623 Computational Finance, Fall 2014.
Lecturer: Sergey Nadtochiy.

Chapter III. Monte Carlo (probabilistic) methods.
Partial information can be found in Chapters 1.1.1–1.1.3, 2.1.1, 2.1.2, 2.2, 2.3, 3.1, 3.2, 3.3.1, 3.4.1, 3.4.2, 6.1, 6.5, 7.1, 7.2.1, 7.2.3 in "Monte Carlo Methods in Financial Engineering" by Glasserman 2004; in Chapters 11.4, 21.1.3, 21.1.7 in "Interest Rate Models – Theory and Practice" by Brigo and Mercurio 2006; and in Chapters 2.2, 2.6, 4.5, 5.1, 5.2, 5.13, 8.1–8.5, 9.1, 9.2, 9.4 in "Credit Risk Modeling" by Lando 2004.

1 Going beyond LV models: motivation

There are several ways to criticize LV models. Here, we split them into two categories.

- **Inconsistency with the dynamics of empirically observed Implied Volatility**. This is best illustrated by the phenomenon of the so called "sticky smile". Assuming a LV model holds, with $r = 0$, consider the LV function $\sigma(K)$ as a function of $x = \log(K/S)$: $\tilde{\sigma}(x,T) = \sigma(Se^x, T)$. And perform the same change of variables in the associated Implied Volatility: $\tilde{\Sigma}(x,T) = \Sigma(Se^x, T)$. Notice that the model implies that σ (as a function) remains that same at all times. However, the function $\tilde{\sigma}(x,T) = \sigma(Se^x, T)$ will change with S. If $\sigma(K,T)$ is regular enough in T (e.g. differentiable), a small change in S will result in (almost) a shift along the x-axis of the function $\tilde{\sigma}(x,T)$. It turns out that, in this case, the change in the implied volatility $\tilde{\Sigma}(x,T) = \Sigma(Se^x, T)$, viewed as a function of x, will also be very similar to a shift along the x-axis (this can be seen, for example, by analyzing the connection between the shot-term local and implied volatilities, given at the end of Chapter II). However, in many markets (equity, FX, fixed income) the observed behavior of $\tilde{\Sigma}(x,T)$ is different: its shape, as a function of x does not change too much with the changes in S, and it typically has a local maximum around $x = 0$. This is often referred to as a "sticky smile". Clearly, this empirical fact is inconsistent with a LV model. Thus, LV models can capture the shape of IV on a given day, but they **fail to produce realistic dynamics of IV**.

- **Inconsistency with the dynamics of empirically observed spot volatility**. The spot volatility in a LV model is given by a function of the risky asset S_t (we ignore the dependence on time in this discussion, as we consider relatively large time horizons, for which one does not have enough options traded to deduce this time dependence from the option prices). However, this assumption is violated by empirical data. Figure 1 shows the S&P 500 index and its spot volatility approximation (given by the VIX index) over the time period from 2004 to 2010. Notice that, during the crisis of 2008, the value of VIX went down to (and then below) its values in 2008, while the volatility was pretty far from its values in 2004. This can be observed in many other examples, and indicates that the spot volatility σ_t, in fact, is not given by a function of S_t. What is apparent from Figure 1 is that there is a **strong negative correlation between the changes in the asset and its volatility, dS_t and $d\sigma_t$**, (as opposed to a monotone decreasing functional dependence between S_t and σ_t themselves), which accounts for the **leverage effect** (and, in turn, for the skew of IV). In addition, the spot volatility seems to exhibit a somewhat **mean-reverting** behavior, always reverting to back its long-term mean. **None of these dynamical features of the spot volatility can be reproduced by a LV model**.

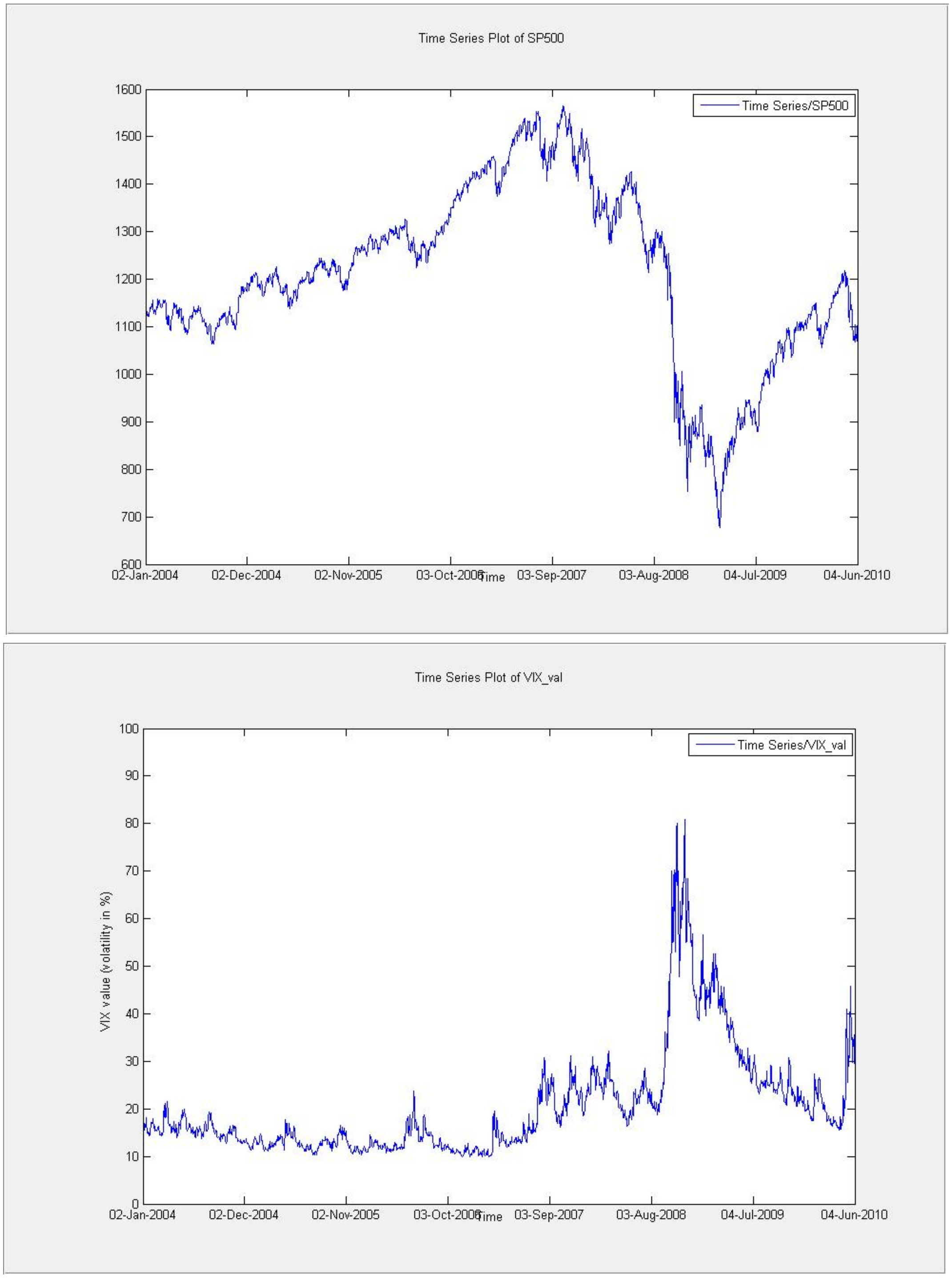

Figure 1: S&P500 index (top) and its spot volatility, approximated by the VIX index (bottom).

1.1 Stochastic Volatility models

Heston model

- "The" stochastic volatility model is due to Heston (1993). Under the physical (or, real-world) measure $\mathbb{P}$, the dynamics of the risky asset S, in this model, are given by

  $$\begin{cases} dS_t = \mu S_t dt + \sqrt{Y_t} S_t dW_t^1 \\ \\ dY_t = a(b - Y_t)dt + c\sqrt{Y_t}(\rho dW_t^1 + \sqrt{1-\rho^2} dW_t^2), \end{cases}$$

  where $\rho \in [-1, 1]$, $\mu \in \mathbb{R}$, $a \geq 0$, $b \geq 0$ and $c > 0$ are some constants. The bank account, as usual, grows at a constant rate r.

- The correlation coefficient ρ is usually assumed to be negative, as it accounts for the negative correlation between dS_t and $d\sigma_t$, indicated by Figure 1.

- It is easy to see that Y is the **Cox-Ingersoll-Ross (CIR)**) process.

- As usual, we stop Y as soon as it hits zero. Then, in order to avoid arbitrage in this model, we have to assume that, either $\mu = r$, or $2ab \geq c^2$ and $Y_0 > 0$ – in the latter case, the process Y stays **strictly positive at all times**.

- Heston model is incomplete, because the number of Brownian motions is 2, whereas the number of traded assets is 1.

  Rem 1. *In fact, the model becomes complete if $|\rho| = 1$. In addition, one can add another tradable asset (e.g. a futures contract on the spot volatility), to complete the market. So, strictly speaking, on has to specify which assets are assumed to be (liquidly) traded before discussing the completeness of a model. Very often, however, this is not stated explicitly and one has to guess which assets are assumed to be liquidly traded from the context. In the classical formulation of the Heston model, we assume that S is the only tradable asset.*

- As the Heston model is incomplete, there are infinitely many pricing measures that we can choose from. It turns out that there exists a pricing measure $\mathbb{Q}$, under which the pair (S, Y) has **the same type of dynamics as under** $\mathbb{P}$. Namely, if $2ab \geq c^2$, then, there exist constants $A, B \geq 0$ and an EMM $\mathbb{Q}$, such that, under $\mathbb{Q}$:

  $$\begin{cases} dS_t = rS_t dt + \sqrt{Y_t} S_t dW_t^{\mathbb{Q},1}, \\ \\ dY_t = A(B - Y_t)dt + c\sqrt{Y_t}\left(\rho dW_t^{\mathbb{Q},1} + \sqrt{1-\rho^2} dW_t^{\mathbb{Q},2}\right), \end{cases}$$

  with $W^{\mathbb{Q},1}$ and $W^{\mathbb{Q},2}$ being independent $\mathbb{Q}$-BMs.

- Once a pricing measure $\mathbb{Q}$ is chosen, the **prices of European options**, in this model, can be computed via

  $$V_t = V(S_t, Y_t, t),$$

  where the function V satisfies the BSPDE

  $$\partial_t V + \frac{1}{2} Y S^2 \partial^2_{SS} V + \frac{1}{2} c^2 Y \partial^2_{YY} V + \rho c Y S \partial^2_{SY} V + rS\partial_S V + A(B - Y)\partial_Y V - rV = 0,$$

  with the appropriate **terminal** and, possibly, **boundary conditions**.

- A *closed form of solution* is not available for the above BSPDE. In principle, it can be solved numerically, but the computational complexity grows very fast with the dimension of the problem (this is known as the **curse of dimensionality**), and, even in the two-factor model, such as Heston, the numerical methods will not be very efficient.

- A *closed form of solution* is available for the **characteristic function of the underlying** (see *Heston (1993)*). Then the prices of European options can be obtained via **numerical inversion of the Fourier transform**.

- Option prices may also be represented by an **asymptotic expansion** (usually with respect to c^2T): this is done, for example, by Fouque et al (2002).

- We can see that the PDE methods (as well as any other analytical methods) are hard to use as the dimension of the model increases. In particular, even in the case of Heston model, we have to resort to the more general (but less precise) **Monte Carlo** methods to compute option prices, hedging ratios, or estimate the desired risk measures.

- Once the European options' prices are computed, one can obtain the Implied Volatility of the Heston model. This IV has a **negative skew, provided** $\rho < 0$, and it **does have the "sticky smile" property**. The downside of this model (as pretty much any stochastic volatility model) is the **lack of flexibility of the model when it comes to calibration**: one cannot calibrate a Heston model to the market implied smile, unless it contains only one maturity and very few strikes.

SABR model

- SABR stands for "stochastic alpha, beta, rho" and it is a stochastic volatility model proposed by Hagan et al (2002). It is popular among practitioners, as it calibrates to the implied smile of a single maturity better than the Heston model, but, yet, it can only handle a smile consisting of very few strikes. Just like the Heston model SABR has the "sticky smile" property.

- In this model, we deal with the discounted price of a tradable asset, F_t (e.g. you may think of it as the **forward price** $F_t = e^{r(T-t)}S_t$), which is modeled under the risk-neutral measure $\mathbb{Q}$ as:

$$\begin{cases} dF_t = \alpha_t F_t^{\beta} dW_t^{\mathbb{Q},1} \\ d\alpha_t = \nu\alpha_t \left(\rho dW_t^{\mathbb{Q},1} + \sqrt{1-\rho^2} dW_t^{\mathbb{Q},2}\right), \end{cases}$$

  where $\rho \in [-1, 1]$, $\beta < 1$, $\nu \geq 0$, $F_0, \alpha_0 \geq 0$.

- Notice that α is a GBM, and F is an extension of the CEV process. As usual, we assume that F is absorbed at zero.

- European option prices in this model are typically computed via asymptotic expansion for $T\nu^2 \to 0$.

1.2 Structural models of credit risk

- Merton (1974) considered the BS model for S

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

but viewed S as the **value** of a firm, rather than its equity. Then, assuming that the debt level of the company is D and given a time horizon T, the default occurs (only) at time T if and only if $S_T \leq D$.

- In this simple model, the probability of the default of a company can be computed in a closed form as

$$\mathbb{P}(S_T \leq D) = \mathbb{P}\left(W_T \leq \frac{\log(D/S_0) - (\mu - \sigma^2/2)T}{\sigma}\right)$$

- Black and Cox (1976) considered the following extension of the model. They allowed the debt holders to trigger the default event whenever the firm's value S_t drops below a given threshold H_t. In other words,

$$\tau = \inf\left\{t \in [0,T] \, : \, S_t \leq H_t\right\}$$

  is the default time. The default threshold can be taken as the present value of the debt: $H_t = De^{-r(T-t)}$, given a constant interest rate r (i.e. the debt holders want to be able to recover their investment by selling the company). However, typically, the debt holders give the company a chance to recover, and $H_t \leq De^{-r(T-t)}$.

- In this model, there is no closed-form expression for the default probability. However, it is not hard to see that the default probability can be viewed as a barrier option on S, hence, it can be computed using the PDE methods developed in Chapter II.

Collateralized Debt Obligations (CDOs)

The idea behind CDO is to give an investor exposure to the credit risk of multiple obligors. This resonates with the idea of diversification and, in principle, may allow one to have a better control over the risk exposure (i.e. of course, only if one has a **good understanding of the product and the input parameters**).

- Consider a pool of companies with values $(S^1, \ldots, S^N)$, which follow the multidimensional BS model:

$$\begin{pmatrix} dS^1_t \\ \vdots \\ dS^N_t \end{pmatrix} = \begin{pmatrix} \mu^1 S^1_t \\ \vdots \\ \mu^N S^N_t \end{pmatrix} dt + \begin{pmatrix} \sigma^{11} S^1_t & \cdots & \sigma^{1M} S^1_t \\ \vdots & \ldots & \vdots \\ \sigma^{N1} S^N_t & \cdots & \sigma^{NM} S^N_t \end{pmatrix} \begin{pmatrix} dW^1_t \\ \vdots \\ dW^M_t \end{pmatrix}$$

- Assume that every company issues a zero coupon bond, with the same face value, and with maturity T.

- Then, the total relative loss process associated with this pool is given by:

$$L_t = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{\tau^i \leq t\}},$$

  where τ^i is the default time of the ith firm:

$$\tau^i = \inf\left\{t \in [0,T] \, : \, S^i_t \leq H^i_t\right\},$$

  with given H^i.

- A CDO tranche, with face value 1 and the attachment points $0 \leq K_1 < K_2 \leq 1$, has the following payoff at time T:

$$1 - \frac{1}{K_2 - K_1} \min\left((L_T - K_1)^+, K_2 - K_1\right)$$

- The higher is K_1, the more protection the tranche offers. Tranches with low $K_1 = 0$ are called "equity tranches"; tranches with slightly higher K_1 (say, 5%) are called "junior" (or "mezzanine"); those with high K_1 are called "senior".
- Notice that, in a structural model, as above, a CDO tranche can be viewed as a barrier option written on $(S^1, \ldots, S^N)$. In principle, we can price such options by writing a BSPDE. However, this PDE has N-dimensional space variable, hence, there is no chance to solve it numerically when N is large (usually, $N > 100$). In this case, we have to apply the **Monte Carlo** techniques.

2 The basic MC method

The Monte-Carlo (MC) method for numerically estimating the expectations (or, integrals) of the form

$$\mathbb{E}X \left(= \int_{\mathbb{R}^n} x f(x) dx \right),$$

based on the **Law of Large Numbers (LLN)** for independent identically distributed (i.i.d.) random variables. To take the simplest example, consider the situation where a fair coin is tossed many times. Then we expect that in a large number of tosses approximately 50% of the tosses will come up heads. This is a particular case of the *Strong Law of Large Numbers*.

Thm 1. **(Strong LLN)** *Let $X_1, X_2, ..,$ be i.i.d. random variables with the same distribution as a random variable X, having a finite expectation. Then*

$$\lim_{N \to \infty} \frac{X_1 + \cdots + X_N}{N} = \mathbb{E}[X] \quad \text{with probability 1.} \tag{1}$$

If we know how to generate a large number of independent copies of the variable X by, say, using a random number generator on a computer, then (1) enables us to numerically estimate $\mathbb{E}[X]$.

Thus there are two ingredients to the MC method:

1. Express the quantity we wish to estimate as the expectation $\mathbb{E}[X]$ of some random variable X.
2. Create a random number generator which efficiently generates large numbers of approximately independent variables with distribution the same as X.

The first part is implemented at the stage when the general result (e.g. formula) is developed. The second part is the "black box" which we typically focus on when we develop the MC method itself.

An important issue in numerical analysis is always to get an idea of the **computational complexity (or efficiency)**: i.e. how many computations are required to estimate a desired quantity with a given precision (or vice versa: what is the precision that can be achieved with a given number of computations). In the case of MC, the **approximation error is random** – it is usually impossible to obtain a good estimate of the error, which would hold for all possible outcomes. This is the main **disadvantage of the MC method**, as compared to the PDE (or other **analytical**) methods. However, we can get an idea about the **distribution of the error** using the **Central Limit Theorem (CLT)**:

Thm 2. (CLT) *Suppose a random variable X has finite mean μ and variance σ^2. Let $X_1, X_2, .., $ be i.i.d. random variables with the same distribution as X. Then*

$$Z_N = \frac{\sqrt{N}}{\sigma}\left[\frac{X_1 + \cdots + X_N}{N} - \mu\right] \quad \text{converges in distribution, as } N \to \infty, \text{ to } Z, \tag{2}$$

where Z is a standard normal random variable with the probability density function (pdf) $\phi(\cdot)$ given by

$$\phi(z) \ = \ \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in \mathbb{R}. \tag{3}$$

The convergence in distribution means that, for any $a \in \mathbb{R}$, we have

$$\lim_{N\to\infty} P(|Z_N| > a) = P(|Z| > a)$$

We have already mentioned in Chapter II that $P(|Z| > 3) < .003$. Hence, CLT implies that, for large N,

$$P\left(\left|\frac{X_1 + \cdots + X_N}{N} - \mu\right| > \frac{3\sigma}{\sqrt{N}}\right) \ < \ .003. \tag{4}$$

We can write (4) alternatively as

$$\left|\frac{X_1 + \cdots + X_N}{N} - \mu\right| \ = \ \text{Error}(N), \text{ and Error}(N) < \frac{3\sigma}{\sqrt{N}} \text{ with high probability.} \tag{5}$$

Motivated by the above, we will say that the **error of the MC method is proportional to the inverse square root of the number of simulations**. In other words, MC sample of size N allows us to estimate $\mathbb{E}[X]$ with the precision $O(1/\sqrt{N})$. However, one needs to remember that an exact bound on the error of the form "const/$\sqrt{N}$" **can only be obtained with a certain (high) probability, but not for all random outcomes**.

To summarize, when implementing the MC method to estimate $\mathbb{E}[X]$, we proceed as follows.

1. Generate $N \simeq 10^4$ values $X_1, .., X_N$ of i.i.d. variables with the same distribution as X.
2. Compute the averages $\bar{X}_n, \ n = 1, 2, ..N$ given by

$$\bar{X}_n \ = \ \frac{1}{n}\sum_{j=1}^{n} X_j \ , \tag{6}$$

   and graph the **convergence diagram** $n \to \bar{X}_n, \ n = 1, .., N$.
3. If the true variance σ^2 is **not known**, compute the sample variance $\hat{\sigma}_N^2$ defined by

$$\hat{\sigma}_N^2 \ = \ \frac{1}{N}\sum_{j=1}^{N} [X_j - \hat{X}_N]^2 \ . \tag{7}$$

4. Report the estimated value $\bar{X}_N$ for $\mathbb{E}[X]$ and the **standard error** $\varepsilon_N = \hat{\sigma}_N/\sqrt{N}$.
5. In addition, you may need to report the **proportional error** which is $\varepsilon_N/\bar{X}_N$.

Rem 1. *Note that the sample variance (7) is* **not an unbiassed estimator** *of the variance σ^2 of X. "Unbiassed" means that the expectation of the estimator is equal to the quantity of interest. To get an unbiassed estimator we need to replace $1/N$ in (7) by $1/(N-1)$.*

We can **compare the complexity of the MC method with deterministic methods**. Let $\Phi : [0,1]^d \to \mathbb{R}$ be a function on the d-dimensional unit "cube", and suppose we wish to estimate its integral. Evidently we have that

$$\int_0^1 \cdots \int_0^1 \Phi(x^1,..,x^d)\, dx^1 \cdots dx^d \;=\; E[\Phi(X^1,..,X^d)]\;, \tag{8}$$

where $X^1,..,X^d$ are i.i.d. variables uniform on the interval $[0,1]$.

The standard Riemann sum algorithm for estimating the integral is to choose N equally spaced points $x_1,..,x_N$ in $[0,1]^d$ which are the centers of disjoint subcubes each of volume $1/N$. Then we have

$$\int_{[0,1]^d} \Phi(x)\, dx \;\approx\; \frac{1}{N}\left[\Phi(x_1) + \cdots \Phi(x_N)\right]\;. \tag{9}$$

If Q_j is the subcube with center x_j then

$$\left|\int_{Q_j} \Phi(x)\, dx - \frac{\Phi(x_j)}{N}\right| \;\le\; \int_{Q_j} |\Phi(x) - \Phi(x_j)|\, dx \;\le\; \frac{\sup_{x\in Q_j} |\Phi(x) - \Phi(x_j)|}{N}\;. \tag{10}$$

If the function $\Phi(\cdot)$ is continuously differentiable then $|\Phi(x) - \Phi(x_j)| \le C_1|x - x_j|$ for a constant C_1.

As the volume of Q_j is $1/N$, the length of an edge of Q_j is $N^{1/d}$. Next notice that

$$\sup_{x\in Q_j} |x - x_j| \le C_2 \cdot (\text{edge of } Q_j) = \frac{C_3}{N^{1/d}}$$

We conclude from (10) that

$$\left|\int_{[0,1]^d} \Phi(x)\, dx - \frac{1}{N}\left[\Phi(x_1) + \cdots \Phi(x_N)\right]\right| \;\le\; \frac{C}{N^{1/d}} \tag{11}$$

for some constant C. In the MC method we have

$$\int_{[0,1]^d} \Phi(x)\, dx \;\approx\; \frac{1}{N}\left[\Phi(X_1) + \cdots \Phi(X_N)\right]\;, \tag{12}$$

$$\left|\int_{[0,1]^d} \Phi(x)\, dx - \frac{1}{N}\left[\Phi(X_1) + \cdots \Phi(X_N)\right]\right| \le \frac{C}{N^{1/2}},\ \text{with high probability,}$$

where $X_j = (X_j^1,..,X_j^d),\ j = 1,2,..,$ and the random variables$\{X_j^i\}_{i=1,...,d,\, j=1,2,..,}$ are i.i.d. uniform on $[0,1]$.

From the above, we see that the MC method is worse than the Riemann sum method if $d = 1,2$, but it is better if $d \ge 3$. In general, the MC method has a big advantage over PDE (or other analytical) methods when the problem is **high-dimensional**.

Ex 1. *We can use MC to estimate the value of a European put option in the BS model. From the general NA argument, we know that the price of the option is given by the expectation*

$$V(S_0, 0) \;=\; e^{-rT} E(K - S_T)^+\;. \tag{13}$$

The random variable S_T is log-normal and can be represented as

$$S_T \;=\; S_0 \exp\left\{(r-\sigma^2/2)T + \sigma\sqrt{T}\xi\right\}\;, \tag{14}$$

*where ξ is a standard normal. In our MC simulation, we generate N values $\xi_1,..,\xi_N$ of i.i.d. standard normal variables (e.g. using the MatLab function "*randn*") and set*

$$X_j \;=\; e^{-rT}\left(K - S_0\exp\left\{(r-\sigma^2/2)T + \sigma\sqrt{T}\xi_j\right\}\right)^+\;,\quad j=1,..,N. \tag{15}$$

Then, we proceed with the algorithm described by the steps $1--5$ above. In particular, we obtain the following estimate of the put price:

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

Ex 2. *In the above example we could use a PDE method to approximate the price of the option, or even compute it using the available approximation of the normal cdf. Let us now consider an example in which the MC method provides the only efficient solution.*

Consider the Merton's model of default risk for multiple companies. We denote each firm's value by S^i, with $i=1,\dots,M$, and, as prescribed by the model, assume that each of them follows a GBM with mean μ^i and volatility σ^i. Denote by D^i the debt level of each firm. Recall that the ith firm defaults at time T if and only if $S^i_T \le D^i$.

Q 1. *What is the* **probability that at least 25% of the companies default***? This question is relevant for investors exposed to the collective credit risk of the given pool of companies (e.g. to a CDO holder).*

The default probability of each company can be computed via MC as in the above example, or using analytical methods. However, if M is **large***, and if*

- *either the companies are* **not homogeneous** *(i.e. μ^i and σ^i are not the same across $i=1,\dots,M$),*
- *or the BMs $\left\{W^i\right\}$ are not independent,*
- *or both,*

computing the joint default probability may be **very difficult using the analytical tools***. Thus, we use the* **MC approach***.*

To estimate the desired probability, we need to generate $N\times M$ matrix $\left\{\xi^i_j\right\}_{i=1,\dots,M,\,j=1,\dots,N}$, where

- *each row $(\xi^1_j,\dots,\xi^M_j)$ has the same distribution as $(W^1_T,\dots,W^M_T)$,*
- *and the rows are independent across all $j=1,\dots,N$.*

Next, we compute the number of defaults n the jth scenario:

$$K_j = \sum_{i=1}^{M} \mathbf{1}_{\left\{S^i_0\exp\left((\mu^i-(\sigma^i)^2/2)T+\sigma^i\sqrt{T}\xi^i_j\right)\le D^i\right\}}\;,\quad j=1,..,N.$$

$$X_j \;=\; \mathbf{1}_{\{K_j\ge 0.25M\}}\;,\quad j=1,..,N.$$

The desired probability is, then, estimated by

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

3 Random number generation

Even though MatLab is able to generate samples from random normal and some other distributions, the number of available distributions is still very limited. In practice, one often has to create a home-made generator for a target distribution, given the basic building blocks (i.e. basic distributions) provided by the chosen computer language.

Generating uniform random variables

In fact, any random number generation starts from the **uniform distribution** on the interval $[0, 1]$. A sample for the uniform distribution is typically generated by computing a sequence of numbers following a recursive relation of the form

$$U_{n+1} = F(U_n),$$

where $F : [0, 1] \to [0, 1]$ is an arithmetic function which can be accurately computed in a relatively small number of steps. Evidently the sequence $U_1, U_2, ..,$ is **deterministic**, but in a good random number generator, the sequence $\{U_j\}_{j=1,2,...}$ is **statistically uniformly distributed and independent** of the variable $\{U_{j+n}\}_{j=1,2,...}$, for any n. That is, this sequence will pass all the standard statistical tests for independence and uniform distribution. This is why such sequences are often called **pseudorandom**.

Most MC methods for uniform distribution are based on the **linear congruential generator**, which produces a sequence $U_1, U_2, \ldots$, as follows:

$$X_{i+1} = aX_i \operatorname{mod} m,$$

$$U_{i+1} = X_{i+1}/m,$$

given the initial value X_0, which is also called a **seed**.

Re-running the sequence with the same seed produces the same sequence $U_1, U_2, \ldots$. The latter feature may be useful in practice: e.g. if we need to compute prices of several derivatives, while preserving the static no-arbitrage conditions.

Notice that, X_i can only take $m - 1$ distinct values. Therefore, sooner or later, the sequence will repeat the same number twice. After that moment, it will repeat itself, forming a cycle. The length of the cycle of a linear congruential generator is called a **period** of the generator. Clearly, one can never get a perfect independence as the sequence (X_n) and, in turn, (U_n), are always **periodic**. However a good random number generator should have a period at least 10^{20}. Then, the sequence looks statistically independent as long as the size of the sample is small relative to the period (e.g. for samples of size $N \approx 10^7$.

A generator with period $m - 1$ is said to have a **full period**. Thus, the numbers m and a are chosen so that the **generator has full period** and m **is large**. Note that m cannot be too large, because of the **overflow** problem: multiplying tow large numbers we can out of the range of possible values allowed by MatLab. So, the choice of m and a is a non-trivial problem. In this course, we do not deal with this problem and simply use the MATLAB command "*rand*(m, n)", which generates an $m \times n$ matrix with entries that are independent uniform random variables. Thus, in what follows, we take a sample of i.i.d. uniform variables as given.

Generating random variables with a discrete distribution

Assume that X takes values in a finite state space $\{x_1, \ldots, x_M\}$ with probabilities $\{p_1, \ldots, p_M\}$ respectively. Define

$$I_1 = (0, p_1], \quad I_n = \left(\sum_{i=1}^{n-1} p_i, \sum_{i=1}^{n} p_i \right], \ n = 2, \ldots, M.$$

Simulate $U \sim Unif(0,1)$, find n, such that $U \in I_n$, and define X to be equal to x_n. It is easy to see that X has the desired distribution (i.e. $\mathbb{P}(X = x_n) = p_n$), and, of course, the independent realizations of U will produce independent realizations of X.

Ex 3. *The above is used in the simulation of Markov Chains. In Finance, Markov Chains are, for example, used to model the* **transition of the credit rating** *of a company throughout the state space* $\{Aaa, Aa, A, Baa, Ba, B, C, D\}$ *(see the book "Credit Risk Modeling", by D. Lando, for more).*

The above method works even if the state space is infinite ($M = \infty$), but **countable**. A simple example is the **Poisson distribution** with rate λ, denoted $Pois(\lambda)$: $p_n = e^{-\lambda}\frac{\lambda^n}{n!}$, for $n = 0, 1, 2, \ldots$.

Ex 4. *A difference between two independent Poisson processes can be used to model the* **total demand** *for an asset in a simple model of market microstructure.*

In addition, the arrival times of a Poisson process can be used a **simple (constant-intensity) model for default times** *in a given pool of companies. We will discuss the* **intensity models** *in more detail later in this chapter.*

Inverse transform method

The above algorithm is just a particular example of the following method. Assume that we need to generate a random variable X, with values in $\mathbb{R}$, whose cdf is given by function F. Consider its inverse F^{-1} (if the definition of F^{-1} is ambiguous (i.e. if F is not strictly monotone), we can, for example, define it as the "right inverse"). Generating $U \sim Unif(0,1)$, we easily check that $X = F^{-1}(U)$ has the desired distribution

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

Ex 5. *Assume we need to generate a sample of independent random exponentials* $X_1, \ldots, X_N \sim Exp(\lambda)$.

- *First, generate i.i.d.* $U_1, \ldots, U_N \sim Unif(0,1)$.
- *Then compute the inverse of the cdf:*

$$F(x) = 1 - e^{-\lambda x}, \quad F^{-1}(y) = -\frac{1}{\lambda}\log(1-y).$$

- *Finally, compute the desired sample:*

$$X_1 = -\frac{1}{\lambda}\log(1-U_1), \ldots, X_N = -\frac{1}{\lambda}\log(1-U_N)$$

Ex 6. *One can also simulate conditional distributions using the same trick. Consider the random variable* X *which takes values in* $\mathbb{R}$ *and has the cdf* F. *Assume that we need to simulate* X *conditional on* $X \in [a,b]$. *Simulating* $U \sim Unif(0,1)$, *we compute*

$$F^{-1}(F(a) + (F(b) - F(a))U),$$

and notice that it has the desired distribution:

$$\mathbb{P}(F^{-1}(F(a) + (F(b) - F(a))U) \leq x) = \mathbb{P}(F(a) + (F(b) - F(a))U \leq F(x))$$

$$= \mathbb{P}(U \leq \frac{F(x) - F(a)}{F(b) - F(a)}) = \frac{F(x) - F(a)}{F(b) - F(a)} = \mathbb{P}(X \leq x \,|\, X \in [a,b])$$

In many cases, it is **impossible to compute F^{-1} explicitly**. However, **if there is an efficient way to compute the values of F**, we can invert this function by the so called **bisection method**.

- Assume that we need to find x, such that $F(x) = y$.
- Start from some small a_0 and large b_0, such that $F(a_0) \leq y \leq F(b_0)$.
- For any $n = 0, 1, 2, \ldots$,
  - if $F((a_n + b_n)/2) \leq y$, define $a_{n+1} = (a_n + b_n)/2$ and $b_{n+1} = b_n$;
  - if $y \leq F((a_n + b_n)/2)$, define $a_{n+1} = a_n$ and $b_{n+1} = (a_n + b_n)/2$.

After N steps, use $(a_N + b_N)/2$ as an approximation for x. It approximates x with the precision $O(2^{-N})$.

Box-Müller method

The cdf of a standard normal is given by

$$\Phi(x) \ = \ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy \ , \quad x \in \mathbb{R}. \tag{16}$$

There is no explicit formula for $\Phi(x)$, and certainly no explicit inverse formula. Nevertheless, there exist very precise analytical approximations of Φ and its inverse. In fact, the standard optimally efficient method for generating i.i.d. standard normal variables does use a sophisticated version of the inverse transform method.

A straightforward, but not particularly efficient, way of generating the standard normal variable from the uniform variable is the Box-Müller method. To see how this works we consider two independent standard normal variables X, Y and consider their joint distribution:

$$\mathbb{P}((X,Y) \in A) = \int_A \mathbb{P}(X \in dx, Y \in dy) = \int_A \frac{1}{2\pi} e^{-(x^2+y^2)/2}\, dxdy \ = \ \int_B \frac{1}{2\pi} e^{-r^2/2}\, rdrd\theta, \tag{17}$$

where we switched to polar coordinates:

$$r = \sqrt{x^2 + y^2}, \quad \theta = \arctan(y/x),$$

$$x(r,\theta) = r\cos\theta, \quad y(r,\theta) = r\sin\theta,$$

and the set B is defined as

$$B = \{(r,\theta)\ :\ (x(r,\theta), y(r,\theta)) \in A\}$$

We can make another change of variables, introducing

$$u = e^{-r^2/2}, \quad v = \theta/(2\pi)$$

$$x(u,v) = \sqrt{-2\log u}\cos(2\pi v), \quad y(u,v) = \sqrt{-2\log u}\sin(2\pi v),$$

to obtain from (17):

$$\mathbb{P}((X,Y) \in A) = \int_B \frac{1}{2\pi} e^{-r^2/2}\, rdrd\theta \ = \ \int_C dudv, \tag{18}$$

where

$$C = \{(u,v)\ :\ (x(u,v), y(u,v)) \in A\}$$

Let us now define the two random variables U, V by

$$U \;=\; \exp[-(X^2+Y^2)/2], \; V \;=\; \frac{1}{2\pi}\arctan(Y/X)\;, \tag{19}$$

and let us show that they are independent and uniformly distributed in $[0,1]$. It is clear that U, V are restricted to the interval $[0,1]$. Consider an arbitrary rectangle $[\bar{u}+\delta u]\times[\bar{v}+\Delta v]$. Then, we have

$$\mathbb{P}((U,V)\in[\bar{u}+\delta u]\times[\bar{v}+\Delta v]) = \mathbb{P}((X,Y)\in A),$$

where $A = \{(x,y)\,:\,(u(x,y),v(x,y))\in[\bar{u}+\delta u]\times[\bar{v}+\Delta v]\}$. From (18), we obtain

$$\mathbb{P}((U,V)\in[\bar{u}+\delta u]\times[\bar{v}+\Delta v]) = \mathbb{P}((X,Y)\in A) = \int_C dudv,$$

where

$$C = \{(u,v)\,:\,(x(u,v),y(u,v))\in A\} = [\bar{u}+\delta u]\times[\bar{v}+\Delta v]$$

by the definition of A. Thus,

$$\mathbb{P}((U,V)\in[\bar{u}+\delta u]\times[\bar{v}+\Delta v]) = \Delta u\,\Delta v,$$

and, hence, U and V are independent and uniformly distributed in $[0,1]$.

It is easy to see that we can run this argument in the opposite direction (since the distribution of a function of a random vector is uniquely determined by the distribution of this vector) and begin with two independent variables U, V uniformly distributed in $[0,1]$. Then on inverting the formulas (19) we see from (17) that the variables X, Y defined by

$$X \;=\; \sqrt{-2\log U}\cos(2\pi V)\;, \quad Y \;=\; \sqrt{-2\log U}\sin(2\pi V)\;, \tag{20}$$

must be independent and standard normal.

The lack of efficiency in the method (20) is to be found in the necessity of computing logarithms and trigonometric functions in its implementation. In designing algorithms for random number generation it is very important that a random number can be computed with a rather small number of computations. However algorithms for computing accurate values of transcendental functions like logarithm, sine and cosine have to use a significant number of computations (relative to the complexity of other methods – for example, the congruential generator only need arithmetic operations which are computationally cheap).

Acceptance-Rejection method

This is a very general method that works in any dimension and does not require inversion of the cdf.

Assume that the target random variable X with values in $\mathbb{R}^n$ has pdf f, which satisfies

$$f(x) \le cg(x),$$

where c is some constant and g is a pdf of a distribution which we know how to simulate from.

- Gneerate X from the distribution given by g.
- Generate an independent $U \sim Unif(0,1)$.
- Generate Bernoulli random variable ξ_X, which is equal to 1 if $U \le f(X)/(cg(X))$, and to zero otherwise.
- Accept X and call it Y, if $\xi_X = 1$, otherwise reject it.

- Repeat the above until accept.

Denote by Y a representative random variable resulting from the above simulation algorithm. Let us show that the pdf of Y is f:

$$\mathbb{P}(Y \in A) = \mathbb{P}(X \in A \,|\, U \le f(X)/(cg(X))) = \frac{\mathbb{P}(X \in A,\, U \le f(X)/(cg(X)))}{\mathbb{P}(U \le cg(X)/f(X))}$$

$$= \frac{\int_A \frac{f(x)}{cg(x)} g(x)dx}{\int_{\mathbb{R}^n} \frac{f(x)}{cg(x)} g(x)dx} = \int_A f(x)dx$$

Ex 7. **Gamma distribution** *with scale parameter $a > 0$ and shape parameter $b > 0$, denoted $\Gamma(a,b)$ is a probability distribution on $(0,\infty)$, given by the pdf*

$$f_{a,b}(x) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-x/b}, \quad x > 0,$$

where Γ is the **Gamma function**. *It has mean ab and variance ab^2.*

If $b = 2$ and $a = \nu/2$, the Gamma distribution becomes the χ^2_ν **(Chi-square) distribution with ν degrees of freedom**. *This distribution appears very often in statistics (e.g. in hypothesis checking, or in regression analysis) as it represents the distribution of a cumulative square error of a statistical estimator in many popular models (e.g. the "R-squared", in the case of regression). This distribution is also important for simulating the CIR process and, in particular, the paths of spot volatility in the Heston model.*

If $X \sim \Gamma(a,1)$ then $bX \sim \Gamma(a,b)$, hence it is sufficient to be able to generate samples from $\Gamma(a,1)$. To do this in the case of $a \le 1$, Ahrens and Dieter propose to use the **acceptance-rejection** *algorithm. As the auxiliary function g they choose*

$$g(x) = \begin{cases} pax^{a-1}, & x \in [0,1], \\ (1-p)e^{-x+1}, & x > 1, \end{cases}$$

where $p = e/(a+e)$. Generating X with density g is easy:

- *generate a marker I which is 1 with prob. p and 2 with prob. $(1-p)$,*
- *produce X_1 with density $ax^{a-1}\mathbf{1}_{[0,1]}(x)$ (this is done via the inverse transform method),*
- *produce X_2 with density $e^{-x+1}\mathbf{1}_{[1,\infty)}(x)$ (this is also done via the inverse transform method),*
- *set $X = X_1$ if $I = 1$ and $X = X_2$ if $I = 2$.*

It only remains to notice that, if $a \le 1$, the following holds for all $x > 0$:

$$\frac{f_{a,1}(x)}{g(x)} \le \frac{a+e}{ae\Gamma(a)} \le 1.39$$

Thus, we can apply the acceptance-rejection method using the density g, to simulate from the distribution given by $f_{a,1}$.

Rem 2. *If $a > 1$, there also exists an efficient way to generate a sample from Gamma distribution. This algorithm is due to Cheng and Feast and it is based on the* **ratio of uniforms** *methods, which, in turn, is closely related to the acceptance-rejection method restricted to the one-dimensional case.*

3.1 Simulating dependent random variables

Using the example of Merton model for credit risk, we have already discussed the need to simulate multivariate random vectors. Of course, one can simply generate a sample of one-dimensional random variables of size k, and combine them together into a vector of dimension k. However, the entries of this vector will be independent, and in many applications, it is **important to model dependence between the entries of a vector**.

There are many ways of measuring the dependence of 2 random variables X and Y, but the most popular way is through the covariance $\mathrm{cov}[X,Y]$ defined by

$$\mathrm{cov}[X,Y] \;=\; E\left[\{X-E[X]\}\{Y-E[Y]\}\right] \;=\; E[XY]-E[X]E[Y]\,. \tag{21}$$

Evidently the variance of a random variable X is just the covariance of X and X, so $\mathrm{var}[X]=\mathrm{cov}[X,X]$. Observe now from (21) that upon using the Cauchy-Schwarz inequality we have that

$$|\mathrm{cov}[X,Y]| \;\le\; E\left[\{X-E[X]\}^2\right]^{1/2} E\left[\{Y-E[Y]\}^2\right]^{1/2} \;=\; \sqrt{\mathrm{var}[X]\mathrm{var}[Y]}\,. \tag{22}$$

From (22) the **coefficient of correlation** $\mathrm{cor}(X,Y)$ between X,Y defined by

$$\mathrm{cor}(X,Y) \;=\; \mathrm{cov}[X,Y]\big/\sqrt{\mathrm{var}[X]\mathrm{var}[Y]}\,, \tag{23}$$

satisfies the inequality $-1 \le \mathrm{cor}(X,Y) \le 1$. If $\mathrm{cor}(X,Y)=+1$ then $X=\lambda Y$ for some scalar $\lambda>0$. If $\mathrm{cor}(X,Y)=-1$ then $X=-\lambda Y$ for some scalar $\lambda>0$. Thus if $|\mathrm{cor}(X,Y)|=1$ then X,Y are perfectly correlated, i.e. the value of X determines the value of Y.

Observe also that if X,Y are **independent** then $\mathrm{cor}(X,Y)=0$. It **does not** however follow that $\mathrm{cor}(X,Y)=0$ implies X,Y independent. Nevertheless, in statistics, one tends to assume that if $|\mathrm{cor}(X,Y)|<<1$ then X,Y are only weakly correlated, so are “close to being independent”. Of course, this logic only works if the dependence between X and Y is linear, or close to linear.

Exercise 1. *Consider $X\sim N(0,1)$ and $Y=\exp(\sigma X)$. Show that*

$$cor(X,Y)\to 0, \quad \text{as } \sigma\to\infty,$$

despite the perfect dependence between X and Y.

For a k-dimensional random vector $X=(X_1,\dots,X_k)^T$, we define its mean $\mathbb{E}X$ via

$$\mathbb{E}X=(\mathbb{E}X_1,\dots,\mathbb{E}X_k)^T,$$

and its covariance matrix $\Sigma\in\mathbb{R}^{k\times k}$ via $\Sigma=\mathbb{E}(X-\mathbb{E}X)(X-\mathbb{E}X)^T$, or, equivalently,

$$\Sigma_{i,j}=\mathrm{Cov}\left((,X)_i\,,X_j\right)$$

If the components of X are uncorrelated, then Σ is diagonal.

- Evidently $\Sigma_{i,j}=\Sigma_{j,i}$, hence Σ is a **symmetric** matrix.
- It is also **nonnegative definite (or, positive semidefinite)**:

$$\sum_{i,j=1}^{k}\lambda_i\Sigma_{i,j}\lambda_j \;\ge\; 0 \quad \text{for all vectors } \lambda=[\lambda_1,..,\lambda_k]\in\mathbb{R}^k\,. \tag{24}$$

To see why (24) holds we observe that

$$\sum_{1\le i,j\le k} \lambda_i \Sigma_{i,j} \lambda_j \;=\; \mathbb{E}\left[\left\{ \sum_{i=1}^{k} \lambda_i X_i \right\}^2 \right] \;\ge\; 0\,. \tag{25}$$

- If Σ satisfies:

  $$\sum_{i,j=1}^{k} \lambda_i \Sigma_{i,j} \lambda_j > 0 \quad \text{for all vectors } \lambda = [\lambda_1, .., \lambda_k] \neq 0,$$

  then we say that Σ is **positive definite**. Any covariance matrix Σ that has **full rank** is positive definite.

Multivariate Gaussian distribution

Def 1. *Random vector X, with values in $\mathbb{R}^k$, is Gaussian with mean $\mu \in \mathbb{R}^k$ and covariance matrix $\Sigma \in \mathbb{R}^{k\times k}$ (denoted $N(\mu, \Sigma)$), if there exists a standard normal vector $\xi = (\xi_1, \ldots, \xi_m)$, with i.i.d. random variables $\xi_1, \ldots, \xi_m \sim N(0,1)$ and a matrix $A \in \mathbb{R}^{k\times m}$, such that*

$$X = \mu + A\xi, \quad \Sigma = AA^T$$

Gaussian distribution is determined uniquely by (μ, Σ) but there may be many different matrices A that satisfy $AA^T = \Sigma$!

Among all such A, we can identify the one that is **lower triangular**.

Thm 3. **(Cholesky decomposition)** *For any non-negative definite $k \times k$ matrix Σ, there exists a $k \times k$ matrix A such that*

- $\Sigma = AA^T$,
- *A is lower triangular,*
- *its diagonal entries are nonnegative.*

If, in addition, Σ is positive definite (i.e. it has full rank), then such matrix A is unique.

One can use the MATLAB function "chol(Σ)" to find the transposed Cholesky matrix A^T corresponding to a given Σ.

To simulate a k-dimensional (or k-variate) random normal X with mean μ and covariance Σ, we

- find the Cholesky decomposition $\Sigma = AA^T$;
- generate a vector of k independent standard normals $\xi = (\xi_1, \ldots, \xi_k)$,
- define $X = \mu + A\xi$.

Ex 8. *Now, we can complete the last example of Section 2. Namely, we can compute the probability that a specified fraction (say, 25%) of companies in a given pool default at time T, according to the Merton model. The only missing part of the algorithm was the simulation of $(W^1_T, \ldots, W^M_T)$, where W^is are mutually dependent Brownian motions. It is easy to see that*

$$\mathit{Law}(W^1_T, \ldots, W^M_T) = \mathit{Law}\left(\sqrt{T}\xi = \sqrt{T}(\xi^1_T, \ldots, \xi^M_T)\right),$$

where $\xi^i \sim N(0,1)$, for every $i = 1,\dots,M$, and, hence $\mu = \mathbb{E}\xi = 0$, but ξ^is may be mutually dependent, with the dependence structure described by the covariance matrix Σ:

$$\Sigma_{ij} = \mathbb{E}\xi^i\xi^j$$

Choosing appropriate Σ we can capture the dependence between default events, and, in particular, the **default contagion** *effect.*

Ex 1. *Consider a spread call option with maturity T and strike K, written on S^1 and S^2:*

$$V_T = (S_T^1 - S_T^2 - K)^+$$

Such options are popular in energy markets. However, if $K \neq 0$, even in the 2-dimensional BS model

$$dS_t^1 = (r - q_1)S_t^1 dt + \sigma_1 S_t^1 dW_t^1,$$

$$dS_t^2 = (r - q_2)S_t^2 dt + \sigma_2 S_t^2 dW_t^2,$$

with $\langle dW_t^1 dW_t^2\rangle = \rho dt$, there is no explicit formula for the price of such option – one has to solve a 2-dim PDE or evaluate a 2-dim integral.

Exercise 2. *Use the change of numeraire technique to compute the price of the spread option with $K = 0$.*

Nevertheless, the MC method still works. Using the historical values of $\Delta S_t^1/S_t^1$ and $\Delta S_t^2/S_t^2$, we can estimate the correlation ρ, then, simulate (W_T^1, W_T^2), and price the option by MC. To simulate $(W_T^1, W_T^2)^T$, we notice that it is a Gaussian vector, with mean $\mu = 0$ and the covariance matrix

$$\Sigma_T = T\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where we interpreted $\langle dW_t^1 dW_t^2\rangle = \rho dt$ as

$$cov(W_t^1, W_t^2) = \rho t$$

To simulate from the distribution $N(0,\Sigma)$, it only remains to find the decomposition $AA^T = \Sigma$:

$$A = \sqrt{T}\begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix}$$

The above expression for A can be explained by the following (useful) observation: if W^1 and W^2 are two Brownian motions with correlation ρ, then, there exist two mutually independent BM's $\tilde{W}^1$ and $\tilde{W}^2$, such that

$$W_t^1 = \tilde{W}_t^1,$$

$$W_t^2 = \rho\tilde{W}_t^1 + \sqrt{1-\rho^2}\tilde{W}_t^2$$

Copulas

In some cases it is convenient to separate the marginal distributions of a random vector $X = (X_1,\dots,X_k)^T$ (i.e. the distributions of every X_i, for $i = 1,\dots,k$) and the dependence structure between the entries of the vector.

Ex 2. *Recall the example of spread option. Assume that we no longer believe in the BS model, but, instead obtain the marginal distributions of S^1_T and S^2_T from the European options written on each of these assets individually (e.g. by calibrating a LV model and computing the associated pdf). Then, to find the joint distribution of (S^1_T, S^2_T), we need to capture the dependence between the two assets, for example, by looking at the historical data. In addition, it is very desirable that the choice of the dependence structure is "separated" from the choice of marginal distributions. Notice that, modeling dependence by estimating the correlation between the two assets (as discussed above) does not satisfy the latter requirement: to estimate the correlation, we also need to know the structure of the volatility of each asset, which, in turn, is closely related to its marginal distribution.*

Ex 3. Credit Default Swap (CDS). *CDS can be viewed as* **insurance against default** *of a given company. It is one of the biggest derivatives markets in the world. There is no money exchanged at the time when the contract is signed: instead the two parties negotiate the size of future payments from the buyer to the seller (measured as a fraction of the notional and per unit of time).*

Denote the (constant) interest rate by r, the notional of the contract by D, its maturity by T, the premium (also known as the **CDS spread***) by c, the times of payments by $\{T_i\}$, and the (random) default time by τ.*

Then, the seller of CDS receives:

$$\mathbb{E}^{\mathbb{Q}}\left(\sum_{T_i\leq T\wedge\tau} e^{-rT_i}c(T_i - T_{i-1})D\right)$$

And the buyer of CDS receives:

$$\mathbb{E}^{\mathbb{Q}}\left(e^{-r\tau}D\mathbf{1}_{\tau\leq T}\right)$$

So, the CDS premium

$$c(T) = \frac{\mathbb{E}^{\mathbb{Q}}\left(e^{-r\tau}\mathbf{1}_{\tau\leq T}\right)}{\mathbb{E}^{\mathbb{Q}}\sum_{T_i\leq T\wedge\tau} e^{-rT_i}(T_i - T_{i-1})}$$

In the continuous time limit:

$$c(T) = \frac{\mathbb{E}^{\mathbb{Q}}\left(e^{-r\tau}\mathbf{1}_{\tau\leq T}\right)}{\int_0^T e^{-rs}\mathbb{Q}(\tau\geq s)ds}$$

Q 2. *How do we compute the CDS premium $c(T)$?*

It is clear that, to compute $c(T)$, we need to model the distribution of the default time τ. The simplest possible model of default is the Merton's model, where $\tau = T$ if the firm's value at time T is below the debt level. However, this model is too restrictive: it assumes that the default can only occur at a chosen maturity T, and, hence, the premium of any CDS with shorter maturity is zero (there is a zero probability of default). This becomes a problem if we have CDS contracts with multiple maturities traded in the market (and we typically do have a lot of them).

As discussed earlier, a simple extension of the Merton model is provided by Black and Cox, who define τ as the first time that the firm's value drops below a given default boundary H:

$$\tau = \inf\{t \geq 0\, :\, S_t \leq H_t\}$$

In this model, the default may occur at any time, and the CDS premium does not degenerate for any maturity T. However, it turns out that the Black-Cox model (as well as any **structural diffusion model***) cannot produce realistic short-term CDS spreads: $c(0)$. The problem is that, in the BLack-Cox model, the probability of default, $\mathbb{Q}(\tau \leq T)$ decays too fast as $T \to 0$, and, as a result, we always obtain zero spread in this model $c(0) = 0$. A more detailed explanation of this phenomenon is given in Section 2.2.2 of "Credit Risk Modeling" by Lando.*

The real market typically has $c(0) > 0$, indicating higher probability of short-term default than the Black-Cox model (or any structural diffusion model). This market phenomenon can be reproduced by the so called **reduced form (or, intensity) models**.

Reduced form (intensity) models.

Introduce the **default intensity** *(also known as the* **hazard rate***) $\lambda(s)$, for $s \in [0, T]$. Reduced form models assume that*

$$\tau = \inf \left\{ t \geq 0 \,:\, \int_0^t \lambda(s) ds \geq e \right\},$$

where $e \sim Exp(1)$ and e is independent of λ. Assuming, for simplicity, that λ is deterministic, we obtain:

$$\mathbb{Q}(\tau \geq t) = \mathbb{Q}(\int_0^t \lambda(s) ds < e) = \exp(-\int_0^t \lambda(s) ds), \tag{26}$$

$$\mathbb{E}^{\mathbb{Q}} \left(e^{-r\tau} \mathbf{1}_{\tau \leq T} \right) = \int_0^T e^{-rs} \mathbb{Q}(\tau \in dt) = \int_0^T \lambda(s) e^{-rs - \int_0^t \lambda(u) du} du$$

Thus

$$c(T) = \frac{\int_0^T \lambda(s) e^{-\int_0^s (r + \lambda(u)) du} ds}{\int_0^T e^{-\int_0^s (r + \lambda(u)) du} ds}$$

We can calibrate λ to the CDS curve $\{c(T)\}_T$, which gives us the distribution of τ. It is not too hard to notice (by expanding the numerator and denominator in the powers of T) that $c(0) = \lambda(0)$. Hence, there is no problem choosing $\lambda(0) > 0$ so that the short-term spread is positive.

Multiple obligors.

Doing this for various companies, we can find the distribution of each τ^i, for $i = 1, \ldots, M$, individually. Then, to find the joint distribution of $(\tau^1, \ldots, \tau^M)$, which may be needed to price a CDO tranche or estimate the total credit risk exposure of an investor (e.g. a bank), we only need to specify the dependence structure between the entries of the vector $(\tau^1, \ldots, \tau^M)$. This dependence structure should be chosen without changing the marginal distributions (given by the market).

Copula offers an approach that allows us to fit the dependence structure between the entries of a vector, separately from its marginal distributions.

Denote the cdf of X_i by F_i. Notice that

$$U = (F_1(X_1), \ldots, F_k(X_k))^T$$

is a vector of uniformly distributed random variables. However, if the entries of X were dependent (i.e. not independent), then, their dependence structure is transferred to U. These arguments can be reversed: we can start from a vector U, and construct X with the dependence structure given by U.

Def 2. *A cumulative distribution function $C(u_1, \ldots, u_k)$ of a vector of random variables $(u_1, \ldots, u_k) \in [0, 1]^k$, such that $u_i \sim Unif(0, 1)$, for $i = 1, \ldots, k$, is called a* **copula**.

We will say that copula C_X is associated with a random vector (or the multivariate distribution of) X if the cdf of the random vector

$$U = (F_1(X_1), \ldots, F_k(X_k))^T$$

is given by C_X, where F_i is the cdf of X_i, for $i = 1, \dots, k$. In this case

$$C_X(u_1, \dots, u_k) = F(F_1^{-1}(u_1), \dots, F_k^{-1}(u_k)),$$

where F is the cdf of X.

Properties of copulas

- Copula of X does not change if we add a number to each entry of X or multiply each entry of X by a number.
- More generally, copula of X does not change if we apply a monotone increasing function ψ_i to X_i, for every $i = 1, \dots, k$.
- Denote by $f(x_1, \dots, x_k)$ the density of X and by $f_i(x_i)$ the density of X_i, for all $i = 1, \dots, k$. Then, the density of the associated copula is given by

$$c(u_1, \dots, u_k) = \partial^k_{u_1 \cdots u_k} C(u_1, \dots, u_k) = \frac{f(F_1^{-1}(u_1), \dots, F_k^{-1}(u_k))}{f_1(F_1^{-1}(u_1)), \dots, f_k(F_k^{-1}(u_k))}, \quad (u_1, \dots, u_k) \in [0,1]^k.$$

Estimating copula from data

Let us assume that we know the marginal distributions of the vector $X = (X_1, \dots, X_k)^T$: denote by F_i the cdf of X_i. Assume, in addition, that we have an i.i.d. sample of the values of X

$$X^{(i)} = \left(X_1^{(i)}, \dots, X_k^{(i)}\right), \quad i = 1, \dots, n.$$

In many cases (e.g. to achieve a better precision), we need to produce a larger MC sample of the values of X. Then, we have to determine the structure of dependence between its components. This can be done by fitting a copula to the historical sample

$$U^{(i)} = \left(F_1(X_1^{(i)}), \dots, F_k(X_k^{(i)})\right), \quad i = 1, \dots, n.$$

To fit a copula to this sample, one can use a **non-parametric (e.g. kernel) regression** and fit the copula to the **empirical cdf** of $\left\{U^{(i)}\right\}_{i=1}^n$. However, due to the very high computational complexity of the latter algorithm, in higher dimensions, the preferred method is to choose a **parametric family** of copulas and estimate the parameters via **Maximum Likelihood**. For example, we may consider the family of Gaussian copulas, given by

$$C_{Gauss,k,\Sigma}(u_1, \dots, u_k) = \Phi_{k,0,\Sigma}(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_k)),$$

where $\Phi_{k,0,\Sigma}$ is the cdf of k-dimensional Gaussian with zero mean and covariance matrix Σ, and Φ is a standard normal cdf. This family is parameterized by the symmetric nonnegative definite matrices Σ, where, without loss of generality, we can assume that the diagonal entries of Σ are equal to one.

We obtain a flexible family of Gaussian copulas by choosing various Σ. This enables us to model dependence between the entries of a random vector, separately from its marginal distributions. One can further reduce this family (restricting the quality of the fit but reducing the computational complexity), for example, by considering only matrices Σ, whose off-diagonal entries are equal to a fixed number $\rho \in [0, 1]$.

MC simulations via copula

Once we have picked a copula that is consistent with the given initial sample, we can generate larger MC sample from this copula:

$$U^{(i)} = \left(U_1^{(i)}, \dots, U_k^{(i)}\right), \quad i = 1, \dots, N,$$

and produce the corresponding sample of X:

$$X^{(i)} = (F_1^{-1}(U_1^{(i)}), \dots, F_k^{-1}(U_k^{(i)})), \quad i = 1, \dots, N,$$

where $N >> n$.

Ex 9. *Consider a pool of companies and denote the default time of each company by τ^j, for $j = 1, \dots, M$. Assume that the market gives us the* **CDS curves** *$(c^j(T))$, for every $j = 1, \dots, M$, and we have calibrated an intensity λ^j to each curve. Then, we can compute the cdf F_j of each τ^i via (26).*

Assume also that we are given a historical sample of the vector

$$\tau^{(i)} = (\tau^{(i),1}, \dots, \tau^{(i),M}), \quad i = 1, \dots, n$$

Then, assuming the above sample is i.i.d. we can fit a **copula** *(say, from the Gaussian family) to the sample*

$$U^{(i)} = \left(F_1(\tau^{(i),1}), \dots, F_k(\tau^{(i),M})\right), \quad i = 1, \dots, n,$$

and, then, follow the above discussion to generate a larger MC sample of the default vectors.

Based on this new sample, we can, for example, estimate the value of a **CDO tranche** *(whose payoff is a function of the random variables $\mathbf{1}_{\tau^j \leq T}$, for $j = 1, \dots, M$).*

Ex 10. Risk management. *A financial company typically has exposure to many different sources of risk, and it is very important to be able to estimate the total level of risk. The overall portfolio of a company can be written as*

$$P = \sum_{j=1}^{M} w^j S^j,$$

where S^js are the assets that the company holds on its books. The risk of such portfolio is given by the negative part of the total P&L (profit and loss)

$$\Delta P = \sum_{j=1}^{M} w^j \Delta S^j, \quad \Delta S^j = S^j_{t+\Delta t} - S^j_t.$$

This risk can be quantified by choosing a so called **risk measure**. *The most popular choice of a risk measure is the* **Value-at-Risk (VaR)***:*

$$\mathit{VaR}_\alpha = \inf\left\{x \,:\, \mathbb{P}(\Delta P + x \leq 0) \leq \alpha\right\} = \inf\left\{x \,:\, F(-x) \leq \alpha\right\},$$

where F is the cdf of ΔP. However, the above measure has certain deficiencies (e.g. it does not take into account the size of potential losses that exceed VaR, and it does not encourage diversification), hence, the **Expected Shortfall** *was recently proposed by the Basel Committee as a substitute for VaR:*

$$ES_\alpha = \mathbb{E}(-\Delta P \,|\, \Delta P \leq -\mathit{VaR}_\alpha) = \frac{1}{\alpha} \int_{-\infty}^{-\mathit{VaR}_\alpha} (-x) dF(x)$$

Q 3. *It is clear that, to evaluate any of the above risk measures, we need to be able to approximate the cdf F. How can we do it?*

Let us assume that we can estimate the marginal distributions of ΔS^i_t (e.g. this is the case if one can observe the prices of European options written on these assets). Then, the problem becomes to determine the structure of dependence between ΔS^i_t's. This can be done by fitting a copula to the historical sample

$$U^{(i)} = \left(F_1(\Delta S^{(i),1}), \dots, F_M(\Delta S^{(i),M})\right), \quad i = 1, \dots, n,$$

where F_j is the cdf of S^j.

Once we have picked a copula that best explains the historical observations, we can generate MC samples from this copula:

$$U^{(i)} = \left(U^{(i),1}, \dots, U^{(i),M}\right), \quad i = 1, \dots, N,$$

and produce the corresponding sample of asset returns and P&L:

$$\Delta S^{(i),j} = F_j^{-1}(U^{(i),j}), \quad j = 1, \dots, M, \quad \Delta P^{(i)} = \sum_{j=1}^{M} w^j \Delta S^{(i),j}, \quad i = 1, \dots, N,$$

where $N >> n$.

Finally, we can estimate the chosen risk measure using the **empirical cdf** *of ΔP:*

$$\hat{F}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{(-\infty,x]}(\Delta P^{(i)})$$

For example, VaR$_\alpha$ is a negative of the α-quantile of $\hat{F}$:

$$\textit{VaR}_\alpha \approx -\min\left\{\Delta P^{(i)} \;:\; \sum_{\Delta P^{(j)} \leq \Delta P^{(i)}} \frac{1}{N} > \alpha \right\},$$

and

$$ES_\alpha \approx \frac{1}{\alpha N} \sum_{\Delta P^{(j)} \leq -\textit{VaR}_\alpha} (-\Delta P^{(j)})$$

4 Numerical Methods for Stochastic Differential Equations (SDEs)

4.1 One-dimensional case

In Chapter I we studied solutions of ODEs and numerical methods for solving them. Here we shall do an analogous study of stochastic differential equations (SDEs). An SDE can be thought of as a **randomly perturbed** ODE. Thus if $b : \mathbb{R}^2 \to \mathbb{R}$ and $\sigma : \mathbb{R}^2 \to \mathbb{R}_+$ are given functions we associate to them the SDE

$$dX_t \;=\; b(X_t, t)dt + a(X_t, t)dB_t, \tag{27}$$

where B is a Brownian motion. The effect of the process B is to give the particle with position X_t a random "kick" at each time t with the kicks being i.i.d. for different times. Just as for ODEs the evolution equation (27) can be solved uniquely for $t > 0$ with the initial data X_0 given. Evidently X_0 can be an arbitrary **random variable**, which as a special case could be chosen deterministically as $X_0 \equiv x_0 \in \mathbb{R}$ with probability 1.

If $(B_t)_{t\geq 0}$ is a BM then a typical path of $t \mapsto B_t$ is very wiggly and in fact not differentiable in the sense that the pointwise derivative

$$\frac{dB_t}{dt} \;=\; \lim_{\Delta t \to 0} \frac{B_{t+\Delta t} - B_t}{\Delta t}$$

never exists with probability 1. We can however define the integral with respect to B as a limit of integrals of simple (piece-wise constant) processes. In particular, under some regularity assumptions on a and X, we have:

$$\int_t^{t+\Delta t} a(X_u, u) dB_u \approx a(X_t, t)(B_{t+\Delta t} - B_t) = a(X_t, t)\sqrt{\Delta t}\xi,$$

where $\xi \sim N(0, 1)$ and it is independent of everything else observed up to time t. Of course, a similar approximation holds for the drift part:

$$\int_t^{t+\Delta t} b(X_u, u) du \approx b(X_t, t)\Delta t.$$

Explicit Euler method

To numerically solve (27) with given initial data we use the **explicit Euler method** as we did for ODEs:

$$X_{t+\Delta t} \;=\; X_t + b(X_t, t)\Delta t + \sigma(X_t, t)[B_{t+\Delta t} - B_t]\,. \tag{28}$$

Thus, choosing a partition of the time interval with $\Delta t = T/M$, and setting $X^m \simeq X(m\Delta t)$, we make use of (28) to obtain the following recurrence relation

$$\hat{X}_{m+1} \;=\; \hat{X}_m + b(\hat{X}_m, m\Delta t)\Delta t + a(\hat{X}_m, m\Delta t)\sqrt{\Delta t}\xi_m, \quad m = 0, 1, 2, ..M - 1, \tag{29}$$

where $\xi_0, \xi_1, .., \xi_{M-1}$ are i.i.d. standard normal.

Suppose now that we wish to estimate the expectation

$$E[F((X_t)_{t\in[0,T]})],$$

for some function F which **may depend on the entire path of** X. Then we choose an integer M such that $M\Delta t = T$ and use the MatLab function *randn* to generate $\xi_0, .., \xi_{M-1}$. Solving the recurrence (29) with initial data x_0 gives us one MC realization of the paths $(\hat{X}_0, \dots, \hat{X}_M) \approx (X_0, X_{\Delta t}, \dots, X_{M\Delta t})$.

We can do this N times, obtaining N independent sample paths

$$\left\{(\hat{X}_0^{(i)}, \dots, \hat{X}_M^{(i)})\right\}_{i=1}^N,$$

which approximate of $(X_0, X_{\Delta t}, \dots, X_{M\Delta t})$.

Then we set

$$E[F((X_t)_{t\in[0,T]})] \;\approx\; \frac{1}{N}\left[\tilde{F}(X_0^{(1)}, \dots, X_M^{(1)}) + \cdots + \tilde{F}(X_0^{(N)}, \dots, X_M^{(N)})\right], \tag{30}$$

where $\tilde{F}$ is the "discretization" of F – we need to use it to pass from function F, which depends on the entire path $(X_t)_{t\in[0,T]}$ to a function that depends only on a finite-dimensional vector $(X_0, X_{\Delta t}, \dots, X_{M\Delta t})$.

Ex 11. *Consider the BS model:*

$$dS_t = rS_t dt + \sigma S_t dB_t$$

The explicit Euler scheme becomes: $\Delta t = T/M$,

$$\hat{S}_{m+1} = \hat{S}_m + r\hat{S}_m \Delta t + \sigma \hat{S}_m \sqrt{\Delta t} \xi_m, \quad m = 0, 1, 2, ..M - 1,$$

with i.i.d. $\xi_m \sim N(0,1)$. *Having generated a large number* N *of independent sample paths*

$$\left\{ (\hat{S}_0^{(i)}, \ldots, \hat{S}_M^{(i)}) \right\}_{i=1}^{N},$$

We can, for example,

- *price a* **European call***:*

$$e^{-rT} \mathbb{E}(S_T - K)^+ \approx e^{-rT} \frac{1}{N} \sum_{i=1}^{N} (\hat{S}_M^{(i)} - K)^+,$$

- *or an* **Asian call** *with floating strike:*

$$e^{-rT} \mathbb{E}(S_T - \frac{1}{T} \int_0^T S_u du)^+ \approx e^{-rT} \frac{1}{N} \sum_{i=1}^{N} (\hat{S}_M^{(i)} - \frac{1}{M} \sum_{j=1}^{M} \hat{S}_j)^+.$$

Notice that, once the sample paths have been generated, pricing Asian (or any path-dependent) option via Monte Carlo is not any more difficult than pricing the European one. Recall that this was not the case for the analytical methods, and it is one of the main advantages of Monte Carlo methods.

Accuracy

Observe that there are three sources of error in the estimate (30):

1. due to the use of $\tilde{F}$ instead of F;
2. due to the use of $\hat{X}$ instead of X (the distribution of $\hat{X}$ is different from the distribution of X, as $\hat{X}$ is only an approximation to the true solution X) – this is called a **discretization error**;
3. and due the computation of sample average instead of the exact expectation – this is a **Monte Carlo (or, simulation) error**.

The first error may be important if F is path-dependent. However, this error depends heavily on the function F itself, and it is hard to carry out any analysis for a general F. Therefore, in what follows, we will typically ignore this error and assume that $\tilde{F} = F$: i.e. for the sake of this analysis, we may assume that F depends only on the terminal value X_T).

Then, the total MC error can be estimated as follows:

$$\left| \mathbb{E}F(X_T) - \frac{1}{N} \sum_{i=1}^{N} F(\hat{X}_M^{(i)}) \right| \leq \left| \mathbb{E}F(X_T) - \mathbb{E}F(\hat{X}_M) \right| + \left| \mathbb{E}F(\hat{X}_M) - \frac{1}{N} \sum_{i=1}^{N} F(\hat{X}_M^{(i)}) \right|$$

The second term is the **MC error** and it has the usual asymptotic behavior:

$$\left|\mathbb{E}F(\hat{X}_M) - \frac{1}{N}\sum_{i=1}^{N} F(\hat{X}_M^{(i)})\right| = O\left(\frac{1}{\sqrt{N}}\right)$$

Assuming that F is Lipschitz-continuous, we estimate the first term:

$$\left|\mathbb{E}F(X_T) - F(\hat{X}_M)\right| \le \text{const}\,\mathbb{E}|X_T - \hat{X}_M| \le \text{const}\sqrt{\mathbb{E}(X_T - \hat{X}_M)^2},$$

and the term

$$\sqrt{\mathbb{E}(X_T - \hat{X}_M)^2}$$

is called the **discretization error**.

The analysis of the asymptotic behavior of the discretization error is more subtle. Recall that, in the Euler method for ODEs, the discretization error is $O(\Delta t)$. Let us estimate the squared error of a single step approximation in the present case.

Thm 4. *Under certain regularity assumptions on a and b, there exists a constant C, such that:*

$$\mathbb{E}(X_{t+\Delta t} - X_t - (\hat{X}_{t+\Delta t} - X_t))^2 \le C\Delta t^2$$

holds for all $0 \le t \le t + \Delta t \le T$.

Proof:

$$\mathbb{E}(X_{t+\Delta t} - X_t - (\hat{X}_{t+\Delta t} - X_t))^2$$

$$\le \left(\sqrt{\mathbb{E}\left(\int_t^{t+\Delta t} b(X_s,s)ds - b(X_t,t)\Delta t\right)^2} + \sqrt{\mathbb{E}\left(\int_t^{t+\Delta t} a(X_s,s)dB_s - a(X_t,t)(B_{t+\Delta t} - B_t)\right)^2}\right)^2 \quad (31)$$

Next, we need to estimate each term separately. Let us start with the first one. Notice that, due to Ito's lemma,

$$b(X_s,s) - b(X_t,t) = \int_t^s \left(\partial_t b(X_u,u) + b(X_u,u)\partial_x b(X_u,u) + \frac{1}{2}a^2(X_u,u)\partial_{xx}^2 b(X_u,u)\right) du + \int_t^s a(X_u,u)\partial_x b(X_u,u)dB_u$$

Then, assuming that a and the derivatives of b are bounded, we obtain:

$$\left(\sup_{s\in[t,t+\Delta t]}\left|\int_t^s \left(\partial_t b(X_u,u) + b(X_u,u)\partial_x b(X_u,u) + \frac{1}{2}a^2(X_u,u)\partial_{xx}^2 b(X_u,u)\right) du\right|\right)^2 \le C_1\Delta t^2$$

Then, using the Doob's martingale inequality and the basic property of stochastic integral, we obtain:

$$\mathbb{E}\left(\sup_{s\in[t,t+\Delta t]}\int_t^s a(X_u,u)\partial_x b(X_u,u)dB_u\right)^2 \le \mathbb{E}\int_t^{t+\Delta t} \left(a(X_u,u)\partial_x b(X_u,u)\right)^2 du \le C_1\Delta t \quad (32)$$

The two estimates above yield:

$$\mathbb{E}\left(\sup_{s\in[t,t+\Delta t]}|b(X_s,s) - b(X_t,t)|\right)^2 \le C_2\Delta t, \quad (33)$$

which, in turn, implies

$$\mathbb{E}\left(\int_t^{t+\Delta t} b(X_s,s)ds - b(X_t,t)\Delta t\right)^2 = \mathbb{E}\left(\int_t^{t+\Delta t}(b(X_s,s)-b(X_t,t))ds\right)^2$$

$$\leq \Delta t^2 \mathbb{E}\left(\sup_{s\in[t,t+\Delta t]} |b(X_s,s)-b(X_t,t)|\right)^2 \leq C_3 \Delta t^3$$

Next, we analyze the second term in the right hand side of (31). First, we notice that (33) holds with a in place of b. Then, using the basic property of stochastic integral, we obtain:

$$\mathbb{E}\left(\int_t^{t+\Delta t} a(X_s,s)dB_s - a(X_t,t)(B_{t+\Delta t}-B_t)\right)^2$$

$$= \mathbb{E}\left(\int_t^{t+\Delta t}(a(X_s,s)-a(X_t,t))dB_s\right)^2 = C_4 \mathbb{E}\int_t^{t+\Delta t}(a(s,X_s)-a(t,X_t))^2 ds$$

$$\leq C_4 \Delta t \mathbb{E}\left(\sup_{s\in[t,t+\Delta t]} |a(X_s,s)-a(X_t,t)|\right)^2 \leq C_5 \Delta t^2$$

Collecting the above, we obtain the desired estimate. ∎

The squared errors sum up. This is a consequence of the fact that, according to our definition, the errors made at each step are independent, hence the variance of the total error is a sum of the individual errors. Recall that the number of steps is equal to $M = T/\Delta t$. Therefore, for the explicit Euler scheme, we obtain:

$$\sqrt{\mathbb{E}(\hat{X}_M - X_T)^2} \leq C\Delta t^{1/2}$$

Def 3. *We say that a scheme is of* **strong order** α *if there exists a constant* C*, such that*

$$\sqrt{\mathbb{E}(\hat{X}_M - X_T)^2} \leq C\Delta t^{\alpha},$$

for all small enough Δt *(i.e. large enough* M*).*

The explicit Euler scheme is of **strong order** $1/2$.

We can see that the discretization error for the Euler scheme for SDEs is $O(\sqrt{\Delta t})$, whereas for the deterministic Euler method it is $O(\Delta t)$. The reason is that $|B_{t+\Delta t} - B_t| \simeq \sqrt{\Delta t}$ with high probability. Recall that $\Delta t = T/M$, to conclude that the total error is given by

$$O\left(\frac{1}{\sqrt{N}}\right) + O\left(\frac{1}{\sqrt{M}}\right)$$

Thus, **to achieve maximum computational efficiency, we need to take the number of MC simulations to be proportional to the size of the partition:** $N \sim M$.

Rem 3. *If* F*,* a *and* b *satisfy some additional smoothness conditions, we obtain*

$$\left|\mathbb{E}F(X_T) - \mathbb{E}F(\hat{X}_M)\right| \leq C\Delta t,$$

where $\hat{X}$ *is produced by the Euler scheme. Such scheme is called* **weak order one**. *However, the additional smoothness assumptions on* F *are not satisfied even for the European call and put options.*

Rem 4. *Using the estimates in the proof of Theorem 4, we can establish the rate of convergence*

$$\sqrt{\mathbb{E}(\sup_{t\in[0,T]}(\hat{X}_{[t/\Delta t]}-X_t)^2)} = O(\sqrt{\Delta t}),$$

where $[t/\Delta t]$ denotes the nearest integer to $t/\Delta t$ from below, and $t \mapsto \hat{X}_{[t/\Delta t]}$ represents the interpolated path, which remains constant between the times $m\Delta t$ and $(m+1)\Delta t$. Then, we can define

$$\tilde{F}(\hat{X}_0,\ldots,\hat{X}_M) = F((\hat{X}_{[t/\Delta t]})_{t\in[0,T]}),$$

and, under certain regularity assumptions on F, we can deduce the rate of convergence of $\mathbb{E}\tilde{F}(\hat{X}_0,\ldots,\hat{X}_M)$ to $\mathbb{E}F(X)$. Of course, these regularity assumptions (e.g. Lipschitz-continuity) may not always be easy to verify when F is a function on the entire path space.

4.2 Milstein's scheme

We can reduce the truncation error by improving the approximation

$$\int_t^{t+\Delta t} a(X_s,s)\,dB_s \;\approx\; a(X_t,t)[B_{t+\Delta t}-B_t]\,, \tag{34}$$

which is used in (28).

Clearly, the source of error in this approximation is in the fact that $a(X_s,s)$ does not stay constant over the time interval $s\in[t,t+\Delta t]$. In turn, a careful examination of the proof of Theorem 4 reveals that the main term in $|a(X_s,s)-a(X_t,t)|$ is due to the increments of the BM (recall (32)). Thus, we introduce a new approximation, which has better accuracy in the B-variable,

$$a(X_s,s) \;\approx\; a(X_t,t)+a(X_t,t)[X_s-X_t] \;\approx\; a(X_t,t)+a(X_t,t)\frac{\partial a(X_t,t)}{\partial x}[B_s-B_t]\,, \tag{35}$$

upon doing the Taylor expansion of $a(x,t)$ in x around $x = X_t$ and approximating the increments of X by the increments of BM.

Using (35), we have that

$$\begin{aligned} X_{t+\Delta t}-X_t \;&=\; \int_t^{t+\Delta t} dX_s \;\approx\; b(X_t,t)\Delta t+\int_t^{t+\Delta t} a(X_s,s)dB_s \\ &\approx\; b(X_t,t)\Delta t + a(X_t,t)[B_{t+\Delta t}-B_t] + a(X_t,t)\frac{\partial a(X_t,t)}{\partial x}\int_t^{t+\Delta t}[B_s-B_t]dB_s\,. \end{aligned} \tag{36}$$

Using the Ito calculus we have that

$$\int_t^{t+\Delta t}[B_s-B_t]dB_s \;=\; \frac{1}{2}[B_{t+\Delta t}-B_t]^2-\frac{\Delta t}{2}\,. \tag{37}$$

If we denote $B_{t+\Delta t}-B_t=\sqrt{\Delta t}\,\xi$, where ξ is standard normal, then the approximation (36) becomes:

$$\hat{X}_{m+1} \;=\; \hat{X}_m + b(\hat{X}_m,m\Delta t)\Delta t + a(\hat{X}_m,m\Delta t)\sqrt{\Delta t}\,\xi_m + a(\hat{X}_m,m\Delta t)\frac{\partial a(\hat{X}_m,m\Delta t)}{\partial x}\frac{\Delta t}{2}[\xi_m^2-1]\,. \tag{38}$$

Evidently (38) is a refinement of the basic Euler algorithm (29) and is known as Milstein's algorithm. Note that at each step of the algorithm we just need to generate a single standard normal variable to implement the algorithm, as is also the case with the Euler scheme.

Repeating the analysis in the proof of Theorem 4, we deduce that the **Milstein's scheme is of strong order** 1.

Ex 4. *We already observed in Chapter I that for geometric Brownian motion* (S_t) *which is a solution to the SDE*

$$dS_t \;=\; \mu S_t dt + \sigma S_t dB_t, \tag{39}$$

we have

$$S_{t+\Delta t} \;=\; S_t \exp[(\mu - \sigma^2/2)\Delta t + \sigma\sqrt{\Delta t}\,\xi]\,, \tag{40}$$

where ξ *is standard normal.*

We can expand the exponential in (40) in its series expansion

$$e^z \;=\; 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \cdots\,, \tag{41}$$

to obtain increasingly accurate approximations for $S_{t+\Delta t}$*.*

If we keep just the first 2 *terms in the expansion (41) we have from (40) that*

$$S_{t+\Delta t} \;\approx\; S_t + (\mu - \sigma^2/2)S_t\Delta t + \sigma S_t\sqrt{\Delta t}\,\xi, \tag{42}$$

which is **NOT the same** *as the Euler scheme:*

$$S_{t+\Delta t} \;\approx\; S_t + \mu S_t\Delta t + \sigma S_t\sqrt{\Delta t}\,\xi\;. \tag{43}$$

If we keep the first 3 *terms in the expansion (41) we have*

$$S_{t+\Delta t} \;=\; S_t + [(\mu - \sigma^2/2)S_t\Delta t + \sigma S_t\sqrt{\Delta t}\,\xi] + \frac{S_t}{2}[(\mu - \sigma^2/2)\Delta t + \sigma\sqrt{\Delta t}\,\xi]^2 + O[(\Delta t)^{3/2}] \tag{44}$$

We can rewrite (44) as

$$S_{t+\Delta t} \;\approx\; S_t + \mu S_t\Delta t + \sigma S_t\sqrt{\Delta t}\,\xi + \sigma^2 S_t\frac{\Delta t}{2}[\xi^2 - 1], \tag{45}$$

which gives us the **Milstein's scheme** *for 39).*

It is easy to see from the above Taylor's expansion why the algorithm (42) is incorrect, while the Milstein's scheme (45) does converge to the true solution. The reason is that the error in (44) is $O(\Delta t^{3/2})$*, which, after* $\approx 1/\Delta t$ *steps becomes* $O(\sqrt{\Delta t})$ *and vanishes as* $\Delta \to 0$*. On the other hand, the error term in (42) is* $O(\Delta t)$*, which, after* $\approx 1/\Delta t$ *steps becomes* $O(1)$*, and, in general, may not vanish as* $\Delta \to 0$*.*

It is not obvious a priori why the Euler scheme (43) converges to the true solution. The Taylor's expansion does not provide enough precision to see why the Euler scheme converges. Indeed, the above Taylor's expansion only gives us the same error estimate as in (42), that is $O(1)$*, which does not guarantee convergence.*

However, we can see that the Euler scheme (43) converges from the fact that it only differs from the Milstein's scheme (45) by the correction term which is proportional to $\Delta t(\xi^2 - 1)$*. Notice that this term has mean* 0 *and variance* Δt^2*. The sum of* $\approx 1/\Delta t$ *such i.i.d. variables will be will be* $\approx \sqrt{\Delta t}$ *by the CLT, and, hence, it vanishes as* $\Delta t \to 0$*.*

The reason that the **Taylor's expansion does not provide sufficiently precise error estimates** *is that it treats the increments in time and in the value of BM in the same way (combining them into "*z*"), despite the fact that* **these increments are of different order** *(*$O(\Delta t)$ *and* $O(\sqrt{\Delta t})$*). These considerations show that there are some extra subtleties involved with numerical algorithms for SDEs beyond those which occur in the case of ODEs.*

Rem 5. *The main downsides of the Milstein's scheme:*

- *It requires the* **computation of** $\partial_x a$*: this is not a problem if a and its derivative are known in closed form, but may increase the computational complexity significantly if one needs to evaluate $\partial_x a$ numerically.*
- *It* **does not work in multiple dimensions**.

4.3 Exact simulation of numerical solutions to SDEs

Motivation

In many cases, we have some a priori constraints on the values of the solution to an SDE. Then, it may be important (and sometimes necessary) to have numerical approximations that satisfy these constraints.

Consider, for example, the GBM

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

Using the Euler scheme, we obtain the approximation:

$$\hat{S}_{m+1} \;\approx\; \hat{S}_m + \mu \hat{S}_m \Delta t + \sigma \hat{S}_m \sqrt{\Delta t}\, \xi_m \;.$$

Notice that, for a fixed $\hat{S}_m$, ξ_m may take a value satisfying

$$\xi_m < \frac{\hat{S}_{m+1} - \hat{S}_m - r\hat{S}_m \Delta t}{\sigma \hat{S}_m \sqrt{\Delta t}},$$

so that $\hat{S}_{m+1} < 0$. This is a problem if one wants to think of $\hat{S}$ as a stock price (e.g. it may lead to arbitrage opportunities that do not actually exist).

In this case, the problem can be easily resolved by the **logarithmic transformation**:

$$X_t = \log(S_t), \quad dX_t = (\mu - \sigma^2/2)dt + \sigma dW_t,$$

$$\hat{X}_{m+1} \;\approx\; \hat{X}_m + (\mu - \sigma^2/2)\Delta t + \sigma\sqrt{\Delta t}\, \xi_m, \qquad \hat{S}_m = \exp(\hat{X}_m).$$

However, it is not always possible to find a suitable transformation. For example, to simulate the squared volatility in Heston model, we need to solve

$$dY_t = \kappa(\theta - Y_t)dt + \sigma\sqrt{Y_t} dW_t$$

The Euler scheme becomes

$$\hat{Y}_{m+1} \;\approx\; \hat{Y}_m + \kappa(\theta - \hat{Y}_m)\Delta t + \sigma\sqrt{\hat{Y}_m}\sqrt{\Delta t}\, \xi_m \;.$$

Recall that, if $2\kappa\theta \geq \sigma^2$ and $Y_0 > 0$, then the true solution Y remains positive. However, even under these conditions, if ξ_m becomes negative, with a large enough absolute value, then $\hat{Y}_{m+1} < 0$ and **it is not even clear how to make the next step!** And there is no simple transformation that will resolve this issue.

Exact simulation

The above problem can be resolved if we can simulate an approximation $\hat{X}$ which **has exactly the same distribution as the discretized true solution** X to the SDE

$$dX_t = b(X_t, t)dt + a(X_t, t)dW_t$$

In other words, we would like to have

$$\text{Law}(\hat{X}_0, \hat{X}_1, \dots, \hat{X}_M) = \text{Law}(X_0, X_{\Delta t}, \dots, X_{M\Delta t})$$

Since our simulation is given by a recursive relation, the above condition can be simplified to the following holds for all $m = 0, \dots, M-1$ and all $(x_0, \dots, x_{M-1})$:

$$\text{Law}(\hat{X}_{(m+1)} \,|\, \hat{X}_0 = x_0, \hat{X}_1 = x_1, \dots, \hat{X}_m = x_m) = \text{Law}(X_{(m+1)\Delta t} \,|\, X_0 = x_0, X_{\Delta t} = x_1, \dots, X_{m\Delta t} = x_m)$$

Then, **the distribution of the simulated vector of** $(\hat{X}_0, \hat{X}_1, \dots, \hat{X}_M)$ **is the same as the distribution of the corresponding vector of the true solution values,** $(X_0, X_{\Delta t}, \dots, X_T)$.

The benefits are:

- The a priori **constraints** on the values of the solutions **are satisfied by the approximation**.
- If the payoff function F depends only on the values of X at a finite number of time instances, there is **no error due to discretization**. Hence there is no need to increase M, beyond what the value that is needed to compute $F(S)$ precisely.

Rem 6. *Notice that the MC error always remains, even in the case of exact simulation. In addition, if F depends on the entire path, for a continuum of times $t \in [0, T]$, then the error due to replacing F by its discretized version $\tilde{F}$ should be taken into account.*

Ex 12. BS model*:*

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

Notice that the true solution satisfies:

$$S_{t+\Delta t} = S_t \exp\left((\mu - \sigma^2/2)\Delta t + \sigma(W_{t+\Delta t} - W_t)\right)$$

Thus, we choose $\Delta t = T/M$ and construct the **exact simulation** *scheme for S as follows:*

$$\hat{S}_{m+1} = \hat{S}_m \exp((r - \sigma^2/)\Delta t + \sigma\sqrt{\Delta t}\xi_m), \quad m = 0, \dots, M-1,$$

where $\{\xi_m\}$ are i.i.d. $N(0,1)$. In this algorithm the conditional distribution of every next value of the approximation (given all the past values) is exactly the same as the conditional distribution of the true solution. Thus, the distribution of $(\hat{S}_0, \hat{S}_1, \dots, \hat{S}_M)$ is the same as the distribution of the true solution values $(S_0, S_{\Delta t}, \dots, S_T)$.

Assume that we are given a function $F(S)$, which depends only on the values of S_t at $t = m\Delta t$:

$$F(S) = F(S_0, S_{\Delta t}, \dots, S_{M\Delta t}),$$

and that we need to compute $\mathbb{E}F(S)$. Then, we use the exact simulation to generate a sample of size N,

$$\left\{\hat{S}_0^{(i)}, \dots, \hat{S}_M^{(i)}\right\}_{i=1}^N,$$

and use the approximation

$$\mathbb{E}F(S) \approx \frac{1}{N}\sum_{i=1}^N F(\hat{S}_0^{(i)}, \hat{S}_{\Delta t}^{(i)}, \dots, \hat{S}_{M\Delta t}^{(i)})$$

The important feature of this approximation is that its **error is only due to the MC but not to the discretization**.

A **lookback option** *has payoff which is a function of S_T and $\sup_{t\in[0,T]} S_t$ and/or $\inf_{t\in[0,T]} S_t$. For example, a lookback put pays*

$$F(S) = (\sup_{t\in[0,T]} S_t - S_T)^+$$

at maturity T.

Notice that F depends on the entire path $(S_t)_{t\in[0,T]}$. However, there exist so called **discretely monitored** *lookback options. For example, for a chosen partition $\{0, \Delta t, \ldots, M\Delta t = T\}$, the discretely monitored lookback put pays*

$$\tilde{F}(S) = \tilde{F}(S_0, S_{\Delta t}, \ldots, S_{M\Delta t}) = (\sup_{m=0,\ldots,M} S_{m\Delta t} - S_{M\Delta t})^+$$

Thus, if we can simulate exactly from the distribution $(S_0, S_{\Delta t}, \ldots, S_{M\Delta t})$, we can price the discretely monitored lookback with no discretization error (i.e. making only the MC error):

$$\mathbb{E}\tilde{F}(S) \approx \frac{1}{N}\sum_{i=1}^{N} \tilde{F}(\hat{S}_0^{(i)}, \hat{S}_{\Delta t}^{(i)}, \ldots, \hat{S}_{M\Delta t}^{(i)})$$

Ex 13. Vasicek model*:*

$$dr_t = \kappa(\theta - r_t)dt + \sigma dW_t$$

It turns out that we can solve this equation explicitly:

$$r_t = \theta + (r_0 - \theta)e^{-\kappa t} + \sigma \int_0^t e^{-\kappa(t-u)} dW_u$$

Exercise 3. *Show that the above process satisfies the SDE by applying Ito's formula.*

Since the SDE is time-homogeneous (its coefficients do not depend on time), we can find the solution starting from any time $s < t$:

$$r_t = \theta + (r_s - \theta)e^{-\kappa(t-s)} + \sigma \int_s^t e^{-\kappa(t-u)} dW_u$$

Notice that

$$\int_s^t e^{-\kappa(t-u)} dW_u \sim N(0, \frac{1}{2\kappa}(1 - e^{-2\kappa(t-s)})),$$

and the integrals over disjoint sets are independent.

Thus, we obtain the following exact simulation scheme:

$$\hat{r}_{m+1} = \theta + (\hat{r}_m - \theta)e^{-\kappa\Delta t} + \sigma\xi_m,$$

where $\{\xi_m\}$ are i.i.d. $N(0, \frac{1}{2\kappa}(1 - e^{-2\kappa\Delta t}))$.

Notice that the above scheme allows us to simulate exactly from the paths of r evaluated at finite number of points. However, we still have a discretization error, for example, if we want to simulate the integral of a path:

$$\int_0^T r_u du,$$

which is, for example, needed to price a (zero-coupon) bond:

$$P(0,T) = \mathbb{E}e^{-\int_0^T r_u du}$$

Of course, this issue is resolved if we assume discrete compounding.

Rem 7. *In fact, since (r_t) is a Gaussian process, its integral in normally distributed:*

$$\int_0^T r_u du \sim N(\mu, \sigma^2),$$

where $\mu = \int_0^T \mathbb{E}r_u du = \int_0^T (\theta + (r_0 - \theta)e^{-\kappa u})du$ and σ^2 can be computed similarly, for the explicit representation of r_t.

Exercise 4. *Compute $\sigma^2 = Var\left(\int_0^T r_u du\right)$.*

Ex 14. CIR model*:*

$$dr_t = \kappa(\theta - r_t)dt + \sigma\sqrt{r_t}dW_t$$

It turns out that, if $2\kappa\theta \geq \sigma^2$ and $r_0 > 0$,

$$Law(r_t \,|\, r_s) \sim \sigma^2 \frac{1 - e^{-\kappa(t-s)}}{4\kappa}\, \chi^2_\nu(\lambda(t-s, r_s)),$$

$$\nu = \frac{4\kappa\theta}{\sigma^2}, \qquad \lambda(t-s, r_s) = \frac{4\kappa e^{-\kappa(t-s)}}{\sigma^2(1 - e^{-\kappa(t-s)})} r_s,$$

where $\chi^2_\nu(\lambda)$ is a **non-central chi-square distribution** *with ν degrees of freedom and noncentrality parameter λ. It can be represented as*

$$\chi^2_\nu(\lambda) \sim (Z + \sqrt{\lambda})^2 + \xi,$$

where $Z \sim N(0,1)$ is independent of $\xi \sim \chi^2_{\nu-1}$.

Thus, the exact simulation scheme becomes:

$$\hat{r}_{m+1} \sim \sigma^2 \frac{1 - e^{-\kappa\Delta t}}{4\kappa}\, \chi^2_\nu(\lambda(\Delta t, \hat{r}_m)).$$

Recall that, if $2\kappa\theta \geq \sigma^2$ and $r_0 > 0$, the solution remains positive. As discussed earlier, at each step, there is always a positive probability that the Euler approximation becomes negative. To ensure positivity, we need to use the exact simulation.

In addition, the Euler scheme applied to the CIR equation, in general, will **not** *have the predicted accuracy (of order $O(\Delta t^{1/2})$), because the diffusion coefficient in this equation is not Lipschitz-continuous (i.e. the derivative of a square root explodes at the origin).*

Rem 8. If we can simulate X exactly*, then we only need to use a time discretization of the SDE when the payoff function $F(S)$ depends on the path of S (i.e. there is no need to discretize time if $F(S) = F(S_T)$).*

If $F(X)$ depends the values of $(X_t)_{t\in[0,T]}$ at M different times, with M being fixed but large, then, of course, we need to simulate a path of X at the desired partition points: $(X_0, \ldots, X_{\Delta tM})$. And it makes sense to compare *the exact simulation to the Euler scheme in this case. Asymptotically, the exact simulation (when it is available) is always more efficient than the Euler scheme, since it does not require increasing M, but only N – hence, to achieve the precision of $1/\sqrt{N}$, the number of required computations is $O(N)$ (as opposed to $O(N^2)$, in the case of Euler scheme).*

However, every single step of the exact scheme may be much more expensive than a step of Euler scheme: generating from the true distribution of X_t may be more difficult than generating a Gaussian r.v.. Therefore, if $F(X)$ depends the values of $(X_t)_{t\in[0,T]}$ at M different times, with M being (fixed but) large, then, for a fixed N comparable to M, the exact simulation may be more expensive than the Euler scheme.

4.4 Multi-dimensional SDEs

The explicit Euler scheme extends to the multi-dimensional SDEs without any changes. Namely, consider

$$dX_t = b(X_t,t)dt + a(X_t,t)dB_t, \tag{46}$$

where X_t takes values in $\mathbb{R}^d$, b is a vector if functions $b : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$, a is a matrix of functions, $a : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d\times n}$, and B is n-dimensional **standard Brownian motion** (i.e. its entries are independent Brownian motions).

In other words,

$$X_t = (X^1,\dots,X^d_t)^T, \quad b(X_t,t) = (b^1(X_t,t),\dots,b^d(X_t,t))^T, \quad a(X_t,t) = (a^{ij}(X_t,t))_{i=1,\dots,d,\, j=1,\dots,n},$$

and the SDE can written as

$$\begin{pmatrix} dX^1_t \\ \vdots \\ dX^d_t \end{pmatrix} = \begin{pmatrix} b^1(X_t,t) \\ \vdots \\ b^d(X_t,t) \end{pmatrix} dt + \begin{pmatrix} a^{11}(X_t,t) & \cdots & a^{1n}(X_t,t) \\ \vdots & \vdots & \vdots \\ a^{d1}(X_t,t) & \cdots & a^{dn}(X_t,t) \end{pmatrix} \begin{pmatrix} dB^1_t \\ \vdots \\ dB^n_t \end{pmatrix}$$

The above form of a multivariate SDE is canonical. Any system of SDEs

$$dX^i_t = b^i(X_t,t)dt + \tilde{a}^i(X_t,t)dW^i_t, \quad < dW^i, dW^j >= \rho^{ij}dt, \quad i,j = 1,\dots,d, \tag{47}$$

can be transformed into an equivalent canonical SDE of the above form.

To transform (47) into the canonical form, proceed as follows

- Construct the matrix Σ, with $\Sigma^{ij} = \rho^{ij}$, for $i,j = 1,\dots,d$.
- Compute the decomposition: $\Sigma = AA^T$ (e.g. we can use the Cholesky decomposition).
- Finally, define $a = \text{diag}(\tilde{a}^1,\dots,\tilde{a}^d)A$, or, in other words,

$$a^{ik} = \tilde{a}^i A^{ik}, \quad i,k = 1,\dots,d.$$

Then, there exists a standard d-dimensional BM B (i.e. $n = d$), such that $W = AB$ and, hence,

$$dX^i_t = b^i(X_t,t)dt + \sum_{k=1}^d a^{ik}(X_t,t)dB^k_t, \quad i = 1,\dots,d,$$

where $\{B^k\}$ are independent BMs.

The **Euler scheme** for (46) becomes

$$\hat{X}_{m+1} = \hat{X}_m + b(\hat{X}_m, m\Delta t)\Delta t + \sqrt{\Delta t} a(\hat{X}_m, m\Delta t)\xi_m,$$

where ξ is a d-dimensional **vector of standard normals**: $\xi_m = (\xi_m^1, \ldots, \xi_m^d)$ and all $\left\{\xi_m^i\right\}$ are i.i.d. $N(0,1)$.

Ex 5. *Heston model.*

$$\begin{cases} dS_t = \mu S_t dt + \sqrt{Y_t} S_t dW_t^1 \\ dY_t = \kappa(\theta - Y_t)dt + \sigma\sqrt{Y_t} dW_t^2, \\ < dW_t^1, dW_t^2 >= \rho dt \end{cases}$$

First, we define

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

and deduce $\Sigma = AA^T$, with

$$A = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix}$$

This means that we have the representation:

$$W_t^1 = B_t^1, \qquad W_t^2 = \rho B_t^1 + \sqrt{1-\rho^2} B_t^2,$$

where B^1 and B^2 are independent BMs. Then, the SDE becomes

$$\begin{cases} dS_t = \mu S_t dt + \sqrt{Y_t} S_t dB_t^1 \\ dY_t = \kappa(\theta - Y_t)dt + \sigma\sqrt{Y_t}(\rho dB_t^1 + \sqrt{1-\rho^2} dB_t^2), \end{cases}$$

where B^1 and B^2 are independent BMs. The above system of equations can be solved uniquely for $t > 0$ with given initial conditions (S_0, Y_0).

To approximate the solution using MC methods, we choose an integer M so that $M\Delta t = T$ and define $\hat{Y}_m, \hat{S}_m,\ m = 0, .., M$ by the recurrence

$$\hat{S}_{m+1} = \hat{S}_m + \hat{S}_m \left[\mu\Delta t + \sqrt{\hat{Y}_m}\sqrt{\Delta t}\,\{\rho\,\xi_m + \sqrt{1-\rho^2}\,\eta_m\}\right], \tag{48}$$

$$\hat{Y}_{m+1} = \hat{Y}_m + \kappa(\theta - \hat{Y}_m)\Delta t + \sigma\sqrt{\hat{Y}_m}\sqrt{\Delta t}\,\xi_m\ , \tag{49}$$

where $\{\xi_m,\ \eta_m\}_{m=0,\ldots,M-1}$ are i.i.d. $N(0,1)$. We then generate $N >> 1$ independent sample paths of $(\hat{S}, \hat{Y})$ to estimate the expectation of a given function of $(S_t, Y_t)_{t\in[0,T]}$ by the sample average.

If $2\kappa\theta \geq \sigma^2$ and $Y_0 > 0$, then the true solution Y remains positive at all times. We might want to (or even need to, for the scheme to work) **ensure that $\hat{Y}$ remains positive** *as well. We have seen that the exact simulation of Y allows us to achieve precisely this. However, in this case, the process Y has to be simulated together with S, due to the presence of correlation ρ (which, for example, represents the leverage effect). In the case of Euler scheme, this correlation is captured automatically by the form of the random innovations in (48)–(49): notice that they are Gaussian with the given correlation. However, it is not immediately clear how to ensure the desired correlation between the increments of $\hat{Y}$ and $\hat{S}$ in the case when $\hat{Y}$ is simulated exactly.*

To resolve this problem, we first re-write the SDE as

$$\begin{cases} dS_t = \mu S_t dt + \sqrt{Y_t} S_t(\rho dB_t^1 + \sqrt{1-\rho^2} dB_t^2) \\ dY_t = \kappa(\theta - Y_t)dt + \sigma\sqrt{Y_t} dB_t^1 \end{cases}$$

Next, we notice that

$$\sqrt{Y_t} dB_t^1 = \frac{1}{\sigma} dY_t - \frac{\kappa}{\sigma}(\theta - Y_t)dt,$$

and re-write the system again:

$$\begin{cases} dS_t = \left(\mu S_t - S_t \frac{\rho\kappa}{\sigma}(\theta - Y_t)\right) dt + S_t \frac{\rho}{\sigma} dY_t + S_t \sqrt{Y_t}\sqrt{1-\rho^2} dB_t^2 \\ dY_t = \kappa(\theta - Y_t)dt + \sigma\sqrt{Y_t} dB_t^1 \end{cases}$$

The main feature of the above representation is that S **is expressed through Y and B^2, which are independent***.*

Finally, we combine the exact simulation and the Euler scheme:

- *First, we compute $\hat{Y}$ via*

  $$\hat{Y}_{m+1} \sim \sigma^2 \frac{1 - e^{-\kappa\Delta t}}{4\kappa} \chi_\nu^2(\lambda(\Delta t, \hat{Y}_m)), \quad m = 0, \dots, M-1,$$

  where

  $$\nu = \frac{4\kappa\theta}{\sigma^2}, \qquad \lambda(\Delta t, \hat{Y}_m) = \frac{4\kappa e^{-\kappa\Delta t}}{\sigma^2(1 - e^{-\kappa\Delta t})} \hat{Y}_m$$

- *Then, having a path of $\hat{Y}$, we implement the usual Euler scheme for $\hat{S}$:*

  $$\hat{S}_{m+1} = \hat{S}_m + \hat{S}_m \left(\mu - \frac{\rho\kappa}{\sigma}(\theta - \hat{Y}_m)\right)\Delta t + \hat{S}_m \frac{\rho}{\sigma}(\hat{Y}_{m+1} - \hat{Y}_m) + \hat{S}_m \sqrt{\hat{Y}_m}\sqrt{1-\rho^2}\sqrt{\Delta t}\eta_m,$$

  where $\{\eta_m\}$ are i.i.d. standard normals, **independent** *of $\left\{\hat{Y}_m\right\}$.*

If $2\kappa\theta \geq \sigma^2$ and $Y_0 > 0$, the approximation $\hat{Y}$ remains positive. Besides, this "semi-exact" scheme is expected to converge faster than the pure Euler scheme.

Ex 6. *Multivariate BS model:*

$$dS_t/S_t = \mu dt + \sigma dB_t,$$

or, more precisely,

$$\begin{pmatrix} dS_t^1/S_t^1 \\ \vdots \\ dS_t^d/S_t^d \end{pmatrix} = \begin{pmatrix} \mu^1 \\ \vdots \\ \mu^d \end{pmatrix} dt + \begin{pmatrix} \sigma^{11} & \cdots & \sigma^{1n} \\ \vdots & \vdots & \vdots \\ \sigma^{d1} & \cdots & \sigma^{dn} \end{pmatrix} \begin{pmatrix} dB_t^1 \\ \vdots \\ dB_t^n, \end{pmatrix}$$

where μ^i and σ^{ij} are constants, and $\left\{B^i\right\}$ are independent BMs.

Every S^i is a GBM with drift μ^i and volatility

$$\tilde{\sigma}^i = \sqrt{\sum_{j=1}^n (\sigma^{ij})^2}$$

The exact simulation scheme for the solution becomes:

$$\hat{S}^i_{m+1} = \hat{S}^i_m \exp\left((\mu^i - (\tilde{\sigma}^i)^2/2)\Delta t + \sqrt{\Delta t} \sum_{k=1}^{n} \sigma^{ik} \xi^k_m \right), \quad i = 1, \dots, d,$$

where $\left\{ \xi^k_m \right\}_{k=1,\dots,n,\, m=0,\dots,M-1}$ *are i.i.d.* $N(0,1)$*.*

Once the scheme is constructed, we can generate a sample of size $N >> 1$*:*

$$\left\{ \hat{S}^{i,(j)}_0, \dots, \hat{S}^{i,(j)}_M \right\}_{i=1,\dots,d,\, j=1,\dots,N}$$

and use it, for example, to estimate the distribution of the cumulative loss process in the Black-Cox model:

$$\tau^i = \inf\left\{ t \geq 0 \, : \, S^i_t \leq H^i_t \right\}, \qquad L_t = \frac{1}{d} \sum_{i=1}^{d} \mathbf{1}_{\tau^i \leq t},$$

using the approximations

$$\hat{\tau}^{i,(j)} = \inf\left\{ m\Delta t \, : \, \hat{S}^{i,(j)}_m \leq H^i_{m\Delta t} \right\}, \qquad \hat{L}^{(j)}_t = \frac{1}{d} \sum_{i=1}^{d} \mathbf{1}_{\hat{\tau}^{i,(j)} \leq t}$$

For example, we can estimate the expected payoff of a (unit face value) CDO tranche:

$$\mathbb{E}\left(1 - \frac{1}{K_2 - K_1} \min\left((L_T - K_1)^+, K_2 - K_1 \right) \right) \approx \frac{1}{N} \sum_{j=1}^{N} \left(1 - \frac{1}{K_2 - K_1} \min\left((\hat{L}^{(j)}_T - K_1)^+, K_2 - K_1 \right) \right)$$

5 Variance Reduction methods

In the basic MC method we are interested in estimating the expectation of a random variable X. As we have seen, the error of the approximation is given by the sample mean

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i,$$

where $\{X_i\}$ are i.i.d. with distribution Law(X). Recall that $\bar{X}_N$ is **unbiased**,

$$\mathbb{E}\bar{X}_N = \mathbb{E}X,$$

and **consistent**,

$$\lim_{N \to \infty} \bar{X}_N = \mathbb{E}X,$$

as follows from LLN. In addition, CLT implies that

$$\|\mathbb{E}X - \frac{1}{N} \sum_{i=1}^{N} X_i\| \leq \frac{C}{\sqrt{N}},$$

and the constant C is proportional to the standard deviation (square root of the variance) of the sample. The standard deviation of the sample mean converges to σ, the standard deviation of X. Recall the confidence interval for the sample mean estimate

$$\mathbb{P}\left(\left|\mathbb{E}X - \frac{1}{N}\sum_{i=1}^{N} X_i\right| \geq \frac{3\sigma}{\sqrt{N}}\right) \leq 0.003$$

In this section, we will study the methods for decreasing the constant C. More precisely, we will show how to modify the sample and the estimate to ensure that the new estimate has smaller variance (while it keeps the consistency property and, in some cases, remains unbiased). As we typically have $X = F(\xi)$, with some function F and random variable (or vector) ξ, there are two general approaches to decrease the variance:

- Modify function F,
- or modify the input sample $\{\xi_i\}$.

The construction of a modified sample, typically, carries some additional computational cost. When possible, we will account for this extra cost.

In this section we will discuss several methods of variance reduction which have applications tin Finance. These methods are:

- Antithetic variates.
- Control variates.
- Stratified sampling.
- Moments matching.
- Importance sampling.

5.1 Antithetic variates

Recall that we need to estimate $\mathbb{E}X$ using MC simulations.

A Random variable $\tilde{X}$ is an **antithetic variate** (for X) if it is negatively correlated with X and has the same distribution.

Assume that we can generate a sequence $(X_1, \tilde{X}_1), \ldots, (X_N, \tilde{X}_N)$, such that the pairs $(X_1, \tilde{X}_1), \ldots, (X_N, \tilde{X}_N)$ are i.i.d. Then, we can consider the antithetic variate estimate:

$$\mathbb{E}X \approx \bar{X}_N^a = \frac{1}{2N}\left(\sum_{i=1}^{N} X_i + \sum_{i=1}^{N} \tilde{X}_i\right)$$

This estimator is clearly **unbiased**:

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

In fact, it can be viewed as a sample mean for the i.i.d. variables $(X_1 + \tilde{X}_1)/2, \ldots, (X_N + \tilde{X}_N)/2$:

$$\bar{X}_N^a = \frac{1}{N}\sum_{i=1}^{N} \frac{X_i + \tilde{X}_i}{2}$$

Therefore, it is also **consistent**:

$$\lim_{N\to\infty} \mathbb{E}\bar{X}_N^a = \mathbb{E}X,$$

and the CLT implies

$$\bar{X}_N^a - \mathbb{E}X \sim \frac{\sqrt{\text{Var}(\frac{X_i+\tilde{X}_i}{2})}}{\sqrt{N}}\xi, \qquad \xi \sim N(0,1).$$

Assume that the **computational effort to generate $\tilde{X}_i$ is roughly the same as the one required to generate X_i**. Then, we can use the same computational effort to produce a sample $X_1, \ldots, X_{2N}$ of i.i.d. random variables, with the same distribution as X, and use the standard estimate:

$$\bar{X}_{2N} = \frac{1}{2N}\sum_{i=1}^{N} X_i,$$

which is also **unbiased**, **consistent** and satisfies

$$\bar{X}_{2N} - \mathbb{E}X \sim \frac{\sqrt{\text{Var}(X_i)}}{\sqrt{2N}}\xi, \qquad \xi \sim N(0,1).$$

The **antithetic variates estimate, $\bar{X}^a$, is advantageous if it has smaller variance than the alternative (standard) estimate**, which is is the case if and only if

$$\text{Var}\left(\frac{X_i + \tilde{X}_i}{2}\right) < \frac{1}{2}\text{Var}(X_i),$$

which, in turn, is equivalent to

$$\text{cov}(X_i, \tilde{X}_i) < 0$$

Let us summarize the properties that are required from $\tilde{X}$.

- Law$(\tilde{X}) = $ Law(X).
- cor$(X, \tilde{X}) < 0$.
- Simulating $\tilde{X}$ should not be much more expensive than simulating X.

Q 4. *How do we choose the appropriate $\tilde{X}$?*

This may not always be easy, but it is possible in the following (quite typical) setting.

- Assume that $X = F(\xi^1, \ldots, \xi^d)$ – i.e. to generate the output X, we need to generate the input vector $\xi = (\xi^1, \ldots, \xi^d)$.

- Assume also that the function F is **monotone** in each variable.
- Then, we need to generate random vector $\tilde{\xi} = (\tilde{\xi}^1, \dots, \tilde{\xi}^d)$ which has the same distribution as ξ, and has the property that, **when ξ^j takes larger values, $\tilde{\xi}^j$ tends to take smaller values**. Of course, the question is how to choose the desired $\tilde{\xi}^j$. In the case where ξ^j takes values in $\mathbb{R}$, we can take $\tilde{\xi}^j = -\xi$. If ξ^j takes values in $[0, 1]$, we can take $\tilde{\xi}^j = 1 - \xi$.
- Finally, setting $\tilde{X} = F(\tilde{\xi})$, we expect that $\operatorname{cov}(X, \tilde{X}) < 0$, and, hence, the antithetic variates estimator
$$\mathbb{E}F(\xi) \approx \frac{1}{2N}\left(\sum_{i=1}^{N} F(\xi_i) + \sum_{i=1}^{N} F(\tilde{\xi}_i)\right)$$
reduces the variance.

The above heuristic argument can be made precise, in the case of one-dimensional ξ, with symmetric distribution.

Thm 5. *Assume that the function $F : \mathbb{R} \to \mathbb{R}$ is monotone, and ξ is a random variable with values in $\mathbb{R}$, such that Law$(\xi) =$ Law$(-\xi)$. Then* $\operatorname{cov}[F(\xi), F(-\xi)] \leq 0$.

Proof:
Assume F is increasing and let ξ' be independent of ξ and have the same distribution. Then it is easy to see that

$$(F(\xi) - F(-\xi'))(F(-\xi) - F(\xi')) \leq 0$$

Therefore,
$$\mathbb{E}\left[(F(\xi) - F(-\xi'))(F(-\xi) - F(\xi'))\right] \leq 0 \tag{50}$$

This implies that
$$0 \geq 2\mathbb{E}(F(\xi)F(-\xi)) - 2\mathbb{E}(F(\xi)\mathbb{E}F(-\xi)),$$

and, therefore,
$$\operatorname{cov}(F(\xi), F(-\xi)) = \mathbb{E}(F(\xi)F(-\xi)) - \mathbb{E}(F(\xi)\mathbb{E}F(-\xi)) \leq 0.$$

■

Ex 7. *Consider the problem of pricing a European put option in BS model:*

$$F(\xi) \; = \; e^{-rT}\left[K - S_0 \exp\left\{(r - \sigma^2/2)T + \sigma\sqrt{T}\xi\right\}\right]^+ \tag{51}$$

Notice that $F(\xi)$ is a decreasing function of $\xi \in \mathbb{R}$ and Law$(-\xi) =$ Law(ξ), hence it makes sense to use the antithetic variable method here:
$$\mathbb{E}F(\xi) \approx \frac{1}{2N}\left(\sum_{i=1}^{N} F(\xi_i) + \sum_{i=1}^{N} F(-\xi_i)\right),$$

for i.i.d. standard normal $\{\xi_i\}_{i=1}^N$.

5.2 Control variates

Recall that we need to estimate $\mathbb{E}X$. A random variable $\tilde{X}$ is a **control variate** for X if it has a strong (positive or negative) correlation with X.

Assume that $\mathbb{E}\tilde{X}$ is known and that we have simulated the i.i.d. pairs $\left\{(X_i, \tilde{X}_i)\right\}$, with $\text{Law}(X_i) = \text{Law}(X)$ and $\text{Law}(\tilde{X}_i) = \text{Law}(\tilde{X})$. Then, the **control variates estimate** is given by

$$\mathbb{E}X = \mathbb{E}(X - b(\tilde{X} - \mathbb{E}\tilde{X})) \approx \bar{X}_N^b = \frac{1}{N}\sum_{i=1}^{N}(X_i - b(\tilde{X}_i - \mathbb{E}\tilde{X})),$$

where the constant b is chosen to minimize the variance:

$$\text{Var}(X - b(\tilde{X} - \mathbb{E}\tilde{X})) = \text{Var}(X) - 2b\text{cov}(X, \tilde{X}) + b^2\text{Var}(\tilde{X})$$

Thus, the optimal value of b is given by

$$b^* = \frac{\text{cov}(X, \tilde{X})}{\text{Var}(\tilde{X})},$$

provided we know $\text{cov}(X, \tilde{X})$ and $\text{Var}(\tilde{X})$. With the above choice of b^*, the variance is improved by the factor

$$\frac{\text{Var}(X - b^*(\tilde{X} - \mathbb{E}\tilde{X}))}{\text{Var}(X)} = 1 - (\text{cor}(X, \tilde{X}))^2$$

Assume that, given a sample $\{X_i\}$, there is almost no additional cost in simulating $\left\{\tilde{X}_i\right\}$. For example, this is the case if $X = F(\xi)$ and $\tilde{X} = \tilde{F}(\xi)$, with the same input ξ and function F that is cheap to evaluate. Then, the **control variate estimator is advantageous (i.e. it reduces variance) whenever** $|\textbf{cor}(X, \tilde{X})| > 0$.

Q 5. *What if any, or all, of $\mathbb{E}\tilde{X}$, cov$(X, \tilde{X})$ and Var$(\tilde{X})$ are not known explicitly?*

A natural solution is to replace these quantities by the sample averages:

$$\mathbb{E}\tilde{X} \approx \bar{\tilde{X}}_N = \frac{1}{N}\sum_{i=1}^{N}\tilde{X}_i, \quad \text{cov}(X, \tilde{X}) \approx \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}_N)(\tilde{X}_i - \bar{\tilde{X}}_N) \quad \text{Var}(\tilde{X}) \approx \frac{1}{N}\sum_{i=1}^{N}(\tilde{X}_i - \bar{\tilde{X}}_N)^2$$

Then b^* can be approximated by

$$\hat{b}_N^* = \frac{\sum_{i=1}^{N}(X_i - \bar{X}_N)(\tilde{X}_i - \bar{\tilde{X}}_N)}{\sum_{i=1}^{N}(\tilde{X}_i - \bar{\tilde{X}}_N)^2}, \tag{52}$$

and the control variate estimator becomes:

$$\mathbb{E}X \approx \bar{X}_N^{\hat{b}^*} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{b}_N^*(\tilde{X}_i - \bar{\tilde{X}}_N)) \tag{53}$$

The **control variate estimator with $\bar{X}_N^{\hat{b}^*}$ has the same asymptotic variance** as the optimal one (i.e. $\bar{X}_N^{b^*}$), as $N \to \infty$. In addiiton, it is **consistent**. However, for a fixed N, the estimator $\bar{X}_N^{\hat{b}^*}$ has certain drawbacks.

- $\bar{X}_N^{\hat{b}^*}$ may be **biased**.
- $\bar{X}_N^{\hat{b}^*}$ may have **higher variance than theoretically predicted**, due to the need to estimate the moments of $\tilde{X}$ and the covariance of X and $\tilde{X}$. As a result, if N is not large enough and if we don't know $\mathbb{E}\tilde{X}$ and Var$(\tilde{X})$ the actual variance of this estimator may not be smaller than the variance of the regular sample average. Hence, although it is hard to expect that we know cov$(X, \tilde{X})$ we typically want to know as much as possible about the moments of $\tilde{X}$.
- $\bar{X}_N^{\hat{b}^*}$ requires additional computational effort, which could be spent on generating a larger sample instead. Hence, the actual variance reduction should be large enough, to justify the extra computational cost.

Unfortunately, there is no general way to quantify the above practical shortcomings of the control variates method. Hence, we have to rely on our intuition and experience, and try different methods before we choose one.

Q 6. *How do we choose the appropriate $\tilde{X}$?*

Typically, we have $X = F(\xi)$, and we attempt to choose $\tilde{X}$ so that the above shortcomings (as well as the problem of low extra cost for simulating $\tilde{X}$) are addressed. Namely,

- we choose $\tilde{X} = \tilde{F}(\xi)$ (to make it cheap to simulate $\tilde{X}$),
- so that there is a **strong (positive or negative) correlation** between X and $\tilde{X}$ (and, hence, the variance reduction would be large enough to justify the above inefficiencies),
- and so that we know as much as possible about the **moments of** $\tilde{X}$ (to reduce the extra variance coming from the estimates of $\mathbb{E}\tilde{X}$ and Var$(\tilde{X})$).

Ex 8. *Using the* **underlying as a control variate**. *Assume we need to price a European call option written on a tradable asset:*

$$X = F(S_T) = e^{-rT}(S_T - K)^+,$$

where we assume that the interest rate r is constant.

Clearly the payoff of a call option is highly correlated with the terminal asset price. Therefore, it is natural to choose S_T as a control variate.

Then $\tilde{X} = S_T$ and we have

$$\mathbb{E}\left(e^{-rT}(S_T - K)^+\right) = \mathbb{E}X = \mathbb{E}(X - b^*(S_T - \mathbb{E}S_T)) = \mathbb{E}(X - b^*(S_T - S_0 e^{rT}),$$

and

$$b^* = \frac{cov(X, S_T)}{Var(S_T)}$$

To price the option,

- *we generate i.i.d. $\left\{S_T^{(i)}\right\}_{i=1}^N$ (e.g. in the case of BS model, this amounts to sampling from standard normal distribution),*
- *compute $X^{(i)} = F(X^{(i)}) = e^{-rT}(S_T^{(i)} - K)^+$,*

- *define*

$$\hat{b}^* = \frac{\sum_{i=1}^N (X^{(i)} - \bar{X})(S_T^{(i)} - S_0 e^{rT})}{\sum_{i=1}^N (S_T^{(i)} - S_0 e^{rT})^2},$$

- *and compute the approximation*

$$\mathbb{E}X = \mathbb{E}\left(e^{-rT}(S_T - K)^+\right) \approx \frac{1}{N}\sum_{i=1}^N (X^{(i)} - \hat{b}^*(S_T^{(i)} - S_0 e^{rT})$$

Notice that this trick works in any model – not necessarily BS. However, the problem with this method is that, for large K, the correlation between X and S_T becomes very small.

Ex 9. *One can use the* **option's payoff in a simple model as a control variate for estimating the price of the same option in a more complicated model**.

Consider for example a European call option in **Heston model***:*

$$X = F(S_T) = e^{-rT}(S_T - K)^+,$$

$$\begin{cases} dS_t = rS_t dt + \sqrt{Y_t} S_t dB_t^1 \\ dY_t = \kappa(\theta - Y_t)dt + \sigma\sqrt{Y_t}(\rho dB_t^1 + \sqrt{1-\rho^2} dB_t^2) \end{cases}$$

On the other hand, we can consider the same option in a BS model with time-dependent (deterministic) volatility:

$$\tilde{X} = F(\tilde{S}_T) = e^{-rT}(\tilde{S}_T - K)^+,$$

$$d\tilde{S}_t = r\tilde{S}_t dt + \sqrt{y(t)}\tilde{S}_t dB_t^1,$$

where $y(t) = \mathbb{E}Y_t$

Notice that

$$\mathbb{E}Y_t = Y_0 + \int_0^t \kappa(\theta - \mathbb{E}Y_u)du,$$

hence,

$$y'(t) = \kappa(\theta - y(t)), \quad y(0) = Y_0,$$
$$y(t) = \theta + (Y_0 - \theta)e^{-\kappa t}$$

Observations:

- *It is clear that S_T and $\tilde{S}_T$ are closely related: they are driven by the same BM B^1. The only difference is that S has stochastic volatility and the volatility of $\tilde{S}$ is determinstic. Then, naturally, we expect that the payoffs of the option in the two models are strongly correlated: cor$(X, \tilde{X}) \approx 1$.*
- *In addition, to simulated a path of S, we need to approximate the path of B^1, using the i.i.d. standard normals. Then, we can re-use the same normal variables (i.e. the same approximation of the path of B^1) to simulate the path of $\tilde{S}$. Thus, given a sample from the (approximate) distribution of S_T, it is relatively cheap to obtain a sample from the distribution of $\tilde{S}_T$.*

- *Finally, we can compute* $\mathbb{E}\tilde{X}$ *via the BS formula:*

$$\mathbb{E}\tilde{X} = C^{BS}\left(S_0, K, T; \sqrt{\theta + (Y_0 - \theta)\frac{1 - e^{-\kappa T}}{\kappa T}}, r\right) \tag{54}$$

Thus, it is reasonable to use $\tilde{X}$ *as a control variate for* X*.*

To sample from the (approximate) distribution of S_T*, we can, for example, use the standard Euler scheme:*

$$\begin{aligned}\hat{S}_{m+1} &= \hat{S}_m + \hat{S}_m\left(r\Delta t + \sqrt{\hat{Y}_m}\sqrt{\Delta t}\xi_m\right),\\ \hat{Y}_{m+1} &= \hat{Y}_m + (\theta - \kappa\hat{Y}_m)\Delta t + \beta\sqrt{\hat{Y}_m}\sqrt{\Delta t}(\rho\xi_m + \sqrt{1-\rho^2}\eta_m),\end{aligned}$$

where $\{\xi_m,\ \eta_m\}_{m=0,\dots,M-1}$ *are i.i.d.* $N(0,1)$*.*

And we can sample from the distribution of $\tilde{S}_T$ *exactly:*

$$\tilde{S}_{(m+1)\Delta t} = \tilde{S}_{m\Delta t}\exp\left(\left(r - \frac{v((m+1)\Delta t) - v(m\Delta t)}{2\Delta t}\right)\Delta t + \sqrt{v((m+1)\Delta t) - v(m\Delta t)}\xi_m\right),$$

reusing the same ξ_m *as in the scheme for* $\hat{S}$*, and with*

$$v(t) = \int_0^t y(u)du = \theta t + (Y_0 - \theta)\frac{1 - e^{-\kappa t}}{\kappa}$$

Thus, we generate a sample of i.i.d. pairs

$$\left\{\left(\hat{S}_M^{(i)}, \tilde{S}_T^{(i)}\right)\right\}_{i=1}^N$$

and define

$$X^{(i)} = F(\hat{S}_M^{(i)}) = e^{-rT}(\hat{S}_M^{(i)} - K)^+, \qquad \tilde{X}F(\tilde{S}_T^{(i)}) = e^{-rT}(\tilde{S}_T^{(i)} - K)^+, \quad i = 1, \dots, N.$$

Finally, we compute

$$\hat{b}^* = \frac{\sum_{i=1}^N (X^{(i)} - \bar{X})(\tilde{X}^{(i)} - \mathbb{E}\tilde{X})}{\sum_{i=1}^N (\tilde{X}^{(i)} - \mathbb{E}\tilde{X})^2},$$

with $\mathbb{E}\tilde{X}$ *given by (54), and construct the control variate estimate*

$$\mathbb{E}\left(e^{-rT}(S_T - K)^+\right) = \mathbb{E}X \approx \frac{1}{N}\sum_{i=1}^N (X^{(i)} - \hat{b}^*(\tilde{X}^{(i)} - \mathbb{E}\tilde{X}))$$

Ex 10. *One can use the control variates method to improve a MC estimate of the price of one derivative, making use of another one, for which the price is known explicitly.*

Let us consider a **discretely monitored Asian call** *with fixed strike:*

$$V_T = \left(\frac{1}{M} \sum_{m=1}^{M} S_{m\Delta t} - K \right)^+ ,$$

where $\Delta t = T/M$ *and S follows a GBM with drift r and volatility* σ*. Our goal is to compute the arbitrage-free price of this option:*

$$\mathbb{E}(e^{-rT} V_T) = ?$$

Then

$$X = F(S) = e^{-rT} \left(\frac{1}{M} \sum_{m=1}^{M} S_{m\Delta t} - K \right)^+ ,$$

and we choose

$$\tilde{X} = \tilde{F}(S) = e^{-rT} \left(\left(\prod_{m=1}^{M} S_{m\Delta t} \right)^{1/M} - K \right)^+$$

It is clear that computing $\tilde{X}$*, once we have simulated* $(S_{\Delta t}, \dots, S_{M\Delta t})$ *is very cheap.*

Another important feature of $\tilde{X}$ *is that it is* **strongly correlated** *with* X*, as is clear intuitively.*

It turns out that, in addition, we can compute $\mathbb{E}\tilde{X}$ *explicitly*

$$S_{m\Delta t} = S_0 \exp((r - \sigma^2/2) m\Delta t + \sigma W_{m\Delta t}),$$

$$\left(\prod_{m=1}^{M} S_{m\Delta t} \right)^{1/M} = S_0 \exp\left(\frac{(r - \sigma^2/2)(M+1)}{2M} T + \frac{\sigma}{M} \sum_{m=1}^{M} W_{m\Delta t} \right),$$

$$\mathbb{E} \left(\sum_{m=1}^{M} W_{m\Delta t} \right)^2 = \sum_{i=j} \mathbb{E}(W_{i\Delta t} W_{j\Delta t}) + 2 \sum_{i<j} \mathbb{E}(W_{i\Delta t} W_{j\Delta t})$$

$$= \sum_{i=1}^{M} i\Delta t + 2 \sum_{j=2}^{M} \sum_{i=1}^{j} i\Delta t = \frac{T(M+1)}{2} + \frac{T}{M} \sum_{j=2}^{M} j(j-1)$$

The above shows that

$$\frac{\sigma}{M} \sum_{m=1}^{M} W_{m\Delta t} \sim N\left(0, T\sigma^2 \left(\frac{M+1}{2M^2} + \frac{1}{M^3} \sum_{j=2}^{M} j(j-1) \right) \right),$$

and, hence,

$$\left(\prod_{m=1}^{M} S_{m\Delta t} \right)^{1/M}$$

is log-normal.

Thus, we can compute $\mathbb{E}\tilde{X}$ *via the Black-Scholes formula:*

$$\mathbb{E}\tilde{X} = C^{BS}\left(S_0, K, T; \tilde{\sigma}, r, q \right),$$

with

$$\tilde{\sigma}^2 = \sigma^2 \left(\frac{M+1}{2M^2} + \frac{1}{M^3} \sum_{j=2}^{M} j(j-1) \right),$$

band the (auxiliary) dividend yield

$$q = r - \frac{(r - \sigma^2/2)(M+1)}{2M} - \tilde{\sigma}^2/2$$

To obtain a control variate estimate of $\mathbb{E}X$, we

- *Compute $\lambda = \mathbb{E}\tilde{X}$ according to the above formula.*
- *Simulate the i.i.d. paths of S exactly, $\left\{ S^{(i)} = (S^{(i)}_{\Delta t}, \dots, S^{(i)}_{M\Delta t}) \right\}_{i=1}^{N}$.*
- *Compute the associated values $\left\{ X^{(i)} = F(S^{(i)}), \tilde{X}^{(i)} = \tilde{F}(S^{(i)}) \right\}_{i=1}^{N}$.*
- *Compute the coefficient*

$$\hat{b}^* = \frac{\sum_{i=1}^{N} (X^{(i)} - \bar{X})(\tilde{X}^{(i)} - \lambda)}{\sum_{i=1}^{N} (\tilde{X}^{(i)} - \lambda)^2},$$

- *Construct the control variate estimate*

$$V_0 = \mathbb{E}X \approx \frac{1}{N} \sum_{i=1}^{N} (X^{(i)} - \hat{b}^*(\tilde{X}^{(i)} - \lambda)$$

It is not so obvious that the geometric option is sufficiently closely correlated with the arithmetic option to justify the additional computations required for the control variate estimate. This however turns out to be the case.

Rem 9. *Notice that the arithmetic average is always at least as large as the geometric average. Hence, we have*

$$\mathbb{E}X \geq \mathbb{E}\tilde{X}$$

Thus, the control variate $\tilde{X}$ is also often used to obtain a lower bound on the price of an Asian call with fixed strike.

Rem 10. *The control variates methods extends easily to multiple variates $\tilde{X} = (\tilde{X}^1, \dots, \tilde{X}^d)$:*

$$\mathbb{E}X = \mathbb{E}(X - b^T \mathbb{E}(\tilde{X} - \mathbb{E}\tilde{X})),$$

where the optimal vector b is given by

$$b^* = \Sigma_{\tilde{X}}^{-1} \Sigma_{X\tilde{X}},$$

where $\Sigma_{\tilde{X}}$ is the covariance matrix of $\tilde{X}$, and

$$\Sigma_{X\tilde{X}} = (cov(X, \tilde{X}^1), \dots, cov(X, \tilde{X}^d))^T$$

Rem 11. *Similarly, one can consider nonlinear control variate estimates:*

$$\mathbb{E}X \approx \bar{X}\frac{\mathbb{E}\tilde{X}}{\bar{\tilde{X}}}, \quad \mathbb{E}X \approx \bar{X}\exp(\tilde{X} - \mathbb{E}\tilde{X}),$$

and, more generally,

$$\mathbb{E}X \approx \mathbb{E}h(\bar{X}, \bar{\tilde{X}}), \quad h(x, \mathbb{E}\tilde{X}) = x$$

Such estimates may be biased, but they are **consistent**, *and the function h may be chosen so that the estimate has asymptotically lower variance. However, it turns out that, for large N, the nonlinear control variate estimator, given by h, becomes equivalent to a linear estimator, with the coefficient b given by the partial derivatives of h. This is why it is sufficient to restrict our analysis to linear control variates.*

5.3 Stratified sampling

Stratified sampling consists in dividing the set of all possible realizations into disjoint subsets $E_1, \ldots, E_k$ and sampling from conditional distributions of X, given E_i. In other words, the stratified sampling is based on the representation:

$$\mathbb{E}X = \sum_{i=1}^{k} p_i \mathbb{E}(X \,|\, E_i),$$

where we assume that $p_i = \mathbb{P}(E_i)$ are known.

For example, if $\tilde{X}$ is another random variable, which takes k possible values $\{1, \ldots, k\}$, and such that

- we know $p_i = \mathbb{P}(\tilde{X} = i)$, for all i,
- and we can sample from $\text{Law}(X \,|\, \tilde{X} = i)$.

Then, we can consider stratified sampling with $E_i = \left\{\tilde{X} = i\right\}$.

To implement a MC method based on the above observation, we split the total size of the sample

$$N = N_1 + \cdots + N_k,$$

so that

$$N_i/N \approx p_i, \quad i = 1, \ldots, k.$$

Then, we sample independently from every $\text{Law}(X \,|\, E_i)$:

$$\mathbb{E}(X \,|\, E_i) \approx \frac{1}{N_i}\sum_{j=1}^{N_i} X_i^j, \quad i = 1, \ldots, k,$$

where all $\left\{X_i^j\right\}_{i,j}$ are independent (but not all identically distributed) and, **for each fixed $i = 1, \ldots, k$, the random variables $\left\{X_i^j\right\}_j$ have the same distribution as X conditional on E_i.**

Finally, we define the **stratified sampling estimate**

$$\mathbb{E}X \approx \bar{X}_N^p = \sum_{i=1}^{k} p_i \frac{1}{N_i}\sum_{j=1}^{N_i} X_i^j \approx \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} X_i^j,$$

where we recall that $N_i/p_i \approx N$ and assume that p_i's are known and that we can sample from Law$(X \mid E_i)$.

Let us analyze the **properties of the resulting estimate**. Denote

$$\mu_i = \mathbb{E}(X \mid E_i), \quad \sigma_i^2 = \text{Var}(X \mid E_i) = \mathbb{E}(X^2 \mid E_i) - \mu_i^2, \quad i = 1, \dots, k.$$

Notice

$$\mu = \mathbb{E}X = \sum_{i=1}^{k} \mathbb{E}(X \mid E_i)\mathbb{P}(E_i) = \sum_{i=1}^{k} p_i \mu_i,$$

$$\text{Var}(X) = \mathbb{E}X^2 - \mu^2 = \sum_{i=1}^{k} p_i \mathbb{E}(X^2 \mid E_i) - \mu^2 = \sum_{i=1}^{k} p_i(\sigma_i^2 + \mu_i^2) - (\sum_{i=1}^{k} p_i \mu_i)^2$$

Then,

$$\mathbb{E}\bar{X}_N^p = \mathbb{E}\left(\sum_{i=1}^{k} p_i \frac{1}{N_i} \sum_{j=1}^{N_i} X_i^j\right) = \sum_{i=1}^{k} p_i \mu_i = \mu = \mathbb{E}X,$$

hence, the estimate is **unbiased**.

It is also easy to see that the **stratified sampling estimate is consistent**:

$$\lim_{N\to\infty} \frac{1}{N_i} \sum_{j=1}^{N_i} X_i^j = \lim_{N_i\to\infty} \frac{1}{N_i} \sum_{j=1}^{N_i} X_i^j = \mu_i,$$

for every $i = 1, \dots, k$, due to LLN. Then,

$$\lim_{N\to\infty} \bar{X}_N^p = \sum_{i=1}^{k} p_i \lim_{N\to\infty} \frac{1}{N_i} \sum_{j=1}^{N_i} X_i^j = \sum_{i=1}^{k} p_i \mu_i = \mu$$

In addition,

$$\text{Var}(\bar{X}_N^p) = \text{Var}\left(\sum_{i=1}^{k} p_i \frac{1}{N_i} \sum_{j=1}^{N_i} X_i^j\right) = \sum_{i=1}^{k} p_i^2 \text{Var}\left(\frac{1}{N_i} \sum_{j=1}^{N_i} X_i^j\right) = \sum_{i=1}^{k} p_i^2 \frac{\sigma_i^2}{N_i} \approx \frac{1}{N} \sum_{i=1}^{k} p_i \sigma_i^2$$

On the other hand, the regular sample mean estimate has variance

$$\text{Var}(\bar{X}_N) \approx \frac{1}{N}\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{k} p_i(\sigma_i^2 + \mu_i^2) - (\sum_{i=1}^{k} p_i \mu_i)^2\right)$$

The above estimates show that the **stratified sampling has lower asymptotic variance if**

$$\sum_{i=1}^{k} p_i \mu_i^2 \geq \left(\sum_{i=1}^{k} p_i \mu_i\right)^2.$$

The above inequality does, indeed, hold and is known as the Jensen's inequality. Assuming that all $p_i > 0$, the equality in the above is only possible if all μ_i are equal.

Thus, the **stratified sampling never increases the variance and, typically, reduces it**.

Ex 11. *Consider the problem of computing*

$$\mathbb{E}X = \mathbb{E}G(\xi),$$

where $G : \mathbb{R} \to \mathbb{R}$ and ξ has cdf F, such that we can approximate F and F^{-1} efficiently (e.g. normal cdf). Then, to decrease the variance of the MC estimate of $\mathbb{E}G(\xi)$, we choose a partition $-\infty = a_0 < a_1 < \cdots < a_{k-1} < a_k = \infty$ and define $\tilde{X}$ to take value i if $\xi \in A_i = (a_{i-1}, a_i]$, for $i = 1, \ldots, k$. We assume that

$$p_i = \mathbb{P}(E_i) = \mathbb{P}(\tilde{X} = i) = \mathbb{P}(\xi \in A_i), \quad i = 1, \ldots, k,$$

are known explicitly or can be computed at a negligible cost.

To implement the stratified sampling method, we need to be able to simulate from Law$(X \mid \xi \in A_i)$, or, alternatively, from Law$(\xi \mid \xi \in A_i)$. The following algorithm addresses this issue, using the **inverse transform method**, *and constructs the stratified sampling estimate for $\mathbb{E}X = \mathbb{E}G(\xi)$.*

- *Choose the overall sample size N and partition it into*

$$N = N_1 + \cdots N_k, \quad N_i \approx p_i N, \quad i = 1, \ldots, k.$$

- *For a fixed $i = 1, \ldots, k$, simulate $\left\{U_i^j\right\}_{j=1}^{N_i}$ – i.i.d. $Unif(0,1)$.*

- *Compute*

$$\xi_i^j = F^{-1}(F(a_{i-1}) + (F(a_i) - F(a_{i-1}))U_i^j),$$

Finally, we construct the stratified sampling estimate

$$\mathbb{E}X = \mathbb{E}G(\xi) \approx \bar{X}_N^p = \frac{1}{N} \sum_{i,j} G(\xi_i^j),$$

which has a lower variance than the standard estimate $\bar{X}_N$.

Rem 12. *An additional advantage of the above choice of "strata", $E_i = \{\xi \in A_i\}$, is that the histograms of ξ_i^j, produced by the stratified sampling, look much closer to the true density of ξ than the histograms of the direct sampling from the distribution of ξ.*

Rem 13. *The above example is included here for the educational purposes only. Notice that it is limited to a one-dimensional ξ, in which case MC may not be the best way to approximate $\mathbb{E}G(\xi)$ at all. In fact, if the cdf F is known, one can simply approximate numerically the following integral:*

$$\mathbb{E}G(\xi) = \int_{\mathbb{R}} G(z) dF(z)$$

Ex 12. *A general (but not always most efficient) method of generating from Law$(X \mid E_i)$ is the following version of* **acceptance-rejection** *method.*

Assume that we need to estimate $\mathbb{E}X$, with $X = G(\xi)$ where ξ takes values in $\mathbb{R}^d$, and that we are given a partition $A_1, \ldots, A_k$ of the state space $\mathbb{R}^d$, such that the probabilities $p_i = \mathbb{P}(\xi \in A_i)$ are known for all $i = 1, \ldots, k$.

For example, consider standard Gaussian vector $\xi = (\xi_1, \xi_2)$, which is used to price a spread option in the BS model, or to price an option that depends on the underlying value at two different moments in time. Then, we know that the state space $\mathbb{R}^2$ can be split into the four orthants: $A_1 = \{(x_1, x_2) : x_1 \geq 0, x_2 \geq 0\}$, $A_2 = \{(x_1, x_2) : x_1 < 0, x_2 \geq 0\}$, $A_3 = \{(x_1, x_2) : x_1 < 0, x_2 < 0\}$, $A_4 = \{(x_1, x_2) : x_1 \geq 0, x_2 < 0\}$. Notice that $\mathbb{P}(\xi \in A_i) = 1/4$.

Assuming that we can sample from Law(X), the stratified sampling can be implemented as follows.

- *Choose the overall sample size N and partition it into*

$$N = N_1 + \cdots N_k, \quad N_i \approx p_i N, \quad i = 1, \ldots, k.$$

- *Introduce the variables $L_1, \ldots, L_k$, whose initial values are set to zero.*
- *Simulate $\xi^{(1)}$ and increase L_i by one, where $i \in \{1, \ldots, k\}$ is such that $\xi^{(1)} \in A_i$.*
- *Repeat the simulation until $L_i = N_i$, for some $i = 1, \ldots, k$. After that, reject every following $\xi^{(j)}$ that falls in A_i.*
- *Repeat this until $L_i = N_i$ for all $i = 1, \ldots, k$.*
- *Construct the stratified sampling estimate*

$$\mathbb{E}X = \mathbb{E}G(\xi) \approx \frac{1}{N} \sum_{j=1}^{N} G(\xi^{(j)})$$

The above method will reduce the variance of the estimate for $\mathbb{E}X$ but at the expense of extra computations due to rejections. Whether or not these extra costs are justified by the variance reduction is a very difficult question that does not have a general answer. Certain specific cases are analyzed, for example, in Glasserman, Heidelberger and Shahabuddin (2000).

5.4 Moments matching

Strictly speaking, **moments matching** is not a variance reduction technique: i.e. it does not always reduce the variance of the target estimate. Instead of reducing the asymptotic variance, it is used to reduce (or eliminate) biases in the (finite) input sample. This is achieved by one of the two methods

- **adjusting the sample**,
- or **adjusting the weights** within the sample.

We present this method here as it is especially important in Finance: it allows us to avoid arbitrage in the simulated data.

Let us focus on the **sample adjustment** first. Assume that we need to estimate $\mathbb{E}X$, where $X = G(\xi)$, and we have simulated a sample of i.i.d. random variables $\left\{\xi^{(i)}\right\}_{i=1}^{N}$. Assume also that we know a priori the first moment of ξ: $\mathbb{E}\xi = \mu$. The sample mean $\bar{\xi}$ is typically different from the true mean: $\bar{\xi} \neq \mu$. Hence, if we want to ensure that the sample mean equals the true mean, we can consider one of the transformed samples instead

$$\left\{\xi^{(i)} \frac{\mu}{\bar{\xi}}\right\}_{i=1}^{N} \quad \text{or} \quad \left\{\mu + (\xi^{(i)} - \bar{\xi})\right\}_{i=1}^{N}$$

Notice that the above transformed samples are not i.i.d. but they have the **correct sample means**.

Of course, there are many other possible transformations that match the first moment of the sample. In addition, if we also know the second moment of ξ, we can compute $\sigma^2 = \text{Var}(\xi)$ and adjust the sample accordingly:

$$\left\{\mu + (\xi_i - \bar{\xi})\frac{\sigma}{\hat{\sigma}}\right\}_{i=1}^N, \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^N (\xi_i - \bar{\xi})^2$$

Q 7. *When is it important to match moments?*

Indeed, by matching the sample moments we ensure that there is no bias in estimating the moments of ξ. But we already know the moments of ξ, so there is no need to estimate them.

However, recall that we are interested in estimating $\mathbb{E}G(\xi)$, rather than the moments of ξ. In financial applications, ξ usually represents the value of the underlying factor(s). A typical example is $\xi = (S_0, \dots, S_{M\Delta t})$, where S is a tradable asset (e.g. this is the case in the problem of pricing equity options). Assume that we have simulated N i.i.d. (discrete) paths of S

$$\left\{(S_0^{(i)}, \dots, S_{M\Delta t}^{(i)})\right\}_{i=1}^N$$

Notice that the **simulated paths themselves define a model of financial market**. In this model, under the pricing measure, the **risky asset S takes one of the possible (simulated) paths, with probability $1/N$ each**. Then, pricing an option by computing the sample mean of the discounted payoff $\left\{G(S^i)\right\}_{i=1}^N$ is the same as computing the discounted expectation (precisely) in the new model. But if we want to use the new model for pricing, it is important to ensure that it is arbitrage-free. This, in turn, translates into a moment matching condition:

$$\bar{S}_{m\Delta t} = S_0 e^{rm\Delta t}, \qquad m = 1, \dots, M$$

To ensure that this condition is satisfied, we consider the adjusted sample instead:

$$\left\{(S_0, S_{\Delta t}^{(i)} \frac{S_0 e^{r\Delta t}}{\bar{S}_{\Delta t}}, \dots, S_{M\Delta t}^{(i)} \frac{S_0 e^{rM\Delta t}}{\bar{S}_{M\Delta t}})\right\}_{i=1}^N, \qquad \bar{S}_{m\Delta t} = \frac{1}{N}\sum_{i=1}^N S_{m\Delta t}^{(i)}, \qquad m = 1, \dots, M.$$

And we use the above sample to price an option (i.e. approximate $\mathbb{E}G(S)$).

Another way to match the sample moments is to **adjust the probabilities** of the simulated paths, rather than the paths themselves. Assume that we have generated the i.i.d. paths

$$\left\{(S_0^{(i)}, \dots, S_{M\Delta t}^{(i)})\right\}_{i=1}^N$$

As mentioned before, these paths yield a model in which the risky asset S may take any one of these paths with equal probability $1/N$. Then, to make sure that, for example,

$$\bar{S}_{M\Delta t} = \sum_{i=1}^N \frac{1}{N} S_{M\Delta t}^{(i)} = \mathbb{E}S_T,$$

instead of adjusting the sample paths, we can choose a set of weights $\{w_1, \dots, w_N\}$, such that

- $w_1 + \cdots + w_N = 1, \quad w_i > 0, \ i = 1, \dots, N,$

- and

$$\sum_{i=1}^{N} w_i S_{M\Delta t}^{(i)} = \mathbb{E} S_T,$$

Of course, we can extend the above constraints to include the moment-matching conditions at all $m = 1, \dots, M$. Nevertheless, the number of conditions is usually much smaller than N – the number of w_i's. Hence, they do not determine all w_i's uniquely. In practice, one may choose a parametric family of $\{w_1(\theta), \dots, w_N(\theta)\}$, so that $w_i(\theta) \geq 0$ and $\sum w_i(\theta) = 1$, and, then, find the parameter θ such that the moment conditions are matched.

We do not examine this approach (which his known as the **weighted MC**) in detail, because it can be viewed as a **change of measure technique**, which is discussed in the next subsection. Indeed, changing $1/N$ to w_i is the same as changing the probability of each simulated paths.

Rem 14. *It is important to avoid arbitrage in pricing for many different reasons. For example, a* **market maker** *needs to ensure that the quotes she provides for various OTC (over-the-counter) products do not allow for arbitrage opportunities (or, at least, that these opportunities do not exceed the bid-ask spread).*

Another example arises in **statistical arbitrage** *or optimal portfolio searches. In such problems, one typically fits a model to the data and then assumes that the parameters of this model remain unchanged for at least some period of time. Then, the prices of various assets are computed according to the model and the potential arbitrage opportunities are detected. If the tradable assets in question include derivatives and a MC method is used to compute their prices, then, it is necessary to use moment matching to ensure that the algorithm does not detect the arbitrage opportunities which do not really exist.*

For example, if the MC estimate produces $\bar{S}_T \neq S_0 e^{rT}$, then the put-call parity may not be satisfied according to the generated sample:

$$e^{-rT}\bar{\mathbb{E}}(S_T - K)^+ - e^{-rT}\bar{\mathbb{E}}(K - S_T)^+ \neq S_0 - e^{-rT}K,$$

where $\bar{\mathbb{E}}$ denotes the expectation taken in the model produced by the MC sample (with every sample path having probability $1/N$). This may seem like an arbitrage opportunity, as we know that the put-call parity has to hold (both in theory and in practice), and may lead one to opening a trade that would exploit this "arbitrage". However, such opportunity may be merely a consequence of the MC error, and it can be eliminated by the moment matching technique.

Rem 15. *The estimates resulting from moments matching are usually designed to be* **consistent***. In particular, the multiplication adjustment of asset prices, described above, is consistent.*

However, such estimates are typically **biased***. A simple example is given by the sample variance:*

$$\mathbb{E}G(\xi) = \mathbb{E}(\xi - \mu)^2 \approx \frac{1}{N}\sum_{i=1}^{N}(\xi^{(i)} - \bar{\xi})^2,$$

where $\mu = \mathbb{E}\xi$ is known, and the right hand side of the above can be viewed as an estimate resulting from the adjusted sample $\left\{\mu + (\xi^{(i)} - \bar{\xi})\right\}$. It is well known that the above estimate is biased.

Rem 16. *Notice that moments matching can, in fact, be viewed as a variance reduction method, but for the expectation of the input, $\mathbb{E}\xi = \mu$, rather than the target expectation $\mathbb{E}G(\xi)$. Indeed, the sample mean of the modified input is always equal to μ, by construction, hence, it is a zero-variance estimate of μ. Of course, this observation is of now practical use, as we assume that μ is known a priori. However, it justifies the fact that moments matching is sometimes included in the "variance reduction methods".*

5.5 Importance sampling

Importance sampling is a variance reduction technique based on **changing the probability measure** from which the sample is generated. Roughly speaking, we change the measure to give **more weight to the more important outcomes** (e.g. the outcomes for which the output is not zero), hence the name.

Consider the problem of estimating

$$\mathbb{E}X = \mathbb{E}h(\xi) = \int_{\mathbb{R}^d} h(x)f(x)dx,$$

where ξ takes values in $\mathbb{R}^d$ and has pdf f.

Consider another distribution on $\mathbb{R}^d$ given by pdf g and assume that $g(x) > 0$ for all $x \in \mathbb{R}^d$. Then

$$\mathbb{E}X = \mathbb{E}h(\xi) = \int_{\mathbb{R}^d} h(x)f(x)dx = \int_{\mathbb{R}^d} h(x)\frac{f(x)}{g(x)}g(x)dx = \mathbb{E}\left(h(\tilde{\xi})\frac{f(\tilde{\xi})}{g(\tilde{\xi})}\right),$$

where $\tilde{\xi}$ has pdf g.

Then, the **importance sampling estimate** of $\mathbb{E}X$ is given by

$$\mathbb{E}X = \mathbb{E}h(\xi) \approx \bar{X}^g = \frac{1}{N}\sum_{i=1}^{N} h(\tilde{\xi}^{(i)})\frac{f(\tilde{\xi}^{(i)})}{g(\tilde{\xi}^{(i)})},$$

where $\tilde{\xi}^{(1)}, \dots, \tilde{\xi}^{(N)}$ are **drawn independently from the distribution with pdf** g.

Notice that

$$\mathbb{E}\bar{X}^g = \tilde{\mathbb{E}}\left(h(\xi)\frac{f(\xi)}{g(\xi)}\right) = \mathbb{E}\left(h(\tilde{\xi})\frac{f(\tilde{\xi})}{g(\tilde{\xi})}\right) = \int_{\mathbb{R}^d} h(x)\frac{f(x)}{g(x)}g(x)dx = \mathbb{E}X,$$

where $\tilde{\mathbb{E}}$ is the expectation under a different probability measure (under which ξ has pdf g), and $\tilde{\xi}$ is a random vector that has pdf g under the original probability measure. Thus, the importance sampling estimate is **unbiased**.

The importance sampling estimate is also **consistent**, as follows from the LLN.

Recall that the standard MC estimate is given by

$$\mathbb{E}X \approx \bar{X} = \frac{1}{N}\sum_{i=1}^{N} h(\xi^{(i)}),$$

where $\xi^{(1)}, \dots, \xi^{(N)}$ are drawn independently from the distribution with pdf f.

To compare the variances of the two estimates, we notice that

$$\mathbb{E}\bar{X}^2 = \mathbb{E}h^2(\xi) = \int_{\mathbb{R}^d} h^2(x)f(x)dx$$

and

$$\mathbb{E}(\bar{X}^g)^2 = \mathbb{E}\left(h(\tilde{\xi})\frac{f(\tilde{\xi})}{g(\tilde{\xi})}\right)^2 = \int_{\mathbb{R}^d} h^2(x)\frac{f(x)}{g(x)}f(x)dx$$

It is clear that depending of the joint behavior if the functions f and f/g, the **importance sampling may or may not have a smaller variance than the standard estimate**.

It is clear that, for a variance reduction, we would like to choose g so that f/g is as small as possible. However, g has to remain positive and integrate to one to be a pdf, hence we cannot make the ratio f/g arbitrarily small. Let us find the optimal choice of g. Assume that $h, f > 0$. Then

$$g(x) = \frac{h(x)f(x)}{\int_{\mathbb{R}^d} h(x)f(x)dx}$$

is a pdf. The importance sampling estimate, then, becomes constant:

$$\bar{X}^g = \frac{1}{N}\sum_{i=1}^{N} h(\tilde{\xi}^{(i)})\frac{f(\tilde{\xi}^{(i)})}{g(\tilde{\xi}^{(i)})} = \int_{\mathbb{R}^d} h(x)f(x)dx$$

In other words, its **variance is reduced to zero**.

The above choice, of course, does not make any sense form a practical point of view: to define g as above, we need to know $\int_{\mathbb{R}^d} h(x)f(x)dx$, which is exactly what we need to estimate. However, it helps us develop the intuition about how to choose g to reduce the variance, which is the **most important part of constructing the importance sampling estimate**. From the above considerations, it is clear that we should aim to **choose g so that it is close to being proportional to fh.**

Ex 13. Changing mean of normal distribution. *Assume that we wish to estimate $\mathbb{E}X = \mathbb{E}[h(\xi)]$, where ξ is standard normal, but $h : \mathbb{R} \to \mathbb{R}$ only takes significant values when $\xi \leq -\alpha$ where $\alpha > 2$.*

An example of this is a **far out of the money** *put option in the BS model. In this case we have*

$$h(\xi) = e^{-rT}\left(K - S_0 \exp\left((r - \sigma^2/2)T + \sigma\sqrt{T}\xi\right)\right)^+,$$

with

$$K < S_0 \exp\left((r - \sigma^2/2)T - 2\sigma\sqrt{T}\right)$$

Thus $h(\xi) = 0$ for $\xi > -2$, as the option is far out of the money. If we use the standard MC procedure then $h(\xi^{(i)}) = 0$ for most of our simulations $\xi^{(i)}$ (in fact, it will be the case for more than 98% of simulations). Thus, we know a priori the answer to most of the MC simulations – i.e. zero. The **importance sampling** *allows us to modify the method in such a way that each new simulation really does give us* **new information**.

To implement the importance sampling, we need to choose function g. Recall that g should give more weight (i.e. probability) to the non-trivial outcomes – i.e. the ones for which $h(\xi)$ is not zero. In the present case, it means that we need more outcomes to be very large negative numbers. It turns out that, in the Gaussian case, we can find the appropriate function g by a simple translation of the variables:

$$g(x) = f(x - \theta) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\theta)^2}{2}},$$

where f is the standard normal pdf.

Then

$$\frac{g(x)}{f(x)} = \frac{e^{-\frac{x^2}{2}+\theta x-\frac{\theta^2}{2}}}{e^{-\frac{x^2}{2}}} = e^{\theta x - \frac{\theta^2}{2}},$$

$$\mathbb{E}h(\xi) \approx \bar{X}^g = \frac{1}{N}\sum_{i=1}^{N} h(\tilde{\xi}^{(i)})\exp\left(-\theta\tilde{\xi}^{(i)} + \frac{\theta^2}{2}\right),$$

where $\left\{\tilde{\xi}^{(i)}\right\}$ *are i.i.d.* $N(\theta, 1)$.

Now we need to decide which value of θ *is best to reduce variance of the MC estimate. It is clear that* θ *should be negative but there is no precise rule on how to choose it. For the far out of the money put option we claim that we should take*

$$\theta = \frac{\log(K/S_0) - (r - \sigma^2/2)T}{\sigma\sqrt{T}} \tag{55}$$

This is not completely obvious since for any negative θ *with large enough* $|\theta|$ *the importance sampling estimate* $\bar{X}^g$ *will typically be non-zero.*

There is actually a trade off here. On the one hand the larger $|\theta|$ *is, the more likely the MC simulation yields non-zero values. On the other hand as* $|\theta|$ *increases the variance of the additional factor (also known as the* **likelihood ratio***),*

$$Var[f(\tilde{\xi})/g(\tilde{\xi})] = \mathrm{Var}[e^{-\theta\tilde{\xi}+\theta^2/2}] = e^{\theta^2}\left(\mathbb{E}(e^{-2\theta\tilde{\xi}}) - (\mathbb{E}e^{-\theta\tilde{\xi}})^2\right) = e^{\theta^2}(1 - e^{-\theta^2}) = e^{\theta^2} - 1,$$

increases rapidly.

To resolve this dilemma, we use the following heuristic arguments. When computing the expected payoff of a deep out of the money option, the most important part of the expectation is due to the values of the underlying that are around the strike (around at the money). The reason is that the other (in the money) values have much smaller probabilities, hence, they are negligible for small K. *Thus, to maximize the amount of information produced by the simulations, we need to generate* S_T *from a distribution that has most of its values around* K. *Since we have restricted ourselves to the BS model, we need* ξ *to have most of its mass around* θ *given by (55). As we have also restricted the possible distributions of* ξ *to a shifted standard normal, whose pdf has maximum around the mean, we end up with (55).*

Ex 14. *The above example is slightly artificial, because, in one dimension, we can approximate the distribution of a standard normal, and its integrals, efficiently using analytical methods instead of MC. However, the same method will work in higher dimensions.*

For example, if we need to price a **spread call option in BS model***, the problem reduces to the computation of*

$$\mathbb{E}h(\xi_1, \xi_2)$$

$$= \mathbb{E}\left(e^{-rT}\left(S_0^1 \exp\left((r - \sigma_1^2/2)T + \sigma_1\sqrt{T}\xi_1\right) - S_0^2 \exp\left((r - \sigma_2^2/2)T + \sigma_2\sqrt{T}(\rho\xi_1 + \sqrt{1-\rho^2}\xi_2)\right) - K\right)^+\right),$$

where ξ_1 *and* ξ_2 *are independent* $N(0, 1)$.

Then, if $K >> S_0$, *we need to shift the means of* ξ_1 *and* ξ_2 *so that, for example,*

$$\mathbb{E}\xi_1 = \theta, \qquad \mathbb{E}\xi_2 = -\frac{\theta\rho}{\sqrt{1-\rho^2}},$$

with some $\theta >> 1$. *The new density* $g(z_1, z_2)$ *is the density of a Gaussian vector, with covariance given by the identity matrix and the above mean. The likelihood ratio can be computed as before, and* θ *can be chosen according to the same argument that we used to derive (55), namely:*

$$\theta = \frac{1}{\sigma_1\sqrt{T}}\left(\log\left(\frac{K + S_0^2 \exp\left((r - \sigma_2^2/2)T\right)}{S_0^1}\right) - (r - \sigma_1^2/2)T\right)$$

6 Hedging and estimating sensitivities via MC

6.1 Hedging

We assume that the model is described by the following (possibly multi-dimensional) diffusion

$$dX_t = b(X_t,t)dt + a(X_t,t)dB_t,$$

where $b(X_t,t) \in \mathbb{R}^d$, $a(X_t,t) \in \mathbb{R}^{d\times n}$, and $B = (B^1,\dots,B^n)^T$ is n-dimensional standard BM.

Denote the price of an option at time t by V_t. Then, if the option's payoff at maturity T is a function of the state process X_T, we have

$$V_t = V(X_t,t)$$

Applying the Ito's formula, we obtain

$$dV(X_t,t) = (\cdots)dt + \sum_{j=1}^{n}\left(\sum_{i=1}^{d} a^{ij}(X_t,t)\partial_{x^i}V(X_t,t)\right)dB_t^j$$

Assume that we would like to hedge this option by trading in asset S, where

$$dS_t = \tilde{\mu}(X_t,t)dt + \sum_{j=1}^{n}\tilde{\sigma}^j(X_t,t)dB_t^j$$

Denote by H_t the time t value of the hedging portfolio, which consists of π_t shares of S and the rest being held in a bank account with **deterministic interest rate**.

Then, we have

$$dH_t = (\cdots)dt + \pi_t dS_t = (\cdots)dt + \sum_{j=1}^{n}\pi_t\tilde{\sigma}^j(X_t,t)dB_t^j$$

- Hedging strategy π is chosen to minimize the difference between H and V.
- This, in turn, is equivalent to minimizing the difference between their increments, dH_t and dV_t.
- Ignoring the dt terms (as they are of a smaller order), we need to choose π_t to minimize

$$\|\sum_{j=1}^{n}\left(\sum_{i=1}^{d} a^{ij}(X_t,t)\partial_{x^i}V(X_t,t)\right)dB_t^j - \sum_{j=1}^{n}\pi_t\tilde{\sigma}^j(X_t,t)dB_t^j\|^2$$

$$= \|\sum_{j=1}^{n}\left(\sum_{i=1}^{d} a^{ij}(X_t,t)\partial_{x^i}V(X_t,t) - \pi_t\tilde{\sigma}^j(X_t,t)\right)dB_t^j\|^2$$

- Since the increments of B^j's are independent normal r.v.'s with mean zero, the choice of "**quadratic norm**" yields the following problem

$$\min_{\pi_t}\sum_{j=1}^{n}\left(\sum_{i=1}^{d} a^{ij}(X_t,t)\partial_{x^i}V(X_t,t) - \pi_t\tilde{\sigma}^j(X_t,t)\right)^2,$$

$$\pi_t = \frac{\sum_{j=1}^{n}\sum_{i=1}^{d}\tilde{\sigma}^j(X_t,t)a^{ij}(X_t,t)\partial_{x^i}V(X_t,t)}{\sum_{j=1}^{n}\left(\tilde{\sigma}^j(X_t,t)\right)^2}, \tag{56}$$

- The above strategy is called **quadratic hedging**. Its first rigorous definition is due to Schweitzer and Föllmer.

In a **complete model**, the strategy given by (56), in fact, eliminates the difference between the dB_t terms in dH_t and dV_t. Hence, with such choice of π, the discounted price of $H - V$ cannot be a martingale unless it is constant. Then, by the no-arbitrage argument, it has to be constant and equal to its initial price, which is zero. Thus, in a complete market, the strategy given by (56) provides perfect hedge (or perfect replication).

Ex 15. *In BS model, we have $d = n = 1$, $X = S$, $\tilde{\mu}(X_t,t) = \mu S_t$, $\tilde{\sigma}(X_t,t) = \sigma S_t$, $b = \tilde{\mu}$, $a = \tilde{\sigma}$,*

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

Assuming that the option is of European type, we have

$$V_t = V(S_t,t),$$

and

$$\pi_t = \frac{\sigma S_t \sigma S_t \partial_S V(S_t,t)}{(\sigma S_t)^2} = \partial_S V(S_t,t),$$

where we recognize the standard **delta-hedging** *formula.*

Note that, if the option is path-dependent, we would still be able to hedge it perfectly in BS model, but we would need to increase the dimension of the state process.

Rem 17. *Of course, the replication in complete models is perfect only in theory. In practice, the rebalancing can only be done at discrete moments in time, hence, the actual hedge contains a discretization error. In theory, the discretization error can be reduced by increasing the trading frequency, but, in practice, this matt become a problem due to transaction costs. So, in real-world applications, perfect replication is never possible, but the theoretical results approximate the reality well if we consider relatively small transaction costs and large trades in the options.*

If the model is **incomplete**, the replication is not perfect. In such markets, the quadratic hedging is not the only possible choice of hedging strategy: a different choice of the objective, inside the associated minimization problem, will lead to a different hedging strategy. Quadratic hedging strategy is one of the most straightforward and popular strategies in incomplete markets.

Ex 16. *In Heston model, we have $d = n = 2$, $X = (S,Y)^T$, and*

$$\begin{cases} dS_t = \mu S_t dt + \sqrt{Y_t} S_t dB_t^1 \\ dY_t = \kappa(\theta - Y_t)dt + \sigma\sqrt{Y_t}(\rho dB_t^1 + \sqrt{1-\rho^2} dB_t^2) \end{cases}$$

Assuming that the option is of European type

$$V_t = V(S_t,t),$$

we have

$$\begin{aligned} \pi_t &= \frac{\sqrt{Y_t}S_t\left(\sqrt{Y_t}S_t\partial_S V(S_t,Y_t,t) + \sigma\sqrt{Y_t}\rho\partial_Y V(S_t,Y_t,t)\right)}{YS_t^2} \\ &= \partial_S V(S_t,Y_t,t) + \frac{\sigma\rho}{S_t}\partial_Y V(S_t,Y_t,t) \end{aligned}$$

The above examples show that, in order to construct hedging strategies, we need to be able to compute the derivatives of a price function $V(X_t, t)$ with respect to the components of the state process, $\partial_{x^i} V(X_t, t)$. Such derivatives can be viewed as the **sensitivities** of option's price to the changes in initial level of the state process X.

6.2 Estimating sensitivities

The basic problem we investigate here is how to approximate $\alpha'(\theta)$, where

$$\alpha(\theta) = \mathbb{E}Z(\theta),$$

is evaluated by a MC method, with $Z(\theta)$ being a random variable depending on the parameter $\theta \in \mathbb{R}$.

Finite-difference approximation

The most straightforward way to approximate α' is to consider the **forward finite-difference approximation**:

$$\alpha'(\theta) \approx \frac{\mathbb{E}Z(\theta + \Delta\theta) - \mathbb{E}Z(\theta)}{\Delta\theta} \approx \frac{\bar{Z}(\theta + \Delta\theta) - \bar{Z}(\theta)}{\Delta\theta},$$

where $\bar{Z}$ is a MC approximation of $\mathbb{E}Z$ (e.g. sample mean).

Notice that such approximation is **biased**:

$$\mathbb{E}\frac{\bar{Z}(\theta + \Delta\theta) - \bar{Z}(\theta)}{\Delta\theta} = \frac{\alpha(\theta + \Delta\theta) - \alpha(\theta)}{\Delta\theta} = \alpha'(\theta) + \frac{1}{2}\alpha''(\theta)\Delta\theta + O(\Delta\theta^2),$$

where we assumed that α is smooth enough and used the Taylor's expansion:

$$\alpha(\theta + \Delta\theta) = \alpha(\theta) + \alpha'(\theta)\Delta\theta + \frac{1}{2}\alpha''(\theta)\Delta\theta^2 + O(\Delta\theta^3)$$

One can use the **central finite-difference approximation**:

$$\alpha(\theta) \approx \frac{\bar{Z}(\theta + \Delta\theta) - \bar{Z}(\theta - \Delta\theta)}{2\Delta\theta},$$

to obtain a smaller asymptotic bias:

$$\mathbb{E}\frac{\bar{Z}(\theta + \Delta\theta) - \bar{Z}(\theta - \Delta\theta)}{2\Delta\theta} = \alpha'(\theta) + \frac{1}{6}\alpha'''(\theta)\Delta\theta^2 + O(\Delta\theta^3),$$

In fact, the above **bias does not vanish as the number of simulations N grows**: even if we could compute $\alpha(\theta)$ precisely, there would still be an error in the finite-difference approximation of $\alpha'(\theta)$. Of course, we can make it vanish by taking $\Delta\theta \to 0$. However, the biggest problem of the finite-difference approximation is that the **variance of the estimate may explode as $\Delta\theta \to 0$.**

Assume, for simplicity, that $\bar{Z}$ is a sample mean, with the sample size N. Then

$$\text{Var}\left(\frac{\bar{Z}(\theta + \Delta\theta) - \bar{Z}(\theta)}{\Delta\theta}\right) = \frac{1}{\Delta\theta^2 N}\text{Var}\left(Z(\theta + \Delta\theta) - Z(\theta)\right) \tag{57}$$

Assume also that

$$Z(\theta) = G(\xi; \theta),$$

where ξ is a random variable (or vector).

Ex 17. *Consider $V(S_t, t)$ – the price of a European call with strike K and maturity T in BS model. The partial derivative $\partial_S V(S_t, t)$ can be understood as $\alpha'(S_t)$, where*

$$\alpha(s) = \mathbb{E}Z(s) = \mathbb{E}G(\xi; s) = \mathbb{E}e^{-r(T-t)}\left(s\exp((r-\sigma^2/2)(T-t)+\sigma\sqrt{T-t}\xi)-K\right)^+,$$

with s being the level of risky asset at time t, and $\xi \sim N(0,1)$.

Thus, $Z(\theta) = G(\xi; \theta)$ and $Z(\theta + \Delta\theta) = G(\xi; \theta + \Delta\theta)$ are computed by simulating the input ξ. Then, there are two possible implementations of the associated estimates $\bar{Z}(\theta + \Delta\theta)$ and $\bar{Z}(\theta)$.

- We can **simulate ξ for $G(\xi; \theta + \Delta\theta)$ and $G(\xi; \theta)$ independently**. In this case, as $\Delta\theta \to 0$, assuming the continuity of Var$(G(\xi; \theta))$ with respect to θ, we obtain:

  $$\text{Var}\left(Z(\theta+\Delta\theta) - Z(\theta)\right) = \text{Var}\left(G(\xi; \theta+\Delta\theta)\right) + \text{Var}\left(G(\xi', \theta)\right) \to 2\text{Var}\left(G(\xi; \theta)\right),$$

  where ξ' is an independent copy of ξ. We see that, in this case, the **variance of the finite-difference approximation, given in (57), explodes as $\Delta\theta \to 0$.**

  Assuming that the bias is $O(\Delta\theta^h)$, the total mean-square error of the finite-difference approximation is given by

  $$O(\Delta\theta^h) + O(N^{-1/2}\Delta\theta^{-1})$$

  For maximum efficiency, the asymptotic behavior of $\Delta\theta$ should be such that the above two terms are the same. This is achieved if

  $$\Delta\theta \sim N^{-\frac{1}{2(h+1)}},$$

  and the total error becomes

  $$O\left(N^{-\frac{1}{2}\left(1-\frac{1}{h+1}\right)}\right),$$

  with $h = 1$ for forward difference, and $h = 2$ for central difference.

  We see that the above **error is always asymptotically larger than the standard error** $O\left(N^{-\frac{1}{2}}\right)$.

- Alternatively, we can **reuse the same ξ for $G(\xi; \theta+\Delta\theta)$ and $G(\xi; \theta)$**. In this case, assuming Lipschitz continuity of $G(\xi; \theta)$ with respect to θ, we obtain:

  $$\text{Var}\left(Z(\theta+\Delta\theta) - Z(\theta)\right) = \text{Var}\left(G(\xi; \theta+\Delta\theta) - G(\xi, \theta)\right) \leq C\Delta\theta^2$$

  In this case, the **variance of the finite-difference approximation, given in (57), remains bounded as $\Delta\theta \to 0$.**

  Assuming that the bias is $O(\Delta\theta^h)$, the total mean-square error of the finite-difference approximation is given by

  $$O(\Delta\theta^h) + O(N^{-1/2})$$

  The maximum efficiency is achieved if

  $$\Delta\theta \sim N^{-\frac{1}{2h}},$$

  and the total **error becomes asymptotically the same as the standard MC error**:

  $$O\left(N^{-\frac{1}{2}}\right),$$

6.3 Pathwise approximation of derivatives

In some cases, it is possible to compute

$$Z'(\theta) = \lim_{\Delta\theta \to 0} \frac{Z(\theta + \Delta\theta) - Z(\theta)}{\Delta\theta},$$

for almost every random outcome (i.e. almost every path).

$Z'(\theta)$ is called a **pathwise derivative** of $Z(\theta)$.

Then, we can make use the following identity:

$$\alpha'(\theta) = \frac{d}{d\theta}\mathbb{E}Z(\theta) = \mathbb{E}Z'(\theta),$$

provided we can interchange the integration and differentiation.

Thm 6. *(Fubini) If there exists $\Delta\theta > 0$, s.t.*

$$\int_{\theta-\Delta\theta}^{\theta+\Delta\theta} \mathbb{E}\left|Z'(z)\right| dz < \infty,$$

then we can interchange the integration and differentiation:

$$\frac{d}{d\theta}\mathbb{E}Z(\theta) = \mathbb{E}Z'(\theta)$$

Now, instead of constructing $\bar{Z}(\theta)$, the MC approximation for $\mathbb{E}Z(\theta)$, we can **approximate** $\mathbb{E}Y(\theta)$**, with** $Y(\theta) = Z'(\theta)$ **directly**. The most straightforward way to do it is by using the **sample mean**:

$$\mathbb{E}Y(\theta) \approx \bar{Y}(\theta) = \frac{1}{N}\sum_{i=1}^{N} Y^{(i)},$$

where $\left\{Y^{(i)}\right\}$ are i.i.d. Law$(Y(\theta))$.

This method has several **advantages**, as compared to the finite-difference approximation.

- It is **unbiased**:
$$\mathbb{E}\bar{Y}(\theta) = \mathbb{E}Y(\theta) = \mathbb{E}Z'(\theta) = \alpha'(\theta)$$
- Assume that the simulation from Law$(Y(\theta))$ requires approximately the same computational effort as the simulation from Law$(Z(\theta))$. Then, the **pathwise approximation method is twice as fast as the finite-difference approximation**, since it only requires simulating at a single value of θ.

Ex 18. *Consider $V(S_t, t)$ – the price of a European call with strike K and maturity T in BS model. As before, we need to compute the partial derivative $\partial_S V(S_t, t) = \alpha'(S_t)$, where*

$$\alpha(s) = \mathbb{E}Z(s) = \mathbb{E}\left(e^{-r(T-t)}\left(S_{T-t}(s) - K\right)^{+}\right),$$

with

$$S_{T-t}(s) = s\exp\left((r-\sigma^2/2)(T-t) + \sigma\sqrt{T-t}\,\xi\right)$$

and s *being the level of risky asset at time* t*, and* $\xi \sim N(0,1)$*. Then*

$$Y(s) = Z'(s) = \frac{d}{ds}\left(e^{-r(T-t)}\left(S_{T-t}(s)-K\right)^+\right) = e^{-r(T-t)}\frac{S_{T-t}(s)}{s}\,\mathbf{1}_{[K,\infty)}\left(S_{T-t}(s)\right)$$

Thus, simulating i.i.d. standard normal $\left\{\xi^{(i)}\right\}$*, we use the above expression to obtain i.i.d.* $\left\{Y^{(i)}\right\}$ *sampled from Law*$(Y(s))$*. Finally,*

$$\partial_S V(s,t) = \alpha'(s) \approx \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y^{(i)}$$

It is possible to develop a general approach to **computing pathwise derivatives with respect to the initial values of the state process**. For simplicity, we consider the following one-dimensional diffusion equation:

$$dX_t = b(X_t,t)dt + a(X_t,t)dB_t \tag{58}$$

Denote by $X(x)$ the solution to the above SDE with initial condition $X_0(x) = x$. Assume that we need to estimate $\alpha'(x)$, where

$$\alpha(x) = \mathbb{E}F\left(X_T(x)\right),$$

and F is differentiable. According to the pathwise approach,

$$\alpha'(x) = \mathbb{E}\partial_x F\left(X_T(x)\right) = \mathbb{E}\left(F'\left(X_T(x)\right)\partial_x X_T(x)\right)$$

Finally, under some regularity assumptions on a and b, we can compute $\partial_x X_T(x)$ by passing the differentiation inside the stochastic integral. In other words,

$$\begin{aligned} d\left(\partial_x X_t(x)\right) &= \partial_x X_t(x)\partial_X b(X_t(x),t)dt + \partial_x X_t(x)\partial_X a(X_t(x),t)dB_t \\ &= \tilde{b}(\partial_x X_t(x), X_t(x), t)dt + \tilde{a}(\partial_x X_t(x), X_t(x), t)dB_t, \qquad \partial_x X_0(x) = 1 \end{aligned} \tag{59}$$

Finally, we solve the system of equations (58), (59) numerically, by choosing the appropriate method (e.g. Euler scheme). This allows us to generate samples of $(X_T(x), \partial_x X_T(x))$ and use them to approximate $\alpha'(x)$.

This approach can be extended to multi-dimensional diffusions X in a straight forward way.

Rem 18. *The above method can be viewed as a practical interpretation of the abstract* **Mallivian calculus***, or its Markovian version known as the theory of* **Stochastic Flows** *(due to Kunita).*

Ex 19. *Consider* $V(S_t, Y_t, t)$ *– the price of a European call with strike* K *and maturity* T *in Heston model*

$$\begin{cases} dS_t(s,y) = \mu S_t(s,y)dt + \sqrt{Y_t(s,y)}S_t(s,y)dB^1_t, \quad S_0(s,y) = s, \\ dY_t(s,y) = \kappa(\theta - Y_t(s,y))dt + \sigma\sqrt{Y_t(s,y)}(\rho dB^1_t + \sqrt{1-\rho^2}dB^2_t), \quad Y_0(s,y) = y \end{cases}$$

Notice that

$$d\left(\partial_s Y_t(s,y)\right) = 0, \quad \partial_s Y_0(s,y) = 0,$$

hence, $Y_t(s,y)$, in fact, does not depend on s, and we will suppress this argument. According to the general formula, we introduce

$$S_t^1(s,y) = \partial_s S_t(s,y), \qquad S_t^2(s,y) = \partial_y S_t(s,y), \qquad Y_t^2(s,y) = \partial_y Y_t(s,y),$$

and obtain

$$\begin{cases} dS_t = \mu S_t dt + \sqrt{Y_t} S_t dB_t^1, \quad S_0 = s, \\ dY_t = \kappa(\theta - Y_t)dt + \sigma\sqrt{Y_t}(\rho dB_t^1 + \sqrt{1-\rho^2} dB_t^2), \quad Y_0 = y \\ dS_t^1 = \mu S_t^1 dt + \sqrt{Y_t} S_t^1 dB_t^1, \quad S_0^1 = 1, \\ dS_t^2 = \mu S_t^2 dt + \left(\sqrt{Y_t} S_t^2 + \frac{Y_t^2}{2\sqrt{Y_t}} S_t\right) dB_t^1, \quad S_0^2 = 0, \\ dY_t^2 = -\kappa Y_t^2 dt + \sigma \frac{Y_t^2}{2\sqrt{Y_t}}(\rho dB_t^1 + \sqrt{1-\rho^2} dB_t^2), \quad Y_0^2 = 1 \end{cases}$$

Then,

$$V(s,y,t) = \mathbb{E}Z(s,y) = \mathbb{E}\left(e^{-r(T-t)}\left(S_{T-t}(s,y) - K\right)^+\right),$$

$$\partial_s Z(s,y) = e^{-r(T-t)} S_{T-t}^1 \mathbf{1}_{[K,\infty)}\left(S_{T-t}\right), \qquad \partial_y Z(s,y) = e^{-r(T-t)} S_{T-t}^2 \mathbf{1}_{[K,\infty)}\left(S_{T-t}\right),$$

and

$$\partial_s V(s,y,t) = \mathbb{E}\partial_s Z(s,y) = e^{-r(T-t)}\mathbb{E}\left(S_{T-t}^1 \mathbf{1}_{[K,\infty)}\left(S_{T-t}\right)\right),$$

$$\partial_y V(s,y,t) = \mathbb{E}\partial_y Z(s,y) = e^{-r(T-t)}\mathbb{E}\left(S_{T-t}^2 \mathbf{1}_{[K,\infty)}\left(S_{T-t}\right)\right)$$

We can sample from the (approximate) joint distribution of $(S_{T-t}^1, S_{T-t}^2, S_{T-t})$, for example, by applying Euler scheme to the above system of SDEs.