CHAPTER II - MONTE CARLO METHODS

JOSEPH G. CONLON

1. The Basic Method

The Monte-Carlo (MC) method for numerically estimating certain quantities is based on the *law of averages* 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*.

Let X be a random variable with finite mean and variance. We have that

$$\text{Var}[X] \;=\; E[\,(X - E[X])^2\,] \;=\; E[X^2] - (E[X])^2\;. \tag{1.1}$$

Since variance is non-negative it follows from (1.1) that $(E[X])^2 \leq E[X^2]$, with equality only if X is *deterministic* i.e. X takes a single value with probability 1. Let $X_1, X_2, ..$, be i.i.d. random variables with the same distribution as X in (1.1). The strong law of large numbers then tells us that

$$\lim_{N\to\infty} \frac{X_1 + \cdots + X_N}{N} \;=\; E[X] \quad \text{with probability 1.} \tag{1.2}$$

If we know how to generate large numbers of independent copies of the variable X by say using a random number generator on a computer, then (1.2) enables us to numerically estimate $E[X]$. Thus there are two ingredients to the MC method:
(a) Express the quantity we wish to estimate as the expectation $E[X]$ of some random variable X.
(b) Create a random number generator which efficiently generates large numbers of approximately independent variables with distribution the same as X.

Now **(b)** is the "black box" which is essential to any MC method. In practice one creates random number generators for the variable X, which is *uniformly distributed* on the interval $0 < X < 1$. These generally work by an algorithm of the form $X_{n+1} = F(X_n)$, where $F : [0,1] \to [0,1]$ is an arithmetic function which can be accurately computed in a short number of steps. Evidently the sequence $X_1, X_2, ..$, is *deterministic*, but in good random number generators the variable $j \to X_j,\ j = 1, 2, ...$, is almost independent of the variable $j \to X_{j+1}, j = 1, 2, ..$, in the *sample probability space*. That is suppose we sample N values of X from the random number generator for some large N. Our probability space is now $\Omega_N = \{1, 2, .., N\}$ and the probability distribution we put on Ω_N is the uniform distribution. Then the variable $j \to X_j,\ j = 1, 2, ...$, is almost independent of the variable $j \to X_{j+1}, j = 1, 2, ..$, on Ω_N. One can never get complete independence even theoretically in the limit $N \to \infty$, because the functions $F : [0,1] \to [0,1]$ are always *periodic*. That is there exists N_0 such that $X_{N_0} = X_0$ for all X_0. However for good random number generators we shall have $N_0 >> N$. For example we typically have $N \simeq 10^7$ whereas $N_0 \simeq 10^{20}$, and that gives very good approximate

independence. In MATLAB the command rand(m,n) generates an $m \times n$ matrix with entries which are independently generated values of the uniform variable. The command randn(m,n) generates an $m \times n$ matrix with entries which are independently generated values of the standard normal variable. The operation randn works by first implementing rand to generate values of i.i.d. uniform variables, and then using a deterministic algorithm to turn these values into values of i.i.d. standard normal variables. Because of our limitation on possible random number generators, we shall be restricted to estimating $E[X]$ in **(a)** using MC only for variables X of the form $X = g(Y^1, ..., Y^k)$, where $Y^1, .., Y^k$ are either i.i.d. uniform or normal variables.

An important issue in numerical analysis is always to get an idea of how many computations are required to estimate a desired quantity to a given degree of accuracy. The answer to this question for the MC method is provided by the *central limit theorem:*

Theorem 1.1. *Suppose a random variable X has finite mean μ and variance σ^2. Let $X_1, X_2, ..,$ be i.i.d. copies of X. Then*

(1.3)
$$Z_N = \frac{\sqrt{N}}{\sigma}\left[\frac{X_1 + \cdots + X_N}{N} - \mu\right] \quad \text{converges in distribution as } N \to \infty \text{ to } Z,$$

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

$$\rho(z) \;=\; \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty. \tag{1.4}$$

Now convergence in distribution means that for any $a \in \mathbf{R}$ then $\lim_{N\to\infty} P(|Z_N| > a) = P(|Z| > a)$. We have already observed in Chapter I that $P(|Z| > 3) < .003$. Hence Theorem 1.1 implies that for large N then

$$P\left(\left|\frac{X_1 + \cdots + X_N}{N} - \mu\right| > \frac{3\sigma}{\sqrt{N}}\right) \;<\; .003. \tag{1.5}$$

We can write (1.5) alternatively as

(1.6)
$$\frac{X_1 + \cdots + X_N}{N} - \mu \;=\; \text{Error}(N), \text{ where Error}(N) < \frac{3\sigma}{\sqrt{N}} \text{ with high probability.}$$

Hence the error in the MC method is proportional to the inverse square root of the number of simulations. We conclude that MC enables us to estimate $E[X]$ with $\simeq N$ computations up to an accuracy $\simeq 1/\sqrt{N}$.

To implement the MC method for estimating $E[X]$ we proceed as follows:

(a) Generate $N \simeq 10^7$ values $X_1, .., X_N$ of i.i.d. variables with the same distribution as X.
(b) Compute the averages $\hat{X}_n,\ n = 1, 2, ..N$ given by

$$\hat{X}_n \;=\; \frac{1}{n}\sum_{j=1}^{n} X_j \;, \tag{1.7}$$

and graph the *convergence diagram* $n \to \hat{X}_n,\ n = 1, .., N$.
(c) Compute the sample variance $\hat{\sigma}_N^2$ defined by

$$\hat{\sigma}_N^2 \ = \ \frac{1}{N}\sum_{j=1}^{N}[X_j - \hat{X}_N]^2 \ . \tag{1.8}$$

(d) Report the estimated value $\hat{X}_N$ for $E[X]$ and the *standard error* $\varepsilon_N = \hat{\sigma}_N/\sqrt{N}$. In addition report the *proportional error* which is $\varepsilon_N/\hat{X}_N$.

Remark 1. *Note that the sample variance (1.8) 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 (1.8) by $1/(N-1)$.*

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

$$\int_0^1 \cdots \int_0^1 \Phi(x^1, .., x^d)\ dx^1 \cdots dx^d \ = \ E[\Phi(X^1, .., X^d)] \ , \tag{1.9}$$

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 estimate

$$\int_{[0,1]^d} \Phi(x)\ dx \ \simeq \ \frac{1}{N}\left[\Phi(x_1) + \cdots \Phi(x_N)\right] \ . \tag{1.10}$$

If Q_j is the subcube with center x_j then

$$\left|\int_{Q_j} \Phi(x)\ dx - \frac{\Phi(x_j)}{N}\right| \ \leq \ \int_{Q_j} |\Phi(x) - \Phi(x_j)|\ dx \ \leq \ \frac{\sup_{x\in Q_j}|\Phi(x) - \Phi(x_j)|}{N} \ . \tag{1.11}$$

If the function $\Phi(\cdot)$ is differentiable then $|\Phi(x) - \Phi(x_j)| \leq C|x - x_j|$ for a constant C. Since the volume of Q_j is $1/N$ the length of an edge of Q_j is $\simeq 1/N^{1/d}$. We conclude from (1.11) that

$$\left|\int_{[0,1]^d} \Phi(x)\ dx - \frac{1}{N}\left[\Phi(x_1) + \cdots \Phi(x_N)\right]\right| \ \leq \ \frac{C}{N^{1/d}} \tag{1.12}$$

for some constant C. In the MC method we have

$$\int_{[0,1]^d} \Phi(x)\ dx \ \simeq \ \frac{1}{N}\left[\Phi(X_1) + \cdots \Phi(X_N)\right] \ , \tag{1.13}$$

where $X_j = (X_j^1, .., X_j^d),\ j = 1, 2, ..$, and each $X_j^r,\ 1 \leq r \leq d, j = 1, 2, ..$, are i.i.d. uniform on $[0,1]$. The error is $\simeq 1/\sqrt{N}$ and hence is worse than the Riemann sum method if $d = 1, 2$ but is better if $d \geq 3$. We see then that the MC method has a big advantage over deterministic methods when we wish to estimate *high dimensional* integrals.

Example 1. *We can use MC to estimate the value of the standard put option. We have already seen in §2 of Chapter I that the BS price of the put option is given by the expectation*

$$V(S_0,0) \;=\; e^{-rT}E\left[\max[K-S(T),0]\;\middle|\;S(0)=S_0\right]\;. \tag{1.14}$$

The variable $S(T)$ is log normal and given by the formula

$$S(T) \;=\; S_0\exp\left\{(r-\sigma^2/2)T+\sigma\sqrt{T}\xi\right\}\;, \tag{1.15}$$

where ξ is standard normal. In our MC simulation we generate using randn N values $\xi_1,..,\xi_N$ of i.i.d. standard normal variables and set

$$X_j \;=\; e^{-rT}\max\left[K-S_0\exp\left\{(r-\sigma^2/2)T+\sigma\sqrt{T}\xi_j\right\},0\right]\;,\quad j=1,..,N. \tag{1.16}$$

Then we proceed as in the MC procedure **(a)**-**(d)** *already given.*

2. Numerical Methods for the Solution of Stochastic Differential Equations (SDE)

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 $b : \mathbf{R}\times\mathbf{R}\to\mathbf{R}$ and $\sigma:\mathbf{R}\times\mathbf{R}\to\mathbf{R}^+$ are given functions we associate to them the SDE

$$\frac{dY(t)}{dt} \;=\; b(Y(t),t)+\sigma(Y(t),t)W(t),\quad \text{where } W(t),\ t\in\mathbf{R},\ \text{is the white noise process.} \tag{2.1}$$

The effect of the process $W(\cdot)$ is to give the particle with position $X(\cdot)$ a random "kick" at each time t with the kicks being i.i.d. for different times. To make this precise let $\xi_j, j=0,\pm1,\pm2,..$, be an infinite set of i.i.d. standard normal variables. Then for $\Delta t>0$ we have

$$\int_{j\Delta t}^{(j+1)\Delta t} W(s)ds \;=\; \sqrt{\Delta t}\;\xi_j\;. \tag{2.2}$$

The actual white noise process is a limit as $\Delta t\to 0$ of the process defined by (2.2). Another way of thinking about the white noise process is as the "derivative" of Brownian motion. If $B(t),\ t\geq 0$, is BM then a typical path of $B(\cdot)$ is very wiggly and in fact not differentiable in the sense that the pointwise derivative

$$\frac{dB(t)}{dt} \;=\; \lim_{\Delta t\to 0}\frac{B(t+\Delta t)-B(t)}{\Delta t} \tag{2.3}$$

never exists with probability 1. We can however write (2.2) alternatively as

$$\int_{j\Delta t}^{((j+1)\Delta t} W(s)ds \;=\; B((j+1)\Delta t)-B(j\Delta t),\quad j=0,1,... \tag{2.4}$$

Taking $\Delta t\to 0$ in (2.4) we have formally the equation $dB(t)/dt = W(t)$. This however needs to be interpreted not in the classical sense (2.3) but in a generalized sense. We therefore can write (2.1) in differential form as

$$dY(t) \;=\; b(Y(t),t)dt+\sigma(Y(t),t)dB(t)\;, \tag{2.5}$$

which is the form we used in Chapter I. Just as for ODEs the evolution equation can be solved uniquely for $t > 0$ with given initial data $Y(0) = Y^0$ given. Evidently Y^0 can be an arbitrary *random variable*, which as a special case could be chosen deterministically as $Y^0 \equiv y^0 \in \mathbf{R}$ with probability 1.

To numerically solve (2.5) with given initial data we use the explicit Euler method as we did for ODEs. Thus on setting $Y^m \simeq Y(m\Delta t)$, then (2.5) yields the recurrence relation

$$Y^{m+1} \;=\; Y^m + b(Y^m, m\Delta t)\Delta t + \sigma(Y^m, m\Delta t)\sqrt{\Delta t}\;\xi_m, \quad m = 0, 1, 2, .., \tag{2.6}$$

where $\xi_0, \xi_1, ..,$ are i.i.d. standard normal. Suppose now we wish to estimate the expectation $E[\Phi(Y(T)) \mid Y(0) = Y^0]$ for some function $\Phi : \mathbf{R} \to \mathbf{R}$. Then we choose an integer M such that $M\Delta t = T$ and use randn to generate $\xi_0, .., \xi_{M-1}$. Solving the recurrence (2.6) with initial data Y^0 gives us one MC realization of $Y^M \simeq Y(T)$. We can do this N times yielding N independent simulations $Y_1^M, .., Y_N^M$ of the random variable $Y(T)$. Then we set

$$E[\Phi(Y(T)) \mid Y(0) = Y^0] \;\simeq\; \left[\Phi(Y_1^M) + \cdots + \Phi(Y_N^M)\right]/N \;. \tag{2.7}$$

Observe that there are two sources of error in the estimate (2.7). There is the MC error as we have seen before which is $O(1/\sqrt{N})$. There is also the discretization error which depends on Δt and disappears as $\Delta t \to 0$. Now the amount of computation required to get the estimate (2.7) is $\simeq MN$, and so we should ask that for a given value of MN say $MN = 10^7$, how we should choose M and N to minimize the error. In the Euler method for ODEs we saw the discretization error is $O(\Delta t) \simeq 1/M$. If the same holds for the SDE then we should choose M so that discretization error is comparable to MC error i.e. $1/M \simeq 1/\sqrt{N}$ or $N = M^2$. In actual fact the discretization error in (2.6) is $O(\sqrt{\Delta t}\;) \simeq 1/\sqrt{M}$, whence we should expect to choose $M \simeq N$ in order to minimize error.

We write the Euler method (2.6) for the SDE (2.5) as

$$Y(t + \Delta t) \;=\; Y(t) + b(Y(t), t)\Delta t + \sigma(Y(t), t)[B(t + \Delta t) - B(t)] \;. \tag{2.8}$$

We can see from this why the discretization error for the SDE 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. In the algorithm (2.8) the coefficient of $B(t+\Delta t)-B(t)$ is $\sigma(Y(t), t)$, but this obviously carries an error comparable to $\sigma(Y(t+\Delta t), t+\Delta t)-\sigma(Y(t), t)$, which is $O(\Delta t)$ or worse. The *truncation error* for (2.8) then is at least $O[(\Delta t)^{3/2}]$, whence the cumulative error is $O(\sqrt{\Delta t})$. We can reduce the truncation error by improving the approximation

$$\int_t^{t+\Delta t} \sigma(Y(s), s)\; dB(s) \;\simeq\; \sigma(Y(t), t)[B(t + \Delta t) - B(t)] \;, \tag{2.9}$$

which is used in (2.8). From (2.5) we have that for $t < s < t + \Delta t$,

$$Y(s) - Y(t) \;=\; \int_t^s dY(s') \;\simeq\; \sigma(Y(t), t)[B(s) - B(t)] \;. \tag{2.10}$$

In this approximation then

(2.11)

$$\sigma(Y(s), s) \;\simeq\; \sigma(Y(t)+\sigma(Y(t), t)[B(s)-B(t)], t) \;\simeq\; \sigma(Y(t), t)+\frac{\partial\sigma(Y(t), t)}{\partial y}[B(s)-B(t)] \;,$$

upon doing the Taylor expansion of $\sigma(y,t)$ in y about $y = Y(t)$. Now from (2.5), (2.11) we have that

(2.12)

$$Y(t+\Delta t) - Y(t) \;=\; \int_t^{t+\Delta t} dY(s) \;\simeq\; b(Y(t),t)\Delta t + \int_t^{t+\Delta t} \sigma(Y(s),s) dB(s)$$

$$\simeq\; b(Y(t),t)\Delta t + \sigma(Y(t),t)[B(t+\Delta t) - B(t)] + \sigma(Y(t),t)\frac{\partial \sigma(Y(t),t)}{\partial y}\int_t^{t+\Delta t}[B(s) - B(t)]dB(s)\;.$$

Using the Ito calculus we have that

$$\int_t^{t+\Delta t}[B(s) - B(t)]dB(s) \;=\; \frac{1}{2}[B(t+\Delta t) - B(t)]^2 - \frac{\Delta t}{2}\;. \tag{2.13}$$

If we write now $B(t+\Delta t) - B(t) = \sqrt{\Delta t}\;\xi$ where ξ is standard normal, then the algorithm (2.12) is the same as

$$\begin{aligned} Y(t+\Delta t) \;=\; & Y(t) + b(Y(t),t)\Delta t \\ & + \sigma(Y(t),t)\sqrt{\Delta t}\;\xi + \sigma(Y(t),t)\frac{\partial \sigma(Y(t),t)}{\partial y}\frac{\Delta t}{2}[\xi^2 - 1]\;. \end{aligned} \tag{2.14}$$

Evidently (2.14) is a refinement of the basic Euler algorithm (2.6) and is known as Milstein's algorithm. Generally we expect the discretization error in (2.14) to be $O(\Delta t)$. 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 algorithm.

We already observed in Chapter I that for geometric Brownian motion $S(t)$ which is a solution to the SDE

$$dS(t) \;=\; S(t)[rdt + \sigma dB(t)]\;, \tag{2.15}$$

we have that

$$S(t+\Delta t) \;=\; S(t)\exp[(r - \sigma^2/2)\Delta t + \sigma\sqrt{\Delta t}\;\xi]\;, \tag{2.16}$$

where ξ is standard normal. We can expand the exponential in (2.16) in its series expansion

$$e^z \;=\; 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \cdots\;, \tag{2.17}$$

to obtain increasingly accurate approximations for $S(t+\Delta t)$. If we keep just the first 2 terms in the expansion (2.17) we have from (2.16) that

$$S(t+\Delta t) \;=\; S(t) + (r - \sigma^2/2)S(t)\Delta t + \sigma S(t)\sqrt{\Delta t}\;\xi\;, \tag{2.18}$$

which we can compare to the Euler algorithm (2.6),

$$S(t+\Delta t) \;=\; S(t) + rS(t)\Delta t + \sigma S(t)\sqrt{\Delta t}\;\xi\;. \tag{2.19}$$

If we keep the first 3 terms in the expansion (2.17) we have

(2.20)

$$S(t+\Delta t) \;=\; S(t) + [(r-\sigma^2/2)S(t)\Delta t + \sigma S(t)\sqrt{\Delta t}\;\xi] + \frac{S(t)}{2}[(r-\sigma^2/2)\Delta t + \sigma\sqrt{\Delta t}\;\xi]^2\;.$$

We can rewrite (2.20) as

$$S(t+\Delta t) \;=\; S(t) + rS(t)\Delta t + \sigma S(t)\sqrt{\Delta t}\;\xi + \sigma^2 S(t)\frac{\Delta t}{2}[\xi^2 - 1] + O[(\Delta t)^{3/2}], \tag{2.21}$$

which gives us the Milstein algorithm for solving (2.15).

We can also see from (2.21) why the Euler algorithm (2.19) is a correct numerical scheme for solving the SDE (2.15) whereas the algorithm (2.18) is incorrect. The reason is that the principal correction term to the Euler algorithm involves the variable $\xi^2 - 1$ which has mean 0. Now the cumulative error will involve a sum of $\simeq 1/\Delta t$ i.i.d. such variables, and hence by the CLT this sum will be $\simeq 1/\sqrt{\Delta t}$. Hence the cumulative error in the Euler method is $\simeq \Delta t/\sqrt{\Delta t} = O(\sqrt{\Delta t}\ )$. By contrast the error in the algorithm (2.18) is $\Delta t/\Delta t = O(1)$ and so does not vanish as $\Delta t \to 0$. It is clear then that there are some extra subtleties involved with algorithms for solving SDEs beyond those which occur in algorithms for solving ODEs.

As an application of the above we consider the Heston stochastic volatility model, given by

$$dY(t) \quad = \quad [\theta - \kappa Y(t)]dt + \beta\sqrt{Y(t)}\ dB(t)\ , \tag{2.22}$$

$$dS(t) \quad = \quad S(t)\left[rdt + \sqrt{Y(t)}\{\rho\ dB(t) + \sqrt{1-\rho^2}\ dZ(t)\}\right]\ . \tag{2.23}$$

In (2.22), (2.23) the parameters θ, κ, β are positive and the parameter ρ satisfies $-1 \le \rho \le 1$. The processes $B(\cdot)$ and $Z(\cdot)$ are independent Brownian motions. We have that

$$\text{Var}\left[\int_t^{t+\Delta t} \rho\ dB(s) + \sqrt{1-\rho^2}\ dZ(s)\right] \ = \ \rho^2\Delta t + (1-\rho^2)\Delta t = \Delta t, \tag{2.24}$$

so the process $\tilde{B}(t) = \rho\ B(t) + \sqrt{1-\rho^2}\ Z(t),\ t \ge 0$, is also a copy of Brownian motion. Hence the stock volatility is $\sqrt{Y(t)}$, which is stochastic since $Y(t)$ evolves by the SDE (2.22).

The set of equations (2.22), (2.23) can be solved uniquely for $t > 0$ with given initial conditions $Y(0) = Y^0, S(0) = S^0$. The value of a call option today with expiration T and strike price K is as in Chapter I given by the formula

$$V(S^0, 0) \ = \ e^{-rT}E\left[\max\{S(T) - K, 0\} \mid S(0) = S^0, Y(0) = Y^0\right]\ . \tag{2.25}$$

Now today's stock price S^0 is easily observed, but it is not so clear how to assign the initial condition Y^0 for the volatility process (2.22). To see what this should be we note that the process is *mean reverting*. This means that $|Y(t)|$ is bounded for all $t \ge 0$ with probability 1, in contrast to the stock price $S(t)$ which tends to grow exponentially. To see this we consider the deterministic situation $\beta = 0$ with initial data y^0, which can be easily solved to obtain

$$y(t) \ = \ e^{-\kappa t}y^0 + \theta[1 - e^{-\kappa t}]/\kappa\ . \tag{2.26}$$

It follows from (2.26) that $y(t)$ converges exponentially fast in time to the value $y^\infty = \theta/\kappa$. In the stochastic case $\beta > 0$ we expect something similar to happen but now the random variable $Y(t)$ converges *in distribution* to a random variable Y^∞, the so called *invariant measure* for the SDE (2.22), which for β small should be concentrated around θ/κ. Hence it makes sense to take $Y(0) = \theta/\kappa$ in (2.25).

To estimate the value (2.25) of the call option using the MC method, we therefore choose an integer M so that $M\Delta t = T$ and use the Euler method to solve (2.22), (2.23) with initial conditions $Y(0) = Y^0 = \theta/\kappa,\ S(0) = S^0$. Thus as in (2.6) we define $Y^m, S^m,\ m = 0, .., M$ by the recurrence

$$Y^{m+1} \quad = \quad Y^m + \ [\theta - \kappa Y^m]\Delta t + \beta\sqrt{Y^m}\sqrt{\Delta t}\ \xi_m\ , \tag{2.27}$$

$$S^{m+1} \quad = \quad S^m + S^m \left[r\Delta t + \sqrt{Y^m}\sqrt{\Delta t}\ \{\rho\ \xi_m + \sqrt{1-\rho^2}\ \eta_m\} \right] \ , \tag{2.28}$$

where the $\xi_m,\ \eta_m,\ m = 0, ..., M-1$ are samplings of i.i.d. standard normal variables. The corresponding value of the call option is from (2.25)

$$\text{value of option } = \ e^{-rT} \max\{S^M - K, 0\} \ . \tag{2.29}$$

We then need to do $N >> 1$ independent samples of this procedure and average over the corresponding values of (2.29) to estimate the MC value of the option.

We need to restrict the parameters θ, β in (2.22) to satisfy $\theta > \beta^2/2$ since if this inequality is violated the solution $Y(t)$ becomes 0 at some random time $\tau > 0$ with probability 1, and then $Y(t) = 0$ for all $t \geq \tau$. Thus the stock eventually has zero volatility, which is the trivial situation of the value of the equity increasing at the risk free rate r. We can see heuristically why if $\theta < \beta^2/2$ then $Y(t)$ becomes zero eventually and remains there for all later times. Suppose that $Y(t) = y > 0$ where y is small, and consider how much time is needed for $Y(s),\ s > t$, to decrease to $y/2$ with probability which is $\simeq 1/2$. From(2.22) we have that

$$-y/2 \ \simeq \theta\Delta t - \beta\sqrt{y\Delta t} \ . \tag{2.30}$$

Now (2.30) is a quadratic in $z = \sqrt{\Delta t}$ given by the equation

$$\theta z^2 - \beta\sqrt{y}z + y/2 = 0, \quad \text{with real solution if } (\beta\sqrt{y})^2 > 4\theta y/2 \ . \tag{2.31}$$

Hence if $\theta < \beta^2/2$ there is a probability $\simeq 1/2$ that $Y(t+\Delta t) = y/2$, where $\Delta t \simeq y$. It is easy to see now that the probability that $Y(t+\Delta t) \leq y/2$ for some $\Delta t \simeq y$ is strictly large than $1/2$ if $\theta < \beta^2/2$. Hence $Y(\cdot)$ with probability 1 hits 0 at some random time τ and then remains there.

The Heston model can be used to explain to some degree the so called volatility "smiles" which are observed in the market. Thus consider quoted options on a particular stock with a range of strike prices K and expiration dates T. These options have observed prices $V(K,T)$, and so we can assign to the underlying stock *implied volatilities* $\sigma(K,T)$ defined by

$$V(K,T) \ = \ \text{Black Scholes price of the option with strike } K, \text{ expiration } T, \text{ and volatility } \sigma = \sigma(K,T) \ . \tag{2.32}$$

Now if the BS theory perfectly matched the market then $\sigma(K,T)$ would be a constant function of K and T. In practice the function $K \to \sigma(K,T)$ is non-constant. For stock options it tends to be a *decreasing function* of K. For foreign currency options it tends to be convex with a minimum when K equals today's currency price, so for the *at the money* option. The qualitative difference in these behaviors can be explained by the value of the coefficient of correlation ρ in (2.23). Now stock price and stock volatility are *positively correlated* if $\rho > 0$ and *negatively correlated* if $\rho < 0$. We expect $\rho < 0$ for stocks since as the price of the stock increases players withdraw from the market, which tends to drive down volatility. Correspondingly as prices decrease stock volatility tends to increase. In contrast for foreign currency we expect $\rho \simeq 0$, in other words little correlation between currency price and currency volatility. The reason for this is that there is more of an even balance between buyers and sellers of the currency than in the case of a stock. If say the British pound increases in value with respect to the US dollar, then dollars have become cheaper in Britain, which encourages *increased* buying of dollars in Britain. If the

British pound decreases in value with respect to the US dollar, then pounds have become cheaper in the US, which encourages *increased* buying of pounds in the US. The Heston model for different ρ does produce "smile" curves, which are qualitatively similar to the curves observed in the market. However there is a significant *quantitative* difference in the effect. The observed market effect is much larger than the effect produced in the Heston model.

3. Correlation

We have already discussed correlation in the context of the Heston model. There are many ways of measuring the correlation of 2 random variables X and Y, but the simplest way is through the covariance $\text{cov}[X,Y]$ defined by

$$\text{cov}[X,Y] \;=\; E\left[\{X-E[X]\}\{Y-E[Y]\}\right] \;=\; E[XY]-E[X]E[Y]\;. \tag{3.1}$$

Evidently the variance of a random variable X is just the covariance of X and X, so $\text{var}[X]=\text{cov}[X,X]$. Observe now from (3.1) that upon using the Schwarz inequality we have that

$$|\text{cov}[X,Y]| \;\le\; E\left[\{X-E[X]\}^2\right]^{1/2} E\left[\{Y-E[Y]\}^2\right]^{1/2} \;=\; \sqrt{\text{var}[X]\text{var}[Y]}\;. \tag{3.2}$$

From (3.2) the *coefficient of correlation* $\rho(X,Y)$ between X,Y defined by

$$\rho(X,Y) \;=\; \text{cov}[X,Y]\big/\sqrt{\text{var}[X]\text{var}[Y]}\;, \tag{3.3}$$

satisfies the inequality $-1\le\rho(X,Y)\le 1$. If $\rho(X,Y)=+1$ then $X=\lambda Y$ for some scalar $\lambda>0$. If $\rho(X,Y)=-1$ then $X=-\lambda Y$ for some scalar $\lambda>0$. Thus if $|\rho(X,Y)|=1$ 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 $\rho(X,Y)=0$. It *does not* however follow that $\rho(X,Y)=0$ implies X,Y independent. Nevertheless one tends to assume that if $|\rho(X,Y)|<<1$ then X,Y are only weakly correlated, so are close to being independent.

We can apply these considerations to the Heston model. The volatility process $Y(\cdot)$ for the Heston model is driven by the Brownian motion $B(\cdot)$ in (2.22), whereas the stock process $S(\cdot)$ is driven by the Brownian motion $\tilde{B}(\cdot)=\rho B(\cdot)+\sqrt{1-\rho^2}Z(\cdot)$ in (2.23), where $Z(\cdot)$ is a BM independent of $B(\cdot)$. We can compute the coefficient of correlation between increments of $B(\cdot)$ and increments of $\tilde{B}(\cdot)$. Thus we have

$$\text{cov}[\{B(t+\Delta t)-B(t)\},\{\tilde{B}(t+\Delta t)-\tilde{B}(t)\}] \;=\; \rho E[\{B(t+\Delta t)-B(t)\}^2] \;=\; \rho\Delta t\;. \tag{3.4}$$

We also have that

$$\text{var}[B(t+\Delta t)-B(t)] \;=\; \text{var}[\tilde{B}(t+\Delta t)-\tilde{B}(t)] \;=\; \Delta t\;. \tag{3.5}$$

It follows from (3.3)-(3.5) that the coefficient of correlation is ρ, so ρ is the coefficient of correlation between the BM driving the stock price and the BM driving the stock volatility.

Correlation ideas are important when it comes to valuing basket options, which are options with a payoff depending on the behavior of several stocks. Consider the case of an option with value depending on the price of d stocks, where $S_1(t),...,S_d(t)$ are the prices of the stocks at time $t\ge 0$. The *arithmetically averaged* basket call option with expiration date $T>0$ and strike price K has payoff given by

$$\text{payoff} \;=\; \max\left[\frac{S_1(T)+\cdots+S_d(T)}{d}-K,0\right]\;. \tag{3.6}$$

The corresponding *geometrically averaged* option has payoff given by

$$\text{payoff} \;=\; \max\Big[\{S_1(T)\cdots S_d(T)\}^{1/d} - K, 0\Big] \;. \tag{3.7}$$

Now a well known inequality is that the geometric mean is smaller than the arithmetic mean, so if $a_1, .., a_d$ are any d positive numbers then

$$\{a_1 \cdots a_d\}^{1/d} \;\leq\; [a_1 + \cdots + a_d]/d \;. \tag{3.8}$$

It is not hard to see why (3.8) holds since it is a consequence of the Jensen inequality

$$E[f(X)] \;\leq f(E[X]) \quad \text{for any concave function } f : \mathbf{R} \to \mathbf{R} \text{ and random variable } X. \tag{3.9}$$

Evidently (3.8) follows from (3.9) by taking the logarithm of (3.8) and using $f(x) = \log x,\ x > 0$, in (3.9). It follows from (3.8) that the value of the arithmetic option is larger than the value of the geometric option.

To estimate the value of basket options on the d stocks, we model the evolution of their prices $S_1(t), ..., S_d(t)$ by geometric BM so

$$\begin{cases} \frac{dS_1(t)}{S_1(t)} \;=\; r\ dt + \sigma_1\ dB_1(t), \\ \frac{dS_2(t)}{S_2(t)} \;=\; r\ dt + \sigma_2\ dB_2(t), \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \frac{dS_d(t)}{S_d(t)} \;=\; r\ dt + \sigma_d\ dB_d(t), \end{cases} \tag{3.10}$$

where $B_1(\cdot), ..., B_d(\cdot)$ are d (in general) correlated Brownian motions. The volatilities of the d stocks are given by $\sigma_1, ..., \sigma_d$. The correlations between the d Brownian motions is determined by a $d \times d$ *covariance matrix* $\boldsymbol{\rho} = [\rho_{i,j}]$. The entries $\rho_{i,j}$ of the matrix $\boldsymbol{\rho}$ are defined by

$$\rho_{i,j}\Delta t \;=\; \text{cov}[\{B_i(t+\Delta t) - B_i(t)\}, \{B_j(t+\Delta t) - B_j(t)\}] \;, \quad 1 \leq i, j \leq d. \tag{3.11}$$

Evidently $\rho_{i,j} = \rho_{j,i}$, $|\rho_{i,j}| \leq 1$ and $\rho_{i,i} = 1$ for $1 \leq i, j \leq d$, so $\boldsymbol{\rho}$ is a symmetric matrix with diagonal entries all equal to 1. The matrix $\boldsymbol{\rho}$ is also *non-negative definite*. This means that

$$\sum_{1 \leq i,j \leq d} \xi_i \rho_{i,j} \xi_j \;\geq\; 0 \quad \text{for all vectors } \xi = [\xi_1, .., \xi_d] \in \mathbf{R}^d \;. \tag{3.12}$$

To see why (3.12) holds we observe from (3.11) that

$$\Delta t \sum_{1 \leq i,j \leq d} \xi_i \rho_{i,j} \xi_j \;=\; E\left[\ \left\{\sum_{i=1}^{d} \xi_i [B_i(t+\Delta t) - B_i(t)]\right\}^2\ \right] \;\geq\; 0 \;. \tag{3.13}$$

Now one way of exhibiting symmetric non-negative definite $d \times d$ matrices is by choosing any $d \times d$ matrix A and setting $\boldsymbol{\rho} = A^T A$, where A^T is the transpose of A. It turns out that the converse is also true, namely that for any non-negative definite $d \times d$ matrix $\boldsymbol{\rho}$ there exists a $d \times d$ matrix A such that $\boldsymbol{\rho} = A^T A$. The matrix A is not uniquely determined. In fact if A is such a matrix then for any orthogonal matrix O i.e. $OO^T = O^T O = I_d$, the matrix OAO^T also works. One can specify a unique matrix A by requiring additional conditions on A. These conditions are:

Cholesky Decomposition : A is lower triangular with diagonal entries all nonnegative. (3.14)

Hence $A = [a_{i,j}]$ with $a_{i,j} = 0$ for $1 \leq i < j \leq d$, and $a_{i,i} \geq 0$ for $1 \leq i \leq d$. One can use the MATLAB function "chol" to find the Cholesky matrix A corresponding

to $\boldsymbol{\rho}$. Now chol($\boldsymbol{\rho}$) gives an upper triangular matrix so we set A to be the transpose of this, whence $A = \text{chol}(\boldsymbol{\rho})'$.

In order to use MC to simulate the evolution of $[S_1(t), ..., S_d(t)]$ in (3.10) it is necessary to decompose the correlated Brownian motions $B_1(t), .., B_d(t)$ into *independent* BMs. We can use the decomposition $\boldsymbol{\rho} = A^T A$ to achieve this. Thus let $Z_1(t), .., Z_d(t)$, be independent BMs and set

$$B_i(t) \;=\; \sum_{r=1}^{d} a_{i,r} Z_r(t) \;, \quad 1 \le i \le d. \tag{3.15}$$

Then we have that

$$\text{cov}[\{B_i(t+\Delta t) - B_i(t)\}, \{B_j(t+\Delta t) - B_j(t)\}] \;= \tag{3.16}$$
$$\Delta t \sum_{r,s=1}^{d} a_{i,r} a_{j,s} \delta_{r,s} \;=\; \Delta t \sum_{r=1}^{d} a_{i,r} a_{j,r} \;=\; \Delta t \{AA^T\}_{i,j} \;=\; \Delta t \rho_{i,j} \;.$$

Thus the BMs in (3.15) have the same covariance structure as the BMs in (3.10). The solution to (3.10) at time T is therefore given by

$$S_i(T) \;=\; S_i(0) \exp\left[(r - \sigma_i^2/2)T + \sigma_i \sqrt{T} \sum_{r=1}^{d} a_{i,r} \xi_r \right] \quad \text{for} \;\; 1 \le i \le d, \tag{3.17}$$

where $\xi_1, .., \xi_d$ are i.i,d standard normal. Once we have simulated $S_1(T), .., S_d(T)$ we can estimate the values of the basket options with the payoffs given by (3.6), (3.7) using the standard procedures.

4. Variance Reduction

In the basic MC method we are interested in estimating the expectation of a function of a random variable which we can easily simulate, such as the standard normal variable. Thus if $\Phi : \mathbf{R} \to \mathbf{R}$ is a function and ξ the standard normal variable then

$$E[\Phi(\xi)] \;\simeq\; \frac{\Phi(\xi_1) + \cdots + \Phi(\xi_N)}{N} \;, \tag{4.1}$$

where $\xi_1, .., \xi_N$ are N independent simulations of the standard normal variable. We have already seen that the error in (4.1) is proportional to the square root of $\text{Var}[\Phi(\xi)]/N$. A *variance reduction method* for estimating $E[\Phi(\xi)]$ in (4.1) is then a function $\Psi : \mathbf{R} \to \mathbf{R}$ such that

$$E[\Phi(\xi)] \;=\; E[\Psi(\xi)] \quad \text{and } \text{Var}[\Psi(\xi)] \;<<\; \text{Var}[\Phi(\xi)] \;. \tag{4.2}$$

Hence we can estimate $E[\Phi(\xi)]$ by

$$E[\Phi(\xi)] \;\simeq\; \frac{\Psi(\xi_1) + \cdots + \Psi(\xi_N)}{N} \;. \tag{4.3}$$

The amount of work involved in computing the RHS of (4.3) is comparable to the amount of work involved in computing the RHS of (4.1). However because $\text{Var}[\Psi(\xi)]$ is much less than $\text{Var}[\Phi(\xi)]$, it follows that the error in (4.3) is smaller than the error in (4.1). In order to implement this method we need to come up with a suitable function $\Psi(\cdot)$, and that of course is not so easy to do. Obvious choices like $\Psi(\xi) = \Phi(\xi) + \lambda\xi$ where λ is a constant satisfy the identity in (4.2), but there is no reason to expect the new function to have reduced variance. In this section we

will discuss three methods of variance reduction which have application to pricing derivatives. These methods are:

(a) Method of antithetic variables.
(b) Control variate method.
(c) Importance sampling.

Method of antithetic variables: Suppose that we wish to estimate $E[\Phi(\xi)]$ where ξ is a *symmetric variable* so that ξ and $-\xi$ have the same distribution, as for the standard normal variable for example. Then we have that

$$E[\Phi(\xi)] \;=\; E[\Psi(\xi)] \quad \text{where } \Psi(\xi) \;=\; [\Phi(\xi)+\Phi(-\xi)]/2\ . \tag{4.4}$$

From (3.1) and (4.4) we see that

$$\mathrm{Var}[\Psi(\xi)] \;=\; \frac{1}{2}\left\{\mathrm{Var}[\Phi(\xi)] + \mathrm{cov}[\Phi(\xi),\Phi(-\xi)]\ \right\}\ . \tag{4.5}$$

Hence if $\mathrm{cov}[\Phi(\xi),\Phi(-\xi)] < 0$ we have a variance reduction of at least 50%. We can be more precise about how much computation is saved in using antithetic variables by comparing $2N$ simulations in the two cases. For the standard MC we have

$$E[\Phi(\xi)] \;\simeq\; \frac{\Phi(\xi_1)+\cdots+\Phi(\xi_{2N})}{2N}\ , \tag{4.6}$$

where $\xi_1,..,\xi_{2N}$ are $2N$ independent simulations of the variable ξ. Using the antithetic method we have

$$E[\Phi(\xi)] \;\simeq\; \frac{\{\Phi(\xi_1)+\Phi(-\xi_1)\}+\cdots+\{\Phi(\xi_N)+\Phi(-\xi_N)\}}{2N}\ . \tag{4.7}$$

Observe that in computing the RHS of (4.7) we need only generate N independent simulations of ξ, whence there is a 50% saving in computation from computing the RHS of (4.6). We have also that

$$\mathrm{var}\left[\ \frac{\Phi(\xi_1)+\cdots+\Phi(\xi_{2N})}{2N}\ \right] \;=\; \frac{\mathrm{var}[\Phi(\xi)]}{2N}\ , \tag{4.8}$$

$$\mathrm{var}\left[\ \frac{\{\Phi(\xi_1)+\Phi(-\xi_1)\}+\cdots+\{\Phi(\xi_N)+\Phi(-\xi_N)\}}{2N}\ \right] \;=\; \frac{1}{2N}\left\{\mathrm{Var}[\Phi(\xi)] + \mathrm{cov}[\Phi(\xi),\Phi(-\xi)]\ \right\}\ . \tag{4.9}$$

Hence if $\mathrm{cov}[\Phi(\xi),\Phi(-\xi)] < 0$ the error in (4.9) is less than the error in (4.8).

The following lemma gives a simple criterion to determine when $\mathrm{cov}[\Phi(\xi),\Phi(-\xi)]$ is negative

Lemma 4.1. *Suppose the function $\Phi : \mathbf{R} \to \mathbf{R}$ is monotonic i.e. always increasing or always decreasing. Then* $\mathrm{cov}[\Phi(\xi),\Phi(-\xi)] \le 0$.

Proof. Assume $\Phi(\cdot)$ is monotonic and let η be a variable independent of ξ but with the same distribution. Then we have that

$$E\left[\ \{\Phi(\xi)-\Phi(\eta)\}\{\Phi(-\xi)-\Phi(-\eta)\}\ \right] \;\le\; 0. \tag{4.10}$$

To see this let us assume that $\Phi(\cdot)$ is increasing. Then for any $\xi,\eta \in \mathbf{R}$ with $\xi \ge \eta$ one has $\Phi(\xi)-\Phi(\eta) \ge 0$ and $\Phi(-\xi)-\Phi(-\eta) \le 0$ so $\{\Phi(\xi)-\Phi(\eta)\}\{\Phi(-\xi)-\Phi(-\eta)\} \le 0$. Since we can similarly argue if $\xi \le \eta$ we conclude that the average on the LHS of (4.10) is negative. □

In Example 1 we used MC to price the standard put option. In that case

$$\Phi(\xi) \;=\; e^{-rT}\max\left[K - S_0\exp\left\{(r-\sigma^2/2)T+\sigma\sqrt{T}\xi\right\},0\right]\;, \tag{4.11}$$

so $\Phi(\xi)$ is an decreasing function of $\xi\in\mathbf{R}$. Hence it makes sense to use the antithetic variable method here.

Control variate method: Suppose we wish to estimate $E[\Phi(\xi)]$ using MC, and we also know precisely the value $g^* = E[g(\xi)]$ of the expectation of $g(\xi)$ for some function $g:\mathbf{R}\to\mathbf{R}$. We can use this fact to reduce variance by considering for $\beta\in\mathbf{R}$ functions $\Psi_\beta(\xi) = \Phi(\xi)+\beta[g^*-g(\xi)]$. Evidently we have that

(4.12)

$$E[\Phi(\xi)] \;=\; E[\Psi_\beta(\xi)]\;,\quad \mathrm{var}[\Psi_\beta(\xi)] \;=\; \mathrm{var}[\Phi(\xi)] - 2\beta\mathrm{cov}[\Phi(\xi),g(\xi)] + \beta^2\mathrm{var}[g(\xi)]\;.$$

The minimum of $\mathrm{var}[\Psi_\beta(\xi)]$ over all $\beta\in\mathbf{R}$ occurs at $\beta=\beta_{\min}$, where

$$\beta_{\min} \;=\; \frac{\mathrm{cov}[\Phi(\xi),g(\xi)]}{\mathrm{var}[g(\xi)]},\quad \min_{\beta\in\mathbf{R}}\mathrm{var}[\Psi_\beta(\xi)] \;=\; \{1-\rho[\Phi(\xi),g(\xi)]^2\}\mathrm{var}[\Phi(\xi)]\;. \tag{4.13}$$

In (4.13) the function $\rho(\cdot,\cdot)$ is the coefficient of correlation (3.3). The method gives good variance reduction then if the function $g(\xi)$, whose expectation we know precisely, is closely correlated to $\Phi(\xi)$, the function whose expectation we wish to estimate. Note that to get an order of magnitude increase in accuracy in the modified MC simulation i.e. one more decimal place accuracy in the MC estimate, we need to have $1-|\rho|<1/100$. This is the case since error is proportional to standard deviation, so to the square root of the variance. We also need to estimate using MC the quantity $\beta_{\min}$ in (4.13).

If we wish to implement this method we need to estimate using standard MC the values of the three quantities $\mathrm{var}[\Phi(\xi)],\mathrm{var}[g(\xi)],\mathrm{cov}[\Phi(\xi),g(\xi)]$. To do this we generate N independent simulations $\xi_1,..,\xi_N$ of the variable ξ and set

$$\hat\Phi_N = \frac{1}{N}\sum_{j=1}^N\Phi(\xi_j),\quad \mathrm{var}[\Phi(\xi)] \;\simeq\; \frac{1}{N}\sum_{j=1}^N\{\Phi(\xi_j)-\hat\Phi_N\}^2\;, \tag{4.14}$$

(4.15)

$$\mathrm{var}[g(\xi)]\simeq\frac{1}{N}\sum_{j=1}^N\{g(\xi_j)-g^*\}^2\;,\quad \mathrm{cov}[\Phi(\xi),g(\xi)]\simeq\frac{1}{N}\sum_{j=1}^N\{\Phi(\xi_j)-\hat\Phi_N\}\{g(\xi_j)-g^*\}\;.$$

The computations (4.14), (4.15) enable us to estimate $\rho=\rho[\Phi(\xi),g(\xi)]$, which needs to be very close to ±1 for the method to yield an improvement over standard MC. If this is the case then we use (4.14), (4.15) to estimate $\beta_{\min}$ in (4.13). Our improved MC estimate for $E[\Phi(\xi)]$ is then

$$E[\Phi(\xi)] \;\simeq \hat\Phi_N + \beta_{\min}\frac{1}{N}\sum_{j=1}^N\{g^*-g(\xi_j)\}\;. \tag{4.16}$$

We can use the control variate method to help improve the accuracy of a MC estimate for the value of an arithmetic basket option with payoff (3.6). Our control variate will be the corresponding geometric basket option with payoff (3.7). We first observe that one can use the Black-Scholes formula to calculate the value of the geometric option. To see this we set $S(T) \;=\; \{S_1(T)\cdots S_d(T)\}^{1/d}$ where (3.10)

implies that

$$S_j(T) \;=\; S_j(0)\exp\left[(r-\sigma_j^2/2)T+\sigma_j B_j(T)\right]\;,\quad 1\le j\le d. \tag{4.17}$$

Thus we have that

$$S(T) \;=\; \{S_1(0)\cdots S_d(0)\}^{1/d}\exp\left[\left(r-\frac{1}{2d}\sum_{j=1}^d\sigma_j^2\right)T+\frac{1}{d}\sum_{j=1}^d\sigma_jB_j(T)\right]\;. \tag{4.18}$$

We write now

$$\frac{1}{d}\sum_{j=1}^d\sigma_jB_j(T) \;=\; \sigma\sqrt{T}\;\xi\quad\text{with }\sigma^2\;=\;\frac{1}{d^2}\sum_{1\le i,j\le d}\sigma_i\rho_{i,j}\sigma_j\;, \tag{4.19}$$

where ξ is a standard normal variable and $\boldsymbol{\rho}=[\rho_{i,j}]$ is the covariance matrix (3.11) for the d Brownian motions. Next we define $S(0)$ by

$$S(0) \;=\; \{S_1(0)\cdots S_d(0)\}^{1/d}\exp\left[\frac{1}{2}\left(\sigma^2-\frac{1}{d}\sum_{j=1}^d\sigma_j^2\right)T\right] \tag{4.20}$$

so that (4.18) is the same as

$$S(T) \;=\; S(0)\exp[(r-\sigma^2/2)T+\sigma\sqrt{T}\;\xi]\;. \tag{4.21}$$

The value of the geometric basket option with payoff (3.7) is therefore given by the BS price of a call option on a stock with strike price K, volatility σ as in (4.19), and today's stock price $S(0)$ given by (4.20). The interest rate r and the expiration time T remain the same.

It is not so obvious that the geometric option is sufficiently closely correlated with the arithmetic option to justify using the geometric option as a control variate. This however turns out to be the case, a fact which helps us resolve the paradox that both stocks and indices on stocks can be accurately modeled by geometric Brownian motion. Stock indices typically measure movements of arithmetic averages of the stocks composing the index. If we assume there are d stocks composing the index which evolve by GBM as in (3.10), then it is clear that the *arithmetic average* $[S_1(t)+\cdots+S_d(t)]/d$ cannot evolve by GBM. In contrast the *geometric average* $\{S_1(t)\cdots S_d(t)\}^{1/d}$ does evolve by GBM. To see this we set

$$S(t) \;=\; \{S_1(t)\cdots S_d(t)\}^{1/d}\exp\left[-\frac{1}{2}\left(\sigma^2-\frac{1}{d}\sum_{j=1}^d\sigma_j^2\right)t\right]\;, \tag{4.22}$$

and observe that

$$d[\log S(t)] \;=\; d\left[\frac{\log S_1(t)+\cdots\log S_d(t)}{d}\right]-\frac{1}{2}\left(\sigma^2-\frac{1}{d}\sum_{j=1}^d\sigma_j^2\right)dt \tag{4.23}$$

$$=\;\left[\frac{\sum_{j=1}^d\{(r-\sigma_j^2/2)dt+\sigma_jdB_j(t)\}}{d}\right]-\frac{1}{2}\left(\sigma^2-\frac{1}{d}\sum_{j=1}^d\sigma_j^2\right)dt\;=\;(r-\sigma^2/2)dt+\sigma dB(t)\,,$$

where σ is given by (4.19) and $B(\cdot)$ is BM. We conclude that

$$\frac{dS(t)}{S(t)} \;=\; rdt+\sigma dB(t)\;, \tag{4.24}$$

so $S(t)$ evolves according to GBM. Thus to be mathematically consistent our stock indices should be geometric averages of stocks. Since the arithmetic and geometric averages are closely correlated in practice, we are justified in modeling movement of stock indices by GBM.

Importance sampling: We consider the situation where we wish to estimate $E[\Phi(\xi)]$, where ξ is standard normal, but $\Phi : \mathbf{R} \to \mathbf{R}$ takes on only significant values when $\xi \leq -\alpha$ where $\alpha > 2$ say. An example of this is a *far out of the money* put option. In this case we have $\Phi(\xi)$ given by (4.11) where

$$K \;=\; S_0 \exp\Big\{(r - \sigma^2/2)T - \sigma\sqrt{T}\alpha\Big\} \;. \tag{4.25}$$

Thus $\Phi(\xi) = 0$ for $\xi > -\alpha$, so if $\alpha > 2$ say the option is far out of the money. If we use the standard MC procedure then $\Phi(\xi) = 0$ for most of our simulations ξ, in fact with $\alpha = 2$ it will be 98% of simulations. We saw in Chapter I that the Gauss elimination method was inefficient for solving large systems of equations $Au = b$ where A is a sparse matrix. The reason was that in the implementation there was a lot of adding of zeros, which we of course know the answer to without having to write a computer code. Here we have a similar situation in that we know apriori the answer to most of the MC simulations i.e. zero. We should try to modify the method in such a way that each new simulation really does give us *new information*. Thinking of this in terms of information theory, we should try to come up with a method which for a given amount of computation yields maximum information on the desired quantity.

In the case here we use a simple translation of the variable, so

$$E[\Phi(\xi)] \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \Phi(\xi) e^{-\xi^2/2}\, d\xi \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \Phi(\eta-\beta) e^{-(\eta-\beta)^2/2}\, d\eta \;. \tag{4.26}$$

Thus we have that

$$E[\Phi(\xi)] \;=\; E[\Psi_\beta(\xi)] \;, \quad \text{where } \Psi_\beta(\xi) \;=\; \Phi(\xi-\alpha) e^{\beta\xi - \beta^2/2} \;. \tag{4.27}$$

Now we need to decide which value of β is best to reduce variance in the MC simulation. For the far out of the money put option we claim that we should take $\beta = \alpha$, where α is defined by (4.25). This is not completely obvious since for $\beta \geq \alpha$ the MC simulation of $\Psi_\beta(\cdot)$ will typically yield something non-zero. There is actually a trade off here. On the one hand the larger β is the more likely the MC simulation yields non-zero values. On the other hand as β increases beyond α then $\mathrm{var}[\Psi_\beta(\xi)]$ tends to rapidly increase. The reason for this is that

$$\mathrm{var}[e^{\beta\xi - \beta^2/2}] \;=\; e^{-\beta^2}\{E[e^{2\beta\xi}] - E[e^{\beta\xi}]^2\} \;=\; e^{-\beta^2}\{e^{2\beta^2} - e^{\beta^2}\} \;=\; e^{\beta^2} - 1. \tag{4.28}$$

Hence we should choose β as small as we can but so that simulations of $\Psi_\beta(\cdot)$ typically give non-zero values, whence we take $\beta = \alpha$. In that case roughly 50% of simulations give non-zero values.

Other variance reduction methods: We touch on some other variance reduction methods. A particularly simple method which is easy to implement is the *method of moment matching*. Thus suppose we generate N independent samples $\xi_1, ..., \xi_N$ of the standard normal variable. We set $\hat{\mu}_N$ and $\hat{\sigma}_N^2$ to be the sample mean and

variance of this set, so

$$\hat{\mu}_N \;=\; \frac{\xi_1 + \cdots + \xi_N}{N} \;, \quad \hat{\sigma}_N^2 \;=\; \frac{1}{N}\sum_{j=1}^{N} \{\xi_j - \hat{\mu}_N\}^2 \;. \tag{4.29}$$

Now we modify the original sample by setting

$$\eta_j \;=\; \frac{\xi_j - \hat{\mu}_N}{\hat{\sigma}_N} \;, \quad 1 \le j \le N, \tag{4.30}$$

and estimate $E[\Phi(\xi)]$ as

$$E[\Phi(\xi)] \;\simeq\; \frac{\Phi(\eta_1) + \cdots + \Phi(\eta_N)}{N} \;. \tag{4.31}$$

We have modified our original sample using a translation and dilation so that the sample mean is zero and the sample variance is 1. Thus our modified sample (4.30) has now the same mean and variance as the standard normal variable.

Finally *stratified sampling* is a generalization of the importance sampling method. In importance sampling there is a region $\{\xi \in \mathbf{R} : a < \xi < b\}$ where the significant values of $\Phi(\xi)$ dominate. If this region is at least 2 standard deviations from the mean, we were able to zero in on it by doing a translation $\eta = \xi - \alpha$ of the normal variable. Alternatively we can write

$$E[\Phi(\xi)] \;\simeq\; E[\Phi(\xi) \;\big|\; a < \xi < b] P(a < \xi < b) \;. \tag{4.32}$$

If we can simulate the standard normal variable ξ conditioned on $a < \xi < b$, then we can estimate $E[\Phi(\xi)]$ by a MC simulation of the RHS of (4.32). Furthermore we can extend this to situations where the significant values of $\Phi(\xi)$ belong to several disjoint regions. So for k regions $\{\xi \in \mathbf{R} : a_j < \xi < b_j\},\ j = 1, .., k$, we set

$$E[\Phi(\xi)] \;\simeq\; \sum_{j=1}^{k} E[\Phi(\xi) \;\big|\; a_j < \xi < b_j] P(a_j < \xi < b_j) \;. \tag{4.33}$$

5. The Brownian Bridge Process

In §2 we saw that if we wish to MC simulate the path $S(t),\ t \ge 0$, of a GBM process which satisfies (2.15), we can use the formula (2.16). The payoff in the continuous time Asian option was given in Chapter I as

$$\text{payoff} \;=\; \left[S(T) - \frac{1}{T}\int_0^T S(t)\; dt \right]^+ \;, \tag{5.1}$$

and for the discrete time Asian option as

$$\text{payoff} \;=\; \left[S(T) - \frac{1}{L+1}\sum_{i=0}^{L} S(t_i) \right]^+ \;. \tag{5.2}$$

In both cases the weight of the stock price $S(T)$ at expiration in the payoff formula is higher than than the stock price at intermediate times. However in the MC simulation method (2.16) the price $S(T)$ is the final stock price that is computed in a path simulation, and thus accumulates more MC error than intermediate stock prices. Since $S(T)$ has the highest weight in the payoff formula, it would seem then that a method which allows us to compute $S(T)$ at the beginning of a path

simulation rather than at the end will give us a more accurate method of estimating the price of the Asian option.

The Brownian bridge process (BB) enables us to do that. Let $B(t),\ t \geq 0$, be BM and for $t' < t < t''$ consider the variable $B(t)$ conditioned on $B(t'), B(t'')$. It is not difficult to see that $B(t)$ is normal with

$$\text{mean } = \frac{t''-t}{t''-t'}B(t') + \frac{t-t'}{t''-t'}B(t'') \ , \quad \text{variance } = \frac{(t''-t)(t-t')}{t''-t'} \ . \tag{5.3}$$

Note that (5.3) tells us that the mean of the conditioned BM $B(t)$ is obtained by linear interpolation of the values of $B(s)$ at $s = t', t''$. We can use (5.3) to simulate the path of BM $B(t),\ 0 \leq t \leq T$. Evidently $B(0) = 0$ and $B(T) = \sqrt{T}\ \xi_0$, where ξ_0 is standard normal. Next we see from (5.3) that $B(T/2)$ conditioned on $B(0), B(T)$ is Gaussian with mean $B(T)/2$ and variance $T/4$. Thus we have that

$$B(T/2) \ = \ B(T)/2 + \sqrt{T/4}\ \xi_1, \quad \text{where } \xi_1 \text{ is standard normal independent of } \xi_0 \ . \tag{5.4}$$

Similarly we have that

$$B(T/4) \ = \ B(T/2)/2 + \sqrt{T/8}\ \xi_3 \ , \quad B(3T/4) \ = \ [B(T/2) + B(T)]/2 + \sqrt{T/8}\ \xi_4 \ , \tag{5.5}$$

where $\xi_1, \xi_2, \xi_3, \xi_4$ are independent standard normal. Evidently we can generate the entire BM path by continuing this method of dyadic decomposition of the interval $[0, T]$. The Asian option payoffs can then be computed by setting

$$S(t) \ = \ S(0) \exp\left[(r - \sigma^2/2)t + \sigma B(t)\right] \ , \quad 0 \leq t \leq T. \tag{5.6}$$

6. Random number generation

In §1 we mentioned the problem of generating i.i.d. copies of a random variable Y with a given distribution $\rho(y),\ y \in \mathbf{R}$. Our basic input is i.i.d copies of the uniform variable X in the interval $0 < X < 1$. If we can easily compute the cumulative distribution function (cdf) $F : \mathbf{R} \to [0, 1]$ for the variable Y,

$$F(x) \ = \ \int_{-\infty}^{x} \rho(y)\ dy, \quad \text{where } F(\cdot) \text{ is increasing and } F(-\infty) = 0,\ F(\infty) = 1, \tag{6.1}$$

then we can generate i.i.d copies of Y from i.i.d. copies of the uniform variable X. This follows by observing that the variable Y defined by $F(Y) = X$ has the distribution $\rho(\cdot)$ if X is uniformly distributed in $0 < X < 1$. In fact

$$P(Y < a) \ = \ P(F(Y) < F(a)) \ = \ P(X < F(a)) \ = \ F(a), \tag{6.2}$$

so the function $F(\cdot)$ is the cdf for Y. To carry this out we need to be able to efficiently invert the function F and that may not be so easy. For the standard normal variable we have that

$$F(x) \ = \ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\ dy \ , \quad -\infty < x < \infty. \tag{6.3}$$

There is no explicit formula for $F(x)$ in (6.3) and certainly no explicit inverse formula. Nevertheless the standard optimally efficient method for generating i.i.d. standard normal variables does use a sophisticated version of this basic cdf 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 the distribution of the polar coordinate variables R, Θ defined by

$$X = R\cos\Theta,\ Y = R\sin\Theta, \quad \text{so } R = \sqrt{X^2+Y^2},\ \Theta = \arctan(Y/X). \tag{6.4}$$

We can find the joint distribution of the variables (R, Θ) by writing the joint pdf of (X, Y) in polar coordinates. Thus we have

$$\frac{1}{2\pi}e^{-(x^2+y^2)/2}\,dxdy = \frac{1}{2\pi}e^{-r^2/2}\,rdrd\theta = d[-e^{-r^2/2}]\,d\left[\frac{\theta}{2\pi}\right] = dudv\ . \tag{6.5}$$

If we define two variables U, V by

$$U = \exp[-R^2/2],\ V = \Theta/2\pi\ , \tag{6.6}$$

then it is clear that U, V are restricted to the interval $0 < U, V < 1$ and from (6.5) it follows that they are independent and uniformly distributed in $[0, 1]$. We can run this argument in the opposite direction and begin with two independent variables U, V uniformly distributed in $[0, 1]$. Then on inverting the formulas (6.6) we see from (6.4) that the variables X, Y defined by

$$X = \sqrt{-2\log U}\cos(2\pi V)\ , \quad Y = \sqrt{-2\log U}\sin(2\pi V)\ , \tag{6.7}$$

are independent and standard normal. The lack of efficiency in the method (6.7) 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.

Finally we return to the discussion of some of the issues covered in §1 concerning the advantages of MC simulation over deterministic methods for estimating integrals in high dimension. We saw that

$$\int_{[0,1]^d}\Phi(x)\,dx \simeq \frac{1}{N}\left[\Phi(X_1)+\cdots\Phi(X_N)\right] + \text{Error}(N)\ . \tag{6.8}$$

If the points $X_1, .., X_N$ are uniformly distributed in the "cube" $[0, 1]^d$ then $\text{Error}(N) \simeq 1/N^{1/d}$, whereas if they are randomly distributed $\text{Error}(N) \simeq 1/N^{1/2}$. Thus if $d \geq 3$ the MC method outperforms (with high probability) the deterministic method. One can try to improve on the deterministic method by choosing $X_1, .., X_N$ to be pseudo-random and hope to get an error which is better than the MC error $1/\sqrt{N}$. Of course our computer "random number generator" really is deterministic, so this might seem like a contradiction in terms. Actually as we pointed out in §1 the random number generator algorithm is designed to minimize correlations between successive values of $X_j, j = 1, 2, ...$ We might then expect that it is possible to come up with other sets of numbers $X_j, j = 1, 2, ..,$ in which successive values do have significant correlation, but for which the error in (6.8) is better than the MC error. Such numbers are called *low discrepancy numbers*. The best of them improve on the MC error in moderate dimensions -up to d=20 say- but for large enough d the MC method still wins out.

The simplest of such sequences of numbers are the *Halton numbers*. In one dimension we define them as follows: We associate to each integer $n \geq 1$ its corresponding base 2 Halton number ξ_n by writing the positive integers to base 2.

Thus

$$n \;=\; \sum_{k=1}^{\infty} a_k 2^{k-1} \text{ with } a_k \in \{0,1\} \quad \text{implies } \xi_n \;=\; \sum_{k=1}^{\infty} a_k 2^{-k} \; . \tag{6.9}$$

The $\xi_n,\ n = 1, 2, ..,$ all lie in the unit interval $[0, 1]$ and are somewhat randomly placed in this interval. Specifically we have $\xi_1 = 1/2, \xi_2 = 1/4, \xi_3 = 1/2 + 1/4 = 3/4, \xi_4 = 1/8, \xi_5 = 1/8+1/2 = 5/8$ etc. To get Halton numbers in the d dimensional cube $[0, 1]^d$ we choose d distinct prime numbers $p_1, .., p_d$ and write an integer n to base p_j, which yields a Halton number $\xi_n^j \in [0, 1],\ j = 1, .., d$. Then the nth Halton point $\xi_n \in [0, 1]^d$ is given by $\xi_n = [\xi_n^1, .., \xi_n^d]$.

University of Michigan, Department of Mathematics, Ann Arbor, MI 48109-1109
E-mail address: conlon@umich.edu