MANAGEMENT SCIENCE
Vol. 53, No. 11, November 2007, pp. 1756–1769
ISSN 0025-1909 | EISSN 1526-5501 | 07 | 5311 | 1756

DOI 10.1287/mnsc.1070.0734

Simulation of Coherent Risk Measures Based on Generalized Scenarios

Vadim Lesnevski
Royal Bank of Scotland, London, United Kingdom, vadim.lesnevski@rbs.com

Barry L. Nelson, Jeremy Staum
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208
{nelsonb@northwestern.edu, j-staum@northwestern.edu}

In financial risk management, coherent risk measures have been proposed as a way to avoid undesirable properties of measures such as value at risk that discourage diversification and do not account for the magnitude of the largest, and therefore most serious, losses. A coherent risk measure equals the maximum expected loss under several different probability measures, and these measures are analogous to "populations" or "systems" in the ranking-and-selection literature. However, unlike in ranking and selection, here it is the *value* of the maximum expectation under any of the probability measures, and not the *identity* of the probability measure that attains it, that is of interest. We propose procedures to form fixed-width, simulation-based confidence intervals for the maximum of several expectations, explore their correctness and computational efficiency, and illustrate them on risk-management problems. The availability of efficient algorithms for computing coherent risk measures will encourage their use for improved risk management.

Key words: simulation; ranking and selection; good deal bounds; coherent risk measures; risk management
History: Accepted by Michael Fu, stochastic models and simulation; received June 2, 2005. This paper was with the authors $1\frac{1}{2}$ months for 3 revisions. Published online in *Articles in Advance* October 19, 2007.

1. Introduction

Both poor risk measures and scarcity of computational resources hamper effective risk management. For example, value at risk (VaR) is currently used by nearly all major financial institutions and is enshrined in the international regulatory framework of the Basel accords. The owner of a portfolio may experience a loss, and the goal of risk measurement is to quantify the risk inherent in this possibility of loss. VaR is a quantile of the distribution of this loss, having the interpretation of the largest likely loss. One of VaR's flaws is that it can discourage diversification, which would reduce risk, while enabling and encouraging business units to hide risks by subdividing portfolios into different accounts, thus making it more difficult for risk managers and regulators to perform their supervisory functions. Another flaw is that VaR fails to take into account the magnitude of the largest losses, which pose the gravest danger. As a result, financial institutions and regulators are considering moving away from VaR toward superior risk measures, primarily coherent risk measures of the type introduced by Artzner et al. (1999), as a suitable basis for financial risk management. Coherent risk measures are also applicable to the problem of pricing derivative securities with good deal bounds. Under some conditions, the resulting bid and ask prices can be expressed in terms of coherent (or convex) risk measures (Jaschke and Küchler 2001, Staum 2004).

The practice of financial risk management and derivative security pricing frequently involves intensive computer simulation. With this application in mind, we develop sequential (multistage) simulation procedures that generate a fixed-width, two-sided confidence interval for a coherent risk measure that is the maximum of several expectations. The availability of efficient algorithms for computing coherent risk measures will facilitate improved risk management.

Any coherent risk measure ρ with suitable continuity properties has a representation of the form

$$\rho(Y) = \sup_{\mathbf{P}\in\mathcal{P}} \mathrm{E}_{\mathbf{P}}[-Y/r], \tag{1}$$

where Y is the value of a portfolio at a future time horizon, $1/r$ is a stochastic discount factor which represents the time value of money, and $\mathcal{P}$ is a set of probability measures (Delbaen 2002, Theorem 3.2). Equations of a similar form exist for the related problems in derivative security pricing. We simplify the problem somewhat by assuming that the set $\mathcal{P}$ has only a finite number k of elements $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_k$. This assumption often holds, for example, when the decision maker designs the coherent risk measure (or the underlying acceptance set, in the case of derivative

security pricing) by specifying k generalized scenarios. The assumption also covers approximation of $\mathcal{P}$ by the convex hull of k probability measures. Let $X := -Y/r$ and $\mu_i := \mathrm{E}_{\mathbf{P}_i}[X]$. The risk measurement (1) involves a single random variable X, which is a negative discounted portfolio value or a discounted loss, viewed under multiple probability measures. For clarity in discussing simulations, let X_i be a random variable whose distribution under the probability measure Pr is the same as that of X under $\mathbf{P}_i$, that is, such that $\Pr\{X_i \le x\} = \mathbf{P}_i[X \le x]$.

Financial simulations typically require large samples, so we assume, for purposes of theoretical analysis, that sample averages of each X_i are approximately normally distributed. Therefore, we study inference for $\max_{i=1,2,\ldots,k} \mu_i$ based on data $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $j = 1, 2, \ldots,$ where the means and variances are all unknown. This problem is the same as that studied in the literature on ranking and selection, in which the primary goal is inference about the identity of the maximum (Bechhofer et al. 1995). Because of this commonality, the results presented here are applicable to the problem of selecting the best system if one is also interested in knowing the mean of the best system, which is different from estimating the mean of the selected system. For convenience, we will refer to "system i" and to μ_i and σ_i^2 as its mean and variance, rather than referring to probability measure $\mathbf{P}_i$ and to the mean and variance of X under it.

The problem of *estimating the maximum* is more difficult than that of *selecting the best*. To see this, we introduce more notation. Define $[i]$ as the index of the ith smallest mean, $\mu_{[i]}$. Thus, $\mu_{[k]} = \max_{i=1,2,\ldots,k} \mu_i$ is the largest mean, which we want to estimate. Let $\hat{\mu}_i$ be an estimator for μ_i. An obvious choice is $\hat{\mu}_i = \overline{X}_i$, the sample average of the random variable X_i. The problem features a natural bias: the most obvious estimator $\max_{i=1,2,\ldots,k} \overline{X}_i$ is an upper bound for, and has a larger expectation than, $\overline{X}_{[k]}$, whose mean is $\mu_{[k]}$. Even maximum likelihood estimation for this problem is not simple and produces remarkable results (Dudewicz 1971). The effect of positive bias in estimating the maximum, applied to risk management, would be overestimation of risk, resulting in excessively conservative oversight and unduly high capital charges for risky activities.

The attraction of the fixed-width confidence-interval approach is that it avoids the need to directly quantify the bias in $\max_{i \in I} \hat{\mu}_i$ as an estimator for $\mu_{[k]}$; instead, we simply take the confidence-interval width L small enough so that the error is negligible relative to the decision that must be made.

Our starting point is a two-stage procedure for forming a fixed-width confidence interval for the largest mean of k independent normal populations due to Chen and Dudewicz (1976). We enhance the Chen-Dudewicz procedure in a number of ways so as to make it useful in the type of risk management simulations we have in mind. Specifically, we use screening ideas from ranking and selection to reduce drastically the number of systems that need to be simulated to estimate the maximum, and we use variance-reduction techniques to sharpen the screening and reduce the total sample size required for estimating the maximum. To sharpen screening, we employ common random numbers (CRN; see Law and Kelton 2000) to induce positive correlation between the systems and thereby reduce the variance of their differences. To reduce the number of replications required for estimation, we employ control-variate estimators (CV; see Law and Kelton 2000) to exploit strong correlation between the response of interest, X, and a collection of random variables with known expectations, called control variates. Control variates are often plentiful in financial simulations where the risks associated with individual components of a portfolio or the values of simple financial instruments are easily computed. The introduction of screening, CRN, and CV required significant methodological advances that we report here.

In Online Appendix B (see the e-companion for all online appendices),[1] we prove the validity of the new procedures, under specified conditions, including normally distributed data. We study the robustness of a more advanced version of our procedures to non-normality in Lesnevski et al. (2006). For the sake of simplicity and convenience, we employ an approximation to a sample size formula which is required for validity. This approximation is quite accurate, as discussed in Nelson and Staum (2006).

Another approach to estimating the largest mean is to use a procedure for selecting the best system, then to estimate its mean using independently generated observations: in Online Appendix C.3, we discuss this approach and explain why it tends to be less efficient.

Section 2 contains an exposition of two examples that motivated our development of new procedures for estimating coherent risk measures by simulation. Section 3 explains our framework for proving bounds on error probabilities of a fixed-width confidence interval for the largest of k means. In §4, we use this framework to develop several procedures for generating such a confidence interval. The reader who is more interested in applying our procedures than in theoretical analysis may be served best by reading Online Appendix A, which contains algorithms, instead of §4. Section 5 provides a computational evaluation of the efficiency and coverage of the procedures developed in §4 when applied to the examples

[1] An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

of §2, leading to conclusions in §6. The e-companion to this paper contains the appendices.

2. Motivating Examples

2.1. Basket Put

We will test the performance of our procedures in pricing a basket put option. This is a derivative security whose payoff at a terminal time T is $\max\{0, K - w'S(T)\}$, where K is a contractually specified strike price, w is a vector of weights, and $S(T)$ is the vector of terminal prices of the securities in the basket. The basket put is the right to sell the basket of securities for the strike price K at time T. If the underlying security price vector S obeys the Black-Scholes model (see, e.g., Shiryaev 1999, Chap. VII, §1b), the basket put's price should be its expected discounted payoff.

Under the Black-Scholes model, the price vector S follows multivariate geometric Brownian motion with drift r, the risk-free interest rate, and with covariance matrix Σ. That is, $\ln S_j(T) = \ln S_j(0) + (r - \|A_j\|^2/2)T + A_j Z\sqrt{T}$, where A is a matrix satisfying $AA' = \Sigma$, $\|A_j\|$ is the Euclidean norm of its jth row, i.e., the volatility of the jth asset, and Z is a multivariate standard normal random vector. The short-term interest rate r is observable, and there are standard methods for calibrating the underlying securities' individual volatilities $\|A_j\|$, whether from historical data or by fitting to observable prices of market-traded options on the underlying securities: see Cont and Tankov (2004, Chaps. 7 and 13) and Shiryaev (1999, Chap. IV). However, estimation of the nondiagonal elements of Σ poses a greater problem. For pricing the basket put, the crucial quantity is $\|w'A\|$, the volatility of the basket, and this depends strongly on the correlations between assets. There may be a range of plausible correlations and thus a range of plausible prices for the basket put.

In this example, the basket is a weighted average of three security prices with weights $w_1 = 0.5$, $w_2 = 0.3$, and $w_3 = 0.2$. The initial security prices are all 100, and the strike price is $K = 85$. The interest rate $r = 5\%$ and the volatilities are $\|A_1\| = 40\%$, $\|A_2\| = 30\%$, and $\|A_3\| = 20\%$. To account for uncertainty about correlations, we use the $k = 4^3 = 64$ probability measures produced by allowing each of the three pairwise correlations to be 0.2, 0.35, 0.55, or 0.75. Although the payoff in this example is far from normally distributed, the sample averages are approximately normally distributed, and the minimum coverage guarantees for the confidence limits held in all our experiments, which include 5,000 independently simulated confidence intervals (§5).

The three control variates used in this example are the discounted payoffs of put options with strike K on each individual asset in the basket. Their means are given by the Black-Scholes pricing formula, based on the known volatilities.

2.2. Options Portfolio

In this example, we assess the risk of a portfolio of European-style call and put options on three assets with initial prices of 100 and terminal prices $S_1(T)$, $S_2(T)$, and $S_3(T)$. All options in the portfolio expire at a terminal time T. We also consider a market index whose terminal level is $S_0(T)$. For each of $j = 0, 1, 2, 3$, $S_j(T)$ follows geometric Brownian motion with drift d_j and volatility σ_j, so $\ln S_j(T) = \ln S_j(0) + (d_j - \sigma_j^2/2)T + \sigma_j W_j \sqrt{T}$, where W_j is standard normal. There is a one-factor model of dependence among the assets: let λ_1, λ_2, λ_3 be constant "factor loadings." Under probability measure $\mathbf{P}$, Z_0, Z_1, Z_2, and Z_3 are independent standard normal random variables, $W_0 = Z_0$, and $W_j = \lambda_j Z_0 + \sqrt{1 - \lambda_j^2} Z_j$ for $j = 1, 2, 3$. In this model, Z_0 corresponds to the market factor common to all assets, while Z_1, Z_2, and Z_3 are idiosyncratic factors corresponding to each individual asset.

The risk measure we consider in this setting is the maximum expected loss incurred while holding the portfolio, where the maximum is taken over $4^4 = 256$ conditional expectations given a generalized scenario. Of the probability measures $\mathbf{P}_i$ in Equation (1), 255 are defined by $\mathbf{P}_i[E] = \mathbf{P}[E \mid A_i]$ for some event A_i of probability $\mathbf{P}[A_i] = 1/20 = 5\%$, while the 256th probability measure is $\mathbf{P}$ itself. This risk measure is similar in spirit to worst conditional expectation (Artzner et al. 1999, §5). We construct generalized scenarios by restricting some of the factors Z_0, Z_1, Z_2, and Z_3. Each of the factors can be "up" (corresponding to a large increase of the asset price), "down" (a large decrease), "middle" (not extreme), or "unrestricted." The probabilities of the restrictions on the restricted factors are always equal. For example, letting Φ be the standard normal distribution function, in the scenario "up-down-unrestricted-unrestricted," Z_0 is sampled conditional on exceeding $\Phi^{-1}(1 - 1/\sqrt{20})$, Z_1 is sampled conditional on being below $\Phi^{-1}(1/\sqrt{20})$, while Z_2 and Z_3 are not restricted. By independence among Z_0, Z_1, Z_2, and Z_3, the probability of this event is 1/20. The time horizon T is one week, and the parameters were calibrated using three years of historical weekly data on the S&P 500 index and shares of Intel (INTC), ExxonMobil (XOM), and Microsoft (MSFT). The result was the annualized volatilities $\sigma_1 = 39.8\%$, $\sigma_2 = 19.3\%$, and $\sigma_3 = 27.0\%$ and the factor loadings $\lambda_1 = 0.617$, $\lambda_2 = 0.368$, and $\lambda_3 = 0.785$ to match the observed correlations. Because one week is such a short period of time that the expected return is negligible, while mean returns are hard to estimate due to a high ratio of volatility to mean, we take each $d_j = 0$. Because we

Table 1 Amounts of Options in the Portfolio

Asset	Option type	Strike price 85	90	95	100	105	110	115
1	Put	−2,000	−2,000	−2,500	1,000	0	0	0
2	Put	2,500	−1,000	1,000	500	0	0	0
3	Put	1,500	1,000	2,500	−1,500	0	0	0
1	Call	0	0	0	−1,000	1,500	−500	−1,000
2	Call	0	0	0	1,500	−2,500	2,000	−2,000
3	Call	0	0	0	−2,000	−1,000	1,000	2,500

do not need to simulate S_0, the parameters d_0 and σ_0 are not relevant.

We investigated the performance of our procedures on several portfolios. The extent of the efficiency improvement depends on the portfolio, so here we present a portfolio yielding results we consider typical. Table 1 lists the number of each type of option in this example portfolio. Each option is the right to buy or sell 100 shares. We do not use control variates in this example.

3. A Framework for Estimating the Maximum

Recall that our goal is to provide a fixed-width confidence interval for $\mu_{[k]}$, the largest mean. Our methods seek a random subset $I \subseteq \{1, 2, \ldots, k\}$, estimators $\hat{\mu}_i$, $i = 1, 2, \ldots, k$, and constants $a, b > 0$ such that

$$\Pr\left\{\mu_{[k]} \geq \max_{i \in I} \hat{\mu}_i - a\right\} \geq 1 - \alpha_a, \tag{2}$$

$$\Pr\left\{\mu_{[k]} \leq \max_{i \in I} \hat{\mu}_i + b\right\} \geq 1 - \alpha_b, \tag{3}$$

and $a + b = L$, where the user specifies the error probability bounds $\alpha_a, \alpha_b \in (0, 1/2)$ and the confidence interval width L. Together, inequalities (2) and (3) imply that

$$\Pr\left\{\max_{i \in I} \hat{\mu}_i - a \leq \mu_{[k]} \leq \max_{i \in I} \hat{\mu}_i + b\right\} \geq 1 - \alpha_a - \alpha_b. \tag{4}$$

The random subset I contains the systems deemed to have a sufficiently high chance of being the best, and will be generated in such a way as to give the best system $[k]$ a high probability of being in I. The systems not in I are "screened out." For an argument that screening is likely to enhance efficiency, see §4.3.

The appropriate error probability bounds α_a, α_b and confidence interval width L depend on the application. In pricing derivatives, we might use an error probability bound $\alpha_b = 0.2\%$ that is very low because offering to sell a derivative security at a low price can lead to large losses, which can be tolerated only very infrequently. We might also consider confidence interval widths L of 0.1% to 1% of the derivative's true value because these widths are comparable to or slightly smaller than typical bid-ask spreads. That is, at greater widths, one would be unable to quote competitive prices. Lesser widths would be unnecessarily precise. A risk-management problem, on the other hand, does not require such high confidence and precision. Risk management is more a matter of decisions internal to a firm, so there are no customers to take advantage of violations of the upper confidence limit when they occur (in at most a fraction α_b of the cases), or whose business is lost when the upper confidence limit is too far above the true value. Moreover, in risk-management problems, X involves the value of a portfolio containing many securities, so it is usually very expensive to generate. If so, then demanding very high confidence or precision could result in an unacceptably large time to run the simulation.

Consider the upper confidence limit, and note that

$$\begin{aligned}\Pr\left\{\mu_{[k]} \leq \max_{i \in I} \hat{\mu}_i + b\right\} &\geq \Pr\{[k] \in I, \mu_{[k]} \leq \hat{\mu}_{[k]} + b\} \\ &\geq 1 - \Pr\{[k] \notin I\} - \Pr\{\mu_{[k]} > \hat{\mu}_{[k]} + b\}. \end{aligned} \tag{5}$$

Thus, if we can guarantee that

$$\Pr\{[k] \notin I\} \leq \alpha_I \quad \text{and} \tag{6}$$

$$\Pr\{\mu_{[k]} > \hat{\mu}_{[k]} + b\} \leq \alpha_b', \tag{7}$$

where $\alpha_I + \alpha_b' = \alpha_b$, then the upper confidence limit will be valid as in inequality (3).

Next, consider the lower confidence limit, and note that

$$\begin{aligned}\Pr\left\{\mu_{[k]} \geq \max_{i \in I} \hat{\mu}_i - a\right\} &\geq \Pr\left\{\mu_{[k]} \geq \max_{i=1,2,\ldots,k} \hat{\mu}_i - a\right\} \\ &= \Pr\{\hat{\mu}_i \leq \mu_{[k]} + a,\ i = 1, 2, \ldots, k\} \\ &\geq \Pr\{\hat{\mu}_i \leq \mu_i + a,\ i = 1, 2, \ldots, k\} \\ &\geq 1 - \sum_{i=1}^{k} \Pr\{\hat{\mu}_i > \mu_i + a\} \end{aligned} \tag{8}$$

by the Bonferroni inequality. Therefore, the lower confidence limit will be valid as in inequality (2) if, for $i = 1, 2, \ldots, k$,

$$\Pr\{\hat{\mu}_i > \mu_i + a\} \leq \alpha_a' = \alpha_a / k. \tag{9}$$

To obtain a fixed-width confidence interval, we need to determine the half-widths a and b, given the width L and the error spending structure, so that $a + b = L$ and inequalities (7) and (9) hold. To verify the validity of the confidence limits for the estimation of the systems' means μ_i, we need to show that there are increasing functions G_a and G_b defined on the positive part of the real line, such that, for all $i = 1, 2, \ldots, k$ and $x > 0$,

$$\begin{aligned}\Pr\{\hat{\mu}_i - \mu_i > x\} &\leq 1 - G_a(cx) \quad \text{and} \\ \Pr\{\hat{\mu}_i - \mu_i < -x\} &\leq 1 - G_b(cx), \end{aligned} \tag{10}$$

where

$$a = \frac{1}{c} G_a^{-1}(1-\alpha_a'), \tag{11}$$

$$b = \frac{1}{c} G_b^{-1}(1-\alpha_b'), \quad \text{and} \tag{12}$$

$$c = \frac{1}{L}(G_a^{-1}(1-\alpha_a') + G_b^{-1}(1-\alpha_b')). \tag{13}$$

This determines the sampling scheme in such a way that it bounds the distribution of $\hat{\mu}_i - \mu_i$ by a function that is free of dependence on i (see §§4.1 and 4.2 for examples).

Proposition 3.1. *Inequalities* (2) *and* (3) *hold if inequalities* (6) *and* (10) *hold, where G_a and G_b are increasing functions defined on the positive part of the real line, satisfying $G_a(0) < 1-\alpha_a' < \lim_{x\to\infty} G_a(x)$ and $G_b(0) < 1-\alpha_b' < \lim_{x\to\infty} G_b(x)$.*

Proof. Because $G_a(0) < 1-\alpha_a' < \lim_{x\to\infty} G_a(x)$ and $G_b(0) < 1-\alpha_b' < \lim_{x\to\infty} G_b(x)$, a and b exist and are positive. For all $i = 1, 2, \ldots, k$, by inequality (10) and Equation (11), $\Pr\{\hat{\mu}_i - \mu_i > a\} \le \alpha_a'$, while by inequality (10) and Equation (12), $\Pr\{\hat{\mu}_i - \mu_i < -b\} \le \alpha_b'$. Thus, for all $i = 1, 2, \ldots, k$, inequality (9) holds, which we already argued implies inequality (2). Inequality (7) holds, and we have already argued that with inequality (6), it implies inequality (3). □

To show that a procedure delivers confidence limits with at least the coverage probabilities specified in inequalities (2) and (3), we will verify that the screening procedure satisfies inequality (6), and exhibit increasing functions G_a and G_b with $G_a(0) = G_b(0) = 1/2$ such that the mean estimators satisfy inequality (10). These results provide a general framework for estimating $\mu_{[k]}$; the remainder of the paper works out details for specific ways to form the subset I and the estimators $\hat{\mu}_i$. The procedures we will discuss all have the following structure.

1. Simulate all systems, possibly over multiple stages, and retain a subset $I \subseteq \{1, 2, \ldots, k\}$.

2. For all systems $i \in I$, compute a terminal sample size N_i and simulate more observations to get a total of N_i.

3. Compute an estimator $\hat{\mu}_i$ of the mean μ_i for each system $i \in I$.

4. Report the confidence interval $[\max_{i\in I} \hat{\mu}_i - a, \max_{i\in I} \hat{\mu}_i + b]$.

We obtain *efficient* procedures in two ways:

1. by reducing $|I|$, the number of means that we estimate, and

2. by employing efficient estimators $\hat{\mu}_i$ of μ_i, so that the means we do estimate require as little computational effort as possible.

In Lesnevski et al. (2004), we reported on two-stage and multistage procedures that fit this framework. These procedures used screening to form the subset I, estimated μ_i using a sample mean, and assumed that the systems were simulated independently. In this paper, we employ CRN to further reduce $|I|$, estimate μ_i using control-variate estimators, and investigate "restarting" the procedure after screening, which allows us, in effect, to tackle a smaller problem.

4. Procedures

In this section, we construct simulation procedures that generate a fixed-width, two-sided confidence interval for a coherent risk measure that is the maximum of k means. Online Appendix A contains algorithms implementing procedures with various combinations of these features. Proofs of the procedures' validity appear in Online Appendix B.

4.1. The Basic Procedure

First, we briefly explain our variant of the procedure of Chen and Dudewicz (1976), which serves as our standard for comparison on examples without control variates. This is a two-stage procedure. The first stage is called stage 0. For each system i, in stage 0, the procedure generates independent replications $X_{i1}, X_{i2}, \ldots, X_{in_0}$, whose common distribution is that of the discounted loss X under probability measure $\mathbf{P}_i$. The replications are used to estimate the variances $\sigma_i^2 := \mathrm{Var}[X_i]$. These are

$$S_i^2 := \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} (X_{ij} - \bar{X}_i)^2,$$

where $\bar{X}_i := \sum_{j=1}^{n_0} X_{ij}/n_0$ are the stage-0 sample averages. Let $\lceil x \rceil$ represent the smallest integer greater than or equal to x. After stage 0, the total sample sizes

$$N_i = \max\{n_0, \lceil c^2 S_i^2 \rceil\} \tag{14}$$

are computed on the basis of the variance estimates S_i^2 and the scaling constant c as defined in Equation (13), where $G_a = G_b = F_{t_{n_0-1}}$, the t distribution with $n_0 - 1$ degrees of freedom. In the second stage, called stage 1, additional replications X_{ij} are simulated for $i = 1, 2, \ldots, k$ and $j = n_0 + 1, n_0 + 2, \ldots, N_i$. The procedure estimates the means μ_i with the cumulative sample averages as of the end of stage 1,

$$\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}.$$

Note that the standard procedure manipulates the marginal distributions of the estimators $\hat{\mu}_i$ individually; their relative values and joint distribution have no impact.

4.2. Controlled Control Variates

In this section, we present an extension (Algorithm A.1) of the Chen-Dudewicz (1976) procedure that incorporates control variates in mean estimation. Although we could also use control variates in screening, we found little added benefit because common random numbers alone were so effective for the financial examples we considered. As control variates in screening introduce technical complications, we defer that topic to Nelson and Staum (2006).

To have a fair comparison of the performance of our procedures on examples using control variates, the standard of comparison will be this variant of the Chen-Dudewicz procedure that uses control variates. For details about the construction of the control-variate estimators, see Online Appendix D. We introduce a q_i-dimensional vector C_i of control variates with known mean ξ_i. Because X_i comes from a portfolio value simulated under $\mathbf{P}_i$, usually C_i represents other financial variables generated simultaneously under $\mathbf{P}_i$. Frequently, the dimension q_i is the same for all i, because the same financial variables are used in each case. In Example 1, the control variates are the payoffs of European put options whose prices are known by the Black-Scholes formula. For more on control variates in financial simulations, see Glasserman (2004, §4.1).

We now allocate error α_C to a bound on the sample variance of the control-variate point estimator, which depends on control-variate observations after the first stage of sampling (see Nelson and Staum 2006), unlike the sample variance of the sample mean, which only depends on first-stage observations. Define $q := \max_{i=1,2,\ldots,k} q_i$, the maximum number of control variates used for any system. The functions that generate the scaling constant c in Equation (13) are given by $G_a(x) = G_b(x) = F_{t_{n_0-q-1}}(x) - \alpha_C$, so

$$\begin{aligned} c &= \frac{1}{L}(G_a^{-1}(1-\alpha_a') + G_b^{-1}(1-\alpha_b')) \\ &= \frac{1}{L}(t_{n_0-q-1,\,1-\alpha_a'+\alpha_C} + t_{n_0-q-1,\,1-\alpha_b'+\alpha_C}), \end{aligned}$$

where $t_{\nu,u}$ represents the u quantile of the t distribution with ν degrees of freedom. This corresponds to decomposing the error bounds as $\alpha_b' = \alpha_C + \alpha_b''$ and $\alpha_a' = \alpha_C + \alpha_a''$, and using the $1-\alpha_a''$ and $1-\alpha_b''$ quantiles of a t distribution. When using control variates, replace in Equation (14) the sample variance S_i^2 of X_i with the sample residual variance $\hat{\tau}_i^2$ of the regression of X_i on the control variates C_i (see Online Appendix D). As in Nelson and Staum (2006, Procedure 4 and Remark B.2), the effect of spending α_C on controlling the dispersion of the control variates' sample average from its expectation is to add $\chi^2_{q_i,\,1-\alpha_C}$, the $1-\alpha_C$ quantile of the chi-squared distribution with q_i degrees of freedom, to the required number of replications:

$$N_i = \max\{n_0, \lceil c^2\hat{\tau}_i^2 + \chi^2_{q_i,\,1-\alpha_C}\rceil\}. \tag{15}$$

This formulation subsumes the case without control variates discussed in the previous section, with $q_i = 0$, $\alpha_C = 0$, and $\hat{\tau}_i^2 = S_i^2$.

4.3. Screening with Common Random Numbers

Algorithm A.2 is a two-stage algorithm with screening. In this section, we describe this algorithm and how to implement it with common random numbers.

Let $U_1, U_2, \ldots$ be a sequence of independent, identically distributed random vectors. Each U_j is interpreted as a vector of random numbers forming the basis for the jth replication in the simulation. For all $i = 1, 2, \ldots, k$, the jth realization of the negative discounted portfolio value $X_{ij} = X_i(U_j)$ and the jth realization of the control-variate vector $C_{ij} = C_i(U_j)$ are generated from the vector of common random numbers U_j, which are common to all systems. The result is that random variables such as X_{hj} and X_{ij} are dependent, but for different replications $j \neq l$, X_{hj} and X_{il} are independent.

For screening, define the stage-0 sample variances of the differences $X_h - X_i$ as

$$S_{hi}^2 := \frac{1}{n_0 - 1}\sum_{j=1}^{n_0}(X_{hj} - X_{ij} - (\bar{X}_h - \bar{X}_i))^2.$$

Construct the set $I := \{i \mid \forall\, h \neq i,\ \bar{X}_i \geq \bar{X}_h - W_{hi}\}$, where the threshold

$$W_{hi} := t_{n_0-1,\,1-\alpha_1/(k-1)}\frac{S_{hi}}{\sqrt{n_0}}.$$

The set I contains those systems that could plausibly be the best, in the sense of not being statistically dominated by some other system at stage 0. Every $i \notin I$ has been screened out.

Does screening increase efficiency? The error spent on screening subtracts from the error that can be spent on estimating systems' means (Equations (6) and (7)), and thus inflates the sample size required for each system that survives screening. If screening does not eliminate enough systems, it will increase the total number of replications that the procedure requires. However, in financial simulations sample sizes are usually large, and therefore the benefits of screening out the inferior scenarios early are usually substantial. Even in situations where some systems have means that are very close to the best, screening will generally be effective. One reason is that the benefits can still exceed the costs even if only a few systems are eliminated. Another reason is that, in financial applications, systems whose means are very similar usually also have high correlation, which makes common

random numbers very effective, so it is often not too hard to screen out a system that is only slightly inferior to another system.

The worst-case efficiency loss due to screening is in fact very limited. If all k means are the same, it would be best to forgo screening and use a procedure such as Procedure 4 of Nelson and Staum (2006). However, as long as the screening budget is less than the required final sample size, the ratio of the sample sizes with and without screening is approximately

$$\left(\frac{\Phi^{-1}(1-\alpha_a/k)+\Phi^{-1}(1-\alpha_b+\alpha_I)}{\Phi^{-1}(1-\alpha_a/k)+\Phi^{-1}(1-\alpha_b)}\right)^2.$$

This follows from Equation (13) for the scaling constant c, which determines the total sample sizes in all of the procedures, and from approximating a t distribution with many degrees of freedom by a normal distribution. When $\alpha_a = 0.8\%$, $\alpha_b = 0.2\%$, $\alpha_I = (0.2)\alpha_b$, and $k = 256$, as in the options portfolio example, this worst-case efficiency loss is only 2%. This worst-case efficiency loss is decreasing in k, and its maximum over k is just 2.65%. For these reasons, screening is very likely to improve efficiency, and even if it does not, it cannot decrease efficiency by much.

The performance of the two-stage procedure depends significantly on the initial sample size n_0 (Lesnevski et al. 2004). When n_0 is small, increasing it tends to lead to improved screening because more information at stage 0 allows more systems to be screened out. If n_0 becomes too large, however, computational resources are wasted on poor systems that could have been screened out earlier and on systems with low standard deviations for which the desired terminal sample size $N_i < n_0$; see Equation (14). It would be preferable to have a procedure that is less sensitive to n_0, and the multistage procedure described in the next section has this property.

4.4. Multistage Screening

In this procedure, there are m screening stages $0, 1, \ldots, m-1$ and one final estimation stage m. Our notation is that a number l in parentheses indicates a quantity that applies to or is estimated after the lth stage. For example, the sample average of X_i over all stages up to l is $\bar{X}_i(l) := \sum_{j=1}^{N(l)} X_{ij}/N(l)$, where $N(l)$ is the total number of replications sampled from each surviving system through screening stage l. Online Appendix C.1 includes an explanation of why the sample size $N(l)$ is the same for each system still in contention.

There are three main aspects of the multistage procedure to resolve. We must specify

1. the screening stage sample sizes $N(l)$ for $l = 1, 2, \ldots, m-1$,

2. the screening thresholds $W_{hi}(l)$ for $l = 0, 1, \ldots, m-1$ and $h, i = 1, 2, \ldots, k$, and

3. the sample size N_i used in constructing the mean estimate $\hat{\mu}_i$ for systems $i \in I(m)$ that survive screening.

We must choose the screening-stage sample sizes and thresholds so that there is an error decomposition satisfying inequality (6) and choose the final sample size so that inequality (10) holds. It turns out that these three issues are intimately related by the way in which simulated data are used to supply variance estimates.

More than one scheme is possible, but here, for simplicity, we set all screening-stage sample sizes $N(0), N(1), \ldots, N(m-1)$ before the simulation begins. We have found experimentally that a good way of choosing these sample sizes is to choose n_0 and a constant growth factor R, and then set $N(l) = \lceil n_0 R^l \rceil$. The intuition behind this is that it makes standard errors likely to decrease by roughly the constant factor $\sqrt{R}$ at each stage. If, for example, sample sizes grew at a constant arithmetic rate instead of a constant geometric rate, later stages would be spending opportunities to look at the data (see point 2 of the list below) with very little chance of screening out a system that had survived the previous stage.

How should the growth factor R be chosen? The maximum number of replications during screening for each system is $N(m-1) = \lceil n_0 R^{m-1} \rceil$. If this number is too large, the number of replications sampled during screening can exceed the number N_i required for the estimate $\hat{\mu}_i$, which is wasteful. Suppose that we choose a maximum screening budget $N(m-1)$ that is not too large. Given this maximum budget, the initial sample size n_0, and the number of screening stages m, the factor $R = (N(m-1)/n_0)^{(1/(m-1))}$. We should choose n_0 and m with the following points in mind.

1. The ends of the m screening stages are the only m opportunities at which systems can be screened out. The fewer these opportunities, the longer the procedure must wait to screen out a system, and the more work is expended on systems that are eventually screened out.

2. On the other hand, the screening thresholds defined in Equation (16) below are increasing in m. Given a fixed amount of data, fewer systems can be screened out when m is larger. The more opportunities there are to screen out a system, the less aggressive the procedure can be at each screening opportunity, if a fixed error probability is to be maintained.

3. It is desirable to have n_0 small, so that extremely poor systems can be screened out quickly. However, if n_0 is too small, then the normal approximation used to justify the confidence limits may break down at early stages (Lesnevski et al. 2006).

Next, we consider the screening thresholds and error decomposition, given that sample sizes are fixed

in advance. After each stage $l = 0, 1, \ldots, m-1$, screening takes place by constructing

$$I(l+1) := \{i \in I(l) \mid \forall h \in I(l), \bar{X}_i(l) \geq \bar{X}_h(l) - W_{hi}(l)\},$$

where $I(0) = \{1, 2, \ldots, k\}$. Define the threshold

$$W_{hi}(l) := t_{N(l)-1,1-\alpha_l/(m(k-1))} \frac{S_{hi}(l)}{\sqrt{N(l)}}, \tag{16}$$

where the stage-l sample variance is

$$S^2_{hi}(l) := \frac{1}{N(l)-1} \sum_{j=1}^{N(l)} (X_{hj} - X_{ij} - (\bar{X}_h(l) - \bar{X}_i(l)))^2.$$

We use fully updated, cumulative sample variances to set the screening thresholds. Typically, multistage screening procedures for ranking and selection use only stage-0 sample variances to simplify inference. In this procedure, it is valid to use updated variance information, and valuable to do so while keeping n_0 very small. Because a large fraction of systems were screened out at stage 0 in our examples, this allows us to decrease the sample size. Updating variance information makes thresholds at later stages smaller because it reduces the t quantile, which allows more screening to take place.

After screening, we must choose a final sample size N_i for estimation of the mean μ_i by $\hat{\mu}_i$. We cover the case with CV, which subsumes that without CV (§4.2). The scaling constant c comes from Equation (13) and $G_a(x) = G_b(x) = F_{t_{N(m-1)-q-1}}(x) - \alpha_C$. Equation (15) becomes

$$N_i = \max\{N(m-1), \lceil c^2 \hat{\tau}_i^2(m-1) + \chi^2_{q_i,1-\alpha_C} \rceil\}, \tag{17}$$

where $\hat{\tau}_i^2(m-1)$ is the sample residual variance of the regression of $X_{i1}, \ldots, X_{i,N(m-1)}$ on the control variates $C_{i1}, \ldots, C_{i,N(m-1)}$.

In stage m, X_{ij} is simulated for $i \in I(m)$ and $j = N(m-1)+1, N(m-1)+2, \ldots, N_i$, and then the confidence limits are constructed around $\max_{i \in I(m)} \hat{\mu}_i$, where each estimate $\hat{\mu}_i$ is based on all replications $j = 1, 2, \ldots, N_i$. That is, $\hat{\mu}_i$ is either the sample average $\bar{X}_i(m) = \sum_{j=1}^{N_i} X_{ij}/N_i$, or this sample average after correction by control variates, as detailed in Online Appendix D.

This works because $N(m-1)$ is a constant. For purposes of mean estimation, it does not matter how we screen, as long as the probability of wrongly screening out the best system satisfies inequality (6) and we finish the screening phase with a variance estimator that has the desired distribution and is independent of the existing sample average $\bar{X}(m-1)$. Fixing the screening stage sample sizes in advance is one way to achieve this.

The situation would be far more delicate if we allowed the screening-stage sample sizes to be random, for example, to depend on sample variances from prior stages. In particular, the arguments above rely on a constant sample size $N(m-1)$ at the end of screening for all systems that survive. This means that we have not entirely solved the n_0 problem faced by a two-stage procedure. Similarly, the multistage procedure could be said to have an $N(m-1)$ problem. If we choose the maximum per-system screening budget $N(m-1)$ too small, not enough screening is done. If we choose $N(m-1)$ too large, then this multistage procedure wastes effort by exceeding the desired final sample size $\lceil c^2 \hat{\tau}_i^2(m-1) + \chi^2_{q_i,1-\alpha_C} \rceil$ in Equation (17) for any system that survives too long.

In the next section, an enhancement to the multistage procedure ameliorates this problem. Nonetheless, even for the procedures described below, there is still some danger of wasting effort by choosing $N(m-1)$ too large. Lesnevski et al. (2006) makes further progress in solving this problem.

4.5. Early Stopping During Screening

In many of our examples, we found that all systems but the best were screened out before the scheduled end of screening; that is, the event $I(l) = \{[k]\}$ often occurred for some screening stage $l < m-1$. Clearly, it makes sense to stop screening once the set I has become a singleton and move immediately to estimation. This helps us to avoid the problem, mentioned at the end of the previous section, that the screening budget $N(m-1)$ might be larger than the desired final sample size: frequently, I becomes a singleton before the screening budget is exhausted and before the desired final sample size is exceeded.

Define the random stage

$$M := \min\{m, \inf\{l \mid |I(l)| = 1\}\},$$

at which we would like to proceed to mean estimation. Unfortunately, invoking our estimation procedure from this random stage alters the distribution of the final estimator in ways that we cannot explicitly evaluate. Where $I(M) = \{i\}$, we might like to use $N_i = \max\{N(M-1), \lceil c^2 \hat{\tau}_i^2(M) + \chi^2_{q_i,1-\alpha_C} \rceil\}$. However, unlike in previous sections, we do not find a chi-squared distribution related to $\hat{\tau}_i^2(M)$. This is because the event $M = l$ of stopping at an early stage l is associated with low values of $S^2_{ih}(M)$ for all systems $h \neq i$: when these sample variances are low, it helps system i to screen out all the others quickly. Low values of $S^2_{ih}(M)$ are associated with low values of $S_i^2(M)$, and low values of $S_i^2(M)$ are associated with low values of $\hat{\tau}_i^2(M)$, so although there is a chi-squared distribution related to $\hat{\tau}_i^2(l)$ for any fixed l, there is not for $\hat{\tau}_i^2(M)$. A remedy for this technical problem is to set the terminal sample size as

$$N_i = \max\{N(M-1), \lceil c^2 \hat{\sigma}_i^2 + \chi^2_{q_i,1-\alpha_C} \rceil\}, \tag{18}$$

where $\hat{\sigma}_i^2$ is a variance estimator with the right distribution. We accomplish this by following a fixed screening schedule for a small number of stages and allowing early stopping only after that.

More precisely, we fix a stage l^* between 1 and $m-1$, and forbid early stopping until after stage l^*, forcing $M \geq l^*$. We only use variance information up through stage l^* to determine the terminal sample size for estimation. That is, $\hat{\sigma}_i^2 = \hat{\tau}_i^2(l^*)$ is the sample residual variance of the regression of $X_{i1}, X_{i2}, \ldots, X_{iN(l^*)}$ on the control variates $C_{i1}, C_{i2}, \ldots, C_{iN(l^*)}$. Because $\hat{\tau}_i^2(l^*)$ is computed over a prespecified constant number $N(l^*)$ of replications, we can find associated chi-squared and t distributions. The scaling constant c comes from Equation (13) with

$$G_a(x) = G_b(x) = F_{t_{N(l^*)-q-1}}(x) - \alpha_C. \tag{19}$$

This yields Algorithm A.3.

4.6. Restarting

The critical values that determine the overall sample size for mean estimation depend on the number of systems k. The sample size increases as k increases to compensate for the greater chance of error when there are more alternatives. Consider the situation when $K(M) := |I(M)|$, the number of systems remaining after screening ends at the random stage $M-1$, turns out to be small. It would then be efficient to pretend that the mean-estimation problem only involved the $K(M)$ systems still in play. Unfortunately, this is invalid when we retain the data obtained up to stage M. This is because of selection bias: when the number k of systems is higher, the sample averages through stage M of any systems that survive tend to be higher (Boesel et al. 2003). If, on the other hand, we "restart" the simulation after screening—that is, throw out all data from the screening stages—then our mean-estimation procedure applied only to the $K(M)$ survivors is valid. If $K(M)$ is small enough, then the reduction in required sample size due to reduced critical values will outweigh the cost of discarding the data from the screening stages.

After screening, we will obtain N_i new replications for each surviving system $i \in I(M)$ and form the estimators $\hat{\mu}_i$ from these replications alone. We will choose the sample size N_i by performing an independent two-stage procedure. In the follow-up experiment's first stage, we simulate n_i replications from system i and form a variance estimate $\hat{\sigma}_i^2$, the sample residual variance of the regression of X_{ij} on C_{ij}, $j = 1, 2, \ldots, n_i$. From $\hat{\sigma}_i^2$, we determine the terminal sample size as

$$N_i = \max\{n_i, \lceil c^2\hat{\sigma}_i^2 + \chi^2_{q_i,\,1-\alpha_C}\rceil\}. \tag{20}$$

The scaling constant c comes from Equation (13) with

$$G(x) = F_{t_{n-q-1}}(x) - \alpha_C, \tag{21}$$

where $n := \min_{i\in I(M)} n_i$ is used to quantify the minimum degrees of freedom in constructing any variance estimate $\hat{\sigma}_i^2$ for a surviving system i. In the second and last stage, we simulate replications $j = n_i + 1$, $n_i + 2, \ldots, N_i$.

This two-stage procedure for fixed-width interval estimation is valid for any value of n_i. By increasing n_i, we increase the degrees of freedom of the t distribution in $G(x)$, which helps to reduce the sample size N_i, as well as its variability. However, if we choose n_i too large, then $n_i > \lceil c^2\hat{\sigma}_i^2 + \chi^2_{q_i,\,1-\alpha_C}\rceil$ and we waste effort. Fortunately, it is valid to choose n_i as a function of $\hat{\tau}_i^2(M-1)$, the residual variance estimator obtained from screening, because all data in the follow-up experiment are independent of the screening data. In particular, we will use this information to form a lower prediction limit for the terminal sample size N_i.

As an approximation, suppose that the conditional distribution of $\hat{\tau}_i^2(M-1)/\hat{\sigma}_i^2$, given M, is F with $N(M-1)-1$ and $n_i - 1$ degrees of freedom. Assuming that n_i is large, the distribution of $(N(M-1)-1)\cdot \hat{\tau}_i^2(M-1)/\hat{\sigma}_i^2$ is approximately $\chi^2_{N(M-1)-1}$. This yields an approximate $(1-\epsilon)100\%$ lower prediction limit for $\hat{\sigma}_i^2$ of $(N(M-1)-1)\hat{\tau}_i^2(M-1)/\chi^2_{N(M-1)-1,\,1-\epsilon}$. Because all n_i and hence $n := \min_{i\in I(M)} n_i$ are large, the t distribution in Equation (21) has many degrees of freedom and is thus approximately a normal distribution. This yields, from Equation (13), $c \approx (\Phi^{-1}(1-\alpha_a'') + \Phi^{-1}(1-\alpha_b''))/L$. Putting these approximations together, we set

$$n_i = \left(\frac{\Phi^{-1}(1-\alpha_a'') + \Phi^{-1}(1-\alpha_b'')}{L}\right)^2 \cdot \frac{(N(M-1)-1)\hat{\tau}_i^2(M-1)}{\chi^2_{N(M-1)-1,\,1-\epsilon}} + \chi^2_{q_i,\,1-\alpha_C}, \tag{22}$$

an approximate lower prediction limit for the desired size $c^2\hat{\sigma}_i^2 + \chi^2_{q_i,\,1-\alpha_C}$ in Equation (20). This yields Algorithm A.5.

5. Experimental Results

We now report selected results of computational experiments to test the efficiency and validity of the procedures developed in §4. We discuss the magnitude of the procedures' efficiency gains in §5.1, as well as the factors that contribute to them. This includes, in §5.2, an assessment of the extent to which efficiency depends on the choice of parameters such as sample sizes and error decomposition. Section 5.3 illustrates the validity of the procedures in practice by analyzing the coverage of the confidence intervals they generate. Before reporting the results, we mention choices of parameters common to the experiments.

In all experiments, one fifth of the error is allocated to the upper confidence limit, and four fifths to the

lower confidence limit. For example, for a 99% confidence interval, the probability that the true maximum mean exceeds the upper confidence limit is nominally guaranteed to be no more than $\alpha_b = 0.2\%$, while the probability that it falls below the lower confidence limit is nominally guaranteed to be no more than $\alpha_a = 0.8\%$.

For ease of interpretation, we specify the fixed confidence interval width L as a percentage of a quantity which provides a natural scale for the example. For the options portfolio example, this quantity is the portfolio's standard deviation. For the basket put example, this quantity is the true value $\mu_{[k]}$, interpreted as an ask price for the basket put. In either case, the scaled quantity is estimated in advance by a very precise simulation. To assign L equal to a fraction of an estimate of $\mu_{[k]}$ after stage 0 would introduce additional complications. In financial applications, there is often a previous problem with similar parameters which can supply a value of L giving approximately the desired relative precision.

Except when otherwise specified, the level of precision is 1%, the confidence level is 99%, and the algorithms' parameters are set to the following default values. The error allocated to screening is $\alpha_I = (0.2)\alpha_b$, there are $n_0 = 30$ replications in the initial stage 0, there are $m = 15$ stages, and the cumulative sample size grows by a factor of $R = 2$ at each stage. This makes the budget available for screening $N(m-1) = n_0 R^{m-1} = 30 \cdot 16{,}384 = 491{,}520$. When using control variates, the error allocated to controlling them is $\alpha_C = (0.01)\min\{\alpha'_a, \alpha'_b\}$. This adds 27 or 31 extra replications (at 95% or 99% confidence, respectively) per system that survives screening; the right panel of Figure 3 shows that this cost is not large relative to the simulation's total cost. For the multistage algorithms with early stopping, stopping is forbidden until after stage $l^* = 5$, yielding $N(l^*) = n_0 R^{l^*} = 30 \cdot 32 = 960$ replications to provide variance information for use in setting the final sample sizes. For the multistage algorithm with restarting, the significance level used in creating the prediction limit for the final sample size that underlies Equation (22) is $\epsilon = 1\%$.

5.1. Efficiency: Procedures and Precision

We report efficiency as a speed improvement relative to the standard procedure. This is the ratio of the average number of samples required by the standard procedure to the average number required by our more advanced procedures. The number of samples required by the standard procedure is $\sum_{i=1}^{k} N_i$, where N_i is defined in Equations (14) or (15), depending on whether control variates are in use. We have ignored overhead costs such as those associated with comparisons during screening or with generating and using control variates. In the financial applications we have in mind, generating a single negative discounted portfolio value X_{ij} is moderately to extremely expensive because it involves simulating over many time steps, underlying risk factors, or securities in the portfolio. Also, the control variates C_{ij} used in such applications are usually cheap to compute once X_{ij} has already been simulated.

Table 2 Efficiency Relative to the Standard Procedure at 99% Confidence and 1% Precision

Procedure			Example	
Stages	CRN	Restarting	Basket put	Options portfolio
15	✓	✓	157	249
15	✓		115	147
15			5.5	146
2	✓		41	103

Table 2 reports the efficiency of four procedures: the multistage procedure with restarting and CRN, the multistage procedure with early stopping and CRN, the multistage procedure with early stopping and without CRN, and the two-stage procedure with CRN. Recall that we use CVs in the basket put example and not in the options portfolio example. In practice, the appropriate levels of precision might be 0.1%–1% for the basket put example because the statistical error surrounding a simulation estimate to be used as a derivative security price should be within the bid-ask spread, and 1% or more for a risk-management problem, such as the options portfolio example. For this reason, we use 1% precision in the table. In most cases, the improvement is dramatic.

For the two-stage procedure, the initial sample size n_0 is 3,000 for the basket put example and 1,000 for the options portfolio example. We chose these values to yield good performance for these examples, at this level of confidence and precision. The values are chosen to be this large to allow the two-stage procedure to screen out many systems at the first stage (see the end of §4.3). Nonetheless, the two-stage procedure's performance is markedly inferior to that of the multistage procedure, primarily because the multistage procedure does less work by screening out some systems earlier than others.

Using CRN is very effective for the basket put example, but has little effect for the options portfolio example at this level of precision and confidence. For the basket put example, the procedure without CRN usually spends a great deal of effort on screening: it tends not to stop early because it does not succeed in eliminating all but one of the systems. Indeed, for low precision, the effort may be more than is needed to estimate each system's mean, resulting in a loss of efficiency relative to the standard procedure. Reducing the total budget available for screening would

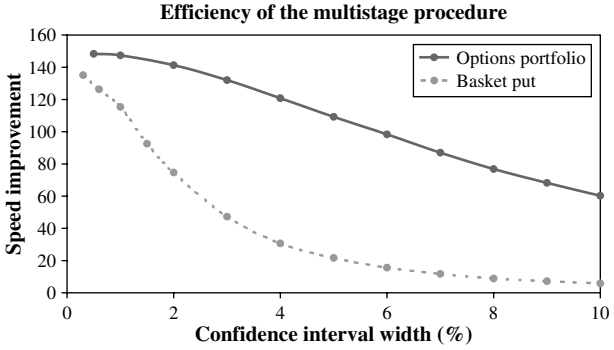
improve the procedure's performance on this example, but to do so would require advance knowledge of the problem. The procedure is not adaptive: for example, it cannot stop screening early when the sample size accumulated during screening reaches a running estimate of the final sample size required for inference about a system's mean. For an adaptive procedure, see Lesnevski et al. (2006). For variations on the multistage procedure that become possible without CRN, see Online Appendix C.1. Here we focus only on the direct impact of correlation among systems induced by CRN, not the indirect impact of changing the procedure to accommodate their use.

Another way to consider the efficiency of the procedures is relative to the maximum possible benefit that might be achieved, which we define as follows. To produce a fixed-width confidence interval for the maximum among k means requires at least as many replications as to produce such a confidence interval for the best system's mean considered in isolation. That is, the minimum sample size is what would be required if we were told in advance which system was best and could ignore the other $k-1$ systems. The ratio of the standard procedure's sample size to this minimum sample size depends on k, the number of systems, and the size of the best system's standard deviation relative to the standard deviations of the other systems. In both examples, the sample size of the multistage procedure with CRN and restarting is within a few percent of this minimum size.

In summary, we recommend using a multistage procedure with CRN. We have found that restarting increases efficiency for most examples. However, in examples where the number of replications required to screen out all but one system is large enough, it is more efficient not to restart.

Having examined the performance of different procedures on the same problems, we now consider the effect of the problem's difficulty on the procedures' efficiency. The same example becomes more difficult when greater confidence or precision is demanded. Greater difficulty is associated with higher efficiency of procedures with screening but without CRN or CV (Lesnevski et al. 2004). This happens because procedures with screening do only enough work on most systems to screen them out, and this is much less than the amount of work the standard procedure must do to estimate means with high confidence and precision. Figure 1 shows the effect of the confidence interval width L on the efficiency of the multistage procedure with early stopping and CRN. The fixed width is expressed as a percentage of a quantity which provides the scale for the example, so that a high percentage indicates that the user asked the procedure to deliver low precision.

Figure 1 Effect of Required Precision on Efficiency of the Multistage Procedure with Early Stopping and CRN Relative to the Standard Procedure at 99% Confidence



In Table 2, we saw that the multistage procedure with early stopping and CRN delivered more than 100-fold efficiency improvement for these examples at 1% precision, which is a reasonable level. From Figure 1, we see that the efficiency improvement is very high for a wide range of precision, and there is substantial improvement even at low precision. We found that the multistage procedures with CRN were more efficient than the standard procedure in every experiment we ran; we recommend using one of them in all simulations of coherent risk measures based on generalized scenarios.

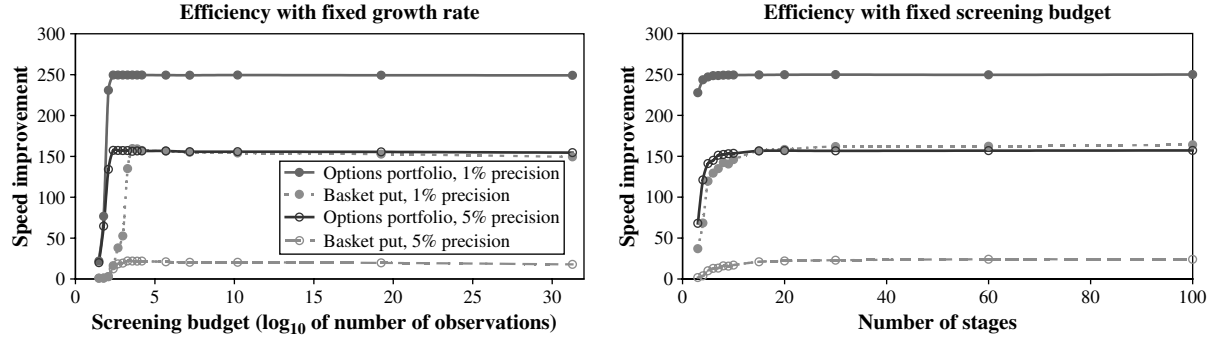
5.2. Efficiency: Parameters

We have selected default values of the procedure parameters based on experimentation to find which values yield good efficiency for a range of problems. Here we present evidence showing that efficiency is fairly robust to the choice of some parameters, indicating that they can be used without further tuning. Analogous results for procedures without CRN or CV agree qualitatively with the results reported here (Lesnevski et al. 2004).

First, we consider the effect of the sample sizes of the screening stages on the efficiency of the multistage procedure with restarting and CRN. The results are easier to interpret than for the multistage procedure with early stopping; changing its screening-stage sample sizes would require an adjustment to l^*, the first stage at which early stopping is allowed.

Recall that the cumulative sample size after l stages is $N(l) = \lceil n_0 R^l \rceil$, where $n_0 = N(0)$ is the stage-0 sample size and R is a constant growth factor. We consider two types of changes to the design of the screening phase. The first type is to vary the number of stages m with R fixed. The primary effect is on the total screening budget, $N(m-1) = \lceil n_0 R^{m-1} \rceil$. The second type is to change the number of stages m with $N(m-1)$ fixed, so that the growth factor R varies inversely with m. The effect is on how often the procedure is allowed to look at a fixed amount of data to screen out poor

Figure 2 Effect of Screening Phase Design on Efficiency of the Multistage Procedure with Restarting and CRN Relative to the Standard Procedure at 99% Confidence



systems. Figure 2 shows how these changes affect the efficiency of the multistage procedure with restarting and CRN.

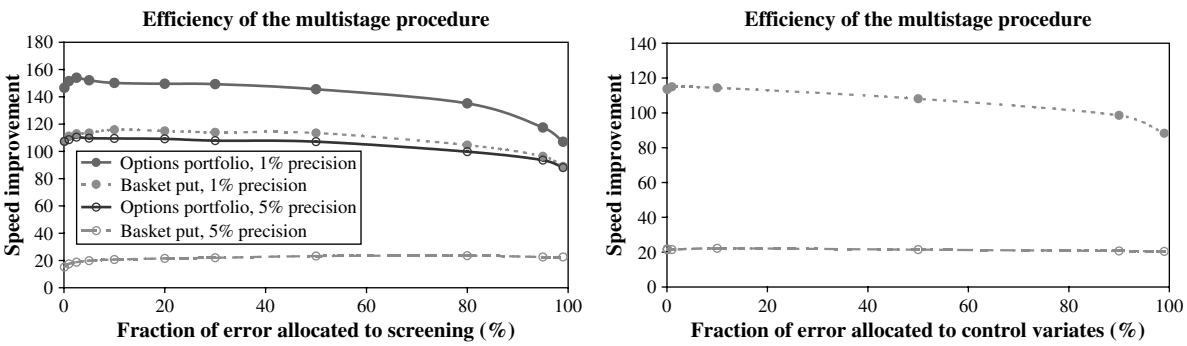
The graphs in Figure 2 show that the procedure's efficiency is gravely limited when the total screening budget $N(m-1)$ or the number of screening stages m are too small. If $N(m-1)$ is too small, not enough screening occurs, and in the final stage, the procedure must estimate an excessive number of systems' means. If m is too small, screening occurs too slowly, and excessive work is done on systems that are eventually screened out. In these examples, choosing m too large does not reduce efficiency by much. There is a statistical price to be paid for looking frequently at the data, but it has a small effect on the efficiency of screening. Having a large screening budget $N(m-1)$ does not mean that it must be used; the procedure restarts once screening has succeeded in eliminating all but one system. In the examples shown in the left panel of Figure 2, the efficiency losses due to occasionally sampling too many replications during screening are detectable but small.

However, a large screening budget poses a danger: as mentioned in the discussion of Table 2, there are examples in which the amount of work required to screen out all but one system exceeds the amount of work required to estimate the system's means. An extremely bad case is when more than one system has the maximum mean. Such ties can easily arise in finance when the discounted portfolio value X has the same distribution under two probability measures. In such cases, making $N(m-1)$ too large is a mistake.

The other parameter controlling the design of the screening phase is the initial sample size n_0. Our experiments showed that choosing n_0 very small maximizes efficiency. The danger in choosing n_0 too small is not a loss of efficiency, but rather a danger that the resulting confidence interval might provide inadequate coverage, due to failure of the normal approximation in the early screening stages causing the best system to be screened out. Results reported in §5.3 show that $n_0 = 30$ yielded adequate coverage for these examples.

Next, we consider the effect of error allocation on the efficiency of the multistage procedure with early stopping and CRN. The user specifies the confidence levels $1-\alpha_a$ and $1-\alpha_b$ associated with the lower and upper confidence limits, respectively, but the procedures have one or two further parameters controlling how the allowable errors α_a and α_b are spent. A portion α_I of α_b must be allocated to screening (inequality (6)). When using control variates, a portion α_C of both lower and upper error must be set aside for controlling them (§4.2). Figure 3 displays the effect of

Figure 3 Effect of Error Allocation on Efficiency of the Multistage Procedure with Early Stopping and CRN Relative to the Standard Procedure at 99% Confidence



changing the fractions α_I/α_b and $\alpha_C/\min\{\alpha'_a, \alpha'_b\}$ on efficiency. It is easy to choose an allocation yielding most of the possible efficiency improvement.

Allocating too much error to screening or control variates degrades the performance of the procedure. Having too little error left to spend on inference about the means of the systems that survive screening inflates the required sample size. However, an implausibly large amount of error must be allocated to screening or control variates before efficiency diminishes much; this mistake is easy to avoid. Likewise, efficiency may decrease if too little error is spent for these purposes, but the procedure's performance is even more robust against deficiency than excess. If α_I is too small, less screening takes place because the thresholds in Equation (16) become larger. However, the behavior of the quantiles of a t distribution (with many degrees of freedom) as a function of tail probability makes this effect small for the examples we considered: with $m = 15$, $N(l^*) = 960$, and $k = 256$, changing α_I from 0.04% to 0.002% changes the relevant t quantile from 5.23 to 5.77. This change corresponds to inflating the threshold by approximately 10%, but screening with CRN eliminates systems so quickly that this has little absolute effect on the efficiency of screening. Similarly, decreasing α_C inflates the chi-squared quantile added to the required final sample size in Equation (15), but α_C can be very small without having much impact. We found that $\alpha_I = (0.2)\alpha_b$ and $\alpha_C = (0.01)\min\{\alpha'_a, \alpha'_b\}$ are reliably good choices.

Finally, there are parameters related to early stopping (§4.5) and restarting (§4.6). After some experimentation, we selected the first stage after which early stopping is allowed as $l^* = 5$. The right choice of l^* depends on the growth structure of the screening stages, as embodied in the initial sample size n_0, the growth factor R, and the number of stages m. We found that choosing l^* too small can substantially degrade performance because of poor variance estimation. Choosing l^* too large has a significant cost only when the maximum screening budget $N(m-1)$ is far too large, as happened to the multistage procedure with early stopping and without CRN on the basket put example, shown in Table 2. For the multistage procedure with restarting and CRN, we found that, over a very wide range of values, efficiency is also rather insensitive to the significance level ϵ used in creating the prediction limit for the final sample size that underlies Equation (22). A good value is $\epsilon = 1\%$.

5.3. Coverage

Our procedures come with coverage guarantees (2) and (3) for their confidence limits, but the guarantees are proved only for normally distributed data X_{ij}.

Table 3 Error Rates of Multistage Procedures with CRN at 95% Confidence and 5% Precision

			With early stopping (%)		With restarting (%)	
Error prob.	Nominal (%)	Estimate	Basket put	Options portfolio	Basket put	Options portfolio
Upper	**1**	UCL	0.90	1.25	1.11	1.18
		Point	**0.64**	**0.94**	**0.82**	**0.88**
		LCL	0.44	0.69	0.59	0.64
Lower	**4**	UCL	0.20	0.07	3.40	4.54
		Point	**0.08**	<**0.01**	**2.90**	**3.96**
		LCL	0.02	<0.01	2.45	3.44

The distribution of a negative discounted portfolio value, especially when it contains derivative securities whose payoffs are nonlinear functions of underlying financial variables, is usually quite far from normal. The coverage guarantees hold in the basket put example for some simpler procedures without CRN or CV (Lesnevski et al. 2004). Table 3 supports the conclusion that the multistage procedures with CRN, either with early stopping or with restarting, also provide confidence limits with the required coverage for both of our examples.

The experiments reported in Table 3 contain 5,000 independent simulations. For each experiment, we report (in bold) the fraction of these 5,000 simulations in which $\mu_{[k]} < \max_{i\in I} \hat{\mu}_i - a$ as a point estimate of the lower error probability $\Pr\{\mu_{[k]} < \max_{i\in I} \hat{\mu}_i - a\}$, and similarly for the upper error probability $\Pr\{\mu_{[k]} > \max_{i\in I} \hat{\mu}_i + b\}$. We also give 95% confidence limits for the error probabilities, based on a binomial distribution for the observed number of errors.

We present experiments at confidence level 95% and precision 5% because this results in relatively low sample sizes. Large sample sizes create sample averages with distributions closer to normal, making it easier for the procedures to attain the nominal coverage. The nominal error probabilities are $\alpha_b = 1\%$ for the upper limit and $\alpha_a = 4\%$ for the lower limit. Entries less than these values show that the procedure is conservative in this case, attaining coverage greater than nominal.

Table 3 shows that the multistage procedure with early stopping and CRN is very conservative. Its conservatism is due to allocating an equal amount of error to each system in inequality (9), even those that are screened out. This was the motivation for the procedure with restarting, which is indeed much less conservative.

6. Conclusions

In this paper, we propose procedures for constructing a two-sided, fixed-width confidence interval for

the maximum of k systems' means. The motivation is financial applications in which the "systems" correspond to generalized scenarios, and we are interested in the mean value of the worst-case scenario. The procedures exploit the advantages that computer simulation provides: the ability to perform sequential experiments and to implement variance-reduction techniques.

Under normal-theory assumptions, our procedures are exact, that is, they deliver at least the nominal coverage probability. Although these assumptions are reasonable in many situations, they are never precisely correct. However, it is comforting to know that our screening procedures, which are usually applied when the sample sizes are small, are protected by the use of very conservative probability inequalities (such as the Bonferroni inequality) in their derivation. Our estimation procedures, on the other hand, will typically require large sample sizes. As we become more demanding, requiring a smaller confidence interval width or higher confidence, the final sample size becomes larger, making normality of mean estimators more plausible. In fact, the procedures provided adequate or even conservative coverage in experiments.

These new procedures are far more efficient than existing ones, and make difficult simulation problems tractable. One might fear that the time to estimate the maximum of k means would be on the order of k times as long as the time to estimate a single mean, and this is true for the standard procedure. Our multistage procedures using screening with CRN improve speed greatly, even when the demand for precision is very low. In examples with $k = 64$ and 256 systems, our procedures take not 64 or 256 times as long to estimate the maximum mean than to estimate a single mean, but usually only about twice as long or less, sometimes only a few percent longer. This makes simulation of coherent risk measures based on generalized scenarios affordable, enabling better risk management and innovative derivative security pricing techniques.

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

Acknowledgments

This material is based on work supported by the National Science Foundation under Grants DMI-0217690 and DMS-0202958, and by the National Security Agency under Grant H98230-04-1-0047. The authors thank the associate editor and two anonymous referees for numerous comments that improved the substance and presentation of this paper.

References

Artzner, P., F. Delbaen, J.-M. Eber, D. Heath. 1999. Coherent measures of risk. *Math. Finance* **9** 203–228.

Bechhofer, R. E., T. J. Santner, D. M. Goldsman. 1995. *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*. John Wiley & Sons, New York.

Boesel, J., B. L. Nelson, S.-H. Kim. 2003. Using ranking and selection to "clean up" after simulation optimization. *Oper. Res.* **51**(5) 814–825.

Chen, H. J., E. J. Dudewicz. 1976. Procedures for fixed-width interval estimation of the largest normal mean. *J. Amer. Statist. Assoc.* **71** 752–756.

Cont, R., P. Tankov. 2004. *Financial Modelling with Jump Processes*. Financial Mathematics Series, Chapman & Hall/CRC, London, UK.

Delbaen, F. 2002. Coherent risk measures on general probability spaces. K. Sandmann, P. J. Schönbucher, eds. *Advances in Finance and Stochastics: Essays in Honour of Dieter Sondermann*. Springer-Verlag, New York, 1–38.

Dudewicz, E. J. 1971. Maximum likelihood estimates for ranked means. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **19** 29–42.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York.

Jaschke, S., U. Küchler. 2001. Coherent risk measures and good-deal bounds. *Finance Stochastics* **5** 181–200.

Law, A. M., W. D. Kelton. 2000. *Simulation Modeling and Analysis*, 3rd ed. McGraw-Hill, New York.

Lesnevski, V., B. L. Nelson, J. Staum. 2004. Simulation of coherent risk measures. R. G. Ingalls, M. D. Rossetti, J. S. Smith, B. A. Peters, eds. *Proc. 2004 Winter Simulation Conf.*, IEEE Press, Piscataway, NJ, 1579–1585.

Lesnevski, V., B. L. Nelson, J. Staum. 2006. An adaptive procedure for estimating coherent risk measures based on generalized scenarios. Working Paper 06-05, Department of Industrial Engineering and Management Science, Northwestern University, Evanston, IL.

Nelson, B. L., J. Staum. 2006. Control variates for screening, selection, and estimation of the best. *ACM Trans. Modeling Comput. Simulation* **16**(1) 52–75.

Shiryaev, A. N. 1999. *Essentials of Stochastic Finance: Facts, Models, Theory*. *Advanced Series on Statistical Science & Applied Probability*, No. 3. World Scientific, Singapore.

Staum, J. 2004. Fundamental theorems of asset pricing for good deal bounds. *Math. Finance* **14**(2) 141–161.