Improving valuation run-times for derivative books
Malcolm Kemp – 25 April 2019

Abstract

This paper proposes a simulation-based approach which it primarily applies to the problem of identifying the fair (i.e. market consistent) valuation of books of derivatives. The proposed approach offers significant run-time improvements relative to more traditional simulation-based approaches when applied to this problem. It is particularly well-suited to derivative books that depend on a relatively small number of underlyings, including ones that in the banking, asset management or insurance worlds arise if customers are guaranteed that future fund-linked payouts will not fall below specified floors. The approach typically involves the preparation of a larger base simulation set (using traditional techniques usually alongside low order moment fitting variance reduction techniques) and creating from this base set a smaller collated simulation set with only the latter then applied to the payoffs being valued. Weights ascribed to the individual collated simulations may vary to optimise valuation accuracy, or they can default to equal weighting if there is no strong reason to diverge from this simpler variant. A refinement involves the valuation algorithm typically being segmented by reference to an indicator (in the example explored in the paper the adjusted ratio of the strike price to the price of the underlying) with different smaller collation sets being used depending on the range within which this indicator lies and/or with weights and values ascribed to different simulation points in the collated simulation set being selected to facilitate more accurate approximation of the payoff function. The paper concludes with an explanation of how the same approach can be applied to problems outside the derivative pricing (or even the broader financial) arena, including to engineering control process problems, if the underlying problem to be solved is amenable to simulation techniques and if run-time efficiency is desirable.

1. Introduction

1.1 Financial markets involve the trading of financial instruments. The market values (market prices) ascribed by markets to these instruments characterise consensus market views, i.e. *market implied* views, on the future behaviours of economic factors that might influence these instruments. The growth of derivatives markets over the last few decades has stimulated considerable research into option pricing theory, i.e. how best to ascribe market or market-like values to options and other derivative instruments that are consistent with prices observed for simpler financial instruments to which the instrument being valued relates.

1.2 Originally, this process tended to focus on identifying prices likely to present profit opportunities for market makers and others active in the marketplace. As the stock of such instruments grew, appropriate ways of establishing market or market-like valuations for such instruments became increasing important to a wider range of financial practitioners, including accountants, actuaries and others responsible for valuing such instruments on a regular basis. In some parts of the financial community the values thus derived are referred to as *fair* values, in other parts it is more common to refer them as *market consistent* values, if the process is not as simple as just lifting an actual (traded) market price from a suitable market feed. An exploration of the concept of market consistency is provided by Kemp (2009).

1.3 Regulators have also become increasingly interested in applying such valuation approaches even when the exposures concerned are not directly traded on financial markets. For

example, the introduction of the Solvency II regulatory framework for EU insurers at the end of 2015 strengthened the need for EU insurers of all sizes to determine market consistent valuations for guarantees given to policyholders, even though no active market involving third parties exists for most individual insurers' guarantees. Solvency II specifically mandates a market consistent approach to the valuation of insurer liabilities (and assets) for regulatory capital purposes. Instruments held in bank trading books or in asset management portfolios also typically need to be fair valued.

1.4 The market standard approach in the insurance industry to carry out market consistent valuations of such exposures involves the use of an *economic scenario generator* (ESG). An ESG is a tool that can deliver a set of market consistent scenarios characterising how economic factors on which such exposures depend (including asset returns) might develop in the future. Smaller firms in this industry thus face the challenge of how to access economic scenario generators that are proportionate to their needs, without incurring excessive costs. Outside the insurance industry different terminologies are common but the underlying concepts and mathematical principles are typically similar, particularly if the payoffs are complicated enough to require simulation-based valuation techniques[1].

1.5 Traditionally, the simulations used in these exercises have been created by:

(a) Developing a credible (economic) model of how the world might behave in the future
(b) Overlaying, where necessary, adjustments to the model to ensure that where practical the model statistically respects the Principle of No Arbitrage
(c) Randomly simulating future evolutions of this model, usually using Monte Carlo simulation techniques (typically but not always choosing the simulations so that they have equal probability weights)
(d) Placing a present value on the payoffs resulting from the relevant exposures by probability-weighting the present values of the payoffs arising in the different simulations. If the simulations are chosen with equal probability weights, this computation involves a simple (unweighted) average across the simulations. If weights are unequal, a weighted average is needed.

We refer to this simulation approach as the 'traditional' (simulation-based) derivative pricing formulation.

1.6 If the underlying economic model in Section 1.5 is chosen in an unconstrained manner and if a naïve approach to discounting is adopted then the end results will rarely respect the Principle of No Arbitrage. For example, suppose we assume that equities will on average outperform fixed income assets, perhaps because we observe that this has typically proved to be the case in the past over long enough time periods, at least for developed western economies that weren't on the losing side of World War II. Suppose we also discount future cash flows from both asset classes using identical discount rates. We will then typically conclude that equities represent better value than fixed income, even though equities with a market worth of €1 now currently trade at the same price as fixed income assets with the same market worth of €1 now, i.e. both trade at €1. A way to square this circle without

[1] Please note that in the insurance industry, ESGs can also be used to create non-market consistent scenarios which insurers can use for purposes other than the market consistent valuation of exposures. These include using the ESG to simulate the future liabilities on a 'real world' basis (that e.g. includes some assumed return differentials between asset classes) but in these simulations then valuing the liabilities at future points in time using a different, usually more market consistent, basis. The approaches described in this paper can also be applied to such problems, see Section 4.23.

altering chosen 'real world' probabilities of outcomes is to adjust the discount factors applied to different future cash flows when determining present values. This is the genesis behind the idea of *stochastic deflators*. Alternatively, we may use a common discount factor and adjust the probabilities used in the model so that they are *risk-neutral*. If suitably chosen, these two approaches will provide the same end-result. For simplicity, we concentrate in this paper on the risk-neutral approach. The stochastic deflator variant is in any case rarely used outside the insurance industry. A focus on the risk-neutral approach is therefore likely to be more intuitive to a broader readership.

1.7 Where the underlying economic model is particularly simple, it is sometimes possible to identify analytical equivalents to the above, i.e. an *analytical formula* that mirrors the limit of the present value that would be determined using the above simulation approach if $n \to \infty$ where n is the number of simulations used. Examples are the celebrated Black-Scholes option pricing formulae and variants, see Appendix A. An 'analytical formula' is here understood to be one that only involves 'standard' mathematical functions. Even so, the term still has a flexible interpretation as there is no single agreed list of 'standard' mathematical functions. Usually the functions are limited to relatively well-known ones that can be computed without much difficulty in standard tools for carrying out numerical calculations such as Microsoft Excel. However, many more obscure mathematical functions can be useful for numerical calculations in specific circumstances. For example, the Maple symbolic algebra system includes the Whittaker M and Whittaker W functions, which are not (at the time of writing) available as built-in functions in Microsoft Excel[2], but can be manipulated and computed numerically (and symbolically) in Maple. In principle, we can create our own analytical functions provided they have well-defined properties, which in this context could merely be that they represent the solution in the limit as $n \to \infty$ to some specified option pricing problem (although this is not the usual way in which the term 'analytical' is interpreted). If we exclude this technical extension to what we understand by 'analytical' then it is generally impractical to identify an analytical formula for the exact valuation of fund-linked guarantees if the provider has some flexibility to modify either the payouts or the structure of the asset portfolio on which they depend. These flexibilities may be codified in the form of assumed *management actions* that the providing firm might adopt in the future.

1.8 In this paper we adopt a rather different philosophical approach, although one that is still formally equivalent in a mathematical sense to the approach described above. We call this the *interpolation formulation*. In this formulation, the valuation approach (and by implication any use of simulation-based techniques within it) operates by:

(a) Identifying a set of economic variables (e.g. asset index levels) that provide the universe of factors that are assumed to influence the future payoffs we are seeking to value
(b) Specifying a set of market instruments, the prices of which are considered to constitute the set of available market observables, and locating the current prices of these instruments
(c) Determining a mathematical function of all available inputs in (a) that simultaneously:
    (1) As far as possible exactly replicates the observed prices of the instruments in (b); and
    (2) Provides a credible way of interpolating or extrapolating from these prices to the deemed prices of all other instruments / payoffs we might wish to value.

[2] Modern web-based tools, such as the Nematrian web function library, see Nematrian (2019a), can also extend the range of functions that are effectively accessible on demand through traditional tools such as Microsoft Excel or corresponding programming environments.

1.9 In the special case as per Section 1.7 where an analytical option pricing model might have been used then it is easy to see that the interpolation formulation is formally identical to other fair (i.e. market consistent) valuation approaches. The natural mathematical function to use in 1.8(c) is this analytical option pricing model (the inputs to which would then be chosen to replicate suitable market observables). Where no analytical option pricing model is available using conventional definitions of 'analytical', the approaches are still equivalent in a formal sense, since we can always select for the mathematical function in 1.8(c) the one that corresponds to the limiting case as $n \to \infty$ of the traditional formulation. Using the terminology of 1.7, this involves manufacturing a new 'analytical' function which has the desired mathematical characteristics. Of course, it isn't usually practical to identify this function using more commonly recognised or accessible mathematical functions. If we really believe in the economic model underlying the traditional formulation, then our goal becomes one of seeking to identify a suitable interpolation methodology that adequately approximates this limiting case without e.g. incurring excessive run-times.

1.10 If the two formulations are formally equivalent then what advantages come from adopting the interpolation formulation? Additional insights it provides include:

(a) It becomes clearer to see that the process of valuing instruments[3] in a market consistent manner involves the mathematical process of integration. Most ways of deriving analytical option pricing formulae involve integration somewhere within the derivation.
(b) Numerical integration requires evaluation of the function being integrated at discrete points in the multi-dimensional space spanned by the set of variables on which the function depends. These points can be viewed as akin to the simulations introduced by the traditional formulation. There is an extensive body of mathematical literature on *quadrature*, i.e. the selection of the best points at which to evaluate a function when it is being integrated numerically[4], which can further inform the interpolation approach. In a formal sense this literature is still applicable to traditional option price techniques, but the link is often not very apparent.
(c) The interpolation formulation highlights that the theoretical correctness of any supposed economic model underlying the valuation approach may not be particularly important. The interpolation formulation involves specifying / identifying sets of payoffs for which we have observed market prices. The implication is that for other payoffs we will typically not have observed prices. The economic model influences how the interpolation should be done, so needs to have some credibility. But ultimately, we have no wholly objective data exactly applicable to the prices to ascribe to payoffs that lie outside the set for which we have observed prices. How we interpolate or extrapolate to such payoffs is inherently subjective. There is no sure way of doing so more robustly except by finding extra market observables closer in nature to the instrument we want to value.

1.11 Insights from quadrature theory include:

(a) The quadrature points (or weights) don't need to be the same for all payoffs. There might be one set of quadrature points that is better suited to valuing one set of payoffs and a different set of points more suitable for another set of payoffs. This is e.g. relevant to how we might incorporate so-called management actions and rebalancing strategies

[3] Or insurance policies in an insurance context.
[4] For the purpose of this paper we include identification of points to use for multi-dimensional integrals (as well as for one-dimensional integrals) within the meaning of 'quadrature'.

into the pricing algorithm (see Section 4) and to some run-time optimisation approaches (see Section 2).

(b) There are established ways of improving the accuracy of numerical integration, by using non-equally weighted quadrature points. An example is Simpson's rule, see Section 2.15. In the traditional formulation this can be thought of as assigning non-equal weights to individual simulations. Although the traditional formulation can use non-equally weighted simulations, it does not commonly do so.

(c) Many other techniques can be used to improve the accuracy of numerical integration for a given run-time. These include importance sampling and variance reduction techniques. Once we switch to a mindset that sees the valuation problem as essentially a problem in (usually numerical) integration, we may become more comfortable employing these sorts of techniques in derivative pricing exercises. Mathematical packages often numerically evaluate common mathematical functions using quadrature approaches but often also using different sets of quadrature points for different ranges of input values. In Sections 2.32 – 2.36 we see that a similar insight can in principle further reduce valuation run-times for some types of payoff.

1.12 Most of the rest of this paper can be thought of as an elaboration of how these insights can be best introduced into simulation-based derivative pricing exercises. "Best" or "good" is here to be interpreted as primarily a mixture of simplicity of implementation and run-time optimisation. It is less about the theoretical robustness of the underlying economic model for the reasons highlighted in Section 1.10(c).

1.13 Of course, run-time optimisation is also a flexible concept. For most of this paper we will concentrate on the case where we have many individual instruments or instrument groupings and each instrument involves a payoff linked to some underlying fund together with a guarantee that the payout to the customer will not fall below some prespecified floor. It is assumed that we wish to place a fair value on these guarantees. Valuations of books of such instruments may group individual ones into model points that are assumed adequately to approximate to the average of all instruments included in the group. We will assume that the level of grouping of individual instruments has been preselected on other criteria, i.e. for the purposes of our analysis can be assumed to be fixed[5].

1.14 Suppose $V_{A,s,n}$ is the value we would place on the derivative book using just n simulations selected according to algorithm A, using (if it involves a random selection of simulations) a random seed, s. Suppose also that if there are m instrument groupings then the overall run-time $T_{A,s,n}$ to derive $V_{A,r,n}$ is of order $O(mn)$. Here "of order $O(x)$" means behaves like kx for some constant k when x is large. We then interpret run-time optimisation to mean identifying a relatively simple algorithm A for which:

(a) The time taken to identify the simulation points is small relative to $T_{A,s,n}$ (which will always be the case if m is large enough relative to the number of underlyings we need to model);

(b) As $n \to \infty$, the algorithm tends to the correct limiting value V_∞ (ignoring rounding or other errors introduced by the machine precision limit applicable to the computer used for the evaluation). V_∞ should be independent of any applicable random seed, see Section 1.7.

[5] One possible 'grouping' approach is, of course, to ascribe each instrument its own model point. This might be because the book contains such a wide range of instruments that there is no clear way of grouping them, or maybe just because we want a process that does not need to worry about whether any grouping of instruments has been done appropriately.

(c) $V_{A,s,n}$ is statistically as close as possible to V_∞ for n as small as possible, which can typically be understood to mean that for a given n, the algorithm aims to keep $Q_{A,n} = variance\left(V_{A,s,n} - V_\infty\right)$ as small as possible.

1.15 As explained in Press et al. (2007), and given that we are assuming evaluation for each instrument grouping needs to be done separately, a naïve (basic) Monte Carlo simulation approach with run-time mn will have $Q_{MC,n} \approx k_{MC} n^{-1/2}$, where k_{MC} is a (near) constant for instruments with similar characteristics. Our goal is to identify an algorithm, A, with a run-time mn that has $Q_{A,n} \approx k_A n^{-\alpha}$ with k_A small as possible (and/or with an α higher than 0.5).

1.16 Most of the remainder of this paper (most of Sections 2 – 5) explores the application of a proposed new approach to simulation in the area of finance (and specifically in derivative pricing and related areas). The proposed new approach does, however, have broader applicability. In Section 6 we explore its applicability to a wider range of problems including some involving engineering control processes and business processes.

2. A single period reinvested asset class index

2.1 In this section we use a simple example to illustrate the formal equivalence but practical mindset differences between the traditional formulation and the interpolation formulation. We also introduce the main innovation in this paper, i.e. a new approach to simulation, and specifically how it might be applied in the field of derivative pricing. We call it the *targeted quantile-spacing* approach. It meets the goal of offering a 'good' way of selecting simulations within the interpolation formulation. Most of the remaining sections of this paper aim to demonstrate that the targeted quantile-spacing approach continues to offer a 'good' solution even for more realistic problems.

2.2 Our example involves a riskless asset and a single risky asset class (i.e. 'fund') that has behaviour over the period from $t = 0$ to $t = T$ aligned with that of a reinvested index that has returns over instantaneous consecutive time periods that are independent identically distributed normal random variables with mean μdt and variance $\sigma^2 dt$ where dt is the length of each such (small) time period and μ and σ^2 are constant. The reinvested index value thus forms a Brownian motion. The rolled-up return index on the riskless asset over the time period $(0, T)$ is assumed to be e^{rT} where r is constant. The rolled-up return index (before expenses) on the risky asset over the time period $(0, T)$ is taken to be S_T and after expenses that are a fraction q_j per unit time of the rolled-up fund value in the meantime is taken to be $S_T e^{-q_j T}$. The opening fund level (the price of the underlying) is S_0. We assume that the we wish to value at time 0 a payoff at time T that is:

$$N_j \max\left(K_j - S_T e^{-q_j T}, 0\right)$$

2.3 This corresponds to the situation where the firm has many customers, with customer j having an investment of N_j units in the underlying (all units being identical, except for a fund-value based annual management charge of q_j which might vary by customer). The firm wants to determine the cost (fair value) of the guarantee that the overall payout will not be less than $N_j K_j$.

2.4 All the usual additional assumptions used in the derivation of the Black-Scholes formulae are assumed to apply (e.g. no arbitrage, no frictions, ability to borrow short, no expenses other than those represented by the q_j, liquid markets etc.). Using standard no-arbitrage arguments, we can analytically value this payoff in a market consistent manner by using the

Black-Scholes formula for a put option (or more precisely, the Garman-Kohlhagen generalisation to a security that pays away a continuously compounded dividend)[6], i.e. the (market consistent) value of the guarantee is $N_j P(K_j, q_j)$ where $P(K_j, q_j)$ is the price of a put option with strike price K_j, underlying S_0 and continuously compounded dividend q_j paid away. Hence:

$$P(K_j, q_j) = -S_0 e^{-q_j T} N(-d_1) + K_j e^{-rT} N(-d_2)$$

where σ, $N(x)$, d_1 and d_2 have the meanings set out in Appendix A.

2.5 Suppose we seek to replicate this valuation using a traditional simulation-based formulation without any variance reduction techniques or other ways of reducing sampling error. Although the cumulative returns (ignoring expenses) in the above example have a 'real world' distribution in line with a lognormal distribution, i.e. $S_T \sim LN(\mu, \sigma^2)$, the risk-neutral probability distribution they follow is a different lognormal distribution with different distributional parameters. This is because the risk-neutral distribution needs to have a mean of $S_0 e^{rT}$ so that its weighted average discounted value (discounting at the risk-free rate) is S_0. A simple Monte Carlo simulation approach to valuing the guarantee might therefore involve the following:

(a) Select n independent uniform random variables, x_k, where $k = 1, \dots, n$
(b) Find $y_k = N^{-1}(x_k)$, where $N^{-1}(x)$ is the inverse (unit) normal distribution function
(c) Estimate the value of the payoff using the following formula:

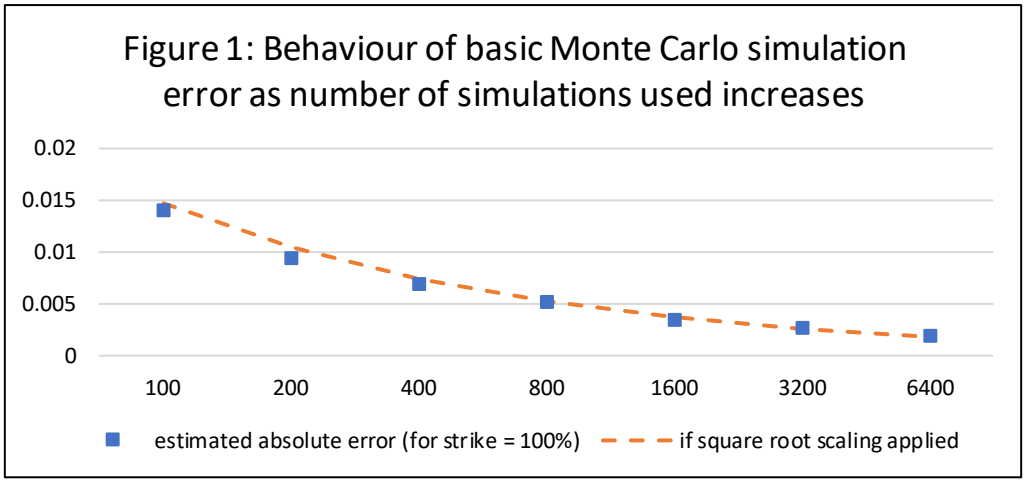
$$V_{MC1,n} = N_j e^{-rT} \sum_{k=1}^{n} \max(K_j - S_0 e^{-q_j T} z_k, 0)$$

where $z_k = G e^{\mu T + \sigma \sqrt{T} y_k}$ is the rolled up cumulative return on the underlying (ignoring expenses) and, for the naïve (basic) Monte Carlo case, we set $G = e^{(r-\mu)T - \sigma^2 T/2}$.

Basic Monte Carlo error dependency

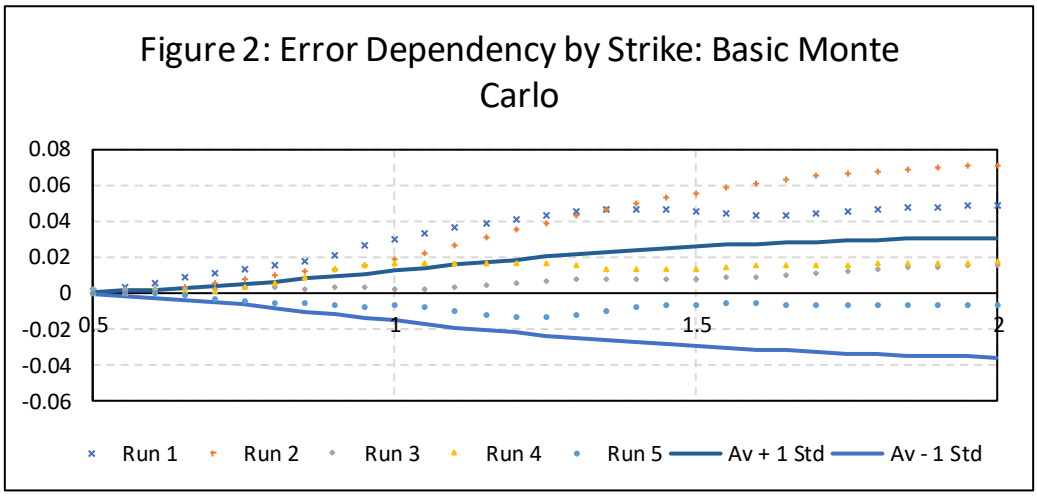
2.6 Different values of V_{MC1} will arise for different seeds used in step 2.5(a). In Figure 1, we plot the resulting sampling error, $Q_{MC1,n}$, for the case where $T = 5$, $\sigma = 0.15$, $N_j = N = 1$, $q_j = 0.01$, $r = 0.02$, $\mu = 0.03$ and $S_0 = 1$, for an at-the-money option (i.e. with $K/S_0 = 1$) and for $n = 100, 200, 400, 800, 1600, 3200, 6400$, estimating $Q_{MC1,n}$ based on 200 separate (basic Monte Carlo) simulation exercises each with its own random seed. It exhibits the $O(n^{-1/2})$ error dependency referred to in Section 1.15.

[6] See e.g. Nematrian (2019b).



2.7 If q_j is the same for all customers (=q) then the only way in which different customers' benefits vary in the situation referred to in Section 2.5 (other than in terms of the numbers of units held) is the guarantee level that is applicable, i.e. the level of K/S_0. It is therefore also helpful to understand how the sampling error might vary for different (relative) strike prices (i.e. for different K/S_0) which we show in Figure 2.

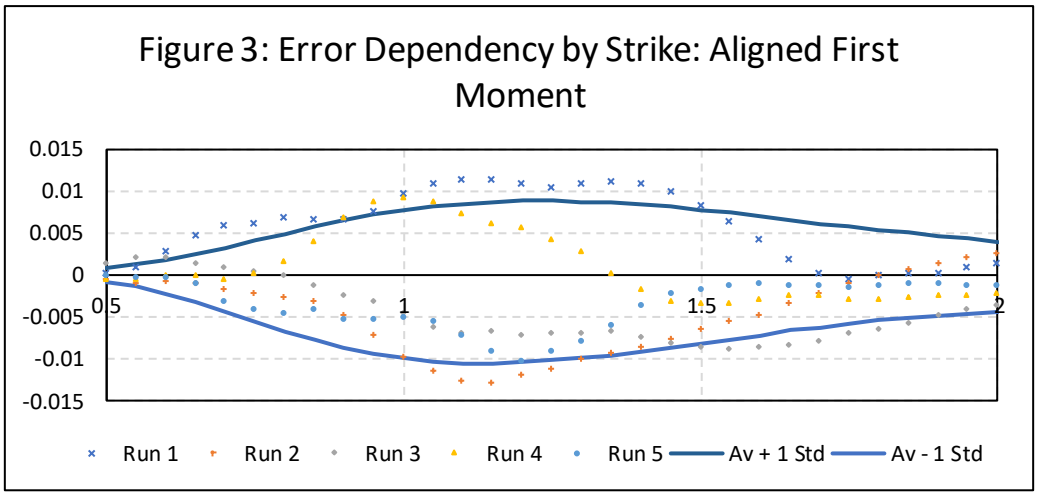
In Figures 2 – 9 we assume that T, σ, N, q, r, μ and S_0 are the same for Figure 1 and we also similarly assume that $n = 100$. The Figures differ merely according to the sampling methodology used. Readers are encouraged to focus on the *y*-axis scale, as this indicates the relative difference in error size for different simulation approaches. A table in Section 2.41 summarises the results for a single strike level. In each of these Figures we again estimate the likely errors present in any single run by carrying out 200 separate simulation exercises each with its own random seed. The *x*-axis is the strike (i.e. the guarantee level, K/S_0), the solid lines indicate the estimated plus or minus one standard deviation range above or below the estimated mean across the 200 different simulation exercises and individual marked points indicate the errors for specific strikes in five specific simulation runs (that each use a different random seed). The main aim of showing the results of a handful of specific simulation runs is to highlight that for any given simulation exercise the errors for different strikes are not uncorrelated. Instead, if the result is biased upwards (or downwards) for a specific strike then it is typically also biased upwards (or downwards) for all nearby strikes, and hence little or no error diversification arises merely by having a book with a range of strikes.



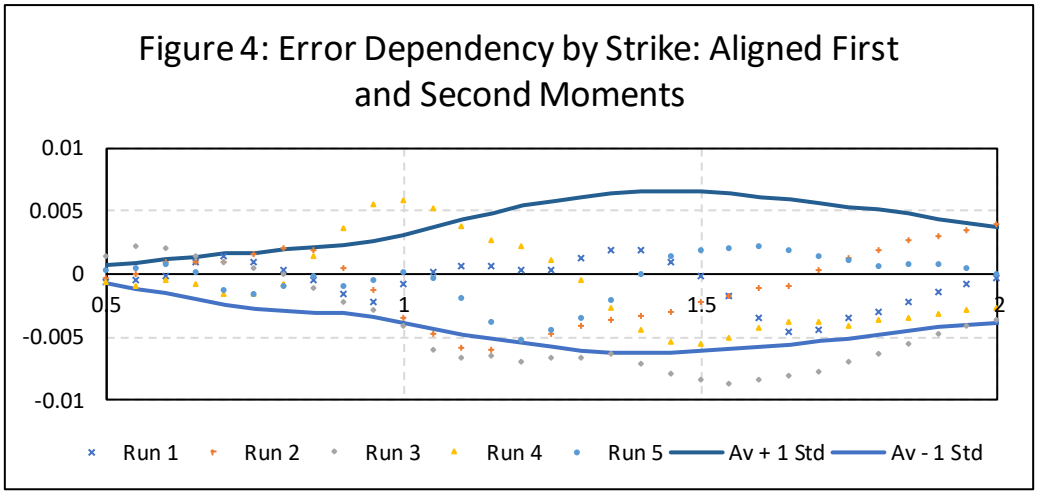
2.8 The sample errors shown in Figure 2 (i.e. for the basic Monte Carlo approach) are very small in absolute terms when the guarantee is almost certain not to bite (i.e. when K/S_0 is small). However, they are much are larger when the guarantee is almost certain to apply (i.e. when K/S_0 is large). This is because a sample set chosen using the basic Monte Carlo approach described above does not (for any given run) have a sample mean for the average rolled-up fund value (ignoring expenses) that is equal to the true (i.e. population) mean of the underlying distribution.

Low order moment fitting approaches

2.9 Suppose we impose the constraint that the sample and population means are aligned, i.e. we set $G = e^{rT}/\left(\frac{1}{n}\sum_k e^{\mu T+\sigma\sqrt{T}y_k}\right)$. The resulting sample errors for different K/S_0 (i.e. different relative strike prices) are shown in Figure 3 and become much closer to zero both when K/S_0 is small and when it is large but are still material for intermediate levels.

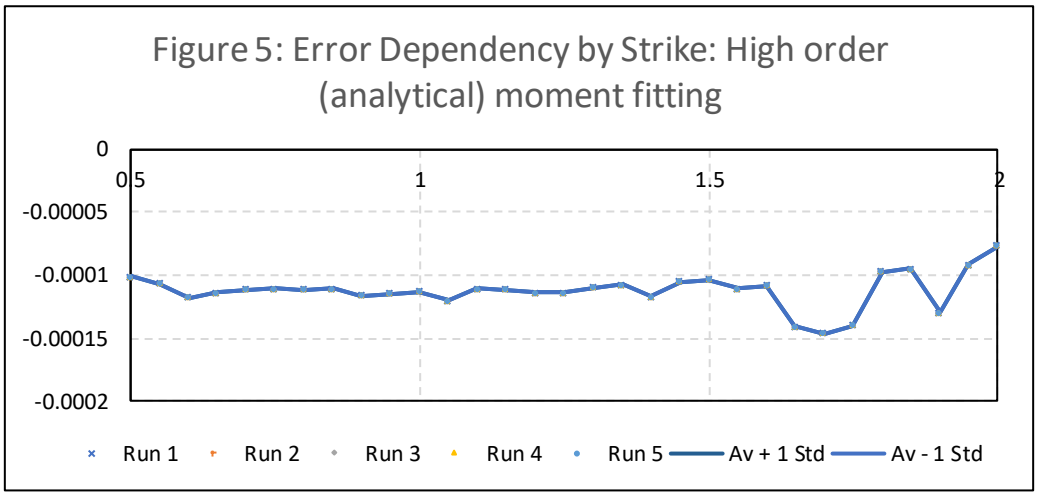


2.10 In Figure 4 we have sampled the distribution subject to the constraint that the sample first moment $\frac{1}{n}\sum z_k$ should be fitted to (i.e. be equal to) its true first moment $E(Z)$. This is an example of a variance reduction technique, i.e. an approach that is used to reduce the variance of the error being introduced by a Monte Carlo simulation. A natural extension is to fit to the second moment as well, or (largely equivalently) to constrain the (sample) standard deviation of the x_k so that it also matches its true population standard deviation, i.e. σ. The resulting sample errors for different K/S_0 are shown in Figure 4. There is an improvement relative to Figure 3, but it is not as marked as between Figures 2 and 3.



High order moment fitting approaches

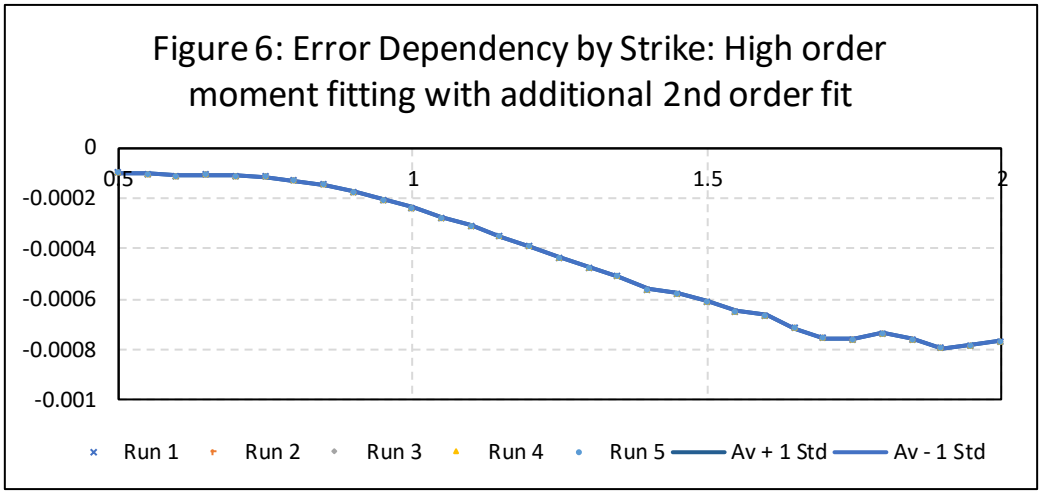
2.11 A much more significant improvement arises if, in effect, we fit to the first n moments, i.e. we move from a low order moment fitting approach (fitting to the mean and standard deviation is just a two-moment fitting approach), to a high order one. One way of achieving this (almost exactly) is by selecting the y_k so that they are a random permutation of equal-quantile spaced points across the whole domain of y, i.e. in this instance by setting the y_k equal to a random permutation of $N^{-1}((k-0.5)/n)$. The resulting sample errors with $G = e^{(r-\mu)T-\sigma^2 T/2}$ (i.e. the same as in the basic Monte Carlo approach) are shown in Figure 5.



For $n = 100$ and $K/S_0 = 1.25$ the resulting errors are roughly equivalent to those that would have arisen had we used 3,600,000 simulations and a basic Monte Carlo simulation approach without any moment fitting, 680,000 simulations with just a first order moment fitting approach or 250,000 simulations with second order moment fitting respectively.

2.12 It is somewhat better *not* to include a further fit to the first two moments when using equal-quantile spaced simulations, see Figure 6 in which we do so versus Figure 5 in which we have not done so[7]. However, for cosmetic reasons linked to the martingale test it may be simpler to explain the process if we do so, see Section 4.15.

[7] The typically increasing error towards the right-hand end of Figure 2.6 appears to be mainly due to the non-linear nature of the payoff function (expressed as a function of equal quantile-spaced points) and to the



2.13 It is typically possible to do even better than Figure 5, even if the number of simulations applied to individual instruments or instrument groups remains fixed. We describe below two different ways of doing so. The first appears to be particularly powerful but struggles to have general applicability. It is the equivalent of using Simpson's rule to carry out numerical integration. The second still offers a significant reduction in percentage terms in the error and more easily generalises to cases where we have e.g. multiple asset categories, see Section 4. It is equivalent to subdividing the function to be integrated into multiple sub-elements, each of which we integrate more accurately.

A Simpson's rule type of approach

2.14 Suppose we want to integrate a function numerically over some finite range and suppose the function to be integrated is sufficiently smooth. The simplest approach is to use the trapezium rule. This is the approach implicit in all the simulation approaches referred to above.

2.15 A typically more accurate numerical estimate for the same integral can be achieved using Simpson's rule[8], provided the function being integrated is sufficiently smooth. In effect, as explained in Press et al. (2007), Simpson's rule for numerically evaluating a closed integral (i.e. one over a finite range), V_∞, involves calculating V_n and V_{2n} (estimates using n equally-spaced and $2n$ equally-spaced quadrature points, with both estimates including the start and end points) and estimating V_∞ by V where:

$$V = \frac{4}{3}V_{2n} - \frac{1}{3}V_n$$

2.16 The underlying insight is that we can expand the error between the integral's sampled value V_n and its true value V_∞, i.e. $V_n - V_\infty$, as a power series in $1/n$. If the function is sufficiently smooth and the points are equally spaced in a domain that can be transformed smoothly to the one that is applicable to the integral then this power series expansion should take the form $V_n - V_\infty = a_2(1/n)^2 + a_3(1/n)^3 + a_4(1/n)^4 + \cdots$ and because of symmetry $a_3 = 0$. So, a quadratic improvement ought to be capable of being achieved by using a formula in which we net off the term in $(1/n)^2$ by estimating V_∞ as above. For 'open' integrals (where

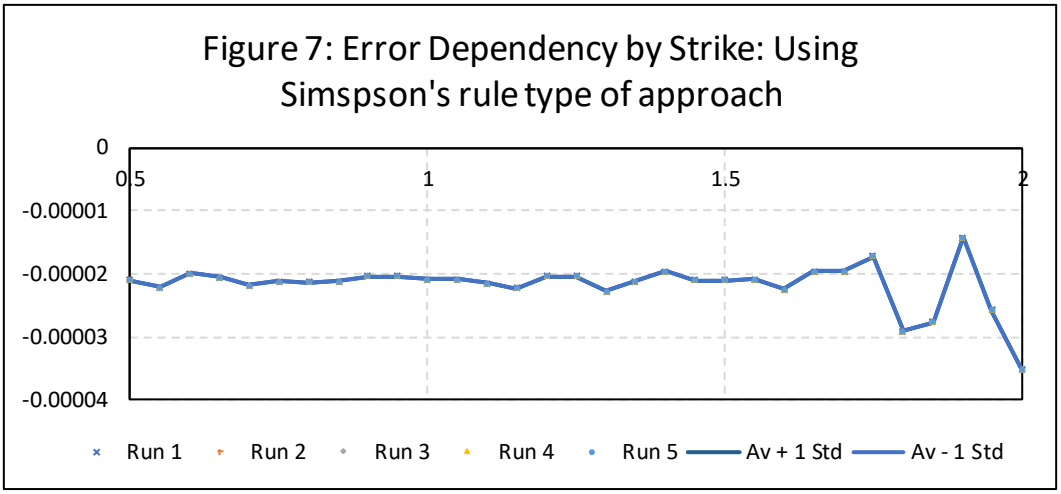
overall guarantee value rising as one moves from left to right (it rises from c. 0.10 for $K/S_0 = 1$ to c. 0.86 for $K/S_0 = 2$). If more simulation points are included the error typically gets smaller.

[8] See e.g. Nematrian (2019c).

the function is ill-defined or infinite at one or both of the range limits, as will usually be the case with derivative pricing problems[9]) a similar improvement is possible if we use $V = (9/8)V_{3n} - (1/8)V_n$ with the V_n and V_{3n} now involving equally-spaced points akin to the equal quantile-spaced points used in Section 2.9.

2.17 For a smooth payoff, e.g. a payoff corresponding to the rolled-up value of the fund, using a Simpson's rule type of approach as above can improve the relative accuracy of the modelled result very dramatically.

2.18 However, for a kinked payoff, such as the guarantees we are considering here, the function to be integrated is no longer smooth, the improvement is not as strong and errors at the end of the integration range appear to become more important. In Figure 7 we show the impact of using the approach set out in Section 2.15 using a combination of a run that uses 100 equally spaced quantiles and a run that uses 300 equally spaced quantiles. We note that there is an improvement versus the 100 equally spaced quantiles, but it is not as dramatic as might have been suggested by the theory in 2.16, particularly bearing in mind that we have used 300 simulations when calculating V_{3n}.



2.19 Crucial to achieving improvements from a Simpson's rule type of approach is for $V_n - V_\infty$ to be capable of being meaningfully expanded as a power series. This is only possible if there is a suitable regularity between how the points are spaced as n changes. Unfortunately, introducing additional asset classes typically disrupts this regularity (even if we use equal quantile-spacing of points for each individual asset class in isolation or if we try multi-dimensional analogues). This makes it impractical in most situations to obtain the level of improvement potentially available in the single asset class case from using a Simpson's rule type of approach.

The targeted quantile-spacing approach

2.20 The other method we introduce for improving the accuracy of the result has more general applicability and is the main novel methodology introduced by this paper. We call this method the *targeted quantile-spacing approach*. At its simplest, if we are generalising from the approach set out in Section 2.11, it involves replacing the equally-spaced quantile values that we would have previously ascribed to the rolled-up fund value in any given simulation

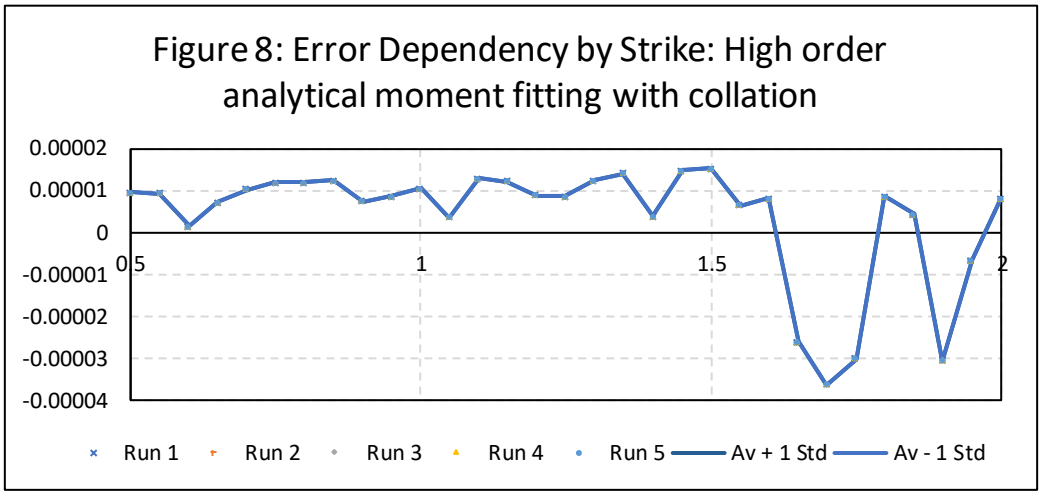
[9] E.g. $N^{-1}(x)$ is ill-defined at $x = 0$ and $x = 1$.

by an average of more finely subdivided quantile values spanning the quantile range corresponding to the simulation that this average is replacing.

2.21 For example, suppose the simulations are ranked so that in the equal (analytical) quantile-spaced approach we would have $z_k = Ge^{\mu T+\sigma N^{-1}((k-0.5)/n)} = Ge^{\mu T}e^{\sigma N^{-1}((k-0.5)/n)}$ (for a lognormal model). Instead of using these z_k we use $z_k = Ge^{\mu T}\sum_{i=1}^{B} p_{k,i}$ where:

$$p_{k,i} = \frac{1}{B}e^{\sigma N^{-1}\left(\frac{k}{n}-\frac{(2B-1)}{2Bn}+\frac{i}{Bn}\right)}$$

2.22 If we ignore any kinking in the payoff etc., this approach in effect allows us to achieve close to the accuracy we would have got with Bn equal quantile-spaced points but only using n simulations. The impact of using this methodology with $n = 100$ and $B = 10$ is shown in Figure 8.



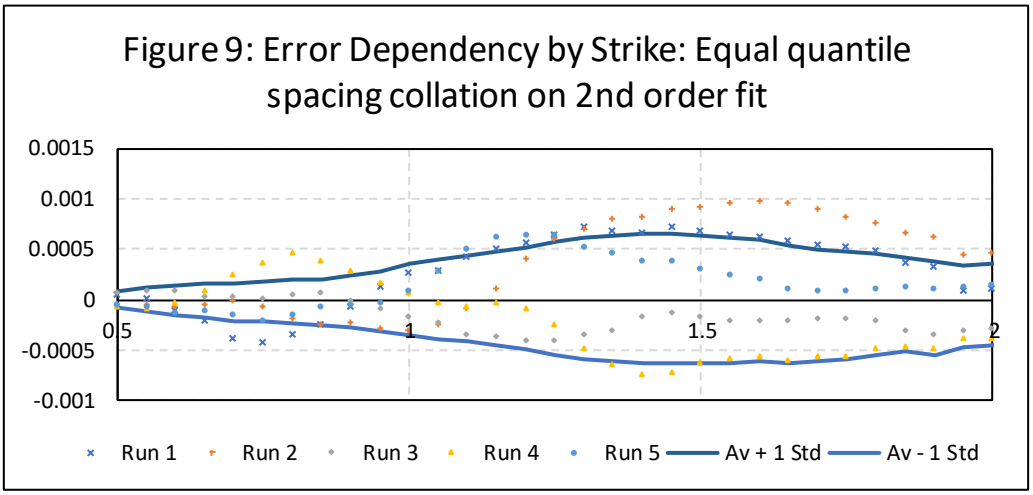
2.23 In the above, we identified the individual $p_{k,i}$ analytically. This is only possible if we know the actual form of the quantile function, or equivalently the distributional form of the underlying risk-neutral distribution. This will rarely be the case in practice. However, we can estimate the $p_{k,i}$ arbitrarily accurately by using a Monte Carlo approach (probably enhanced with a low moment fitting approach) using a large enough number of simulations n^*. Therefore, unlike the Simpson's rule type of approach identified above, this approach *does* successfully generalise to the multi-asset class problem. In effect the algorithm is as follows:

(a) Carry out a Monte Carlo exercise (usually enhanced with a low moment fitting approach) but with a high number of simulations, e.g. $n^* = Bn$;
(b) Carry out a similar Monte Carlo exercise but using just n simulations;
(c) Separately rank the simulated outcomes for the relevant fund in (a) and (b);
(d) Replace (for each asset class) each simulated outcome in (b) by the average of the simulated outcomes (or of quantiles of them) in (a) that (in quantile rank terms in (a)) most closely accord with the outcome in (b) (in quantile rank terms in (b)) being replaced.

In practice, (b) could merely be a subset (e.g. the first n) of the simulations in (a), as the main goal of (b) is to provide a pattern (e.g. through time) that in quantile rank terms is plausible. For the one-period example being considered here, all possible outputs of (b) produce the same end-result. However, for multi-period or multi-asset class fund-linked

underlyings, (b) also provides a structure for co-dependency of simulations through time or across asset classes.

2.24 The reason this approach can be expected to assist with valuation run-times for the sorts of guarantees referred to in Section 2.3 is because we assumed in Section 1.14 that the overall run-time was dominated by the number of instrument groupings present. We still only need to carry out the same number of instrument evaluations, i.e. mn, however large B is. The Bn 'base' simulation points are transformed into a smaller number, n, of *collated* simulation points that can be passed to the engine carrying out the instrument grouping-level valuation. We show in Figure 9 the errors if the base simulation set is found using a (non-analytical) two moment fitted Monte Carlo approach with $B = 100$ and is collated as above into a collated set that only has $n = 100$ elements, the remaining parameters being as before.



The key run-time gain is that it is only the 100 simulations in the collated set that are then used in the remainder of the valuation process, not the 10000 original simulations. As we might expect, the sampling error is around ten times smaller than that shown in Figure 4, i.e. roughly what it would have been had we used $Bn = 10000$ simulations when preparing Figure 4 (given the $O\left(n^{-1/2}\right)$ error dependency of a traditional Monte Carlo simulation approach). In effect the collated set 'inherits' most of the additional accuracy implicit in the larger size of the base set, but only needs to pass the smaller set onto the individual instrument or instrument grouping stage to do so.

2.25 Of course, if B is too large this run-time assumption can break down. A further reason for not going overboard with a large B is that the accuracy of an approach involving just a limited number of collated simulation points is inherently constrained when payoffs are kinked, however large B is.

2.26 This is because the simulation approach inherently contains approximations depending on where a kink lies relative to the selected simulation points. In effect, each simulation point groups together specific 'nearby' trajectories and assumes that the average payoff across these nearby trajectories is linear in the strike. However, if there is a kink in the payoff (as a function of the underlying) within this range then this assumption will no longer hold.

2.27 A similar problem can arise if the collated simulation set does not include any simulations extreme enough to mature in-the-money for a well out-of-the-money option that happens

to have a very high face value. Although such an option may be very unlikely to bite, its value could be material if the option nominal is large enough.

2.28 To minimise errors introduced by kinks in the payoffs being valued etc., we may therefore need to collate simulations in the base set into a smaller number that are not necessarily *equally-weighted* but instead are positioned to best target the (multiple) payoffs we are aiming to value (which then means that the individual simulations in the collated set will need to be non-equally weighted).

2.29 The final component of the targeted quantile-spacing approach is therefore to gather information about the spread of strike prices or guarantee levels in the book of exposures we want to value and to select the weightings ascribed to the individual collated simulations accordingly. For example, suppose we wished to collate the original simulations into just 6 collated simulations which (when ranked) we wanted to have weights 0.1, 0.2, 0.2, 0.2, 0.2, 0.1 respectively. Then the lowest 10% = $0.1/(0.1 + 0.2 + 0.2 + 0.2 + 0.2 + 0.1)$ of the original simulations (by rank) would be collated into the lowest ranked collated simulation, the next 20% of the original simulations (by rank) in the next lowest ranked collated simulation etc. A special case is, of course, to equally weight the collated simulations, which is the approach illustrated in Section 2.23.

2.30 More precisely, the final component is to gather information about the effective spread of strike prices applicable in all variants of the valuation we want to carry out simultaneously. For example, as well as valuing the guarantee, we may want to quantify its likely value in adverse (stressed) conditions. Suppose we want not just to calculate the guarantee cost using the stated value of S_0 but also how much this cost might rise if S_0 were 20% lower. Then we might select our collated simulations so that they would include a suitable range of outcomes that practically contributed to the cost of the guarantee if S_0 were 20% lower, even if we would ideally have fewer such simulations this far into the tail of the distribution if our only aim had been to value the guarantees using the base (unadjusted) S_0.

2.31 However, any gathering and processing of such information reduces the 'straight-through' automated nature of the process being followed, which is likely to introduce other costs into the valuation process. An alternative, if simple equal-weighting is considered inappropriate but an automated approach is still wanted, is to select weights formulaically in a manner that is expected to be robust across the plausible spectrum of exercises that might be carried out. For example, the selection of the collated simulations could be structured so that it automatically selected a few points with low weights and with quantiles corresponding to points in the tails of the relevant distribution, with the remainder being more uniformly spread across the spectrum of more likely outcomes. Such a quadrature approach targets greater accuracy in the tails than a pure equally-weighted approach and so will be more robust to exercises in which the distribution of outcomes might in some cases be shifted significantly towards such tails.

Segmenting or otherwise further optimising the valuation algorithm

2.32 We also observe that if our goal is to be able to value a payoff equal to the rolled-up value of the index within a given (small) range of strikes then we only need one collated simulation for that range, i.e. the average rolled up index value, i.e. we would have a probability of occurrence of p_1 and an outcome Q_1 subject to the constraint that $p_1 = 1$ and with Q_1 equal to the (risk-neutral) rolled-up index value. If our goal is also to be able to value a European-style put or call option with a strike within a given (small) range of strikes then we basically

only need two collated simulations, as this introduces four unknowns, p_1, p_2, Q_1 and Q_2 and we can solve for these by requiring $p_1 + p_2 = 1$ (which will ensure that a zero coupon bond is valued correctly), with $p_1 Q_1 + p_2 Q_2$ equal to the risk-neutral rolled-up index value and with the remaining two degrees of freedom used to ensure that, say, two call options, one with a strike at the bottom of the strike range being considered and one with a strike at the top of the strike range respectively are priced correctly.

2.33 This means that if our aim is solely to minimise run-times then a further refinement is possible although it complicates the valuation process. As noted previously, we don't have to use the same collated simulations and weights for every single payoff maturing at a given time. Instead, we could use different weighting schemas for different strikes. In the extreme, we might for a specific range of strikes be able to value them accurately using just a handful of collated simulations (but with this handful varying for different ranges). For example, if we slice up the whole range of possible strikes into small sub-ranges and for each range identify the p_1, p_2, Q_1 and Q_2 needed to price (to a linear accuracy) any vanilla call or put where the strike lies within the given range then we will in effect have captured a piecewise linear approximation to the pricing of European-style vanilla options of the relevant term. Quadratic or higher order interpolation can in effect be achieved by adding further points to the (typically small) collation set used for a given range of strikes and selecting their weights and values appropriately.

2.34 The trick here is to realise that it is usually relatively quick in computing terms to identify where a value would be placed in an ordered list or to otherwise subdivide up possible input values into distinct ranges for which different algorithms can be used[10]. For example, for an ordered list we can do a binary search, each time identifying whether the entry is above or below the half-way ranked point in a range within which we have concluded the entry should lie. This can have an $O(\log q)$ computation cost if there are q such ranges. If the evaluation of the relevant sub-algorithm applicable if the entry falls within a certain range is sufficiently fast (relative to the speed of one that would need to be used if we didn't subdivide the algorithm in this manner) then it becomes computationally efficient to go to the trouble of creating multiple sub-algorithms and choosing the one to use based on some suitable indicator value (e.g. where the ratio of the guarantee level to the opening level of the underlying lies in some suitable table).

2.35 To maximise the efficiency of such a segmented algorithm approach in our example we need to measure the strikes relative to the underlying in a consistent manner across different instruments. In the example used above, the payoff for each instrument (with a given maturity date on a given underlying) is a linear multiple of $\max\left(K_j - S_T e^{-q_j T}, 0\right)$. This means that the relative strikes for such instruments should ideally be measured consistently across these instruments as $K_j / \left(S_T e^{-q_j T}\right)$ or some constant multiple or fraction of this indicator.

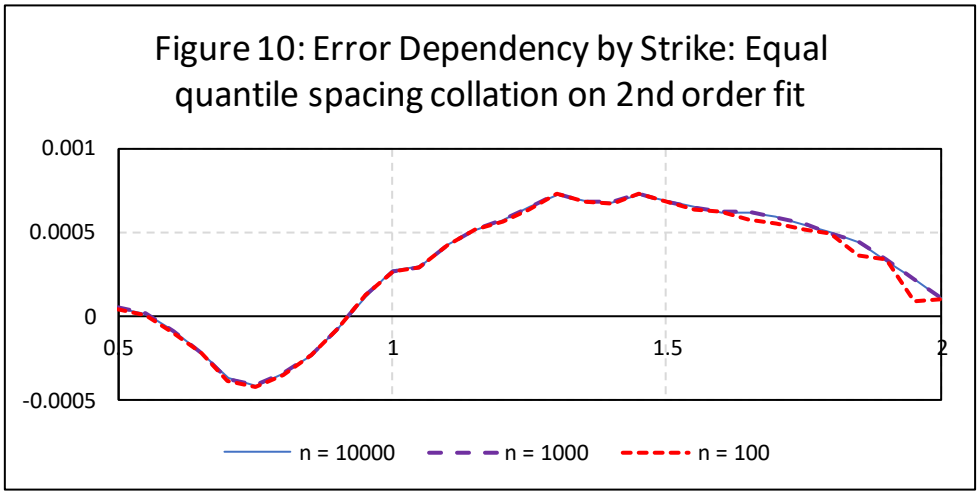
2.36 It should be noted that there is a trade-off here between extent of segmentation and the order of interpolation fit implemented in any given sub-range. Indeed, at one extreme we might not actually segment the range at all but might instead seek a high enough interpolation order for the entire range of possible strikes by using a collation set that is large enough to achieve a suitably accurate representation of the payoff function. However,

[10] As noted previously, mathematical algorithms for calculating standard mathematical functions commonly adopt such a subdivision, e.g. a relatively efficient way of calculating the inverse normal function $N^{-1}(p)$ is to use separate algorithms for three different ranges p might fall within, see Nematrian (2019d), which is based on work originally developed by P.J. Acklam.

many higher-order interpolation approaches can work poorly in the tails of the overall range that an indicator like this might take. To mitigate this risk, we might if planning to adopt a higher-order interpolation approach want to have at least three segments (a high, a low and an in-between segment), and to adopt interpolation approaches that are lower order in the outer two segments than in the in-between one.

Selecting collated set size

2.37 The merits of adopting any refinements as per Sections 2.32 – 2.36 will be influenced by how effective simpler non-segmented (and/or non-higher order interpolated) approaches are at inheriting the increased accuracy of the base simulation set. We show in Figure 10 the error versus the analytical result for a common base simulation set size, $Bn = 10000$, for different collated simulation set sizes, $n = 100, 1000, 10000$ for just the first individual run analysed in the earlier Figures.



2.38 Figure 10 suggests that collating down to just 100 simulations provides very little degradation in simulation accuracy versus collating down to 1000 or 10000 simulations (the latter being equivalent to not collating at all), for this range of strikes and other option parameters.

2.39 For the run analysed in Figure 10, collating down to say just 10 simulations results in much larger absolute errors, particularly at higher strikes. For this run, the highest absolute errors arise for high strikes (with a maximum at $K/S_0 = 1.8$ when the error is -0.008). In relative terms this error is still only around 1.2% of the overall option value, which might be within the margin of error arising from uncertainties in the true values to use for input assumptions (e.g. implied volatilities). In relative terms the error for this run and this collation set size and approach is greatest for low strikes. There turn out to be no simulations low enough to contribute to the valuation at strikes below around 0.6, resulting in the option price below this point being modelled as 0, and therefore having a relative error of -100%.

2.40 This analysis suggests that a collation set size no larger than about 100 is likely to be adequate in nearly all circumstances for this type of derivative, particularly if it is combined with approaches as described in Section 2.31 that aim automatically to include within the collated simulation set (but given reduced weights within it) a handful of outlier simulations.

Summary

2.41 We set out below in Table 1 a summary of the results for the different methods analysed above for a single strike, namely $K/S_0 = 1.25$. With a high enough B, the errors arising from the targeted quantile-spacing approach appear to become similar to ones arising from some analytically selected simulation methods, although depending on the system being used run-time errors such as "Out of Memory" errors may become more likely[11].

Table 1: Estimated errors for different approaches considered above				
Figure	Description	Number of simulations for which payoff evaluated	Av + 1 Std	Av – 1 Std
2	Basic Monte Carlo	100	0.020288	-0.023417
3	Aligned first moment	100	0.008787	-0.010095
4	Aligned first and second moments	100	0.005805	-0.005701
5	High order (analytical) moment fitting	100	-0.000114	-0.000114
6	High order (analytical) moment fitting with additional 2nd order fit	100	-0.000434	-0.000434
7	Simpson's rule type of approach (with *B*=5) (analytical)	300	-0.000020	-0.000020
8	Analytical equal quantile-spacing approach (with *B*=10)	100	0.000009	0.000009
9	Equal quantile spacing collation applied to 2nd order fit (with *B*=100)	100	0.000575	-0.000551
Not shown	Equal quantile spacing collation applied to 2nd order fit (with *B*=1000)	100	0.000391	-0.000434

3. A multi period reinvested asset class index and the martingale test

3.1 The traditional derivative pricing approach to modelling the movement of an asset class through time is to subdivide the whole projection period being analysed into individual sub-periods or time steps and to evolve stochastically an index characterising the asset class return through time step by step, starting at the valuation date[12].

3.2 As in the single period case, it is possible to model using 'real world' probability distributions coupled with deflators or using risk-neutral probability distributions and simulation independent discount factors. As before, we focus here on the risk-neutral approach. For the Principle of No Arbitrage to apply, the present value at outset of a payment of 1 (in the numeraire in which the zero coupon bond is expressed) at time T needs to be the current value of a risk-free zero coupon bond of outstanding term T, so the simulation independent discount factors used need to align with the yield curve that corresponds to the prices of such instruments.

[11] The error exhibited by a given non-analytical run may also exceed the +/- one standard deviation estimate shown in Table 1.

[12] Sometimes it is desirable to subdivide the total return between income and capital return (particularly if there is a difference between the taxes paid on these two types of return), but we do not explore this topic further in this paper.

3.3 We note that modelling the stochastic evolution through time of fixed income securities presents additional challenges, which are outside the scope of this paper. These arise because such instruments, if they do not default in the meantime, will generally mature at par at a given point in the future and stochastic modelling of them needs to incorporate this 'pull to par'[13]. Credit-sensitive instrument also present additional challenges, as they in effect require the introduction of further yield curves or factors that in a risk-neutral world capture similar effects (which may themselves have term structures). Individual credit-sensitive instruments can also undergo price jumps, if they default or become seen as likely to default. It suffices for the purposes of this paper to note that the targeted quantile-spacing approach can also be applied to problems with such elements. Indeed, the approach can be applied whenever it is possible to simulate the future evolution of the underlyings (which typically will always be the case for a well-posed derivative pricing problem).

3.4 For other types of securities, e.g. equities, a traditional derivative-pricing algorithm will typically assume that the rolled-up value of the asset evolves at each step according to a stochastic process that exhibits an expanding funnel of doubt as time progresses. Usually, at least in a risk-neutral formulation, returns in different time steps will be assumed to have no autocorrelation[14], although there may be some assumed heteroscedasticity[15], typically characterised as some assumed term structure to implied volatility. Within these constraints, returns are often assumed to follow some form of Brownian motion, in the sense that the shorter the time step the smaller is the practical range of returns that is observed (with asset prices not exhibiting sudden jumps except because of one-off dividend payments or the equivalent). However, jumps can be included if the modeller thinks this is important. In any case, the underlying valuation model will typically discretise time and the range of returns that might be observed over the resulting time step will not then be infinitesimal in size. Any jump behaviour we care to model can therefore in principle be catered for, provided we choose a suitable one-period distributional form for the behaviour of the asset class over the relevant time step and allow appropriately for heteroscedasticity.

3.5 Suppose the ends of each time step are at times $T_1, \ldots, T_P = T$ (so there are P steps, with the first one starting at time $t = 0 = T_0$). The types of guarantees set out in Section 2.3 are European-style[16] in nature, i.e. have a set maturity date. This means that the present values of all guarantees maturing at time T depend just on the distribution of the cumulative returns between $t = 0$ and $t = T$. Instead of modelling the evolution over individual time steps in isolation, it is therefore generally better (in the sense introduced in Section 1.12) to model the evolution in the following fashion:

(a) Set the rolled-up asset class index at time $t = 0$ to be equal to 1 in all simulations, i.e. $I_{0,k} = 1$;

[13] As the strength of this pull to par varies by how far away maturity is, it may also be necessary to model differently instruments with different durations, but in a manner that constrains how in aggregate they might evolve through time (so that the overall yield curve evolves in a suitable manner and not just individual instruments in isolation dependent on it).

[14] E.g. no tendency for particularly positive returns in one period to be followed by particularly positive returns in the next period (or for particularly negative returns to appear in consecutive periods)

[15] I.e. some tendency for high absolute magnitudes of returns in one period to be followed by higher than average absolute magnitudes of returns in the following period

[16] A European-style option is one that can only be exercised at a specific predefined maturity date. An American-style option is one where the holder can exercise it at any time prior to or at a specific predefined maturity date.

(b) Iteratively identify the sampled distributional form for $I_{j,k}$ by selecting a way of combining the simulated index values $I_{j-1,k}$ at the start of a given time step and the random innovations, $y_{j,k}$ assumed to arise during the time step some so that the $I_{j,k}$ have in aggregate suitable distributional characteristics;
(c) If we need to, to back out the returns over the individual time step, $r_{j,k}$, using

$$r_{j,k} = I_{j,k}/I_{j-1,k} - 1$$

3.6 If we assume that implied volatility has a constant strike structure[17] and a specified term structure then step 3.5(b) can usually be implemented for any given time step by solving a quadratic equation (the coefficients of which depend on the variances and covariance of $I_{j-1,k}$ and $y_{j,k}$ and on the target implied volatility for $I_{j,k}$). The result is the multi-period analogue of the two-moment order fitting approach set out in Section 2.10.

3.7 More accurate, however, is to use a targeted quantile-spacing approach as described in Section 2. For vanilla European-style options, the improvement available will mirror whatever we would have seen in Section 2 had we used the same cumulative returns for the period from $t = 0$ to $t = T$. For other types of option, similar run-time gains are also possible, provided the procedure is adapted to cater for any path-dependent characteristics they might exhibit, see Section 4.23.

Adding a (non-constant) term structure of implied volatility

3.8 In the fitting approach described above, the sampled distributional form for $I_{j,k}$ (where j indexes the time period) is fitted to its assumed 'true' distributional form. The distributional parameters don't need to be constant through time. The approach can therefore easily cater for implied volatilities measured as at $t = 0$ that are not constant over time, i.e. vary with respect to T. Implied volatilities on e.g. equities typically exhibit such a term structure, so it is convenient to include such a feature in a fair (i.e. market consistent) valuation if this term structure is readily observable for the asset class in question.

Adding a (non-constant) strike structure of implied volatility

3.9 Market prices of options often also include a (non-constant) strike structure for implied volatility, i.e. different implied volatilities apply to options of the same term but different strikes. In other words, implied volatility has a two-dimensional *surface*, dependent on term and strike level simultaneously.

3.10 For long-dated guarantees, it may be difficult to identify observables that can reliably guide assumptions about the shape of the option term-strike structure, particularly its strike dependency. Moreover, for funds containing a mix of assets, what we are really interested in is the corresponding structure for the whole fund, not for its individual elements. Diversification both across asset classes and across time may reduce the tendency for (market implied) fat-tailed behaviour and hence option skews to be exhibited by the overall fund.

3.11 However, it will sometimes be important to include such effects, particularly for more sophisticated clients or for derivatives with shorter maturities. If suitable market

[17] By 'strike structure' we mean how implied volatility varies according to the strike price of the option being priced. A constant strike structure is here interpreted as meaning that the same implied volatility is applicable whatever the strike price.

observables can be found to inform the selection of market implied distributions, the targeted quantile-spacing approach described above can be adapted as follows to accommodate such market dynamics:

(a) A market implied risk neutral distribution for the rolled-up index at time T is found that is consistent with market observables (and any expert judgement superimposed on the market observables).
(b) This distribution is used instead of the one that arises from a flat strike structure in Sections 3.5 – 3.6. Thereafter the algorithm is the same, i.e. we create a larger base simulation set, which we collate into a smaller number of collated simulations that suitably average the base simulations within given quantile ranges, the quantile ranges and their weights being chosen to best handle non-smooth elements in the underlying instrument payoffs.

3.12 In practice, this requires the selection of a suitable distributional form from which the distribution in (a) is selected, the fitting of this distributional form to market observables and then sampling from the selected distribution. All are potentially facilitated by using suitable statistical function libraries[18] or mathematical packages with similar functionality.

3.13 However, usually packages favour selecting suitable distributional parameters using e.g. maximum likelihood or other similar fitting criteria, whereas we would ideally like fitting methods that directly target fits of prices of instruments that we can observe in practice. Given the limited availability of market observables, it is also typically necessary to include views on how the distribution behaves beyond regions isolatable from readily available market observables. Distributional fitting and sampling from more complex distributions is also quite complicated and typically quite slow, making it more complex and time consuming to create a large base simulation set. It may then be necessary to limit the size of the base simulation set and hence the effective level of accuracy capturable by using a collated version of it.

4. Multiple asset classes, rebalancing and management actions

4.1 In practice, the funds to which guarantees in the example described in Section 2 may apply will not necessarily invest in a single asset class. Instead, they may invest in a range of assets. It then becomes important for the simulations to be coherent across asset classes, which in general requires us to simulate jointly the returns on multiple asset classes simultaneously. We can conceptually distinguish between two ways in which the asset class mix within a fund may change through time. These are:

(a) Freely drifting asset class mixes. In this case, the fund starts with a specified asset mix at $t = 0$ and any returns on individual asset classes are reinvested exclusively in the same asset class from which they derive.
(b) All other situations, including ones where asset class mixes are regularly rebalanced back to some specified asset mix and/or *management actions* are assumed to apply that will result in the target asset mix being changed in specified circumstances (e.g. the amount in a given asset class may be reduced if its market level falls too far relative to other asset classes within the portfolio). In practice, management actions can involve a decision tree, with second or subsequent management actions getting triggered in some circumstances.

[18] E.g. the probability distributions component of the Nematrian function library, see Nematrian (2019e).

4.2 We explore below the refinements that may be needed to address both (a) and (b), as well as ways in which correlations between asset classes might be incorporated in the valuation framework. Certain steps need to be followed to adhere to the interpolation formulation but once these are addressed the inclusion of the targeted quantile-spacing approach into the simulation exercise is straightforward.

Freely drifting asset mixes

4.3 The simplest case is if the asset class mix within the fund drifts as markets move, with returns on individual asset classes being exclusively reinvested into the asset class from which they derive. The return[19] on the overall fund over the period from $t = 0$ to $t = T$ is then an exact weighted average of the returns on each asset class in isolation, the weights being the opening weights of the fund at time $t = 0$. If we therefore construct simulations for each asset class in the manner described in Section 3 (and include at least a first order moment fitting approach when doing so) then the simulation average for each asset class in isolation will be aligned to the target risk-neutral rolled-up fund return for the combined period. As this target is the same for all asset classes, the average across simulations for the whole fund will also align with this same target return.

Coping with correlations

4.4 However, even when asset classes are freely drifting without rebalancing or other management actions, we still need to arrange for the individual asset classes to co-move in some suitable fashion.

4.5 The simplest way of measuring co-movement is to use correlation coefficients (or, largely equivalently from our perspective, to use covariances). Typically, we might then specify the correlations between asset classes as well as the variances / standard deviations of the relevant asset class returns. The task then becomes identifying a way in which simulations of returns on individual assets can be jointly selected whilst exhibiting the desired correlations.

4.6 This is typically achieved using a Cholesky decomposition[20]. This in effect involves identifying a lower diagonal matrix L with non-negative diagonal elements which when applied to a vector $x = (x_1, \dots, x_n)^T$ of n independent identically distributed zero mean unit normal random variables (so $x \sim N(0, I)$ where I is the identity matrix) results in a vector $y = Lx = (y_1, \dots, y_n)^T$ which is distributed according to $y \sim N(0, A)$ where A is the desired covariance matrix that we wish the variables to exhibit. If A has real entries, is symmetric and is positive definite then it can be decomposed as $A = LL^T$, where L^T is the transpose of L. The entries of L are:

$$L_{j,j} = \sqrt{A_{j,j} - \sum_{k=1}^{j-1} L_{j,k}^2} \qquad L_{i,j} = \frac{1}{L_{j,j}} \left(A_{i,j} - \sum_{k=1}^{j-1} L_{i,k} L_{j,k} \right) \quad for\ i > j$$

4.7 However, from our perspective there is a wrinkle. We ideally want to fit at least the first two moments of the joint distribution, i.e. here to arrange for the sample mean and sample

[19] I.e. the amount that an investment of 1 in a given currency at time $t = 0$ would reach in that currency at time $t = T$.
[20] See e.g. Nematrian (2019f).

covariance matrix to (exactly) match the targeted means and covariances. It is relatively easy to get the means to align by shifting the simulations, i.e. if the simulations for the i'th asset class are $x_{i,k}$ then work out the sample mean as $m_i = \sum_k x_{i,k}$ and if the target mean is M_i replace each $x_{i,k}$by $x_{i,k}^{new} = x_{i,k} + M_i - m_i$. However, it is not so obvious how to get the covariances all to align exactly.

4.8 One solution is to carry out a principal components analysis (PCA)[21] of the original simulations, i.e. from the original sets of $x_{i,k}$ for each of the n different asset classes derive new sets $z_{i,k}$ corresponding to the principal components of the $x_{i,k}$ and then to rescale the $z_{i,k}$ so that for each i they have exactly unit variance. PCA is typically applied to correlated data series, the aim being to find factors that explain as much as possible of the variation arising in the data. It returns a set of orthogonal (i.e. uncorrelated) data series, the first principal component explaining the highest amount of the variation in the original data series, the second principal component the next highest etc. We are not here interested in identifying the fraction of variation explained by each principal component (which we expect to be relatively similar for all identified principal components given that the data is a sample from independent identically distributed variables). We are more interested in the ability of PCA to provide data series that are exactly orthogonal.

4.9 Another solution is to use pseudo-random (also called sub-random) series, e.g. Halton sequences[22]. These are, in some suitable sense, designed to sample as uniformly as possible the multi-dimensional space spanned by the relevant input series. They typically don't provide as accurate a second order fit as a PCA based approach but may provide a smoother fit depending on the asset mix in question[23].

4.10 However, both approaches described above struggle to cater in general for all possible ways in which returns might be related (either with each other or with themselves through time). We typically won't have enough free variables available to get accurate fits to all possible ways in which correlations between different asset classes and between different time periods might be set. Also, we may have rejected the use of correlations and decided to use more complicated ways of characterising co-dependencies between asset classes such as copulas[24]. There are typically no simple ways of establishing the equivalent of exactly uncorrelated simulation series with such approaches.

4.11 In contrast, underlying the targeted quantile-spacing approach is some model which is assumed to have some appropriate coherent approach to the relevant co-dependencies. The large number of simulations used in the base simulation set allows the simulation set to achieve a good fit to this underlying model, which in a suitable sense is then 'inherited' by the collated data set, see Section 2.24.

Adding rebalancing and/or other management actions

[21] See e.g. Nematrian (2019g).
[22] See e.g. Nematrian (2019h).
[23] Results of using such approaches are often not as good as researchers expect given the apparently very helpful features they should exhibit. This appears at least in part to be because to get maximum benefit from the sub-randomness these sequences exhibit you need to have the axes being used by the sub-random sequences to align with axes that are meaningful in the context of the problem at hand. As the asset mix that might be applicable is arbitrary, it is essentially impossible to ensure that this alignment applies for all asset mixes simultaneously.
[24] See e.g. Nematrian (2019i).

4.12 Any modelling of rebalancing / management actions needs to codify the asset mix changes involved by such actions.

4.12 Once codified, the inclusion of rebalancing and/or other management actions might be expected to be conceptually straightforward. For any individual simulation we might expect to be able to derive the overall fund return in any individual time step by calculating the appropriate weighted average of the returns on different asset classes in that simulation for that time step, the weights being as per the codification described above.

4.13 However, if we follow such an approach without modification, we find that the fund's cumulative return averaged across different simulations gradually diverges in a risk-neutral world from the corresponding individual asset class cumulative returns. As time progresses, the average cumulative return for the fund across all simulations typically diverges further and further away from its targeted risk-neutral value. Eventually, the averaged cumulative return can become quite out of line with the corresponding target risk-neutral cumulative return. The divergence typically grows more quickly the larger the amount of asset reallocation involved relative to the freely drifting situation described previously.

4.14 This is a problem, because it means that the simulated fund returns created purely as above will not be *martingales* and will no longer therefore respect the Principle of No Arbitrage. One way of interpreting the effect is to note that inclusion of such rebalancing or other management actions is akin to the carrying out of some form of dynamic hedging that might otherwise be used to hedge some form of option. However, such dynamic hedging typically needs to be done instantaneously, whereas in the above approach it only happens at finite time steps. The corresponding option-like component (and how the instantaneous hedging theoretically needed diverges from the finite time step hedging modelled as above) is however being ignored when we derive the simulated fund simply as above. More formally, a strategy that includes rebalancing and/or other management actions has in effect a non-zero option gamma and thus contributes an additional drift term to the behaviour of the fund, which is ignored if we merely use the approach described in Section 4.12. In practice we can only avoid the divergence (without some adjustment to fund-level returns) by adopting a strategy that has zero option gamma (i.e. by adopting the freely drifting situation described earlier).

Enforcing the martingale criterion

4.15 The potential for such a problem to arise is typically tested for within a risk-neutral valuation exercise by applying a *martingale test*. Typically, this is understood to involve the identification of a suitable test statistic based on the movement in the present value of the rolled up index value (or this index value conditioned on some other variable) and rejecting the assumption that the modelled index series forms a martingale if the relevant weighted average (across the simulation set) of this movement diverges too far (statistically) from zero.

4.16 In the interpolation formulation, the martingale test is automatically enforced for individual asset categories, provided at least a first order moment fitting approach has been included. The sample cumulative return is then forced via the fitting process to align exactly with its desired theoretical value. The corresponding martingale test statistic should then be zero. At issue is that merely enforcing this for individual asset categories in isolation does not ensure that it is also satisfied for combinations of asset categories, if the fund does not just allow the asset mix to drift freely as above.

4.17 The solution is to include for each fund a further step that enforces the martingale property for that fund, by applying a further first order moment fitting to each fund's simulated cumulative returns. This typically involves adding to all simulations for a given fund and time step a (constant across simulations) term that aligns the sample averaged cumulative return for that fund with the targeted risk-neutral cumulative return.

4.18 Although not strictly accurate, such an adjustment can typically be thought of as akin to some uncertainty in how and when within the relevant time step the asset reallocation needed might take place.

Other issues relating to non-infinitesimal time steps

4.19 For funds that are merely rebalanced back to fixed asset weights the drift adjustment required as per 4.17 is typically quite small for individual time steps and reduces in size as the time step being used reduces. Of course, the shorter the time step, the longer the valuation run-time (as calculations are needed at each time step for each instrument or instrument grouping included in the computation). Some suitable compromise is likely to be needed between shorter time steps and more precise rebalancing[25].

4.20 Other forms of management actions may involve larger step changes to asset allocations if some trigger occurs. An issue then becomes how quickly management would in practice react to such triggers. In theory, we might aim to have the time step size aligned with the speed at which the management action might be implemented, but this might result in excessively small step sizes. In practice, refinements may be included in the time step projection to capture the impact of intra-period movements. Similar issues arise with barrier options (and with many other types of path-dependent options).

4.21 A problem that can in principle arise if the step size is too long (and if the fund is quite volatile and overall projection period is too long) is that the across-simulation adjustments needed in 4.17 can be too large to keep the simulated index values non-negative in all circumstances. Two possible solutions are:

(a) Apply a suitable lower bound on any given simulated fund index value, in which case the overall average return to the end of that period will be a little inflated but likely corrects itself relatively soon afterwards
(b) Apply the required correction only to some (typically higher return) simulations, which can allow us to keep the average correct but might possibly create some bias in the volatility.

Applying the targeted quantile-spaced approach

4.22 However, no fundamental refinement is in principle needed to apply the targeted quantile approach once some suitable solution has been found in a more traditional simulation framework to address issue such as those noted in Sections 4.15 to 4.21. It again involves preparing a larger base simulation set using whatever model we have identified is appropriate for the derivative instrument type in question and creating a smaller collated set from this larger set as described previously.

[25] We've also here assumed that rebalancing happens effectively instantaneously, whereas in practice it may merely happen say weekly, monthly, quarterly or even less frequently.

Nested simulations

4.23 Run-times when valuing and analysing derivative books can become particularly large using traditional Monte Carlo techniques if the valuation requires *nested* simulations. These can arise in a variety of situations including:

(a) Exercises where we are projecting forward the book on a non-market consistent basis but need then to value the book at one or more future points in time on a market consistent basis
(b) Options that are path-dependent, e.g. American-style, barrier, lookback or other path-dependent options.

4.24 For example, suppose we wish to evaluate the one year expected shortfall or tail value-at-risk of a book of derivatives. This would involve evaluating the (market consistent) value of the book in one years' time in a variety of (adverse) situations, and then averaging (integrating) over the probabilities of occurrence of these situations happening.

Likewise, to identify the value of an American-style option we in theory need to identify whether at a given point in the projection it is then optimal to exercise the option, which in principle requires valuing the option at points in time further into the future assuming that exercise is deferred, to see if the option is then worth exercising.

Sometimes there can be more than two nesting levels, e.g. if we want to calculate an expected shortfall of a book that includes path-dependent options.

4.25 The targeted quantile-spacing algorithm again offers the potential for significant run-time improvements for such exercises. For example, suppose we need to calculate the one-year expected shortfall of a derivative book consisting of European-style options on one or more underlyings. We would prepare a large base simulation set that captured simulations of the underlyings. We would first sub-divide this base set into sub-base sets that are conditional on (i.e. group together cases where) the underlyings (jointly) reach specific levels in one year's time. We would then create collated sets from each of these sub-base sets, and potentially further collate down these sets (which would differ according to the levels the underlyings had reached in one year's time) into a smaller number of collated sets that best characterised a smoother progression of levels reached then. The collated simulation set data would typically be more complicated than in the non-nested case[26], but the potential run-time efficiency gains would likely be larger (versus the traditional Monte Carlo approach) given the extra complexity of the problem being analysed. The precise form of the problem would typically influence exactly how the collation was best done. For example, the most important contributions to an expected shortfall calculation tend to come from large up or down movements in one or more of the underlyings, so the collation might preferentially target ability to value these scenarios appropriately by including greater numbers of simulation points in the relevant extremities of the simulation distribution (with appropriately adjusted probabilities of occurrence).

5. Further comments on use within the financial sector

[26] For example, if evolution of interest rates is not independent of evolution of the underlying(s) then we would also need to capture stochastic forward discount rates (i.e. simulated one-year forward zero-coupon bond prices) so that we could value instruments appropriately at the one-year valuation point.

5.1 In absolute terms, the run-time and hence cost benefits of the targeted quantile-spacing approach and the interpolation formulation on which it builds offer are likely to be larger for bigger entities with more complex valuation requirements. However, the approach appears to have broad applicability and is reasonably simple to implement whenever a more traditional simulation-based technique could otherwise be applied. It should therefore also offer run-time and cost benefits even to smaller entities.

5.2 For reasons explained in Kemp (2009), it seems likely that there will be a continuing regulatory desire to expand the use of fair (i.e. market consistent) valuations in some financial sectors. For example, the European Insurance and Occupational Pensions Authority (EIOPA) is currently encouraging EU pension funds (also called institutions for occupational retirement provision or *IORPs*) to apply a 'common assessment framework' to their liabilities, e.g. via its specification in EIOPA mandated industry-wide stress tests. This framework targets market consistency and therefore expects IORPs to use appropriate techniques to value option-like exposures that may be present within their balance sheets. The liability structures of many defined benefit ('DB') pension funds include option-like components[27]. Option-like components can also arise from the benefit security mechanisms prevalent for such entities[28]. An approach such as the targeted quantile-spacing approach that makes it more practical to determine such values to a suitable level of accuracy may therefore be appealing to entities and regulators in sectors where such regulatory pressures are present, provided the exposures involved are complicated enough to require valuation using simulation techniques.

5.3 We have concentrated above mainly on straightforward underlyings. The targeted quantile-spacing approach can also be used to price more complicated underlyings, e.g. payoffs that depend on an aggregate underlying that is the greater of (or some other combination of) multiple other underlyings with or without floors or ceilings. However, the more complex the underlying the fewer instruments there are likely to be in the book that depend on it.

5.5 It is not necessary to use the targeted quantile-spacing approach only when there are multiple instruments sharing the same underlying, although this does increase the run-time gains the approach offers. Whether an overall run-time gain arises if we are applying the approach to, say, just a single instrument will depend on the run-time per simulation required for the instrument versus the run-time required to create and then collate a set of simulations together[29].

5.5 We might also use different B's (i.e. ratios of sizes of base to collated simulation sets) and/or different levels of algorithm segmentation / interpolation order (as per Sections 2.32 – 2.36)

[27] E.g. pensions in payment may be subject to increases linked to inflation but these increases may be subject to annual or multi-year ceilings and floors.

[28] For example, some DB pension funds have conditional benefit structures, where benefit payments can be reduced if markets perform poorly. These can be modelled as akin to 'management actions' as per Section 4. Others may rely on sponsor support should their assets perform poorly. To quantify the market consistent value of this sponsor support we may therefore need to identify how much support might be called upon in adverse circumstances and then to incorporate an allowance for the credit risk that the sponsor might not be able to afford to provide the support when needed. Pension funds with a weak or non-existent *sponsor covenant* may also be protected by national pension protection schemes but these may not guarantee the whole of the applicable pension benefits, again introducing some option-like elements into the liability valuation, or at least into how the valuation might be split between different component parts

[29] Even if there is no such run-time gain for a specific instrument type, we might still benefit from applying the technique, e.g. if the nature of the book is not well analysed in advance but we expect that in most cases the approach will lead to efficiency gains.

for different parts of the book (or different elements of a nested simulation exercise), depending on where it is expected to be most beneficial to target run-time efficiencies.

5.6 We have not so far discussed in detail what any valuations derived as per earlier parts of this paper might be used for. One motivation for preparing such valuations is to assist in the making of decisions on how best to manage the (market or other) risks encapsulated in the relevant book of derivatives. In this context, calculations that aim to establish the sensitivity of the valuation to changes in input parameters (e.g. here prices of the underlying, implied volatilities, …) can be particularly run-time intensive, because a common way of calculating such sensitivities[30] is to repeatedly re-run the analysis incorporating small change to the inputs. Such exercises are amenable to algorithmic differentiation techniques, but only typically if the original problem can be solved analytically (which can then allow derivation of analytical expressions for the relevant partial derivatives). Firms wanting to actively manage such exposures may also want to estimate these sensitivities or carry out other stress and scenario simulations relatively frequently, increasing the incentive towards run-time efficient valuation methodologies.

6. Broader applicability

6.1 The main novel approach introduced in this paper is the targeted quantile-spacing approach (together with refinements as set out in Sections 2.32 – 2.36 that can further enhance its run-time properties). We have so far in this paper concentrated on how this approach can be applied in the financial arena, particularly to derivative pricing problems and related exercises.

6.2 However, provided a problem is amenable in the first place to more traditional simulation-based numerical integration techniques, it should also be amenable to the use of enhancements like the targeted quantile-spacing approach and its refinements. Whenever run-time is a constraint and can be improved upon by the targeted quantile-spacing approach, it should potentially offer meaningful benefits. This means that it should be applicable to many problems outside the derivative pricing (or even finance) arena that might otherwise be addressed using more traditional simulation-based techniques including ones that in effect involve numerical integration.

6.3 Integration techniques (including numerical integration techniques) are used in a very diverse range of engineering and physical sciences fields, from areas such as estimation of protein folding characteristics to artificial intelligence algorithms, automotive and other engineering control processes and climate change modelling. A link that motivates many of these applications is the close association between optimisation and integration of some suitable control function characterising the optimisation problem. Solutions to differential equations can often be expressed as integrals and therefore numerical approaches to solving them can also share similar characteristics.

6.4 Consider, for example, an engineering control process that involves a physical system containing the following elements. The overall structure we articulate below is generic and characterises many different control processes in many different fields. In brackets we have included what the elements might look like in a specific example drawn from aeronautics in which the system involves an aircraft wing and the aim is e.g. to avoid excessive vibration, where ‘excessive’ is defined in some manner that here might be relatively granular (as we

[30] For options or books of options, these sensitivities are often called the option ‘greeks’.

might here want to avoid excessive vibration anywhere in the wing and not just for the wing as a whole) or to further some other engineering goal for the system:

(a) One or more sensors measuring the state of the system at one or more points (in space, e.g. pressures or other physically relevant measures at different points along the wing surface, or in time, e.g. the level of vibration at recent past time points, or both)
(b) Features of the system on which higher or lower utility values can be placed (e.g. extent of vibration of specific parts of the wing, where a penalty could be imposed for higher vibration there linked to the impairment of the air-flow dynamics across the wing, the fuel inefficiency it might lead to or the level of passenger discomfort it might cause)
(c) One or more control processes (implemented in hardware, software or both) that interpret the outputs of the sensor(s) in (a) and select appropriate actions bearing in mind the utility values specified to outcomes in (b)
(d) One or more actuators and/or warning indicators driven by the process(es) in (c), that either alter the system being controlled (e.g. alter the speed of the aircraft or alter configurable elements of the wing surface) or that flag up the need for some other remediation action by a system controller.

6.5 We might in the control process (i.e. step (c)) seek to simulate how the system might evolve in the future conditional on the current and recent past sensor readings as per (a), placing different utilities on different states of the system as per (b) with the control outputs from step (c) being used either to automatically apply actions as in step (d) that aim to target some suitable overall outcome for the system, or that indicate to a system (e.g. human) controller that some reaction may be desirable. Suppose the control process needs to operate rapidly (e.g. in real-time) and suppose any simulations it uses are reasonably complex to process (as might be the case here if the aim is to simulate how the airflow across the wing might evolve within the time it might take for any control action to make a difference). A premium may then be placed on simulation approaches that are relatively rapid as well as relatively accurate.

6.6 Such a characterisation directly parallels the derivative pricing problem we have described in Sections 2-4. The sensor readings in (a) can be viewed as analogous to the prices of the underlyings (and other market observables) since they provide the opening values to be fed into the simulations of how elements of the system might then evolve. The elements of the system that influence contributions to utility can be viewed as analogous to individual derivative payoffs. The overall utility is analogous to the valuation of the whole book of derivatives. If rapid accurate estimation of the overall change in utility arising from different possible control responses is helpful, and if modelling of the evolution of factors influencing this utility requires or benefits from simulation-based techniques then use of an approach like the proposed targeted quantile-spacing approach that is faster than more traditional simulation approaches becomes appealing.

6.7 However, there is a wrinkle within the characterisation given in Section 6.6. This is the implicit assumption that practical process control approaches will often include simulation elements. Many more sophisticated practical process control approaches currently adopt the *model predictive control* (MPC[31]) formalism (or simpler equivalents), but often only classical (deterministic) types of MPC. The MPC formalism has been used in e.g. chemical plants and oil refineries since the 1980's and more recently has been extended to e.g. power system management and power electronics. The classical MPC formalism involves or

[31] Also known as 'receding-horizon control', given the finite number of past state measurements incorporated at any given point in time in the control process.

embeds control algorithms that identify in a deterministic manner target changes to make to specific variables (e.g. pressure, flow, temperature) in order to return some output to close to its target (usually whilst also respecting applicable constraints on some variables e.g. that pressures should not exceed specified critical levels)[32].

6.8 Traditional (classical) MPC approaches (and simpler variants) are effectively deterministic in nature, i.e. the control output is essentially deterministically derived from the inputs. Implicit in such approaches is the assumption that the system dynamics can be adequately modelled deterministically. Use of simulation techniques then gets focused on the design stage, i.e. the specification of the structure and characteristics of the process control, rather than in its live running. Often the MPC is linear (possibly with the inputs being transformed in a manner that converts the problem into a linear one), as it can then usually be conveniently codified using matrix algebra. The feedback mechanism within the MPC is implicitly assumed to compensate adequately for structural mismatches between the model and the process being modelled.

6.9 The deterministic modelling of a physical system in the manner implicit in a classical MPC can in the context of this paper be thought of as analogous to trying price a derivative using merely a single deterministic projection of how the underlying might move in the future. For some derivatives, namely linear ones like futures and forwards, a correct price can be obtained using a deterministic approach. However, for non-linear derivatives such as options, the output of a deterministic model can be quite misleading and a stochastic approach (or some analytical equivalent as per Section 1.7) is preferable.

6.10 The need to include randomness in the modelling process is perhaps particularly evident in the derivatives field, given the existence of clearly random contributions to the future evolution of asset prices and the underlying rationale for many derivative transactions, which often involve players wanting to manage or hedge exposures to risks arising from this randomness. It is maybe not quite so evident in the engineering process control field, at least within the paradigm underlying (deterministic) MPCs, which implicitly assumes a deterministic nature to the system of which the MPC is a part.

6.11 Some modern engineering process control techniques do explicitly recognise that the systems concerned do not evolve in purely deterministic ways, that randomness is an inherent feature of the real world and that in any case errors may exist in any state measurements being inputted into the applicable control processes. Mesbah (2016) highlights the exceptional performance of MPC for high-performance control of complex systems but also highlights its inadequacies for systematically dealing with such uncertainties. He highlights the development in recent years of stochastic optimal control techniques and focuses on *stochastic model predictive control* (SMPC) which has the aim of systemically incorporating the probabilistic descriptions of uncertainties into a stochastic control problem. He notes that the ability to regulate the probability distribution of system states/outputs is important for the safe and economic operation of complex systems when the control cost function is asymmetric. His paper includes references describing prior applications of SMPC (both linear and non-linear) to at least the following fields: air traffic control, automotive applications, building climate control, microgrids, networked control

[32] Formulated in this generalised manner, most other simpler control processes can be viewed as special cases of a MPC, including e.g. a proportional-integral-derivative (PID) control which continuously calculates an error between the desired setpoint for a given process variable and its current measured value and applies a correction based on proportional, integral and derivative terms (but often without placing practical constraints on the size of the resulting target control adjustment).

systems, operation research and finance, process control, robot and vehicle path planning, telecommunication network control and wind turbine control. He argues that SMPC allows for the systemic trade-off between fulfilling control objectives and guaranteeing a probabilistic constraint satisfaction due to uncertainty.

6.12 A key challenge Mesbah (2016) notes that currently constrains the use of SMPC in non-linear systems is the efficient propagation of stochastic uncertainties through the system dynamics. The analogous challenge when pricing non-linear derivatives is the efficient propagation of the impact of stochastic uncertainty in future asset price movements to the end valuation result (and hence to the sensitivity of this result to adjustments such as hedging strategies that we might want to introduce into the derivative book). This is precisely the challenge that earlier parts of this paper have sought to address. Incorporation of run-time efficient simulation approaches such as the targeted quantile-spacing approach described in this paper (perhaps particularly if refined as in Sections 2.32 to 2.36 for a control process that is used continuously in real-time) therefore potentially make application of SMPC or other stochastic optimal control approaches more practical and beneficial in a wide range of fields, including ones already noted in Section 6.11.

6.13 Not all SMPC techniques make use of simulation-based approaches. For example, Garifi et al. (2018) explore the use of SMPC for demand response management in a home energy management system. They propose use of chance constrained MPC-based optimisation, in part, it seems, because they consider that the use of Monte Carlo sampling for representing uncertainties in various parameters such as outdoor temperature and renewable energy source generation would be computationally restrictive. This suggests that approaches such as those described earlier in this paper that mitigate these computational challenges will likely also make simulation-based approaches more applicable in such fields. Such a development would mirror experience in the finance field, where increased availability of computing resource has seen increased use of simulation-based approaches including in areas where previously such approaches would have been considered impractical.

References

Garifi, K., Baker, K., Rouri, B. and Christensen, D. (2018). Stochastic Model Predictive Control for Demand Response in a Home Energy Management System. *National Renewable Energy Laboratory*

Kemp (2009). *Market consistency: Model calibration in imperfect markets*. John Wiley and Sons

Mesbah, A. (2016). Stochastic Model Predictive Control: An Overview and Perspectives for Future Research. *IEEE Control Systems Magazine*, Vol 36, Issue 6, Dec 2016

Nematrian (2019a). Types of functions included in the Nematrian Function Library. http://www.nematrian.com/FunctionLists.aspx. *Nematrian*, viewed April 2019

Nematrian (2019b). Formulae for prices and Greeks for European (vanilla) puts in a Black-Scholes world. www.nematrian.com/BlackScholesGreeksVanillaPuts.aspx. *Nematrian*, viewed April 2019

Nematrian (2019c). Accelerated Convergence Techniques. www.nematrian.com/AcceleratedConvergenceTechniques.aspx. *Nematrian*, viewed April 2019

Nematrian (2019d). Inverse Normal: Function Description. http://www.nematrian.com/MnInverseNormal. *Nematrian*, viewed April 2019

Nematrian (2019e). Probability distributions. http://www.nematrian.com/ProbabilityDistributionsIntro. *Nematrian*, viewed April 2019

Nematrian (2019f). Enterprise Risk Management Formula Book. http://www.nematrian.com/ERMFormulaBookMonteCarloMethods. *Nematrian*, viewed April 2019
Nematrian (2019g). Blending Independent Components and Principal Components Analysis. http://www.nematrian.com/IndependentComponentsAnalysis. *Nematrian*, viewed April 2019
Nematrian (2019h). Halton Sequence: Function Description. http://www.nematrian.com/MnHaltonSequence. *Nematrian*, viewed April 2019
Nematrian (2019i). Copulas – A short introduction. http://www.nematrian.com/CopulasIntro. *Nematrian*, viewed April 2019
Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (2007). Numerical Recipes: The Art of Scientific Computing. et al. (2007).

Appendix A: Analytical option pricing formulae

Examples of analytical option pricing formulae include the Black-Scholes (technically the Garman-Kohlhagen) formulae for put (P) and call (C) options:

$$P = -Se^{-qT}N(-d_1) + Ke^{-rT}N(-d_2)$$
$$C = Se^{-qT}N(d_1) - Ke^{-rT}N(d_2)$$

where:
T = time to maturity
S = price of underlying
K = strike price
q = dividend rate on underlying (continuously compounded)
r = interest rate on a risk-free ZCB expiring at time T (continuously compounded)
σ = implied volatility (annualised) of underlying (relative to a ZCB expiring at time T, as this is the instrument that in conjunction with the underlying needs to be used to hedge the option)
$N(x)$ = cumulative unit normal distribution function
and:

$$d_1 = \frac{\log(S/K) + \left(r - q + \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}} \quad and \quad d_2 = \frac{\log(S/K) + \left(r - q - \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}$$