Monte-Carlo pricing and fitting of path-dependent options under various models

Olaf Willocx

1 Background

Let P be the probability measure associated to some Wiener process W_t^P. If P is the risk-neutral measure, the Black-Scholes-Merton model describes a spot price

$$\frac{dS}{S} = (r-q)dt + \sigma dW_t^P$$

under P, so that $\int dP\, S_t = S_0 \exp((r-q)t)$. r is the risk-free interest rate of the currency, q is a continuous dividend yield paid for holding the product.

A 'contingent claim' V_T is measurable by the filtration of S_t^P at time $t = T$. Its 'arbitrage price process' V_t dictates its value at earlier times, determined by the model. We generally write $t = 0$ for now, so consider V_0. Under the Black-Scholes-Merton model and measure P,

$$V_0 = e^{-rT} \int dP\, V_T. \tag{1}$$

The expectation values over dP are infinite-dimensional integrals on a Hilbert space, in physics we would call them path integrals.

There are many ways to calculate V_0. Some work for quite general models (beyond Black-Scholes-Merton) and options (beyond the usual vanillas), some do not. Some are highly efficient, some are not. Some are simple, some complicated. Some can be implemented with low latency, some cannot. Some scale one way with numerical parameters, some another.
There are e.g. PDE methods, like the famous Black-Scholes PDE for path-independent options. Such methods can be generalized to certain path-dependent

options also. PDE methods require the construction of the PDE from the model, which is not always reasonable. And they may need to be adjusted for path-dependence, so nontrivially depend on the option. The dimensionality of the space the PDE acts on depends on the number of stochastic processes in the model; up to two stochastic processes is numerically tractible.
There exist also Fourier methods utilizing the characteristic function of the model distribution, by Carr and Madan. These do not suffer from the dimensionality issues, and characteristic functions are generally quite simple. But these are similarly constrained mostly to path-independent options.
Then there are tree methods, like the binomial tree model, under which path-dependent options can easily be priced, e.g. American options. Here the difficulty is correctly constructing the tree model in accordance to the stochastic process, which may be unreasonable.
It is of clear importance to have a golden standard, a method that is conceptually very simple so that very little can go wrong, but it is also generally applicable to any option or model that can be simulated, directly from the stochastic process and definition of the option price on the path, and where the error of the estimate is simply behaved and analyzable. The results from other more practicle (more performant, lower latency) methods can be checked for correctness against such a golden standard. This golden standard is of course Monte Carlo integration.

In Monte Carlo integration, the infinite-dimensional Hilbert space is approximated by a high dimensional vector space carrying the same statistics. (Fortunately a discretized Wiener process trivially carries the same endpoint statistics. Still, discretization leads to disretization errors if there is path-dependence.) In this high-dimensional space we can sample paths W_t^P under the measure P, evaluate V_T as a function of said path, and use each such V_T as an estimate for the true V_T via the mean; this is one way to evaluate V_0 as in (1).
There are some important subtleties around the world 'sampling'. Sampling can be random or deterministic, and have low or high discrepancy. A generalized cubic lattice is a deterministic low discrepancy uniform sampling of the Euclidian space, i.e. the space between the samples is generally low, but the entire lattice must be sampled before the estimator is correct. This leads to Riemann integration, whereas random samples lead to Monte Carlo integration. I.i.d. uniform random samples have higher discrepancy, are still equidistributed in the Euclidian space, and have the property that each sample in and of itself is a correct estimator also. Low-discrepancy sequences exist somewhere in between, they generate sequences that are not 'random' in

a meaningful sense, but have low discrepancy, are equidistributed, and some partial results are correct estimators. This leads to 'quasi-Monte Carlo' integration.
If the samples are i.i.d. random, the population mean tends to the sample mean with a rate $\propto n^{-1/2}$, i.e. the standard error scales with $n^{-1/2}$. In Riemann integration, the error scales with n^{-1}, which is superior. Really this is due to the low discrepancy of the lattice sequence. It is important that i.i.d. uniform random samples in some sense form the highest discrepancy sequence that is still equidistributed; it requires memory to form a low discrepancy sequence, and i.i.d. random sequences satisfy the Markov property. But Riemann integration suffers from the curse of dimensionality, the number of samples it takes for a given lattice parameter is exponential in the number of dimensions. The number of dimensions for stochastic processes is so large, no spacing at all is reasonable, Riemann integration breaks down completely. Also, Riemann integration is restricted to uniform samples. We can't sample P, we must sample uniformly and weigh appropriately, e.g. via rejection sampling. The same is true for low-discrepancy sequences. For quasi-Monte Carlo methods that errors should theoretically scale as $n^{-1} \ln^d n$ in d dimensions. But quasi-Monte Carlo methods have difficulties reaching actual good convergence due to the high dimensionality. Discontinuous integrands are also a problem, which do sometimes appear in option pricing. This is however addressable, see e.g. papers by Caflisch and Moskowitz in 1995 and 1996 using Brownian bridges, and newer results via randomized quasi-Monte Carlo. In principle, the best tool is randomized quasi Monte Carlo with recursive Brownian bridge sampling to solve the issue of dimensionality, so as to have a low discrepancy on coarse properties of the paths. But we won't get carried away too far and limit our scope to pure Monte Carlo pricing. Importantly, in terms of relevancy, the optimizations we will use will also be applicable to Brownian bridge randomized quasi-Monte Carlo integration.

There are a number of ways to decrease the variance $\langle V_T^2 \rangle - \langle V_T \rangle^2$ of the estimator V_T (conditioned to $t = 0$) in Monte Carlo integration. We understand the issue with i.i.d. random samples is the high discrepancy of the sequence. The discrepancy can be lowered by sampling pairs $(W_t^P, -W_t^P)$, which is trivial since they have an equal probability of occuring. This is the 'method of antithetic variates', a kind of quasi- or 'hybrid' Monte Carlo method. Instead of having n samples with an error that scales as $n^{-1/2}$, we have $n/2$ that scale as $(n/2)^{-1/2}$ and 2 that scales as $1/2$. So the standard error is improved

(multiplied) by a factor

$$\frac{(n/2)^{-1/2}/2}{n^{-1/2}} = 2^{-1/2}.$$

Quasi-Monte Carlo methods suffer from the curse of dimensionality in general (since they can be seen as Monte Carlo-Riemann sum hybrids). But the method of antithetic variates is low-hanging fruit, no issue with dimensionality since we consider a pair, which is a fixed number of elements.
In stratified sampling, the region of integration is subdivided and the number of samples in each stratum is fixed in some way. More samples may be taken in strata of otherwise high variance, reducing overall variance, this is typical adaptive stratified sampling. But the subdivision itself also reduces variance, as it is a kind of correlation, a hybrid scheme. But this is less well suited for stochastic processes with definite initial conditions.
In importance sampling, we sample according to a measure Q instead of P and use

$$V_0 = e^{-rT} \int dQ \frac{dP}{dQ} V_T, \tag{2}$$

with dP/dQ the Radon-Nikodym derivative, the 'weight' used in importance sampling. We choose Q such that the Radon-Nikodym derivative is reasonable to calculate, and such that $\langle V_0^2 \rangle - \langle V_0 \rangle^2$ is lowered. Orders of magnitude improvement in standard error can be achieved in this way (with twice as many orders of magnitude savings in the number of samples to be used otherwise), it is highly relevant, but also 'high trust', so useful in a golden standard Monte Carlo pricer.

2 Importance sampling of some stochastic processes

2.1 Black-Scholes-Merton model

Let P still be the risk-neutral measure of the Black-Scholes-Merton model. Let Q be such that $W_t^Q = W_t^P + \lambda t$. The meaning is that under P, W_t^P is a Wiener process, and under Q, W_t^Q is a Wiener process, but under P, W_t^Q appears as having positive drift, and under Q, W_t^P appears as having negative drift.

The Radon-Nikodym derivative is described by the Cameron-Martin-Girsanov theorem. The result is

$$\frac{dP}{dQ} = \exp\Big(-\frac{\lambda^2 T}{2} + \lambda W_T^Q\Big),$$

which importantly is a martingale. (W_t^Q is also a martingale, so when exponeniating, there is a 'martingale correction' $\lambda^2 t/2$, just as there is in the geometric Brownian motion (Black-Scholes-Merton model). It is equivalent also to the way in which $d\ln S_t \neq dS_t/S_t$, Itô's lemma for the Itô process for dS_t.) It is more interesting (and generalizable) to consider the earlier Cameron-Martin formula, which is simply the calculation of this result. Now dP/dQ is easy enough to calculate as $(dP/dx)/(dQ/dx)$ where e.g. dP/dx is the probability density in the Hilbert space of Wiener paths, or just of the values $W_t = x$. We may consider the paths in the entire Hilbert space. Let Φ be the Gaussian measure.

$$\frac{dQ}{dx} = \lim_{\Delta t\to 0}\prod_{t=\Delta t}^{T}\frac{d\Phi}{dx}\Big(\frac{W_t^Q - W_{t-\Delta t}^Q}{\sqrt{\Delta t}}\Big) \tag{3}$$

$$\frac{dP}{dx} = \lim_{\Delta t\to 0}\prod_{t=\Delta t}^{T}\frac{d\Phi}{dx}\Big(\frac{W_t^P - W_{t-\Delta t}^P}{\sqrt{\Delta t}}\Big) \tag{4}$$

$$= \lim_{\Delta t\to 0}\prod_{t=\Delta t}^{T}\frac{d\Phi}{dx}\Big(\frac{W_t^Q - W_{t-\Delta t}^Q}{\sqrt{\Delta t}} - \lambda\sqrt{\Delta t}\Big) \tag{5}$$

$$\frac{dP}{dQ} = \lim_{\Delta t\to 0}\prod_{t=\Delta t}^{T}\exp\Big(-\frac{\lambda^2\Delta t}{2} + \lambda(W_t^Q - W_{t-\Delta t}^Q)\Big) \tag{6}$$

$$= \lim_{\Delta t\to 0}\exp\sum_{t=\Delta t}^{T}\Big(-\frac{\lambda^2\Delta t}{2} + \lambda(W_t^Q - W_{t-\Delta t}^Q)\Big) \tag{7}$$

$$= \exp\Big(-\frac{\lambda^2 T}{2} + \lambda W_T^Q\Big) \tag{8}$$

This result is independent of Δt, so taking $\Delta t = T$ it is also just

$$\frac{d\Phi/dx\left(\frac{W_T^Q}{\sqrt{T}} - \lambda\sqrt{T}\right)}{d\Phi/dx\left(\frac{W_T^Q}{\sqrt{T}}\right)}, \tag{9}$$

which is what is usually calculated for the Cameron-Martin formula, but we are interested in functions not of $W_T = x$, but of the entire path $\{W_t\}_0^T$. It is

however the same result, which is true also for Girsanov's theorem for Poisson point processes or gamma processes, as used later.

Another piece of the puzzle is choosing λ, which can be done either analytically or numerically.

Analytically, taking λ as minimizing $\langle V_0^2\rangle - \langle V_0\rangle^2$ is not tractible. But we are mostly interested in the region around $S_T^P = K$. The requirement $\langle \ln S_T^P\rangle_Q = \ln K$ is quite solvable for λ.

$$\ln S_t^P - \ln S_0 = \Big(r - q - \frac{\sigma^2}{2}\Big)t + \sigma dW_t^P \tag{10}$$

$$= \Big(r - q - \frac{\sigma^2}{2} - \lambda\sigma\Big)T + \sigma dW_t^Q \tag{11}$$

$$\langle \ln S_T^P\rangle_Q - \ln S_0 = \Big(r - q - \frac{\sigma^2}{2} - \lambda\sigma\Big)T = \ln K - \ln S_0 \tag{12}$$

$$\lambda\sigma T = \ln\frac{S_0}{K} + \Big(r - q - \frac{\sigma^2}{2}\Big)T \tag{13}$$

$$\lambda = \frac{\ln S_0/K}{\sigma T} + \frac{r-q}{\sigma} - \frac{\sigma}{2} \tag{14}$$

This result derived for a European option is often reasonably efficient also for similar path-dependent options, like American options or barrier options.

Numerically, we can measure the random $\langle V_T^2\rangle - \langle V_T\rangle^2$ (using m samples) and adjust λ such that it decreases, e.g. via stochastic gradient descent. This is a kind of 'adaptive Monte-Carlo method'.
Having tried a number of algorithms, what appears to work is the following. Consider steps $\Delta\lambda$ in λ. For $\lambda - \Delta\lambda$ and $\lambda + \Delta\lambda$, take m samples to estimate $\langle V_T^2\rangle - \langle V_T\rangle^2$. If $\lambda - \Delta\lambda$ produces a lower variance, consider next $\lambda - 2\Delta\lambda$ and λ. Similarly, if $\lambda + \Delta\lambda$ produces a lower variance, consider next λ and $\lambda + 2\Delta\lambda$. If neither side is better, e.g. if the option never expires ITM, perform a random walk in λ by updating to either $\lambda \pm \Delta\lambda$ with equal probability.
When parallelizing, λ must be synchronized between threads, and parallelization must occur on the level of the $2m$ samples. m must also be sufficiently large for stability reasons ($\langle V_T^2\rangle - \langle V_T\rangle^2$ must be sufficiently precise), take at least $m \sim 1000$. But for efficient parallelization, each thread should have at least a few hundred paths to sample. So for N threads, pick also $m \gtrsim 100N$.

2.2 Heston model

The Cox-Ingersoll-Ross model for stochastic volatility

$$d\nu_t^P = \kappa(\eta - \nu_t^P)dt + \theta\sqrt{\nu_t}dW_t^P \tag{15}$$

is well studied and used in the Heston model

$$\frac{dS_t^P}{S_t^P} = (r-q)dt + \sqrt{\nu_t^P}dW_t^{1,P} \tag{16}$$

$$d\nu_t^P = \kappa(\eta - \nu_t^P)dt + \theta\sqrt{\nu_t^P}dW_t^{3,P} \tag{17}$$

where $W_t^{3,P}$ and $W_t^{1,P}$ have correlation ρ, i.e.

$$W_t^{3,P} = \rho W_t^{1,P} + \sqrt{1-\rho^2}W_t^{2,P} \tag{18}$$

with $W_t^{1,P}$ and $W_t^{2,P}$ independent. Once again, P is the risk-neutral measure, so that the interest rate of the stock is $r-q$ with r the risk-free interest rate; the stochastic volatility requires no additional martingale correction.
Adjust now

$$dW_t^{1,Q} = dW_t^{1,P} + \lambda_1 t \tag{19}$$

$$dW_t^{2,Q} = dW_t^{2,P} + \lambda_2 t. \tag{20}$$

P and Q are bivariate measures. Let e.g. P_1 be the measure of $W_t^{1,P}$ independently. Since (P_1, Q_1) and (P_2, Q_2) are independent, the Radon-Nikodym derivative separates as

$$\frac{dP}{dQ} = \frac{dP_1}{dQ_1}\frac{dP_2}{dQ_2} \tag{21}$$

$$= \exp\Big(-\frac{(\lambda_1^2+\lambda_2^2)T}{2} + \lambda_1 W_T^{1,Q} + \lambda_2 W_T^{2,Q}\Big). \tag{22}$$

Again S_t^P must be evaluated under the measure Q.

$$\frac{dS_t^P}{S_t^P} = \Big(r - q - \lambda_1\sqrt{\nu_t^P}\Big)dt + \sqrt{\nu_t^P}dW_t^{1,Q} \tag{23}$$

$$d\nu_t^P = \Big(\kappa(\eta - \nu_t^P) - \theta\sqrt{\nu_t^P}(\rho\lambda_1 + \sqrt{1-\rho^2}\lambda_2)\Big)dt + \theta\sqrt{\nu_t^P}dW_t^{3,Q} \tag{24}$$

$$W_t^{3,Q} = \rho W_t^{1,Q} + \sqrt{1-\rho^2}W_t^{2,Q} \tag{25}$$

$$= W_t^{3,P} + (\rho\lambda_1 + \sqrt{1-\rho^2}\lambda_2)t \tag{26}$$

We notice that $\lambda_2 = 0$ is always ideal, so we may reduce this degree of freedom. But we may desire for the squared volatility process to be unaffected by importance sampling, this corresponds to the special case where $\lambda_2 = -\lambda_1\rho/\sqrt{1-\rho^2}$, and

$$\frac{dP}{dQ} = \exp\Big(-\frac{\lambda_1^2 T}{2(1-\rho^2)} + \lambda W_T^{1,Q} - \lambda\frac{\rho}{\sqrt{1-\rho^2}}W_T^{2,Q}\Big). \tag{27}$$

The gives slightly increased standard error compared to $\lambda_2 = 0$, but now the Cox-Ingersoll-Ross process is not modified, so the Feller condition $2\kappa\eta \geq \theta^2$ is unchanged, and we rather not violate the Feller condition for the sake of importance sampling.

2.3 Bilateral gamma model

Consider first a Poisson point process rather than a gamma process. A Poisson point process N_t with rate (mean) λ is such that

$$N_{t+\Delta t} - N_t \sim \text{Pois}(\lambda\Delta t), \tag{28}$$

and almost surely $N_0 = 0$. It may be simulated efficiently by sampling N_T and uniformly placing N_T events in $(0, T]$, this event independence is the defining property of the Poisson point process.
Consider N_t with a rate λ_P under P and λ_Q under Q. Then

$$P_P(N_t = k) = \frac{(\lambda_P t)^k}{k!}e^{-\lambda_P t} \tag{29}$$

$$P_Q(N_t = k) = \frac{(\lambda_Q t)^k}{k!}e^{-\lambda_Q t} \tag{30}$$

$$\frac{dP}{dQ} = \Big(\frac{\lambda_P}{\lambda_Q}\Big)^{k_Q} e^{-(\lambda_P-\lambda_Q)t}. \tag{31}$$

For a gamma distribution with shape α and scale θ, i.e. mean $\mu = \alpha\theta$ and variance $\nu = \alpha\theta^2$, the associated gamma process Γ_t is such that

$$\Gamma_{t+\Delta t} - \Gamma_t \sim \Gamma(\alpha\Delta t, \theta) = \Gamma(\mu^2\Delta t/\nu, \nu/\mu). \tag{32}$$

This is discretized and evaluated similar to a Wiener process.
Consider Γ_t with shape α and scale θ_P under P and θ_Q under Q. Then

$$\Gamma_t(\alpha, \lambda_P) \sim \Gamma(\alpha t, \lambda_P) \tag{33}$$

$$\Gamma_t(\alpha, \lambda_Q) \sim \Gamma(\alpha t, \lambda_Q) \tag{34}$$

$$\frac{dP}{dQ} = \Big(\frac{\lambda_P}{\lambda_Q}\Big)^{\alpha t} \exp\Big(-(\lambda_P - \lambda_Q)x_t^Q\Big). \tag{35}$$

If two samples are distributed according to different gamma distributions, the difference is in general bilateral gamma distributed. And accordingly, the difference of two gamma process is a bilateral gamma process.

$$\Gamma^+_{t+\Delta t} - \Gamma^+_t \sim \Gamma(\alpha^+ \Delta t, \theta^+) \tag{36}$$
$$\Gamma^-_{t+\Delta t} - \Gamma^-_t \sim \Gamma(\alpha^- \Delta t, \theta^-) \tag{37}$$
$$X_t = \Gamma^+_{t+\Delta t} - \Gamma^-_{t+\Delta t} \tag{38}$$
$$X_{t+\Delta t} - X_t \sim \Gamma(\alpha^+ \Delta t, \lambda^+; \alpha^- \Delta t, \lambda^-) \tag{39}$$

Such a process has a drift, a martingale correction ξ can be applied.

$$\xi = -\alpha^+ \ln \frac{\lambda^+}{\lambda^+ - 1} - \alpha^- \ln \frac{\lambda^-}{\lambda^- + 1} \tag{40}$$

The market model is $S_t/S_0 = \exp((r - q + \xi)t + X_t)$. The Radon-Nikodym derivative adjusting θ^+ and θ^- in the usual way is again just the product of the individual ones, since $\Gamma^\pm_t$ are independent.

$$\frac{dP}{dQ} = \left(\frac{\lambda^+_P}{\lambda^+_Q}\right)^{\alpha^+ t} \left(\frac{\lambda^-_P}{\lambda^-_Q}\right)^{\alpha^- t} \exp\left(-(\lambda^+_P - \lambda^+_Q) x_t^{+Q} - (\lambda^-_P - \lambda^-_Q) x_t^{-Q}\right) \tag{41}$$