n}(t) = \left[1 + \frac{itm}{n} + o\left(\frac{1}{n}\right)\right]^n \to e^{itm} \quad \text{for all } n > N.$$

The function $t \mapsto e^{itm}$ is the characteristic function of a random variable $Z = m$ almost surely. From Theorem 2.8 we know that convergence of characteristic functions is equivalent to convergence in distribution. In general convergence in distribution does not imply convergence in probability. However, in the special case when $S_n/n \to Z = m$ as $n \to \infty$ in distribution, we can conclude that the convergence is in probability too. □

Theorem 2.10 (Central limit theorem)**.** *Let $(X_k)_{k\in\mathbb{N}}$ be independent and identically distributed with $\mathbb{E}(X_k) = \mu$ and $Var(X_k) = \sigma^2$. Then*

$$\bar{X}_n := \frac{S_n}{n} := \frac{1}{n}\sum_{k=1}^{n} X_k$$

satisfies

$$\xi_n := \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \quad as \quad n \to \infty, \tag{5}$$

where $Z \sim N(0,1)$.

Proof. Let φ be the the characteristic function of $X_1 - \mu$ i.e. $\varphi(t) = \mathbb{E}\left[e^{it(X_1-\mu)}\right]$ and let φ_n be the the characteristic function of ξ_n. Observe that

$$\xi_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{n\left(\frac{S_n}{n} - \mathbb{E}\left(\frac{S_n}{n}\right)\right)}{\sigma\sqrt{n}}.$$

[2] We say that $f(n) = o(g(n))$ if for any $\varepsilon > 0$ there is N such that

$$|f(n)| \leq \varepsilon|g(n)| \quad \text{for all } n > N.$$

Hence, due to the same independence type calculation as in (3), we get

$$\varphi_n(t) := \mathbb{E}\left[\exp\left(it\xi_n\right)\right] = \mathbb{E}\left[\exp\left(it\frac{\sum_{k=1}^n (X_k-\mu)}{\sigma\sqrt{n}}\right)\right] = \prod_{k=1}^n \mathbb{E}\left[\exp\left(it\frac{(X_k-\mu)}{\sigma\sqrt{n}}\right)\right]$$
$$= \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

From Theorem 2.1 and (4) we get

$$\varphi(t) = 1 - \frac{\sigma^2 t^2}{2} - \frac{t^2}{2}E_2(t).$$

Hence

$$\varphi_n(t) = \left[1 - \frac{\sigma^2 t^2}{2\sigma^2 n} - \frac{t^2}{2\sigma^2 n}E_2(t)\right]^n = \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n \to e^{-\frac{t^2}{2}} \quad \text{as } n\to\infty.$$

The function $t \mapsto e^{-\frac{t^2}{2}}$ is the characteristic function of a $N(0,1)$ random variable and Theorem 2.8 tells us that convergence of characteristic functions is equivalent to convergence in distributions. □

The proof can also be found in Grimmett and Stirzaker [3, Chapter 5, Section 10].

Proposition 2.11. *Let us take $(X_n)_{n\in\mathbb{N}}$ and $\bar{X}_n$ as in the Central limit theorem. Let Φ denote the distribution of a standard normal random variable. Then for any $\delta > 0$ we have*

$$\mathbb{P}\left(\bar{X}_n - z_{\delta/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + z_{\delta/2}\frac{\sigma}{\sqrt{n}}\right) \to 1-\delta \ \text{ as } \ n\to\infty,$$

where $z_{\delta/2}$ is a number such that $1-\Phi(z_{\delta/2}) = \delta/2$.

Proof. Since

$$\Phi(x) = \int_{-\infty}^x \phi(z)dz = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz$$

we see that Φ is continuous. Hence, due to (5) and the definition of convergence in distributions, we know that for all $x\in\mathbb{R}$

$$\mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \le x\right) \to \Phi(x) \quad \text{as} \quad n\to\infty.$$

Thus, taking x equal to φ and to $-\psi$ above, with $0\le\varphi,\psi<\infty$, we get

$$\mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \le \varphi\right) \to \Phi(\varphi) \ \text{ and } \ \mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} < -\psi\right) \to \Phi(-\psi) \ \text{ as } \ n\to\infty.$$

Therefore

$$\mathbb{P}\left(\bar{X}_n - \varphi\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + \psi\frac{\sigma}{\sqrt{n}}\right) = \mathbb{P}\left(-\psi \le \frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \le \varphi\right)$$
$$= \mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \le \varphi\right) - \mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \le -\psi\right) \to \Phi(\varphi) - \Phi(-\psi) \quad \text{as} \quad n\to\infty.$$

For any $\delta > 0$ we can choose φ and ψ such that $\Phi(\varphi) - \Phi(-\psi) = 1 - \delta$. In particular letting $z_{\delta/2}$ be a number such that $1 - \Phi(z_{\delta/2}) = \delta/2$ we see that

$$\Phi(z_{\delta/2}) - \Phi(-z_{\delta/2}) = 1 - \frac{\delta}{2} - \frac{\delta}{2} = 1 - \delta.$$

Hence

$$\mathbb{P}\left(\bar{X}_n - z_{\delta/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\delta/2}\frac{\sigma}{\sqrt{n}}\right) \to 1 - \delta \text{ as } n \to \infty.$$

□

Roughly speaking this means that the estimator $\bar{X}_n$ is a correct estimate for μ up to an error of $z_{\delta/2}\frac{\sigma}{\sqrt{n}}$, with probability $1 - \delta$. That is, with n sufficiently large, we can halve the error by quadrupling the number n. Another way of looking at this is that to be able to say that $\bar{X}_n$ is correct up to an error of $\epsilon > 0$ we need to take $n > z_{\delta/2}^2 \sigma^2 \epsilon^{-2}$.

Of course σ would typically be unknown in Monte-Carlo simulations and so this does not give us a usable error estimate. Nevertheless it is a constant and so we can still say that the Monte-Carlo method would converge with order 1/2 (we halve the error by quadrupling N).

Definition 2.12 (Estimator for $\mathbb{E}X$). *Let us take $(X_n)_{n\in\mathbb{N}}$ and $\bar{X}_n$. We will call $\bar{X}_n$ the* estimator for *$\mathbb{E}X_n = \mu$.*

Note that

$$\mathbb{E}\left(\bar{X}_n\right) = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(X_k) = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(X) = \mathbb{E}(X).$$

An estimator with the property that the expectation of the estimator (recall it is a random variable) is equal to the parameter we are estimating is called *unbiased*. An estimator that does not have this property is called *biased*.

Example 2.13. Let us now return to the setting of Example 1.3. At time t, when the risky asset is worth S, we have the option price $v(t, S)$ given by (2).

As before, we take N independent samples $(z_i)_{i=1}^N$ from $N(0, 1)$. The Monte Carlo approximation is given by

$$v_N(t, S) := \frac{1}{N}\sum_{i=1}^{N} e^{-r(T-t)}\left[S\exp\left(\left(r - \frac{1}{2}\sigma^2\right)(T - t) + \sigma\sqrt{T - t}z_i\right) - K\right]_+ .$$

To use the central limit theorem let us take

$$X_i := e^{-r(T-t)}\left[S\exp\left(\left(r - \frac{1}{2}\sigma^2\right)(T - t) + \sigma\sqrt{T - t}Z_i\right) - K\right]_+$$

where $(Z_i)_{i=1}^N$ are independent and identically distributed standard normal variables. Of course in the actual Monte Carlo experiment we will have $(z_i)_{i=1}^N$ *samples* from the standard normal distribution. But to do the mathematical analysis we have to replace those by random variables. The the expectation of X_i does not depend on i and the same for the variance and we have $\mu = v(t, S) = \mathbb{E}(X_i)$ and $\sigma_{v_N} = \mathrm{Var}(X_i)$. Clearly both these quantities are unknown (unless we calculate $\mu = v(t, S)$ using the

Black–Scholes formula; but that is not the point here). Using central limit theorem we get that

$$\frac{\sqrt{N}(v_N(t,S) - \mu)}{\sigma} \xrightarrow{d} Z \quad \text{as} \quad n \to \infty.$$

In particular we have the same asymptotic estimate as above. That is, for large N we know that $v(t,S) \in (v_N(t,S) - z_{\delta/2}\sigma_{v_N} N^{-1/2}, v_N(t,S) + z_{\delta/2}\sigma_{v_N} N^{-1/2})$.

As we said before σ_{v_N} is an unknown number but it is a constant.

The central limit theorem gives us the order of convergence of a Monte Carlo algorithm. Reducing the variance will reduce the error in the approximation for a fixed number of samples. Hence finding ways that reduce the variance of Monte Carlo simulations is an area of active research interest.

3 Generating random samples

To use Monte Carlo methods we need to generate random samples from various distributions. Of course a computer algorithm will never generate truly random numbers, but there are ways of generating sequences of numbers that “look” random, unless we actually know the algorithm that generated them. We will say that such sequences are *pseudorandom.*

From those we can easily get samples from $U(0,1)$, the uniform distribution on the interval 0 to 1. Now we would like to be able to generate random samples from any distribution efficiently. We will present several methods: inversion, acceptance-rejection method and the Box–Muller method for generating normally distributed random samples.

3.1 Linear congruential pseudorandom number generators and generating uniformly distributed random samples

One of the most commonly used methods for generating pseudorandom numbers is the *linear congruentiual pseudorandom number generator.* Given a random “seed” x_0 we generate pseudorandom numbers $x_1, x_2, \dots$ using the recurrence relation

$$x_{i+1} = (ax_i + c) \mod m,\ i = 0, 1, \dots$$

with a a given multiplier, c a given increment and m a given modulus.

The period of the generator is the smallest $p \in \mathbb{N}$ such that $x_i = x_{i+p}$ for any $i = 0, 1, \dots$. The period of the generator will never exceed m (i.e. $p \le m$). See Knuth [4, Section 3.2.1]. Clearly the smallest value x_i can have is 0 and the maximum is $m-1$.

Example 3.1. Take $x_{i+1} = (7x_i + 8) \mod 15$.

With seed $x_0 = 1$ we get $x_1 = 15 \mod 15 = 0$, $x_2 = 8 \mod 15 = 8$ etc. The period is $p = 12$ because $x_{12} = 1 = x_0$ and for all $i = 1, 2, \dots, 11$ we have $x_i \neq x_0$.

With seed $x_0 = 3$ we get $x_1 = 29 \mod 15 = 14$, since $1 \cdot 15 + 14 = 29$.

If we use linear congruential pseudorandom number generator then we can generate a sequence $(u_i)_{i=1,\dots,m}$ of samples from $U(0,1)$ by taking $u_i = x_i/(m-1)$, where x_i are the numbers produced by the generator with maximum period m.

3.2 Inversion method

From now onwards we will assume that we can generate not just pseudorandom but truly random samples from the uniform distribution.

The inversion method is a method for generating samples from distributions of random variables that take values in $\mathbb{R}$. Say we want to generate samples following the distribution $F : \mathbb{R} \to [0,1]$. Assume that the inverse F^{-1} of F exists.

Recall that we say that the random variable $X : \Omega \to \mathbb{R}$ has the distribution F if $P(X \le x) = F(x)$ for any $x \in \mathbb{R}$. If F is continuous and strictly increasing then for each $u \in (0,1)$ there is $F^{-1}(u)$, given by the usual inverse of the strictly increasing continuous function F. If F is a general distribution function then we define

$$F^{-1}(u) := \inf\{x : F(x) \ge u\} \qquad \text{for } 0 < u < 1.$$

Let $U \sim U(0,1)$. Consider $X := F^{-1}(U)$. Then

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

But U has got uniform distribution and so $\mathbb{P}(U \leq u) = u$ for any $u \in \mathbb{R}$. Hence

$$\mathbb{P}(X \leq x) = F(x).$$

Thus X has the distribution F.

This means that if $(u_i)_{i \in \mathbb{N}}$ are samples from $U(0,1)$ then $(x_i)_{i \in \mathbb{N}}$ given by $x_i = F^{-1}(u_i)$ are samples from the distribution F.

Example 3.2. Say that we would like to generate N random samples from the exponential distribution with parameter $\lambda > 0$. First we would like to invert F. To that end we solve $y = 1 - e^{-\lambda x}$ for x:

$$x = \frac{-\ln(1-y)}{\lambda}$$

and hence

$$F^{-1}(y) = \frac{-\ln(1-y)}{\lambda}.$$

So we can generate N samples from $U(0,1)$, denote them by $(u_i)_{i=1}^N$ and get

$$x_i := F^{-1}(u_i) = \frac{-\ln(1-u_i)}{\lambda}.$$

Of course if $U \sim U(0,1)$ then $1 - U \sim U(0,1)$ and so we can equally well take

$$x_i := F^{-1}(1-u_i) = \frac{-\ln(u_i)}{\lambda}.$$

Exercise 3.3. We would like to sample from the double exponential distribution, also known as the Laplace distribution, which has the density given by

$$f(x) = \frac{\exp(-|x|)}{2}.$$

a) Show that the distribution given by the above density is

$$F(x) = \begin{cases} \frac{1}{2}e^{x} & \text{if} \quad x \leq 0, \\ 1 - \frac{1}{2}e^{-x} & \text{if} \quad x > 0. \end{cases}$$

b) Show that the inverse of F is given by

$$F^{-1}(x) = \begin{cases} \ln(2x) & \text{if} \quad x \leq \frac{1}{2}, \\ -\ln(-2x+2) & \text{if} \quad x > \frac{1}{2}. \end{cases}$$

c) Say we have generated N random samples $(u_i)_{i=1}^N$ distributed uniformly in $[0,1]$. How to generate $(x_i)_{i=1}^N$ samples from the distribution given by the Laplace density?

Very often we would like to generate random samples from the normal distribution. We know that for normally distributed random variables we can write their distribution in terms of the density

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} \phi(y)dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{y^2}{2}\right) dy.$$

Since $\phi(x)$ is strictly positive for all $x \in \mathbb{R}$ we see that F is a strictly increasing function of x and hence its inverse $F^{-1} : (0, 1) \to \mathbb{R}$ exists. Nevertheless there is no "closed form" formula for F^{-1}. This would suggest that one can not use the inversion method for generating normally distributed random numbers. This is not the case. We can either approximate F^{-1} or we can use Newton's method to find the inverse of F numerically.

3.3 Acceptance rejection method

This is a method for generating random samples from a continuous distribution with density f. To use it we have to assume that we can sample from $U(0, 1)$ and also from another distribution with a density g. Finally, we have assume that there is $c > 0$ such that

$$f(x) \leq cg(x) \quad \forall x \in \mathbb{R}. \tag{6}$$

To generate a sample from the distribution with density g we can use the following algorithm:

```
1. Generate a sample $u$ from $U(0,1)$.
2. Generate a sample $y$ from distribution with density $g$.
3. If $u \leq \frac{f(y)}{cg(y)}$ then $x = y$ is a random sample from a distribution with
   density $f$ and we stop.  Otherwise go to step 1.
```

Exercise 3.4. We know how to generate samples from a Laplace (or double exponential) distribution

$$g(x) = \frac{\exp(-|x|)}{2}, \tag{7}$$

see Exercise 3.3. We wish to use this to generate random variables with normal distribution, that is with density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right). \tag{8}$$

To that end we would first have to find $c > 0$ such that $\phi(x) \leq cg(x)$ for all $x \in \mathbb{R}$. Only if such c exists can we use the acceptance-rejection algorithm.

a) Show that the function $\xi : \mathbb{R} \to \mathbb{R}$ given by $\xi(x) = f(x)/g(x)$ is symmetric about $x = 0$.

b) Show that ξ has a maximum at $x = 1$.

c) Hence show that the inequality (6) is satisfied with

$$c = \sqrt{\frac{2e}{\pi}}. \tag{9}$$

From Exercise 3.4 we know that the condition (6) is satisfied for the normal density and double exponential density with c given by (9). To understand what the acceptance rejection algorithm actually does, let us look at Figure 1. In steps one and two we sample u from the uniform distribution and x from the “proposal distribution” (that is, the distribution we already know how to sample from), in this case the double exponential distribution. The value of x gives us the x-coordinate of each point in the plot. To get the the y-coordinate of each point we take u and scale it by $cg(x)$ Here g is the known density given by (7). Now we check whether $y = cug(x)$ is smaller or larger than $\phi(x)$, which is given by (8). If y lies on or under $\phi(x)$ it is accepted, while if it lies above $\phi(x)$ it is rejected.

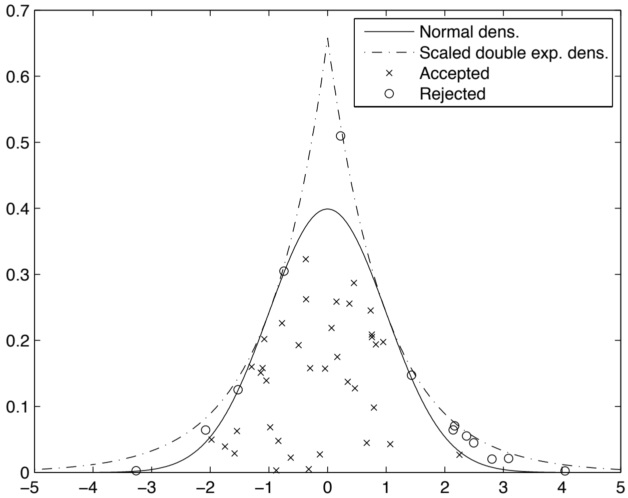


Figure 1: Acceptance-rejection used to generate samples from normal density

Looking at the algorithm we see that unless the generated u and x satisfy $u \leq \frac{f(x)}{cg(x)}$ we will be repeating steps one to three forever. So a natural question is what is the probability that the algorithm terminates at step three?

Proposition 3.5. *Assume that $U \sim U(0,1)$ and that $Y : \Omega \to \mathbb{R}^d$ is a random variable with density g. Let f be a density function. Let there be $c > 0$ such that (6) holds. Then*

$$\mathbb{P}\left(U \leq \frac{f(Y)}{cg(Y)}\right) = \frac{1}{c}.$$

Proof. As U, Y are independent with known densities, we have their joint density:

$$\mathbb{P}\left(U \leq \frac{f(Y)}{cg(Y)}\right) = \int\int_{\{(u,y)\in(0,1)\times\mathbb{R}: u \leq f(y)/cg(y)\}} g(y)\, du\, dy$$
$$= \int_{\mathbb{R}}\int_0^{f(y)/cg(y)} g(y)\, du\, dy = \int_{\mathbb{R}} \frac{1}{c} f(y)\, dy\,.$$

□

Thus the sample generated in step two is accepted with probability $1/c$. So the algorithm will need to generate random samples u and x exactly K times with the probability

$$P(K=k) = \left(1-\frac{1}{c}\right)^{k-1}\frac{1}{c}, \quad k \in \mathbb{N}.$$

Clearly the algorithm will be the most efficient if c is very close to 1.

So far we have only given an algorithm without justifying why the generated random sample has the desired distribution.

Proposition 3.6. *Assume that $U \sim U(0,1)$ and that $Y : \Omega \to \mathbb{R}^d$ is a random variable with density g. Let f be a density function. Let there be $c > 0$ such that (6) holds. Let X be the random variable with distribution given by the distribution of Y conditional on $U \leq f(Y)(cg(Y))^{-1}$. That is, for $A \subset \mathbb{R}^d$,*

$$\mathbb{P}(X \in A) := \mathbb{P}\left(Y \in A \middle| U \leq \frac{f(Y)}{cg(Y)}\right).$$

Then X has the density f.

Proof. Let A be a measurable subset of $\mathbb{R}^d$. To prove the proposition we need to show that

$$\mathbb{P}\left(Y \in A \middle| U \leq \frac{f(Y)}{cg(Y)}\right) = \int_A f(y)dy\,. \tag{10}$$

First we note that

$$\mathbb{P}\left(Y \in A \middle| U \leq \frac{f(Y)}{cg(Y)}\right) = \frac{\mathbb{P}\left(\{Y \in A\} \cap \left\{U \leq \frac{f(Y)}{cg(Y)}\right\}\right)}{\mathbb{P}\left(U \leq \frac{f(Y)}{cg(Y)}\right)}.$$

This means that, due to Proposition 3.5,

$$\mathbb{P}\left(Y \in A \middle| U \leq \frac{f(Y)}{cg(Y)}\right) = c\mathbb{P}\left(\{Y \in A\} \cap \left\{U \leq \frac{f(Y)}{cg(Y)}\right\}\right)$$
$$= c\int_A \int_0^{f(y)/cg(y)} g(y)\,du\,dy = \int_A f(y)\,dy.$$

But this is exactly (10), which concludes the proof. □

3.4 Box–Muller method for generating normally distributed samples

Very often we need to sample from the standard normal distribution. We have seen that we can use the acceptance-rejection method to that end or even the inversion method if we either approximate the normal density or use a numerical method for finding the inverse of the distribution function.

The Box–Muller method is a method designed to produce samples from standard normal distribution efficiently. It is based on the following observation.

Proposition 3.7. *The random variables X and Y are normally distributed and independent with mean 0 and variance 1 if and only if the random variables*

$$R := \sqrt{X^2+Y^2} \quad \textit{and} \quad \Theta := \arctan\left(\frac{Y}{X}\right) \tag{11}$$

are such that R^2 is exponentially distributed with parameter $1/2$ and Θ is uniformly distributed over the interval $[0, 2\pi]$ and R and Θ are independent.

Proof. Assume that X and Y are independent standard normal random variables. The joint density of the random variables X and Y is then given by

$$f_{X,Y}(x,y) := \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right),$$

since for independent continuous random variables their joint density is just the product of densities. We wish to calculate in the joint density of R and Θ. Recall that those are given by (11). We will now carry out essentially just the change of integration variables from cartesian to polar. Notice that with

$$g(x,y) := \sqrt{x^2+y^2} \quad \text{and} \quad h(x,y) = \arctan\left(\frac{y}{x}\right)$$

we have $R = g(X,Y)$ and $\Theta = h(X,Y)$. Furthermore, letting J denote the Jacobian of the transformation,

$$\det J = \det\begin{pmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{pmatrix} = \frac{\partial g}{\partial x}\frac{\partial h}{\partial y} - \frac{\partial g}{\partial y}\frac{\partial h}{\partial x}.$$

Now

$$\frac{\partial g}{\partial x} = \frac{x}{\sqrt{x^2+y^2}} \quad \text{and} \quad \frac{\partial g}{\partial y} = \frac{y}{\sqrt{x^2+y^2}}.$$

Recall that $\frac{d}{dx}\arctan(x) = (1+x^2)^{-1}$. Hence

$$\frac{\partial h}{\partial x} = -\frac{y}{x^2}\frac{1}{1+\frac{y^2}{x^2}} = \frac{-y}{x^2+y^2} \quad \text{and} \quad \frac{\partial h}{\partial y} = \frac{1}{x}\frac{1}{1+\frac{y^2}{x^2}} = \frac{x}{x^2+y^2}.$$

Altogether, letting $r = g(x,y)$,

$$\det J = \frac{x}{r}\frac{x}{r^2} - \frac{y}{r}\frac{(-y)}{r^2} = \frac{1}{r}.$$

Then, letting $\theta = h(x,y)$, the joint density of R and Θ is

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(x,y)(\det J)^{-1} = \frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)r.$$

Note that this is a standard calculation for the joint density of a pair of random variables that are given as functions of another pair of random variables. See e.g. Ross [6, Chapter 6, Section 7]. Let $f_R(r) := re^{-r^2/2}$ and $f_\Theta(\theta) = (2\pi)^{-1}$. We see that

$$f_{R,\Theta}(r,\theta) = f_R(r)f_\Theta(\theta).$$

Hence the random variables are independent. The random variable Θ already has the required distribution. The random variable R has the Raleigh distribution but we are more interested in the distribution of R^2. We see that for $x \leq 0$ we immediately have $\mathbb{P}(R^2 \leq x) = 0$. For $x > 0$:

$$\mathbb{P}(R^2 \leq x) = \mathbb{P}(R \leq \sqrt{x}) = \int_0^{\sqrt{x}} re^{-\frac{1}{2}r^2}dr = 1 - e^{-\frac{1}{2}x}.$$

Thus R^2 has exponential density with parameter $1/2$.

To prove the implication in the other direction we could start with the joint density $f_{R,\Theta}(r,\theta)$, carry out a change of variables, and derive the joint density $f_{X,Y}$. □

Armed with this knowledge we can give the Box–Muller algorithm for generating a pair of independent samples from the joint density of two independent standard normal random variables X and Y.

```
1. Sample $d$ from the exponential distribution $1 - e^{-x/2}$.
2. Sample $\theta$ from the uniform distribution on $(0, 2\pi)$.
3. Let $r = \sqrt{d}$ and $x = r\cos\theta$ and $y = r\sin\theta$.
```

Then x and y are the required samples. Note that we can use the inversion method to sample from the exponential distribution.

3.5 Generating correlated normally distributed samples

We will define a multivariate normal distribution as follows. Let $\mu \in \mathbb{R}^d$ be given and let Σ be a given symmetric, invertible, positive definite $d \times d$ matrix (it is also possible to consider positive semi-definite matrix Σ but for simplicity we ignore that situation here).

A matrix is positive definite if, for any $x \in \mathbb{R}^d$ such that $x \neq 0$, the inequality $x^T \Sigma x > 0$ holds (and positive semi-definite if we only have $x^T \Sigma x \geq 0$). From linear algebra we know that this is equivalent to:

1. There are d eigenvalues of a positive definite matrix Σ are all strictly positive (for positive semi-definite matrix they are non-negative) and the d corresponding eigenvectors are orthonormal.
2. There is a unique (up to multiplication by -1) lower-triangular matrix B such that $BB^T = \Sigma$. This is given by Cholesky decomposition.

For our purposes the matrix B s.t. $BB^T = \Sigma$ doesn't need to be lower triangular and we can use another method[3] to find it: let $(u^{(i)}, \lambda_i)_{i=1}^d$ be the eigenvectors and eigenvalues of Σ. Let $\Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ and U be the matrix of the eigenvectors i.e. $U := (u^{(1)}, \ldots, u^{(n)})$. Since the eigenvectors are orthonormal $UU^T = I$. Moreover, we have $\Sigma U = U\Lambda$. Hence $\Sigma = U\Lambda U^T$. Define $\Lambda^{1/2} := \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d})$. Then

$$\Sigma = U\Lambda^{1/2}(\Lambda^{1/2}U)^T = BB^T \quad \text{with } B = U\Lambda^{1/2}.$$

Let B be a $d \times d$ matrix such that $BB^T = \Sigma$.

Let $(X_i)_{i=1}^d$ be independent random variables with $N(0,1)$ distribution. Let $X = (X_1, \ldots, X_d)^T$ and $Z := \mu + BX$. We then say $Z \sim N(\mu, \Sigma)$ and call Σ the covariance matrix of Z.

Exercise 3.8. Show that $\operatorname{Cov}(Z_i, Z_j) = \mathbb{E}((Z_i - \mathbb{E}Z_i)(Z_j - \mathbb{E}Z_j)) = \Sigma_{ij}$. This justifies the name "covariance matrix" for Σ.

It is possible to show that the density function of $N(\mu, \Sigma)$ is

$$f(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}((x-\mu)^T \Sigma^{-1}(x-\mu))\right). \tag{12}$$

Note that if Σ is symmetric and invertible then Σ^{-1} is also symmetric.

[3] This is sometimes referred to as Principal Component Analysis (PCA).

Exercise 3.9. You will show that $Z = BX$ defined above has the density f given by (12) if $\mu = 0$.

i) Show that the characteristic function of $Y \sim N(0,1)$ is $t \mapsto \exp(-t^2/2)$. In other words, show that $\mathbb{E}(e^{itY}) = \exp(-t^2/2)$. *Hint.* complete the squares.

ii) Show that the characteristic function of a random variable Y with density f given by (12) is
$$\mathbb{E}\left(e^{i(\Sigma^{-1}\xi)^T Y}\right) = \exp\left(-\frac{1}{2}\xi^T\Sigma^{-1}\xi\right).$$
By taking $y = \Sigma^{-1}\xi$ conclude that
$$\mathbb{E}\left(e^{iy^T Y}\right) = \exp\left(-\frac{1}{2}y^T\Sigma y\right).$$
Hint. use a similar trick to completing squares. You can use the fact that since Σ^{-1} is symmetric $\xi^T\Sigma^{-1}x = (\Sigma^{-1}\xi)^T x$.

iii) Recall that two distributions are identiacal if and only if their characteristic functions are identical. Compute $\mathbb{E}\left(e^{iy^T Z}\right)$ for $Z = BX$ and $X = (X_1, \dots, X_d)^T$ with $(X_i)_{i=1}^d$ independent random variables such that $X_i \sim N(0,1)$. Hence conclude that Z has density given by (12) with $\mu = 0$.

You can now also try to show that all this works with $\mu \neq 0$.

To generate N independent samples from $N(\mu, \Sigma)$ (with $\mu \in \mathbb{R}^d$ and Σ a $d \times d$ matrix) we propose the following algorithm:

```
1. Use PCA or Cholesky decomposition to find $B$ such that $\Sigma = BB^T$.
2. Generate $N \times d$ samples from $N(0,1)$ and collect them in $N$ vectors each
   with $d$ components, labelled $\left(x^{(i)}\right)_{i=1}^N$, with $x^{(i)} \in \mathbb{R}^d$ for each $i = 1, \dots, N$.
3. For each $i = 1, \dots, N$ let $z^{(i)} := \mu + Bx^{(i)}$. Now $\left(z^{(i)}\right)_{i=1}^N$ are independent
   samples from $N(\mu, \Sigma)$.
```

3.6 Summary

- We have seen that linear congruential generators can be used to give sequences of pseudorandom natural numbers.
- These can be used to generate samples from the uniform distribution.
- We can then use the inversion method or the acceptance-rejection method to generate samples from other distributions.
- For generating samples from the normal density the Box–Muller algorithm is generally sufficiently efficient.
- The Ziggurat algorithm which is based on acceptance rejection method optimized for efficient implementation is what is used by state of the art numerical libraries (and Matlab).

4 Variance reduction

We will discuss variance reduction techniques in this section. These are techniques which allow us to get a better estimate, on average, without increasing the sample size.

4.1 Antithetic variates

The idea is to reduce variance by introducing negative dependence in pairs of replications. Intuitively, an extremely large draw from a distribution can be compensated by an extremely low one and so the variance of the average will be reduced.

Example 4.1. We can use antithetic variates when sampling from the following distributions:

a) If $U \sim U(0,1)$ then $1 - U \sim U(0,1)$.

b) If $Z \sim N(0,1)$ then $-Z \sim N(0,1)$.

c) If $U \sim U(0,1)$ then $F^{-1}(U)$ and $F^{-1}(1-U)$ both have distribution F.

The method then consists of considering N pairs $(X_1, \tilde{X}_1), \dots, (X_N, \tilde{X}_N)$ that are independent and identically distributed but such that for each i the random variables X_i and $\tilde{X}_i$ are identically distributed but not independent. Assume that there are random variables X and $\tilde{X}$ with the same distribution as X_i and as $\tilde{X}_i$ respectively and such that $\mathbb{E}(X_i) = \mathbb{E}(X)$ and $\operatorname{Var}(X_i) = \operatorname{Var}(X)$ and $\mathbb{E}(\tilde{X}_i) = \mathbb{E}(\tilde{X})$ and $\operatorname{Var}(\tilde{X}_i) = \operatorname{Var}(\tilde{X})$ for $i = 1, \dots, N$.

Definition 4.2 (Antithetic variates estimator for $\mathbb{E}X$)**.** *Let*

$$\bar{X}_N^{AV} := \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^N X_i + \frac{1}{N}\sum_{i=1}^N \tilde{X}_i\right)$$

be the antithetic variates estimator for $\mathbb{E}(X)$*.*

It is easy to check that this estimator is unbiased.

We would like to apply central limit theorem to X_N^{AV}. Of course we can not use the sequence $X_1, \tilde{X}_1, X_2, \tilde{X}_2, \dots, X_N, \tilde{X}_N$ since those random variables are not independent. But the random variables

$$\frac{X_1 + \tilde{X}_1}{2}, \frac{X_2 + \tilde{X}_2}{2}, \dots, \frac{X_N + \tilde{X}_N}{2}$$

are independent and

$$\bar{X}_N^{\mathrm{AV}} = \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^N X_i + \frac{1}{N}\sum_{i=1}^N \tilde{X}_i\right) = \frac{1}{N}\sum_{i=1}^N \frac{X_i + \tilde{X}_i}{2}.$$

Hence due to central limit theorem

$$\frac{\sqrt{N}\left(X_N^{\mathrm{AV}} - \mathbb{E}\left(\frac{X+\tilde{X}}{2}\right)\right)}{\sigma_{\mathrm{AV}}} \xrightarrow{d} N(0,1) \quad \text{as} \quad N \to \infty.$$

Of course X and $\tilde{X}$ are identically distributed and so $\mathbb{E}\left(\frac{X+\tilde{X}}{2}\right) = \mathbb{E}(X)$. Here

$$\sigma_{\text{AV}} = \sqrt{\text{Var}\left(\frac{X+\tilde{X}}{2}\right)}.$$

Now we calculate σ_{AV}. The question of course is how it compares to $\text{Var}(X)$. First

$$\text{Var}(X+\tilde{X}) = \text{Var}X + \text{Var}\tilde{X} + 2\text{Cov}(X,\tilde{X}) = 2\text{Var}X + 2\text{Cov}(X,\tilde{X}).$$

Thus

$$\sigma_{\text{AV}}^2 = \frac{1}{2}\left(\text{Var}(X) + \text{Cov}(X,\tilde{X})\right).$$

So the method will decrease variance provided that $\text{Cov}(X,\tilde{X}) < 0$.

The problem is that, typically, $\mathbb{E}(X)$ is unknown and so $\text{Var}(X)$ is unknown and also $\text{Cov}(X,\tilde{X})$ is unknown. One way to overcome this is to test experimentally, that is estimate $\text{Cov}(X,\tilde{X})$ itself using Monte Carlo.

There is also a theoretical result that may help in some situations. Say that for example $(Z_i)_{i=1}^d$ are independent and distributed according to $N(0,1)$. Let $X := f(Z_1, Z_2, \dots, Z^d)$ for some increasing f which is an increasing function of all its arguments. Then with $\tilde{X} := f(-Z_1, -Z_2, \dots, -Z^d)$ we have $\mathbb{E}(X\tilde{X}) \leq E(X)\mathbb{E}(\tilde{X})$ and hence $\text{Cov}(X,\tilde{X}) \leq 0$. The same is also true if we replace Z_i with U_i and Z_i with $1-U_i$.

Example 4.3. We will employ antithetic variates in a simple situation. Imagine that we would like to use Monte Carlo method to estimate $v(t,S)$ given by (2). We would then use

$$X_i := e^{-r(T-t)}\left[S\exp\left(\left(r-\frac{1}{2}\sigma^2\right)(T-t) + \sigma\sqrt{T-t}Z_i\right) - K\right]_+$$

where $(Z_i)_{i=1}^N$ are independent and identically distributed standard normal variables together with

$$\tilde{X}_i := e^{-r(T-t)}\left[S\exp\left(\left(r-\frac{1}{2}\sigma^2\right)(T-t) - \sigma\sqrt{T-t}Z_i\right) - K\right]_+.$$

An estimator for $v(t,S)$ is then

$$v_N(t,S) := \frac{1}{N}\sum_{i=1}^N \left(\frac{X_i + \tilde{X}_i}{2}\right).$$

4.2 Control variates

This is another variance reduction technique. Recall that our aim is to reduce the variance of our Monte Carlo estimate (and thus the improve the estimate), while keeping the number of samples fixed. The number of random samples used by a Monte Carlo method is a good proxy for the computational effort required. Thus improving accuracy while keeping the number of samples fixed mean we are improving accuracy while keeping the computational effort fixed.

The main idea behind control variates is to use a random variable with a known expectation that is highly correlated with the random variable whose expectation we seek to correct our estimate. For example while $\mathbb{E}(e^{-rT}[S_T - K]_+)$ is unknown (unless we use the Black–Scholes formula) the quantity $\mathbb{E}(e^{-rT}S_T) = S$, since the evolution of the discounted risky asset is a Martingale.

In what follows we will use X to denote the random variable whose expectation is known. We will use Y to denote the random variable whose expectation we wish to estimate.

Assume we have $(X_i, Y_i)_{i=1}^N$ independent and identically distributed with the same distribution as (X, Y). As always, these will be used in the analysis instead of specific samples $(x_i)_{i=1}^N$ and $(y_i)_{i=1}^N$.

Definition 4.4 (Control variates estimator with parameter $b \in \mathbb{R}$ for $\mathbb{E}Y$)**.** *Let* $Y_i(b) := Y_i - b(X_i - \mathbb{E}X)$ *and let*

$$\bar{Y}_N(b) := \frac{1}{N}\sum_{i=1}^N Y_i(b)$$

be the control variate estimator with parameter b *for* $\mathbb{E}(Y)$*.*

Let $\bar{Y}_N$ and $\bar{X}_N$ be the estimators for $\mathbb{E}(Y)$ and $\mathbb{E}(X)$ respectively. Recall that $\mathbb{E}X$ is assumed to be known. Note that

$$\mathbb{E}\left(\bar{Y}_N(b)\right) = \mathbb{E}\left(\bar{Y}_N - b(\bar{X}_N - \mathbb{E}(X))\right) = \bar{Y}_N = \mathbb{E}(Y)$$

and so the control variates estimator with parameter b is (for $b \in \mathbb{R}$) is unbiased. Let $\sigma_b := \sqrt{\mathrm{Var}(Y_i(b))}$. From the central limit theorem we know that

$$\frac{\sqrt{N}\left(\bar{Y}_N(b) - \mathbb{E}\left(\bar{Y}_N(b)\right)\right)}{\sigma_b} \xrightarrow{d} N(0,1) \quad \text{as} \quad N \to \infty.$$

Of course, as $\mathbb{E}\left(\bar{Y}_N(b)\right) = \mathbb{E}Y$ we get the same asymptotic error bounds as for the ordinary estimator, see Proposition 2.11 but with σ replaced by σ_b.

Proposition 4.5. *Let* $\sigma_Y := \sqrt{Var(Y)}$ *and* $\sigma_X := \sqrt{Var(X)}$*. Let*

$$\rho_{XY} := \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

Then there is $b^* \in \mathbb{R}$ *such that* $\sigma_{b^*}^2 = \sigma_Y^2(1 - \rho_{XY}^2)$*.*

Before we proceed to prove this result let us make some observations about what this implies.

Remark 4.6. We can conclude the following.

a) The higher the correlation between X and Y the higher variance reduction can be achieved.

b) The sign of the correlation is not important.

c) With an ordinary estimator we would need

$$\frac{N}{1-\rho_{XY}^2}$$

samples to achieve the same asymptotic error bound (i.e. accuracy) as the control variates estimator.

This means that if we can get X_i without increasing the computational effort then control variate estimator always performs better than ordinary estimator. In practice producing X_i and the slightly more complicated calculation costs something in terms of computing time and so one would not use control variates unless ρ_{XY} is “reasonably high”. What this means must almost always be determined experimentally.

d) If we assume that we get X_i for “free” then a correlation of 0.95 produces a speedup of factor 10 (we can use ten times smaller sample size while maintaining accuracy). Correlation of $2^{-1/2} \approx 0.7$ produces only a speedup of factor 2.

Proof of Proposition 4.5. Recall that, for a general random variable Z we have

$$\mathrm{Var}(Z) = \mathbb{E}\left((Z-\mathbb{E}Z)^2\right) = \mathbb{E}(Z^2) - \mathbb{E}(Z)^2.$$

Further recall that

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left((X-\mathbb{E}X)(Y-\mathbb{E}Y)\right).$$

Hence

$$\begin{aligned}\sigma_X\sigma_Y\rho_{XY} &= \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \mathbb{E}\left((X-\mathbb{E}X)(Y-\mathbb{E}Y)\right)\\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

Now recall that $Y_i(b) = Y_i - b(X_i - \mathbb{E}X)$ and so $\mathbb{E}Y_i(b) = \mathbb{E}Y_i$ as $\mathbb{E}X_i = \mathbb{E}X$. So, if we use the fact that X_i and Y_i have the same distribution as X and Y respectively we obtain

$$\begin{aligned}\mathrm{Var}(Y_i(b)) &= \mathbb{E}\left(Y_i(b)^2\right) - \mathbb{E}(Y_i)^2 = \mathbb{E}\left(Y_i^2 - 2bY_i(X_i-\mathbb{E}X) + b^2(X_i-\mathbb{E}X)^2\right) - \mathbb{E}(Y_i)^2\\ &= \mathbb{E}(Y^2) - 2b\mathbb{E}(YX) + 2b\mathbb{E}(Y)\mathbb{E}(X) + b^2\mathbb{E}\left((X-\mathbb{E}X)^2\right) - \mathbb{E}(Y)^2\\ &= \mathrm{Var}(Y) - 2b\mathrm{Cov}(X,Y) + b^2\mathrm{Var}(X) = \sigma_Y^2 - 2b\sigma_X\sigma_Y\rho_{XY} + b^2\sigma_X^2.\end{aligned}$$

Our aim is to minimise the variance. So we must choose b that minimises the above expression. Hence we seek b^* such that

$$0 = \frac{d}{db}\left(\sigma_Y^2 - 2b\sigma_X\sigma_Y\rho_{XY} + b^2\sigma_X^2\right) = -2\sigma_X\sigma_Y\rho_{XY} + 2b\sigma_X^2.$$

So $b^* = \sigma_Y\rho_{XY}\sigma_X^{-1}$. Then

$$\sigma_{b^*}^2 = \mathrm{Var}(Y_i(b^*)) = \sigma_Y^2 - \sigma_Y^2\rho_{XY}^2.$$

□

There is one "small" problem remaining. Remember that we are trying to estimate $\mathbb{E}Y$. But if we do not know this then it is rather unlikely that we will actually know $\sigma_Y = \mathrm{Var}(Y)$ and $\mathrm{Cov}(X, Y)$ and hence ρ_{XY}. So while the above result of variance reduction of the control variates method is correct it is not usually usable in practice. What one can do though, is to take an estimate for b^*, using the samples generated during the Monte Carlo method and use

$$\hat{b}_N^* := \frac{\sum_{i=1}^N (x_i - \bar{x}_N)(y_i - \bar{y}_N)}{\sum_{i=1}^N (x_i - \bar{x}_N)^2}. \tag{13}$$

Let $\hat{B}_N^*$ denote the random variable we obtain if in the above equation we use X_i and Y_i in place of x_i and y_i and $\bar{X}_N$ and $\bar{Y}_N$ in place of $\bar{x}_N$ and $\bar{y}_N$. Then

$$\begin{aligned}\mathbb{E}(\bar{Y}_N(\hat{B}_N^*) &= \mathbb{E}\left(\bar{Y}_N - \hat{B}_N^*\left(\bar{X}_N - \mathbb{E}X\right)\right) = \mathbb{E}Y - \mathbb{E}\left(\hat{B}_N^*\bar{X}\right) + \mathbb{E}\left(\hat{B}_N^*\right)\mathbb{E}(X)\\ &= \mathbb{E}Y + \left(\mathbb{E}\left(\hat{B}_N^*\right)\mathbb{E}(X) - \mathbb{E}\left(\hat{B}_N^*\bar{X}\right)\right).\end{aligned}$$

We see that $\mathbb{E}(\bar{Y}_N(\hat{B}_N^*)$ is no longer an unbiased estimator for $\mathbb{E}Y$. We have bias equal to $\mathbb{E}\left(\hat{B}_N^*\right)\mathbb{E}(X) - \mathbb{E}\left(\hat{B}_N^*\bar{X}\right)$. It can be shown, though we do not do it here, that the bias is of order $1/N$. Since for a Monte Carlo method the error is of order $1/\sqrt{N}$ we can say that for large N this is not significant. Hence in practice one would use control variates with the estimate given by (13).

Example 4.7. We consider a call option price in the Black–Scholes framework (so we know the exact price as it is given by the Black–Scholes formula). In the risk neutral measure the evolution of the risky asset is given by

$$dS_u = rS_u du + \sigma S_u dW_u, \;\; S_t = S.$$

Here $(W_t)_{t\in[0,T]}$ is a Wiener process in the risk neutral measure.

We have shown before that in the risk neutral measure the process $(e^{-r(T-u)}S_u)_{u\in[t,T]}$ is a martingale and hence $\mathbb{E}\left(e^{-r(T-t)}S_T)\right) = S_t = S$. Now we would like to use control variates to estimate

$$v(t, S) = \mathbb{E}\left(e^{-r(T-t)}[S_T - K]_+\right).$$

We take $Y = e^{-r(T-t)}[S_T - K]_+$ and so we are estimating $\mathbb{E}Y$. We take $X = e^{-r(T-t)}S_T$ as our control since we know that $\mathbb{E}X = S_t = S$.

Now we generate N samples from standard normal distribution and denote them by $(z_i)_{i=1}^N$. We then get an estimate

$$\bar{x}_N = \frac{1}{N}\sum_{i=1}^N e^{-r(T-t)}S\exp\left(\left(r - \frac{1}{2}\sigma^2\right)(T-t) + \sigma\sqrt{T-t}z_i\right)$$

for $\mathbb{E}X = \mathbb{E}e^{-r(T-t)}S_T$. We also calculate

$$\bar{y}_N = \frac{1}{N}\sum_{i=1}^N e^{-r(T-t)}\left[S\exp\left(\left(r - \frac{1}{2}\sigma^2\right)(T-t) + \sigma\sqrt{T-t}z_i\right) - K\right]_+.$$

This is, by itself, an estimate for $\mathbb{E}Y$. But we would like to use control variates and so we use (13) as an estimate for $\hat{b}_N^*$. Then our control variates estimate for $\mathbb{E}Y$ is given by

$$\bar{y}(\hat{b}_N^*) = \bar{y}_N - \hat{b}_N^*(\bar{x}_N - S).$$

This is a number we know how to calculate.

Notice that we did not have to generate another set of random samples to use control variates in this case. This means that whatever reduction in variance we are achieving, it is achieved at the cost of evaluating exp and few additions N times.

4.3 Multiple control variates

It is also possible to generalize the control variate method to multiple controls. Imagine you have a random variable Y and you wish to estimate $\mathbb{E}Y$. Assume that you have $X^{(k)}$ with $\mathbb{E}X^{(k)}$ known for $k = 1, \ldots, m$. Let Σ_X be the $m \times m$ covariance matrix for X, and let Σ_{XY} be the $m \times 1$ covariance matrix for $(X^{(k)}, Y)$, i.e.

$$(\Sigma_X)_{jk} := \mathrm{Cov}(X^{(j)}, X^{(k)}), \quad (\Sigma_{XY})_{j1} := \mathrm{Cov}(X^{(j)}, Y)$$

and as before $\sigma_Y^2 = \mathrm{Var}(Y)$ is a scalar. Hence we have the correlation matrix

$$\begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^T & \sigma_Y^2 \end{pmatrix}$$

for the $\mathbb{R}^{m+1}$ valued r.v. $(X^{(1)}, \ldots, X^{(m)}, Y)$. Define, for $b \in \mathbb{R}^m$,

$$Y(b) := Y - b^T(X - \mathbb{E}X).$$

Proposition 4.8. *For $b \in \mathbb{R}^m$ we have*

$$Var(Y(b)) = \sigma_Y^2 - 2b^T\Sigma_{XY} + b^T\Sigma_X b.$$

The $b^* \in \mathbb{R}^m$ which minimizes $Var[Y(b)]$ is given by

$$b^* = \Sigma_X^{-1}\Sigma_{XY}.$$

4.4 Summary

- We have shown that the appropriate convergence concept for Monte Carlo methods is convergence in distributions.
- We have used central limit theorem to derive an asymptotic error bound for a Monte Carlo approximation in terms of number of samples and variance.
- We have seen that convergence is always of order 1/2.
- We have seen that reducing variance improves the estimate.
- We have discussed two techniques for variance reduction: antithetic variates and control variates.
- We have seen that both provide a tangible improvement only in specific situation and hence one must analyse the problem before deciding whether to use a specific variance reduction technique.
- There are other variance reduction techniques that we have not discussed.

4.5 Further reading

This material is based in particular on Glasserman [1]. See in particular Chapter 2 and Chapter 4, Section 1 and 2.

For more details on variance reduction techniques and various applications see again Glasserman [1].

5 Some applications

5.1 Asian options

Asian options differ sligtly from the European options: the payoff is not based only on the price of the risky asset at the exercise date T but on the average price of the risky asset over several dates before the exercise date.

5.1.1 Geometric Asian option

Let $(S_t)_{t\in[0,T]}$ denote the price of the risky asset at time t. Let

$$\bar{S}_G := \left(\prod_{i=1}^{n} S(T_i)\right)^{1/n},$$

where $(T_i)_{i=1}^{N}$ are some dates that are fixed in the option contract and are such that

$$t < T_1 < T_2 < \ldots < T_N = T.$$

The option contract also specifies the strike $K > 0$. The option payoff at the expiry time T is given by $[\bar{S}_G - K]_+$. Assume we work in the Black–Scholes framework. Then the price, denoted by v_{A_A}, is given by

$$v_{A_G}(t,S) = \mathbb{E}\left(e^{-r(T-t)}[\bar{S}_G - K]_+\right),$$

where the expectation is, as always, in the risk neutral measure.

Exercise 5.1. Show that the Black–Scholes formula can be used with expiry time $\bar{T} - t$, where, $\bar{T} := n^{-1}\sum_{i=1}^{n} T_i$, risk free rate r, strike K, volatility $\bar{\sigma}$ given by

$$\bar{\sigma} = \frac{1}{\sqrt{\bar{T}-t}}\frac{\sigma}{n}\sqrt{\sum_{i=1}^{n}(2i-1)(T_{n+1-i}-t)}.$$

and spot price $Se^{\gamma\bar{T}-t}$, where $\gamma := (1/2)(\bar{\sigma}^2 - \sigma^2)$.

5.1.2 Arithmetic average option

Let $(S_t)_{t\in[0,T]}$ denote the price of the risky asset at time t.
Let

$$\bar{S}_A := \frac{1}{n}\left(\sum_{i=1}^{n} S(T_i)\right),$$

where $(T_i)_{i=1}^{N}$ are some dates that are fixed in the option contract and are such that

$$t < T_1 < T_2 < \ldots < T_N = T.$$

The option contract also specifies the strike $K > 0$. The option payoff at the expiry time T is given by $[\bar{S}_A - K]_+$. Assume we work in the Black–Scholes framework. Then the price, in the risk neutral measure, denoted by v_{A_A} is given by

$$v_{A_A}(t,S) = \mathbb{E}\left(e^{-r(T-t)}[\bar{S}_A - K]_+\right).$$

There is no known formula that would tell us the price of this option.

We can use Monte Carlo to estimate the value of the option. In order to do that we need to simulate the prices of the risky asset at times T_i with $i = 1, \ldots, n$. Since we are working in the Black–Scholes framework we know that if $S_t = S$ then

$$S(u) = S \exp\left(\left(r - \frac{1}{2}\sigma^2\right)(u - t) + \sigma(W_u - W_t)\right) \quad \forall u \in [t, T].$$

Hence for any T_i we know that

$$S(u) = S_{T_i} \exp\left(\left(r - \frac{1}{2}\sigma^2\right)(u - T_i) + \sigma(W_u - W_{T_i})\right) \quad \forall u \in [T_i, T].$$

So, setting $T_0 := t$, we get

$$\begin{aligned} S(T_i) &= S_{T_{i-1}} \exp\left(\left(r - \frac{1}{2}\sigma^2\right)(T_i - T_{i-1}) + \sigma(W_{T_i} - W_{T_{i-1}})\right) \\ &\overset{d}{=} S_{T_{i-1}} \exp\left(\left(r - \frac{1}{2}\sigma^2\right)(T_i - T_{i-1}) + \sigma\sqrt{T_i - T_{i-1}}Z^i\right), \quad i = 1, \ldots, N. \end{aligned} \tag{14}$$

where $Z^i \sim N(0,1)$ are independent standard normal random variables and where $\overset{d}{=}$ is used to denote that two random variables have the same distribution.

For each $i = 1, 2, \ldots n$ we can take N samples from $N(0,1)$ (and thus in total we have $n \cdot N$ samples) and denote them z_j^i, $i = 1, \ldots, n$ and $j = 1, \ldots, N$. Let us define $s_j(T_0) := S$ and define, for $j = 1, \ldots, N$,

$$s_j(T_i) := s_j(T_{i-1}) \exp\left(\left(r - \frac{1}{2}\sigma^2\right)(T_i - T_{i-1}) + \sigma\sqrt{T_i - T_{i-1}}z_j^i\right), \quad i = 1, \ldots, N.$$

Let

$$x_j := e^{-r(T-t)} \left[\frac{1}{n}\sum_{i=1}^n s_j(T_i) - K\right]_+ .$$

Our Monte Carlo approximation is then

$$v_{A_A,N}(t, S) = \frac{1}{N}\sum_{j=1}^N x_j.$$

If we wanted an approximation for the error we would need to use Proposition 2.11. To that end let for each $i = 1, \ldots, n$ let there be N independent standard normal random variables denoted $(Z_j^i)_{j=1}^N$. Let $S_j(T_0) := S$ and

$$S_j(T_i) := S_j(T_{i-1}) \exp\left(\left(r - \frac{1}{2}\sigma^2\right)(T_i - T_{i-1}) + \sigma\sqrt{T_i - T_{i-1}}Z_j^i\right), \quad i = 1, \ldots, N.$$

Let

$$X_j = e^{-r(T-t)} \left[\frac{1}{n}\sum_{i=1}^n S_j(T_i) - K\right]_+ .$$

Note that $\mathbb{E}X_j = v_{A_A}(t, S)$ is the quantity we wish to estimate. Let us use $\sigma_v := \text{Var}(X_j)$. Then

$$\bar{X}_N := \frac{1}{N}\sum_{j=1}^N X_j$$

is an unbiased estimator for $v_{A_A}(t,S)$ and $\forall \delta > 0$

$$\mathbb{P}\left(\bar{X}_N - z_{\delta/2}\frac{\sigma_v}{\sqrt{N}} \leq v_{A_A}(t,S) \leq \bar{X}_N + z_{\delta/2}\frac{\sigma_v}{\sqrt{N}} \right) \to 1 - \delta \quad \text{as} \quad N \to \infty.$$

Remark 5.2. Note that in Section 5.1.2 the option price depends on the *path* of the of the process used to model the risky asset at times $T_1, T_2, \ldots, T_n$ and not just T. Since we know the solution to the equation

$$dS_u = \mu S_u du + \sigma S_u dW_u, \quad S_t = S,$$

we know exactly the distributions the random variables $S(T_1), S(T_2), \ldots, S(T_n)$ have. It is given by (14).

If we use a general stochastic differential equation to model the risky asset we have

$$dX_u = b(u, X_u)du + \sigma(u, X_u)dW_u, \quad X_t = x.$$

Under appropriate assumptions on b and σ we would know that a solution of such equation exists but we would not necessarily know what it is. In this case we could *approximate* $X(T_1) \approx X^1, X(T_2) \approx X^2, \ldots, X(T_n) \approx X^n$ using, for example the Explicit Euler scheme

$$X^i = X^{i-1} + b(T_{i-1}, X^{i-1})(T_i - T_{i-1}) + \sigma(T_{i-1}, X^{i-1})(W_{T_i} - W_{T_{i-1}}).$$

If we now proceed as before there would be two sources of error in our approximation. One would arise, as always, from the use of a Monte Carlo method and can be estimated using the Central limit theorem. The other error would arise from the approximation of $(X_u)_{u\in[t,T]}$ by $(X^i)_{i=1}^n$ and is a type of *discretization* error. In practice one would subdivide $[t,T]$ into more subintervals than just those required for the arithmetic average option in order to decrease the discretization error.

Example 5.3 (Control variates for arithmetic average option)**.** We can use the price of the geometric average option as a control variate in a Monte Carlo method when estimating the arithmetic average option price. Of course we could also use the discounted evolution of the risky asset as in Example 4.7 but it can be shown (at least experimentally) the the correlation between the payoff of the geometric average option and the payoff of the arithmetic average option are higher.

This method is very useful because it works in many situations where a simple model leads to a price given by a formula that we can then use to improve our Monte Carlo method in a more realistic model. Other examples of use are e.g. option pricing with stochastic volatility (with the price given by Black–Scholes formula given as a control).

5.2 Options on several risky assets

We can use Monte Carlo methods to price options on more than one risky asset. In order to do that we need some models for the evolution of the risky assets. The basic one extends the idea of Geometric brownian motion to several dimensions using a d-dimensional Wiener process.

5.2.1 Wiener process in $\mathbb{R}^d$

It will occasionally be more convenient to write $X(t)$ instead of X_t for some stochastic process $\{X_t\}_{t\geq 0} = \{X(t)\}_{t\geq 0}$. This is just a matter of notation.

A process $\{W(t)\}_{t\geq 0}$ is a Wiener process on $\mathbb{R}^d$ if $W(0) = 0$ almost surely, it has independent increments, the function $t \mapsto W(t)$ is almost surely continuous and

$$W(t) - W(s) \sim N(0, (t-s)I),$$

where I is a $d \times d$ identity matrix.

Note that if $\{W_1(t)\}_{t\geq 0}$, $\{W_2(t)\}_{t\geq 0}$, $\ldots$, $(W_d(t))_{t\geq 0}$ are independent, 1-dimensional Wiener processes, then the process given by $W(t) = (W_1(t), W_2(t), \ldots, W_d(t))^T$ satisfies the above definition and so is a Wiener process on $\mathbb{R}^d$.

5.2.2 Multi-dimensional geometric Brownian motion

Let $\{W(t)\}_{t\geq 0}$ be a Wiener process on $\mathbb{R}^k$. Then

$$X(t) := BW(t),$$

is a process with covariance Σ. In fact

$$X(t) - X(s) \sim N\left(0, (t-s)\Sigma\right).$$

If, for a vector z we write $z = (z_1, z_2, \ldots, z_d)^T$ then then

$$X_i(t) = B_{i1}W_1(t) + B_{i2}W_2(t) + \cdots + B_{ik}W_k(t), \; i = 1, 2, \ldots d. \tag{15}$$

Let $\mu, \sigma \in \mathbb{R}^d$. We can model d correlated risky assets using

$$dS(u) = \operatorname{diag}(S(u))\left(\mu du + \operatorname{diag}(\sigma) dX(u)\right), \; S(t) = S.$$

Here $\operatorname{diag}(z)$ denotes a $d \times d$ matrix with the diagonal equal to z and all off-diagonal elements equal to zero. The above equation is equivalent to

$$dS_i(u) = S_i(u)\left(\mu_i du + \sigma_i dX_i(u)\right), \; S_i(t) = S, \; i = 1, 2, \ldots, d.$$

which is in turn

$$dS_i(u) = S_i(u)\left(\mu_i du + \sigma_i \sum_{j=1}^{k} B_{ij} dW_j(u)\right), \; S_i(t) = S, \; i = 1, 2, \ldots, d.$$

We would like to obtain an explicit solution to the stochastic differential equaiton just like in the 1-dimensional case. For this we need the multi-dimensionla Itô formula, see Section A.1. In fact we only need the following: if $\{W_i(t)\}$ and $\{W_j(t)\}$ are independent Wiener processes on $\mathbb{R}$ whenever $i \neq j$ then

$$\langle dW_i(t), dW_j(t)\rangle = \delta_{ij}dt = \begin{cases} dt & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y_i(t) = \ln S_i(t)$. Then using the Itô formula we get

$$dY_i(u) = \mu_i du + \sigma_i dX(u) - \frac{1}{2}\sigma_i^2 \langle dX_i(u), dX_i(u)\rangle.$$

Using (15) we obtain

$$\langle dX_i(u), dX_i(u)\rangle = B_{i1}^2 du + B_{i2}^2 du + \cdots + B_{id}^2 du.$$

Hence

$$dY_i(u) = \left(\mu_i - \frac{1}{2}\sigma_i^2 \sum_{j=1}^{d} B_{ij}^2\right) du + \sigma_i \sum_{j=1}^{d} B_{ij} dW_j(u).$$

Thus

$$S_i(u) = S_i \exp\left(\left(\mu_i - \frac{1}{2}\sigma_i^2 \sum_{j=1}^{d} B_{ij}^2\right)(u-t) + \sigma_i \sum_{j=1}^{d} B_{ij}(W_j(u) - W_j(t))\right).$$

Let the matrix C be such that $C_{ij} = B_{ij}^2$ and let the vector ν be such that $\nu_i = \sigma_i^2$. Then, in a matrix form the above equation reads,

$$S(u) = \text{diag}(S)\exp\left((u-t)\left(\mu - \frac{1}{2}\text{diag}(\nu)C1_d\right) + \text{diag}(\sigma)B(W(u) - W(t))\right),$$

where $1_d := (1, \ldots, 1)^T \in \mathbb{R}^d$.

5.2.3 European style options

Let $T > 0$ be fixed. Let a function $g : \mathbb{R}^d \to \mathbb{R}$ be given. This specifies the option payoff: the holder has gets $g(S_T)$ at time T, where $S_T \in \mathbb{R}^d$ is the price of the risky asset at time T. Assume that B such that $BB^T = \Sigma$ is a $d \times d$ matrix. Then, as in the one dimensional case (though we do not show this), it is possible to find a measure $\mathbb{Q}$ which is equivalent to $\mathbb{P}$ such that $\{e^{-rt}S(t)\}_{t\in[0,T]}$ is a $\mathbb{Q}$-Martingale with evolution given by

$$dS_i(u) = S_i(u)\left(rdu + \sigma_i d\bar{X}_i(u)\right), \; S_i(t) = S, \; i = 1, 2, \ldots, d. \tag{16}$$

Here $\{\bar{X}(t)\}_{t\in[0,T]}$ is given by $\bar{X}(t) = B\bar{W}_t$, where $\{\bar{W}(t)\}_{t\in[0,T]}$ is a d-dimensional $\mathbb{Q}$-Wiener process.

Furthemore it can be shown that the option price is, at time t with spot asset price $S \in \mathbb{R}^d$,

$$v(t, S) = e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}(g(S(T))|S(t) = S).$$

We apply the calculation from Section 5.2.2 to (16) to see that

$$v(t, S) = e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}\Bigg[g\Bigg(\text{diag}(S)\exp\Bigg((T-t)\left(r - \frac{1}{2}\text{diag}(\nu)C1_d\right) + \text{diag}(\sigma)B(W(T) - W(t))\Bigg)\Bigg)\Bigg].$$

In order apply Monte Carlo methods it is useful to note that if $Z \sim N(0, (T-t)I)$ then

$$\begin{aligned} v(t,S) = e^{-r(T-t)}\mathbb{E}\Bigg[g\Bigg(\operatorname{diag}(S)\exp\Bigg((T-t)\left(r - \frac{1}{2}\operatorname{diag}(\nu)C1_d\right) \\ +\sqrt{T-t}\operatorname{diag}(\sigma)BZ\Bigg)\Bigg)\Bigg]. \end{aligned} \tag{17}$$

Remark 5.4. If $d = 1$ then the most efficient method of evaluating the above expression numerically is most likely some form of numerical integration (e.g. trapezium rule) using the normal density. However for $d > 3$ such numerical method will suffer from the “curse of dimensionality”. This means that where for example the trapezium rule needs to evaluate the integrand at $2^1 = 2$ points, if $d = 1$, it will need $2^2 = 4$ points if $d = 2$ and $2^3 = 8$ points if $d = 3$. The consequence of this that while the numerical integration method may have higher order of convergence than a Monte-Carlo method this advantage gets lost quickly as d grows.

What are the possible payoffs of options that are traded? We give two examples.

Basket option Let a vector of “weights” be given and let

$$g(S) = \left[\sum_{i=1}^d w_i S_i - K\right]_+ .$$

Outperformance option Let a vector of “weights” be given. The payoff is

$$g(S) = \left[\max_{i=1,\ldots,d}(w_i S_i) - K\right]_+ .$$

See Glasserman [1, Chapter 3, Section 2.3] for more examples.

Example 5.5. Assume we wish to use Monte Carlo methods to price a basket option. Assume that there are d risky assets, we wish to use Geometric brownian motion as a model. We are give a vector σ for the volatilities of the assets and matrix Σ, which is positive definite, for their correlations.

To use N samples in our Monte-Carlo method we would need N samples from $N(0, I)$, where I is the $d \times d$ identity matrix. This is because we have d risky assets. Call them $(z^i)_{i=1}^N$ (each $z^i = (z^i_1, \ldots, z^i_d)^T \in \mathbb{R}^d$). Then we can approximate (17) by

$$\begin{aligned} e^{-r(T-t)}\frac{1}{N}\sum_{i=1}^N\Bigg[g\Bigg(\operatorname{diag}(S)\exp\Bigg((T-t)\left(r - \frac{1}{2}\operatorname{diag}(\nu)C1_d\right) \\ +\sqrt{T-t}\operatorname{diag}(\sigma)Bz^i\Bigg)\Bigg)\Bigg]. \end{aligned}$$

6 Approximating sensitivity on parameters

6.1 Background

The price of an option (or indeed any derivative), given by any model, depends on market data and the model parameters. The difference between market data and model parameters is important. All market participants will use the same market data (we are assuming such information is available to all). They will not all use the same model parameters. Different market participants will have different models and different models rely on different parameters. Even if two different market participants use the same model they may use different model parameters because they use different calibration procedure! If market data or a model parameter changes, so does the price of the derivative calculated by the model (and one would hope that so does the market price of the derivative. This is of course by no means guaranteed. Very often the derivatives are illiquid which means that their prices cannot be observed very often).

Mathematically the sensitivities are just partial derivatives of the option price with respect to market data or a model parameter.

Example 6.1. Consider a call option modelled in the Black–Scholes framework. The option price depends on S, the spot price of the risky asset. This is market data. It also depends on T the maturity and K the option strike. These do not change during the life of an option. Finally σ the volatility parameter and r the risk free interest rate are model parameters. We can estimate σ from historical data, but we can choose to go back 1 month, or 6 months or 1 year in time, or indeed any other period in time. Or we can choose this to match observed option prices in the market. It is easy to see that different market participants will not necessarily agree on this.

Denoting the option price by v we can list the sensitivities:

$$\Delta := \frac{\partial v}{\partial S}, \ \ \theta := -\frac{\partial v}{\partial t}, \ \ \nu := \frac{\partial v}{\partial \sigma}, \ \ \rho := \frac{\partial v}{\partial r}.$$

These are sometimes referred to as the *greeks* and they tell us how to hedge various risks of a portfolio. For example, we have seen, when deriving the Black–Scholes partial differential equation, that by holding Δ units of the risky asset the risk coming from the fluctuations of the risky asset are completely eliminated. An option market making desk in a bank will try to have all the sensitivities across their portfolio close to 0 at all time by matching couterparties who want various options on the same asset. This ensures that if σ changes from one day to the next then the value of their portfolio remains roughly unchanged. Of course this never works exactly but if the model used is a good approximation of reality then it should be mostly the case.

It is important to remember that the sensitivities *are model dependent.* That is, taking a different model will result in different set of sensitivities. Even if one has e.g. a volatility parameter in two different models then the sensitivity to volatility is different across the two models.

A “good” numerical method in mathematical finance will provide the sensitivities (greeks) with little extra computation required. Of course it is not always possible to design such a method but calculation of greeks should be taken into account when choosing a numerical method.

6.2 General Setting

In general we have the following situation. We consider a random variable X which depends on some parameter $\theta \in \mathcal{O} \subseteq \mathbb{R}$. So now $X : \Omega \times \mathcal{O} \to \mathbb{R}^d$ and we assume it is measurable with respect to $\mathcal{F} \otimes \mathcal{B}(\mathcal{O})$. We will write $X = X(\theta)$ when we wish to emphasise this dependence.

We are interested in finding not only an estimate for

$$v(\theta) = \mathbb{E}[f(X(\theta))] \tag{18}$$

(where f is some measurable function such that $\mathbb{E}[|f(X)|] < \infty$)) but we are also interested in estimating

$$\frac{\partial v}{\partial \theta} = \frac{\partial}{\partial \theta}\mathbb{E}[f(X(\theta))]$$

whenever the partial derivative exists.

6.3 Finite Difference Approach

This is conceptually the easiest approach. We just replace the partial derivatives with their corresponding finite difference approximations. Let $(X_k(\theta))_{k=1}^N$ be iid with the same distribution as $X(\theta)$ for all $\theta \in \mathcal{O}$. Let

$$v_N(\theta) := \frac{1}{N}\sum_{k=1}^N f(X_k(\theta)).$$

We will consider the forward finite difference: given $h > 0$,

$$\frac{\partial v}{\partial \theta} \approx \frac{v_N(\theta+h) - v_N(\theta)}{h} =: \Delta_N^F \;\text{ and }\; \frac{\partial v}{\partial \theta} \approx \frac{v_N(\theta+h) - v_N(\theta-h)}{2h} =: \Delta_N^C.$$

Note the two sources of error in the approximation: one arises from taking finitely many samples N, the other from approximating the partial derivative with $h > 0$.

We note that both estimators are biased even though the estimator $v_N(\theta)$ is unbiased. Indeed

$$\mathbb{E}[\Delta_N^F] = \mathbb{E}\left[\frac{v_N(\theta+h) - v_N(\theta)}{h}\right] = \frac{\mathbb{E}[v_N(\theta+h)] - \mathbb{E}[v_N(\theta)]}{h} \neq \mathbb{E}\left[\frac{\partial}{\partial \theta} f(X(\theta))\right]$$

and similarly for Δ_N^C. To better analyse the bias assume that $v = v(\theta)$ given by (18) is $n+1$ times continuously differentiable on $[\theta, \theta+h]$. Then Taylor's theorem tells us that for some $\xi \in [\theta, \theta+h]$

$$v(\theta+h) = v(\theta) + h\partial_\theta v(\theta) + \frac{h^2}{2!}\partial_\theta^2 v(\theta) + \cdots + \frac{h^n}{n!}\partial_\theta^n v(\theta) + \frac{h^{n+1}}{(n+1)!}\partial_\theta^{n+1} v(\xi).$$

This means that for some $\xi \in [\theta, \theta+h]$

$$\partial_\theta v(\theta) = \frac{v(\theta+h) - v(\theta)}{h} - \frac{h}{2}\partial_\theta^2 v(\xi)$$

and for some $\xi \in [\theta - h, \theta + h]$

$$\partial_\theta v(\theta) = \frac{v(\theta+h) - v(\theta-h)}{2h} - \frac{h^2}{3}\partial_\theta^3 v(\xi).$$

In other words the forward difference Δ^F has bias of $o(h)$ while Δ^C has bias of $o(h^2)$ provided that $v = v(\theta)$ *is smooth enough.* Note that Δ^C is "more expensive" in the sense that we would typically need to calculate be calculating $v_N(\theta)$ anyway, but the central difference doesn't use this information.

So it appears that taking $h > 0$ as small as possible (within machine precision of our computer) is optimal. But what about variance?

$$\mathrm{Var}[\Delta^F_N] = \frac{1}{h^2}\mathrm{Var}[v_N(\theta+h) - v(\theta)].$$

How to choose N and h optimally? For simplicity write $Y := \partial_\theta f(X(\theta))$ and $\bar{Y} := \Delta^F_N$.

$$\begin{aligned}\mathbb{E}\left[(\bar{Y} - \mathbb{E}[Y])^2\right] &= \mathbb{E}\left[(\bar{Y} - \mathbb{E}[\bar{Y}] + \mathbb{E}[\bar{Y}] - \mathbb{E}[Y])^2\right]\\ &= \mathbb{E}\left[(\bar{Y} - \mathbb{E}[\bar{Y}])^2 + 2(\mathbb{E}[\bar{Y}] - \mathbb{E}[Y])(\mathbb{E}[\bar{Y} - \mathbb{E}[\bar{Y}]]) + (\mathbb{E}[\bar{Y}] - \mathbb{E}[Y])^2\right]\\ &= \mathbb{E}\left[(\bar{Y} - \mathbb{E}[\bar{Y}])^2\right] + (\mathbb{E}[\bar{Y}] - \mathbb{E}[Y])^2\\ &= \mathrm{Var}[\bar{Y}] + (\mathrm{Bias})^2.\end{aligned}$$

Fully independent samples: If we have $X_k(\theta)$ and $X_k(\theta+h)$ independent then

$$\mathrm{Var}[\Delta^F_N] = \frac{1}{h^2}\left(\mathrm{Var}[v_N(\theta+h)] + \mathrm{Var}[v_N(\theta)]\right) \le \frac{1}{h^2 N}\sup_\xi \mathrm{Var}[f(X(\xi))].$$

Recall $Y := \partial_\theta f(X(\theta))$ and $\bar{Y} := \Delta^F_N$.

$$\mathbb{E}\left[(\bar{Y} - \mathbb{E}[Y])^2\right] = \mathrm{Var}[\bar{Y}] + (\mathrm{Bias})^2 \le \frac{1}{h^2 N}\sup_\xi \mathrm{Var}[f(X(\xi))] + Ch^2,$$

where C depends on $v = v(\theta)$.

Dependent samples and smooth uniform dependence: If $f(X(\theta)) = g(\theta, \tilde{X})$ for some function g and random variable $\tilde{X}$ then we can consider $(\tilde{X})_{k=1}^N$ and take

$$f(X_k(\theta)) = g(\theta, \tilde{X}_k) \text{ and } f(X_k(\theta+h)) = g(\theta+h, \tilde{X}_k).$$

So now we are using the "same source of randomness" for both $X_k(\theta)$ and $X_k(\theta+h)$. If g is sufficiently smooth as a function on θ, uniformly in the second parameter, then

$$g(\theta+h, \cdot) = g(\theta, \cdot) + h\partial_\theta g(\theta, \cdot) + o(h^2)$$

then

$$\begin{aligned}\mathrm{Var}[\Delta^F_N] &= \mathrm{Var}\left[\frac{1}{h}\left(\frac{1}{N}\sum_{k=1}^N f(X_k(\theta+h)) - \frac{1}{N}\sum_{k=1}^N f(X_k(\theta))\right)\right]\\ &= \mathrm{Var}\left[\frac{1}{N}\sum_{k=1}^N(\partial_\theta f(X_k(\theta)) + o(h^2))\right]\\ &= \frac{1}{N}\mathrm{Var}\left[\partial_\theta f(X_k(\theta))\right].\end{aligned}$$

So

$$\mathbb{E}\left[(\bar{Y} - \mathbb{E}[Y])^2\right] = \mathrm{Var}[\bar{Y}] + (\mathrm{Bias})^2 \le \frac{1}{N}\mathrm{Var}\left[\partial_\theta f(X_k(\theta))\right] + Ch^2,$$

where C depends on $v = v(\theta)$.

Example 6.2. Let us consider the situation of Example 1.3. We wish to use Monte Carlo to approximate the European call option price, this time together with the Greeks “delta” and ”vega” (i.e. $\partial_S v$ and $\partial_\sigma v$).[4]

We have

$$v_N(S,\sigma) := \frac{1}{N}\sum_{i=1}^N e^{-rT}\left[S\exp\left(\left(r-\frac{1}{2}\sigma^2\right)T+\sigma\sqrt{T}Z_i\right)-K\right]_+,$$

where $(Z_i)_{i=1}^N$ are N independent samples from $N(0,1)$. Let $h>0$ be a small number. Then

$$\frac{\partial v_N}{\partial S}\approx\frac{v_N(S+\delta,\sigma)-v_N(S,\sigma)}{h}\quad\text{and}\quad\frac{\partial v_N}{\partial\sigma}\approx\frac{v_N(S,\sigma+h)-v_N(S,\sigma)}{h}.$$

Similarly, we can approximate the other sensitivities (Greeks).

Notice that $x\mapsto[x]_+$ is not differentiable at $x=0$. Nevertheless we have a

$$\mathbb{P}\left(S\exp\left(\left(r-\frac{1}{2}\sigma^2\right)T+\sigma\sqrt{T}Z_i\right)=K\right)=0$$

and so one may hope that the mean square error is of order $1/N+h^2$ in this case.

6.4 Calculating Sensitivities Pointwise in Ω

Recall that we wish to approximate

$$\partial_\theta v=\partial_\theta\mathbb{E}[f(X(\theta))]$$

whenever the partial derivative exists. Imagine for a moment that

$$\partial_\theta\mathbb{E}[f(X(\theta))]=\mathbb{E}\left[\partial_\theta f(X(\theta))\right]. \tag{19}$$

If this is the case then the unbiased estimator of $\partial_\theta v$ is simply

$$\frac{1}{N}\sum_{k=1}^N\partial_\theta f(X_k(\theta))$$

where $(X_k(\theta))_{k=1}^N$ are iid with the same distribution as $X(\theta)$ for all $\theta\in\mathcal{O}$. Its variance is then calculated as usual. Note that this approach is sometimes referred to *pathwise* calculation of sensitivities. This is because to evaluate

$$\mathbb{E}\left[\partial_\theta f(X(\theta))\right]=\int_\Omega\partial_\theta f(X(\omega,\theta))\,d\mathbb{P}(\omega)$$

we need to calculate $\partial_\theta f(X(\omega,\theta))$ for (almost all) $\omega\in\Omega$. When working with stochastic processes fixing ω fixes the path of the process and hence the term *pathwise.*

The question now is: when does (19) hold? We know that by definition (19) is equivalent to

$$\lim_{h\to0}\mathbb{E}\left[\frac{f(X(\theta+h))-f(X(\theta))}{h}\right]=\mathbb{E}\left[\lim_{h\to0}\frac{f(X(\theta+h))-f(X(\theta))}{h}\right]. \tag{20}$$

We will be allowed to claim that (20) holds if we check that we can apply Lebesgue’s theorem on dominated convergence and if the limit inside the expectation on the right-hand side of (20) exists almost surely. Note that we are already assuming that the limit on the left exists since otherwise the partial derivative is not defined.

[4] Of course, since we have the Black–Scholes formula we do not need to use Monte Carlo methods to get the Greeks but this is an illustrative example.

A1: The limit

$$\lim_{h\to 0} \frac{X(\omega,\theta+h)-X(\omega,\theta)}{h}$$

exists for almost all $\omega \in \Omega$.

A2: Let

$$D_f := \{x : \partial_{x_i} f(x) \text{ exists for all } i = 1,\dots,d\}$$

be in $\mathcal{B}(\mathbb{R}^d)$ and such that $\mathbb{P}(X(\theta) \in D_f) = 1$ for all θ.

A3: The function $f = f(x)$ is Lipschitz continuous[5] if x i.e. there exists $L > 0$ such that

$$|f(x) - f(y)| \leq L|x-y| \quad \forall x, y \in \mathbb{R}^d.$$

A4: There is a random variable $M \geq 0$ s.t. $\mathbb{E}[M] < \infty$ and

$$|X(\theta) - X(\theta')| \leq M|\theta - \theta'| \quad \forall \theta, \theta' \in \mathcal{O}.$$

Proposition 6.3. *Assume* **A1-4**. *Then* (19) *holds.*

Proof. Conditions **A1** and **A2** ensure that $\lim_{h\to 0} \frac{f(X(\theta+h))-f(X(\theta))}{h} = \partial_\theta f(X(\theta))$ exists with probability 1. Moreover, due to **A3** and **A4** we have that

$$|Y_h| := \left|\frac{f(X(\theta+h))-f(X(\theta))}{h}\right| \leq \frac{L}{h}|X(\theta+h)-X(\theta)| \leq \frac{L}{h}Mh = LM$$

and $\mathbb{E}[LM] = L\mathbb{E}[M] < \infty$. So we can apply Lebesgue's theorem on dominated convergence to conclude that

$$\lim_{h\to 0}\mathbb{E}[Y_h] = \mathbb{E}\left[\lim_{h\to 0} Y_h\right]$$

or in other words that (19) holds. □

Example 6.4. Our aim is the same as in Example 6.2.We wish to use Monte Carlo to approximate the European call option price, this time together with "delta" i.e. $\partial_S v$. We have

$$v_N(S,\sigma) := \frac{1}{N}\sum_{i=1}^{N} e^{-rT}\left[S\exp\left(\left(r-\frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}Z_i\right) - K\right]_+ .$$

First of all our $f : \mathbb{R} \to \mathbb{R}$ is $f(x) = e^{-rT}[x-K]_+$ and so $D_f = \mathbb{R}\setminus\{K\}$. Our

$$X(S,\sigma) = S\exp\left(\left(r-\frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}Z\right)$$

with $Z \sim N(0,1)$ and

$$\partial_S X(S,\sigma) = \exp\left(\left(r-\frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}Z\right) = \frac{X(S,\sigma)}{S},$$

[5] Note that this condition implies that f is differentiable almost everywhere in $\mathbb{R}^d$ due to Rademacher's theorem. But this does not, in general, imply **A2**.

exists for all $\omega \in \Omega$. We also note that $\mathbb{P}(X(S,\sigma) = K) = 0$ and so **A1** and **A2** above hold. We can check that $|f(x) - f(y)| \leq e^{-rT}$ and that

$$|X(S,\sigma) - X(S',\sigma)| = \frac{X(S,\sigma)}{S}|S - S'|$$

and so **A3** and **A4** hold. Thus

$$\begin{aligned}\partial_S v(S,\sigma) &= \partial_S \mathbb{E}\left[f(X(S,\sigma)\right] = \mathbb{E}\left[\partial_S f(X(S,\sigma))\right] = \mathbb{E}\left[f'(X(S,\sigma))\partial_S X(S,\sigma)\right] \\ &= \mathbb{E}\left[e^{-rT}\mathbb{1}_{\{X(S,\sigma)>K\}}\frac{X(S,\sigma)}{S}\right] = \mathbb{E}\left[e^{-rT}\mathbb{1}_{\{S_T>K\}}\frac{S_T}{S}\right].\end{aligned}$$

Note that if we wanted to do obtain "vega" then we can start by noting that

$$\partial_\sigma X(S,\sigma) = (-\sigma T + \sqrt{T}Z)S\exp\left(\left(r - \frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}Z\right) = (-\sigma T + \sqrt{T}Z)X(S,\sigma)$$

exists for all $\omega \in \Omega$. If we can then also verify **A4** then we would have

$$\begin{aligned}\partial_\sigma v(S,\sigma) &= \partial_\sigma \mathbb{E}\left[f(X(S,\sigma)\right] = \mathbb{E}\left[\partial_\sigma f(X(S,\sigma))\right] = \mathbb{E}\left[f'(X(S,\sigma))\partial_\sigma X(S,\sigma)\right] \\ &= \mathbb{E}\left[e^{-rT}\mathbb{1}_{\{X(S,\sigma)>K\}}(-\sigma T + \sqrt{T}Z)X(S,\sigma)\right] \\ &= \mathbb{E}\left[e^{-rT}\mathbb{1}_{\{S_T>K\}}(-\sigma T + W_T)S_T\right].\end{aligned}$$

More examples are found in Glasserman [1, Ch. 7, Sec. 2]

6.5 The Log-likelihood Method

Recall that we wish to approximate

$$\partial_\theta v = \partial_\theta \mathbb{E}[f(X(\theta))]$$

whenever the partial derivative exists. Assume that $X(\theta)$ has density $g(\cdot;\theta)$ and that $\partial_\theta g(\cdot;\theta)$ exists for any $\theta \in \mathcal{O}$ Then

$$\mathbb{E}[f(X(\theta))] = \int_{\mathbb{R}^d} f(x)g(x;\theta)\,dx$$

and, assuming that we can exchange the integral and the derivative, we have

$$\begin{aligned}\partial_\theta \mathbb{E}[f(X(\theta))] &= \partial_\theta \int_{\mathbb{R}^d} f(x)g(x;\theta)\,dx = \int_{\mathbb{R}^d} f(x)\partial_\theta g(x;\theta)\,dx \\ &= \int_{\mathbb{R}^d} f(x)\frac{\partial_\theta g(x;\theta)}{g(x;\theta)}g(x;\theta)\,dx\,.\end{aligned}$$

We define the "log-likelihood" or "score" function

$$L(x;\theta) := \frac{\partial_\theta g(x;\theta)}{g(x;\theta)} = \partial_\theta\left[\ln g(x;\theta)\right].$$

Then

$$\partial_\theta \mathbb{E}[f(X(\theta))] = \int_{\mathbb{R}^d} f(x)L(x;\theta)g(x;\theta)\,dx = \mathbb{E}\left[f(X(\theta))L(X(\theta);\theta)\right].$$

7 Solutions to some exercises

Solution (to Exercise 3.3)**.** a) Assume first $x \leq 0$. Then

$$G(x) = \int_{-\infty}^{x} g(y)dy = \frac{1}{2}\int_{-\infty}^{x} e^{-|y|}dy = \frac{1}{2}\int_{-\infty}^{x} e^{y}dy = \frac{1}{2}e^{x}.$$

On the other hand, if $x > 0$ then, using also the above calculation for $G(0)$ we get,

$$\begin{aligned} G(x) &= \int_{-\infty}^{x} g(y)dy = \int_{-\infty}^{0} g(y)dy + \int_{0}^{x} g(y)dy = G(0) + \frac{1}{2}\int_{0}^{x} e^{-|y|}dy \\ &= \frac{1}{2} + \frac{1}{2}\int_{-\infty}^{x} e^{-y}dy = \frac{1}{2} - \frac{1}{2}e^{-x} + \frac{1}{2} = 1 - \frac{1}{2}e^{-x}. \end{aligned}$$

b) Assume first $x \leq 0$. If we solve $y = G(x)$ we get $x = \ln(2y)$ but only for y such that $x \leq 0$. That is, only for y such that $\ln(2y) \leq 0$. This means $y \leq 1/2$.

Now consider $x > 0$. If we solve $y = G(x)$ we get $x = -\ln(2-2y)$ for x such that $x > 0$. This means that this is only valid for y such that $\ln(2-2y) < 0$ which is exactly $y > 1/2$.

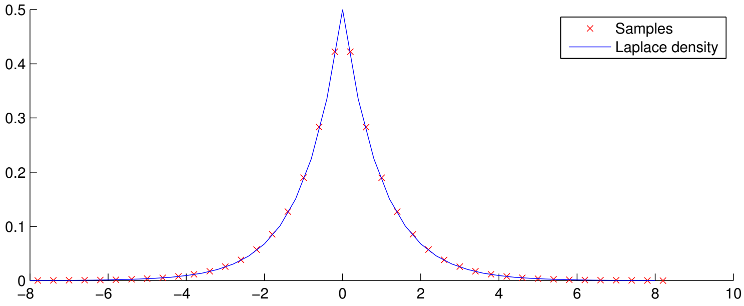
c) We can generate samples from Laplace density using $x_i = G^{-1}(u_i)$.

```
N = 5*10^7;
u = unifrnd(0,1,N,1);

% we must identify those smaller or equal to 1/2 and those bigger 1/2
smaller = logical(u<= 1/2);
bigger = logical(u > 1/2);
x = zeros(N,1);

% now we can apply the inverse
x(smaller) = log(2*u(smaller));
x(bigger) = -log(2-2*u(bigger));

plotLen=8;
step=0.4;
bins = -plotLen:step:plotLen;
Nbins = length(bins);
counts = histc(x,bins);
hold on;
plot(bins+step/2,(Nbins/(2*plotLen))*counts./N,’xr’);
plot(bins,0.5*exp(-abs(bins)));
hold off;
legend(’Samples’,’Laplace density’);
```



Solution (to Exercise 3.4)**.** For all $x \in \mathbb{R}$ we have $f(-x) = (2\pi)^{-1/2}e^{-x^2} = f(x)$ and that $g(-x) = (1/2)\exp(-|x|) = g(x)$. Hence also $\xi(-x) = \xi(x)$.

To find the maximum of ξ it is thus sufficient to consider $x \geq 0$. Then

$$\xi(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}+x} \text{ and so } \xi'(x) = \frac{2}{\sqrt{2\pi}}(1-x)e^{-\frac{x^2}{2}+x}.$$

This means the maximum is achieved at $x = 1$ (and by symmetry also at $x = -1$). Now $\xi(1) = \sqrt{(2e)/\pi}$. Hence $f(x) \leq \sqrt{(2e)/\pi} g(x)$ for all $x \in \mathbb{R}$.

Solution (to Exercise 5.1)**.** First we note that if the evolution of the risky asset is given by

$$dS_t = rS_t dt + \sigma S_t dW_t,$$

where $(W_t)_{t\in[0,T]}$ is a Wiener process in the risk neutral measure then we already know that

$$S_u = S\exp\left(\left(r - \frac{1}{2}\sigma^2\right)(u-t) + \sigma(W_u - W_t)\right), \quad S_t = S.$$

Thus we may calculate

$$\bar{S}_G = \left(\prod_{i=1}^n S_{T_i}\right)^{1/n} = S\exp\left(\left(r - \frac{1}{2}\sigma^2\right)\left(\frac{1}{n}\sum_{i=1}^n T_i - t\right) + \sigma\left(\frac{1}{n}\sum_{i=1}^n W_{T_i} - W_t\right)\right).$$

Next we observe that $\sum_{i=1}^n (W_{T_i} - W_t)$ has the same distribution as $\sum_{i=1}^n Z_i$, where $Z_i \sim N(0, T_i - t)$ and their correlation is given by the variance-covariance matrix $\Sigma = \sigma_{ij} = \min\{T_i, T_j\} - t$. Indeed

$$\begin{aligned}\mathbb{E}\left((W_{T_i} - W_t)\left(W_{T_j} - W_t\right)\right) &= \mathbb{E}\left(W_{T_i}W_{T_j} - W_tW_{T_i} - W_tW_{T_j} + W_t^2\right)\\ &= \min\{T_i, T_j\} - t.\end{aligned}$$

We know that, in general, if $X \sim N(\mu, \Sigma)$ then $AX \sim N(A\mu, A\Sigma A^T)$. In our case we take $A = (1, 1, \dots, 1) \in \mathbb{R}^n$. Then, letting $Z = (Z_1, Z_2, \dots, Z_n)^T$ we get $AZ = \sum_{i=1}^n Z_i$ and

$$A\Sigma A^T = (2n-1)(T_1 - t) + (2n-3)(T_2 - t) + \cdots + (T_n - t) = \sum_{i=1}^n (2i-1)(T_{n+1-i} - t).$$

We can thus conclude that

$$\sum_{i=1}^n Z_i \sim N\left(0, \sum_{i=1}^n (2i-1)(T_{n+1-i} - t)\right).$$

We know that the option payoff must be equal to

$$\begin{aligned}v_{A_G}(t,S) &= e^{-r(T-t)}\mathbb{E}\left([\bar{S}_G - K]_+\right)\\ &= e^{-r(T-t)}\mathbb{E}\left(\left[S\exp\left(\left(r - \frac{1}{2}\sigma^2\right)\left(\frac{1}{n}\sum_{i=1}^n T_i - t\right) + \frac{\sigma}{n}\sum_{i=1}^n Z_i\right) - K\right]_+\right).\end{aligned}$$

Let $\bar{T} := n^{-1}\sum_{i=1}^n T_i$. Notice that if we let $X \sim N(0,1)$ then $\sum_{i=1}^n Z_i$ has the same distribution as the random variable

$$X\sqrt{\sum_{i=1}^n (2i-1)(T_{n+1-i} - t)}.$$

Now we choose $\bar{\sigma}$ such that

$$\bar{\sigma}\sqrt{\bar{T}-t} = \frac{\sigma}{n}\sqrt{\sum_{i=1}^{n}(2i-1)(T_{n+1-i}-t)}.$$

Then

$$\begin{aligned} v_{A_G}(t,S) &= e^{-r(T-t)}\mathbb{E}\left(\left[S\exp\left(\left(r-\frac{1}{2}\sigma^2\right)(\bar{T}-t)+\bar{\sigma}\sqrt{\bar{T}-t}X\right)-K\right]_+\right) \\ &= e^{-r(T-t)+r(\bar{T}-t)}\mathbb{E}\left(e^{-r(\bar{T}-t)}\left[Se^{\gamma(\bar{T}-t)}\exp\left(\left(r-\frac{1}{2}\bar{\sigma}^2\right)(\bar{T}-t)+\bar{\sigma}\sqrt{\bar{T}-t}X\right)-K\right]_+\right), \end{aligned}$$

where we have taken $\gamma := (1/2)(\bar{\sigma}^2-\sigma^2)$. But now the expectation on the right hand side is simply just the price of a European call option in the Black–Scholes framework with spot price $Se^{\gamma(\bar{T}-t)}$, time-to-expiry $\bar{T}-t$, volatility $\bar{\sigma}$, risk free rate r and strike K. This we can calculate using the Black–Scholes formula.

A Appendix

A.1 Multi-dimensional Itô's formula

Let us introduce $\{W(t)\}_{t\geq 0}$, a $\mathbb{R}^m$ valued stochastic process where

$$W(t) = (W_1(t), W_2(t), \dots, W_m(t))^T$$

and $\{W_i(t)\}_{t\geq 0}$ are independent Wiener processes for $i = 1, 2, \dots, m$. So $\{W(t)\}_{t\geq 0}$ is a m-dimensional Wiener process with respect to $\{F_t\}_{t\geq 0}$.

Definition A.1 (Multi-dimensional Itô process)**.** *Let $\{X(t)\}_{t\in[0,T]}$ be a process taking values in $\mathbb{R}^n$. That is, let $X(t) = (X_1(t), X_2(t), \dots, X_n(t))^T$ where the evolution of the ith compoment of the process is given by*

$$dX_i(t) = u_i(t)dt + v_{i1}(t)dW_1(t) + v_{i2}(t)dW_2(t) + \dots + v_{im}(t)dW_m(t).$$

*Assume that $\{u_i(t)\}_{t\geq 0}$ satisfy the conditions placed on $\{U(t)\}_{t\geq 0}$ in the definition of an Itô process for $i = 1, 2, \dots, n$ and $\{v_{ij}(t)\}_{t\geq 0}$ satisfy the conditions placed on $\{V(t)\}_{t\geq 0}$ in the definition of an Itô process for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$. Then $\{X(t)\}_{t\in[0,T]}$ is a n-*dimensional Itô process.

Let us define the vector $u = (u_1, u_2, \dots, u_n)$ and the matrix $v = (v_{ij})$ with $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$. Then we can write

$$dX(t) = u(t)dt + v(t)dW(t). \tag{21}$$

Theorem A.2 (Multi-dimensional Itô formula)**.** *Let the evolution of the stochstic process $\{X(t)\}_{t\in[0,T]}$ be given by (21). Let $g : [0,\infty] \times \mathbb{R}^n \to \mathbb{R}^d$ be given by $g(t,x) = (g_1(t,x), g_2(t,x), \dots, g_d(t,x))$ with $g_k : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ such that all first and second partial derivatives with respect to the function's arguments exist and are continuous.*

Then the process $Y(t) := g(t, X(t))$ is a d-dimensional Itô process and its kth component is given by

$$\begin{aligned} dY_k(t) = \frac{\partial g_k}{\partial (t)}(t, X(t))dt &+ \sum_{i=1}^{n} \frac{\partial g_k}{\partial x_i}(t, X(t))dX_i(t) \\ &+ \tfrac{1}{2} \sum_{i,j=1,\dots,n} \frac{\partial^2 g_k}{\partial x_i \partial x_j}(t, X(t))\langle dX_i(t), dX_j(t)\rangle. \end{aligned}$$

Furthermore $\langle dX_i(t), dX_j(t)\rangle = \delta_{ij}dt$ and $\langle dX_i(t), dt\rangle = \langle dt, dX_i(t)\rangle = 0$.

For proof see e.g. [5, Theorem 4.2.1].

Now we get the product rule for Itô processes.

Corollary A.3. *Let $\{X(t)\}_{t\geq 0}$ and $\{Z(t)\}_{t\geq 0}$ be Itô processes given by*

$$\begin{aligned} dX(t) &= \mu(t)dt + \sigma(t)dW(t),\ X(0) = x_0, \\ dZ(t) &= \phi(t)dt + \psi(t)dW(t),\ Z(0) = z_0. \end{aligned}$$

Then

$$d(X(t)Z(t)) = Z(t)dX(t) + X(t)dZ(t) + \langle dX(t), dZ(t)\rangle.$$

Proof. We have $g : \mathbb{R}^2 \to \mathbb{R}$ given by $g(x, z) = xz$. Clearly

$$\frac{\partial g}{\partial x} = z,\ \frac{\partial g}{\partial z} = x,\ \frac{\partial^2 g}{\partial x^2} = \frac{\partial^2 g}{\partial z^2} = 0,\ \frac{\partial^2 g}{\partial x \partial z} = 1.$$

Using the multi-dimensional Itô formula we get, for $Y(t) = g(X(t), Z(t)$, that

$$d(X(t)Z(t)) = dY(t) = Z(t)dX(t) + X(t)dZ(t) + \frac{1}{2}(0 + 1 + 1 + 0)\langle dX(t), dZ(t)\rangle.$$

□

References

[1] P. Glasserman. *Monte Carlo Methods in Financial Engineering.* Springer 2004.

[2] A. Gut. *An Intermediate Course in Probability.* Springer 2009.

[3] G. Grimmett and D. Stirzaker. *Probability and Random Processes.* Third Edition. Oxford University Press 2001.

[4] D. E. Knuth. *The Art of Computer Programming.* Volume 2: Seminumerical Algorithms, Third Edition. Addison-Wesley 1997.

[5] B. Øksendal. *Stochastic Differential Equations.* Springer 2003.

[6] S. Ross. *A First Course in Probability.* Fifth Edition. Prentice–Hall 1998.

[7] A.N. Shiryaev. *Probability.* Second Edition. Springer 1996.