NORTHWESTERN UNIVERSITY

Simulation of Coherent Risk Measures Based on Generalized Scenarios

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Industrial Engineering and Management Sciences

By

Vadim Lesnevski

EVANSTON, ILLINOIS

December 2006

ABSTRACT

Simulation of Coherent Risk Measures Based on Generalized Scenarios

Vadim Lesnevski

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.

Table of Contents

ABSTRACT 2

Chapter 1. Introduction 5

1.1. Basket Put 9

1.2. Options Portfolio 10

Chapter 2. Basic Procedures with Screening 13

2.1. Procedures with Guaranteed Coverage 13

2.2. Computational Results 18

2.3. Conclusions 23

Chapter 3. Procedures with Common Random Numbers, Control Variates and Dynamic Stopping 25

3.1. A Framework for Estimating the Maximum 25

3.2. Procedures 30

3.3. Experimental Results 45

3.4. Conclusions 58

Chapter 4. An Adaptive Procedure 60

4.1. Adaptive Multi-stage Procedure 61

4.2. Performance of the Adaptive Multi-stage Procedure 74

Chapter 5. Robustness of the Adaptive Procedure 81

5.1. Robustness to Non-normality 81

5.2. Empirical Analysis of Rare Errors 83

5.3. Conclusions 86

References 87

Appendix A. Algorithms 90

A.1. The Standard Algorithm 90

A.2. A Two-Stage Algorithm with Screening 92

A.3. A Multi-Stage Algorithm with Early Stopping 94

A.4. A Multi-Stage Algorithm with Restarting 96

A.5. An Adaptive Multi-Stage Algorithm 99

Appendix B. Proofs 102

B.1. Validity of the Basic Procedures with Screening 102

B.2. Validity of the Procedures with CRN, CV and Dynamic Stopping 105

B.3. Validity of the Adaptive Procedure 111

Appendix C. Variants 114

C.1. Common Random Numbers 114

C.2. Multi-stage Procedures without CRN 115

C.3. Error Spending 116

Appendix D. Control Variate Estimators 119

CHAPTER 1

Introduction

Both poor risk measures and scarcity of computational resources hamper effective risk management. For instance, 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 towards 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 (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 (multi-stage) 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

$$\rho(Y) = \sup_{\mathbf{P}\in\mathcal{P}} \mathrm{E}_{\mathbf{P}}[-Y/r], \tag{1.1}$$

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, Thm. 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 instance, 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 (1.1) 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 $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$ 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 some 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 $\bar{X}_i$ be a sample average of the random variable X_i. The problem features a natural bias: the most obvious estimator $\max_{i=1,2,\ldots,k} \bar{X}_i$ is an upper bound for, and has a larger expectation than, $\bar{X}_{[k]}$, 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 $\max_{i \in I} \widehat{\mu}_i$ as an estimator for $\mu_{[k]}$; instead, we simply take the confidence-interval width L small enough 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 this 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 (Chapter 2). We use variance-reduction techniques to sharpen the screening and reduce the total sample size required for estimating the maximum (Chapter 3). 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.

In Chapter 4 we introduce an adaptive multi-stage procedure that does not require any previous knowledge about the problem at hand in order to be efficient. In Chapter 5 we explore robustness of the adaptive procedure to non-normality and perform empirical analysis of the rare errors when the confidence interval does not contain the true value.

In Appendix A we specify the algorithms used in the experiments, and provide proofs of the validity of the presented procedures in Appendix B. We discuss possible variants of the procedures in Appendix C, and in Appendix D we give definitions and notation for the text's use of control variates.

1.1. 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, 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, Chs. 7, 13) and Shiryaev (1999, Ch. IV). However, estimation of the non-diagonal 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.

The three control variates used in this example in Chapters 3-5 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.

1.2. 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: under a 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 (1.1), 255 are defined by $\mathbf{P}_i[E] = \mathbf{P}[E|A_i]$ 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 $d_j = 0$. Since we 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

Table 1.1. Amounts of Options in the Portfolio

Asset	Option Type	Strike Price 85	90	95	100	105	110	115
1	put	-2000	-2000	-2500	1000	0	0	0
2	put	2500	-1000	1000	500	0	0	0
3	put	1500	1000	2500	-1500	0	0	0
1	call	0	0	0	-1000	1500	-500	-1000
2	call	0	0	0	1500	-2500	2000	-2000
3	call	0	0	0	-2000	-1000	1000	2500

yielding results we consider typical. Table 1.1 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.

CHAPTER 2

Basic Procedures with Screening

Our point of departure is the theorem of Chen and Dudewicz (1976) providing a fixed-width, two-sided confidence interval for the maximum $\mu_{[k]}$, based on a two-stage sampling plan. We also draw on results of Nelson et al. (2001) to analyze a multi-stage simulation with screening: those systems which are very likely not to be the best are discarded so that thereafter computational resources can be devoted to simulating systems that are more likely to be the best.

2.1. Procedures with Guaranteed Coverage

We use as a standard the two-stage procedure of Chen and Dudewicz (1976). In the first stage, it samples n_0 observations from each system. It then estimates the standard deviation of each system, and uses this to determine how many additional observations are required for each system to attain a minimum coverage guarantee for the confidence interval. In the second stage, it samples this additional data.

For simplicity of presentation, we first consider a two-stage procedure with screening. It is a modification of the Chen-Dudewicz procedure, in which we screen out those systems which prove sufficiently uncompetitive in the first stage. We sample only from the

remaining systems in the second stage. Subsequently we present a multi-stage procedure, in which screening takes place between every stage.

To facilitate consistency of notation, henceforth let the first stage be denoted the 0th and the second be denoted the 1st. Let

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

be the stage 0 sample average and sample variance. Then $\bar{X}_{[i]}$ is the stage 0 sample average associated with the population whose mean is $\mu_{[i]}$. Let the total number of samples from system i taken by the end of the stage 1 be N_i; this is specified later in Equation (2.6). Define the stage 1 sample average

$$\bar{X}_i := \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}.$$

Finally, let F_ν be the t distribution with $\nu := n_0 - 1$ degrees of freedom.

We want a two-sided confidence interval of the form

$$\left(\bar{X}_{(k)} - a, \bar{X}_{(k)} + b\right) \tag{2.1}$$

with error bounds

$$\Pr\left[\mu_{[k]} \leq \max_{i\in I} \bar{X}_i - a\right] \leq \alpha \tag{2.2}$$

and

$$\Pr\left[\mu_{[k]} \geq \max_{i \in I} \bar{X}_i + b\right] \leq \beta \tag{2.3}$$

and having fixed width $L := a + b$. The reason for specifying the confidence for the lower and upper confidence limits separately is the asymmetry of the financial problem. It may be considered worse to underestimate risk than to overestimate it; or worse to set the price of a derivative security too low, thus incurring losses, than to set it too high, thus failing to make sales. If so, one would choose $\beta < \alpha$.

2.1.1. A Two-Stage Procedure

To begin with, choose a width L and confidence levels $1 - \alpha$ and $1 - \beta$. There is also freedom to choose the first-stage sample size n_0 and to decompose the upper confidence level as $1 - \beta = (1 - \beta_0)(1 - \beta_1)$ where β_0 is the error bound allocated to screening and β_1 is the error bound allocated to mean estimation. Let

$$a = L \frac{F_\nu^{-1}((1-\alpha)^{1/k})}{F_\nu^{-1}((1-\alpha)^{1/k}) + F_\nu^{-1}(1-\beta_1)} \tag{2.4}$$

and

$$b = L \frac{F_\nu^{-1}(1-\beta_1)}{F_\nu^{-1}((1-\alpha)^{1/k}) + F_\nu^{-1}(1-\beta_1)}. \tag{2.5}$$

Take the stage 0 sample of X_{ij} for $i = 1, \ldots, k$ and $j = 1, \ldots, n_0$. Compute the sample averages $\bar{X}_i$ and variances S_i^2.

Construct the set

$$I := \left\{ i | \forall j \neq i, \bar{X}_i \geq \bar{X}_j - W_{ij} \right\}$$

where

$$W_{ij} := F_\nu^{-1} \left((1 - \beta_0)^{1/(k-1)} \right) \sqrt{(S_i^2 + S_j^2)/n_0}.$$

This is the set of systems which are not too unlikely to 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.

For all $i \in I$, let the sample size by the end of stage 1 be

$$N_i := \max \left\{ n_0, \left\lceil \left(\frac{S_i F_\nu^{-1}(1 - \beta_1)}{b} \right)^2 \right\rceil \right\} \tag{2.6}$$

and sample X_{ij} for $i \in I$, $j = n_0 + 1, \ldots, N_i$. Compute the stage 1 sample averages $\bar{X}_i$, choose the greatest, and from it compute the confidence interval as in (2.1).

There is a tension in choosing n_0. If it is too large, then excessive resources are spent, as one may wish to have $N_i < n_0$, which is impossible. If it is too small, then there is insufficient information to screen out poor systems. This motivates the introduction of a multi-stage procedure, which provides multiple opportunities to screen out poor systems.

2.1.2. A Multi-Stage Procedure

In this procedure, there are m screening stages and one final estimation stage. The upper confidence level decomposes as $1-\beta = \prod_{\ell=0}^{m}(1-\beta_\ell)$ where β_m is for the final estimation stage and $\beta_0, \ldots, \beta_{m-1}$ are for the m screening stages.

Stage 0 is the same as in the previous subsection, with sample size n_0 for each system. Construct in the same way the set I of systems that are not screened out. We need at this point to compute the total sample sizes $N_i(\ell)$ for system i achieved by the end of each stage $\ell > 0$. There is substantial freedom to do this.

We choose to do so on the following principles. First, the standard error of the sample average should be equal for all systems that have not been screened out. Second, this standard error should decrease by a constant factor C between each stage $1, \ldots, m$. Third, the final sample size should be (much as in the previous subsection)

$$N_i(m) = \max\left\{n_0, \left\lceil \left(\frac{S_i F_\nu^{-1}(1-\beta_m)}{b}\right)^2 \right\rceil\right\}.$$

To satisfy these, use

$$N_i(\ell) = \left\lceil n_0 \left(C^{\ell-1}\frac{S_i}{\min_{j\in I} S_j}\right)^2 \right\rceil$$

where

$$C = \left(\frac{F_\nu^{-1}(1-\alpha_m)\min_{j\in I} S_j}{b\sqrt{n_0}}\right)^{1/(m-1)}.$$

After each stage $\ell = 1, \ldots, m$, compute the sample averages $\bar{X}_i(\ell) := \sum_{j=1}^{N_i(\ell)} X_{ij}/N_i(\ell)$ for those systems i that have not been screened out, i.e. $i \in I(\ell - 1)$ where the screening procedure is defined by

$$I(\ell) := \left\{ i | \forall j \in I(\ell - 1) \setminus \{i\}, \bar{X}_i(\ell) \geq \bar{X}_j(\ell) - W_{ij}(\ell) \right\}$$

where $I(0) = I$ and

$$W_{ij}(\ell) := F_\nu^{-1}\left((1 - \beta_\ell)^{1/(k-1)} \right) \sqrt{\frac{S_i^2}{N_i(\ell)} + \frac{S_j^2}{N_j(\ell)}}.$$

For ease of theoretical analysis, the preceding formula uses stage-1 sample variances; they are not updated for purposes of computing screening thresholds.

In the end, the confidence interval is as in (2.1), with final sample average $\bar{X}_i = \bar{X}_i(m)$.

2.2. Computational Results

We test the performance of our procedures in pricing a basket put option. Although the payoff in this example is far from normally distributed, the sample averages were approximately normally distributed, and the minimum coverage guarantees for the confidence limits held in all of our computational experiments, which include 300 independently simulated confidence intervals.

We report in Tables 2.1 and 2.2 efficiency improvements for this example, expressed as the ratio of the average number of samples required by the procedure of Chen and

Dudewicz (1976) to the average number required by our procedures. The results are reported for the two-stage procedure with various choices of n_0, the initial (stage 0) sample size, and for the multi-stage procedure with 30 stages and $n_0 = 1000$. For each of four choices of confidence interval width, the best efficiency improvement of a two-stage procedure is highlighted in bold type.

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, in the results of Table 2.1 for a 99% confidence interval, the probability that the true maximum mean exceeds the upper confidence level is guaranteed to be no more than $\beta = 0.2\%$, while the probability that it falls below the lower confidence level is guaranteed to be no more than $\alpha = 0.8\%$.

For ease of interpretation, we specify the confidence interval width L relative to the true value $\mu_{[k]}$, as 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 simulation with similar parameters, which can supply a value of L giving approximately the desired relative precision.

Table 2.1 uses levels of confidence and precision appropriate for a derivative pricing problem. The error probability bound $\beta = 0.2\%$ is very low because offering to sell a derivative security at a low price can lead to large losses, which can be tolerated only infrequently. We consider confidence interval widths of 0.1% to 1% of the true value, which are comparable to or slightly smaller than typical bid-ask spreads. That is, at

Table 2.1. Efficiency Improvement, 99% Confidence

Width of CI	0.1%	0.2%	0.5%	1%
2-stage, $n_0 = 50000$	9.7	9.1	9.1	7.8
2-stage, $n_0 = 100000$	13	14	12	**8.3**
2-stage, $n_0 = 200000$	22	17	**14**	6.3
2-stage, $n_0 = 500000$	**39**	**29**	9.8	3.4
2-stage, $n_0 = 1000000$	35	22	6.0	1.7
multi-stage, $m = 30$	**43**	**42**	**36**	**27**

greater widths, one would be unable to quote competitive prices. Lesser widths would be unnecessarily precise.

Table 2.2 is appropriate for a risk management problem requiring lower 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 in the 1% of cases where it occurs, or whose business is lost when the upper confidence limit is too far above the true value.

These tables both show that the performance of the two-stage procedure depends significantly on the initial sample size n_0. When n_0 is small, increasing it tends to lead to improved screening, as more information at stage 0 allows more suboptimal systems to be

Table 2.2. Efficiency Improvement, 95% Confidence

Width of CI	0.5%	1%	2%	5%
2-stage, $n_0 = 5000$	2.6	2.7	2.6	2.3
2-stage, $n_0 = 10000$	4.0	4.1	3.8	**2.5**
2-stage, $n_0 = 20000$	6.3	5.8	**4.8**	2.0
2-stage, $n_0 = 50000$	9.6	**7.8**	4.1	1.0
2-stage, $n_0 = 100000$	**14**	7.6	2.7	0.5
2-stage, $n_0 = 200000$	12	5.0	1.5	0.3
multi-stage, $m = 30$	**36**	**24**	**13**	**4.5**

screened out. If n_0 becomes too large, computational resources are wasted on poor systems that could have been screened out earlier and on systems with low standard deviation, for which one would have liked to set $N_i < n_0$ if this were possible—see Equation (2.6). However, because many financial simulations are repeated with parameters only slightly different from those at the previous repetition, a good value of n_0 may well be known in advance.

Nonetheless, the performance of the multi-stage procedure is entirely superior in the examples here. It overcomes limitations of the two-stage procedure by using a small initial sample size $n_0 = 1000$, but continuing screening at subsequent stages. There seems to be little problem in choosing the multi-stage procedure's parameters for an entirely unfamiliar simulation, which makes it superior to the two-stage procedure. The following investigation of the sensitivity of the multi-stage procedure's performance to its parameters is done at 95% confidence and for a confidence interval width of 5%.

Figure 2.1 shows that the efficiency of the multi-stage procedure has low local sensitivity to the number of stages m.

Figure 2.2 shows that the impact of initial sample size n_0 on the procedure's efficiency is not negligible, but is not as dramatic as it is for the two-stage procedure. Varying n_0 from 200 to 2000 caused efficiency to change by less than 5%. However, $n_0 = 1000$ is not very close to optimal, but noticeably too large, if the required precision is low and the variances are much smaller (say, one tenth as large) relative to the differences in expectations. Still, the n_0 problem is much less severe than for the two-stage procedure:

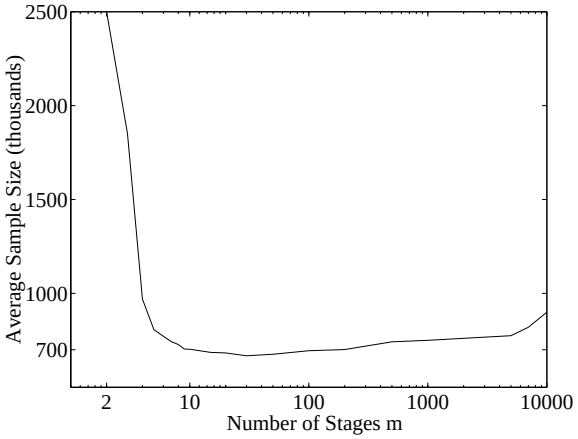

Figure 2.1. Number of Stages and Efficiency

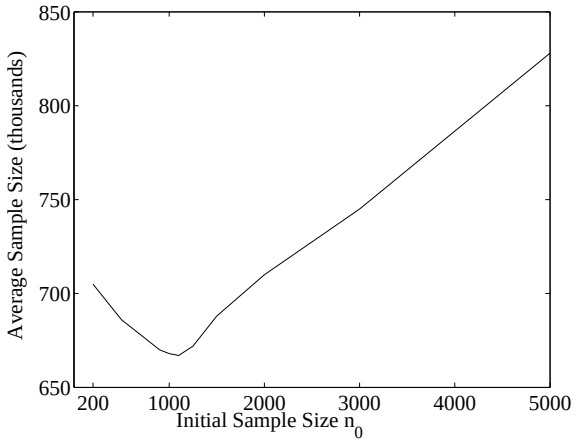

Figure 2.2. Initial Sample Size and Efficiency

$n_0 = 1000$ is close to optimal for a fairly wide range of variances and confidence interval widths L.

The performance of the multi-stage procedure also has little local sensitivity to the decomposition of the upper confidence level $1 - \beta$ into confidence $1 - \beta_m$ for estimation and $1 - \beta_\ell$ for screening at stage $\ell = 0, \ldots, m - 1$. In the examples reported here, we

have chosen $\beta_0 = \cdots = \beta_{m-1}$ and $\prod_{\ell=0}^{m-1}(1-\beta_\ell) = 1-\beta/5$, but as Figure 2.3 shows, there is little change in performance for nearby values of the overall screening confidence level. Allocating too little of the error to screening makes it very difficult to screen out systems; allocating too little of the error to estimation inflates the required final sample size $N_i(m)$ for a system $i \in I(m-1)$ that is never screened out.

2.3. Conclusions

We have introduced a multi-stage screening and selection procedure for producing a simulated confidence interval for the maximum of several expectations. To choose good values of the procedure's parameters (number of stages, initial sample size, and error allocation) does not require precise knowledge of the problem's characteristics; this and superior efficiency are advantages of the multi-stage procedure over the two-stage procedure. For the financial application of simulating a coherent risk measure of a basket

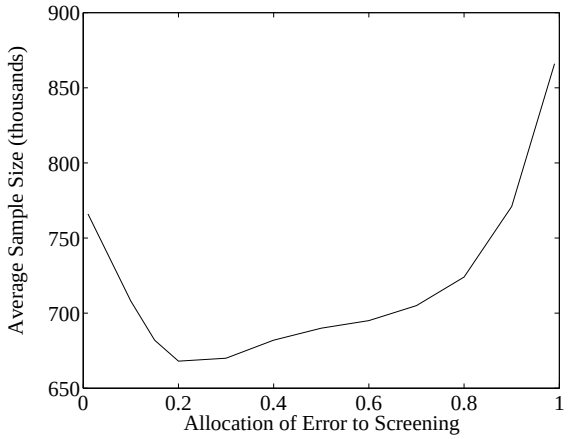

Figure 2.3. Error Allocation and Efficiency

put option, this procedure was between 4.5 and 43 times faster than the procedure of Chen and Dudewicz (1976). The efficiency improvement is greater when the required levels of confidence and precision are higher, in which case it is possible for substantial screening to occur while the procedure runs.

CHAPTER 3

Procedures with Common Random Numbers, Control Variates and Dynamic Stopping

In the previous chapter we introduced procedures that use screening ideas from ranking and selection to reduce drastically the number of systems that need to be simulated to estimate the maximum of several expectations. In this chapter we employ variance reduction techniques to sharpen screening and reduce the total sample size required for estimating the maximum. The multistage procedures that we develop proceed from screening to estimation dynamically as soon as all systems but the best are screened out.

3.1. 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 $\widehat{\mu}_i, i = 1, 2, \ldots, k$, and constants $a, b > 0$ such that

$$\Pr\left\{\mu_{[k]} \geq \max_{i \in I} \widehat{\mu}_i - a\right\} \geq 1 - \alpha_a, \tag{3.1}$$

$$\Pr\left\{\mu_{[k]} \leq \max_{i \in I} \widehat{\mu}_i + b\right\} \geq 1 - \alpha_b, \tag{3.2}$$

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 (3.1) and (3.2) imply that

$$\Pr\left\{\max_{i\in I}\widehat{\mu}_i - a \le \mu_{[k]} \le \max_{i\in I}\widehat{\mu}_i + b\right\} \ge 1 - \alpha_a - \alpha_b. \tag{3.3}$$

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 Section 3.2.3.

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 notice that

$$\begin{aligned} \Pr\left\{\mu_{[k]} \le \max_{i\in I} \widehat{\mu}_i + b\right\} &\ge \Pr\left\{[k] \in I, \mu_{[k]} \le \widehat{\mu}_{[k]} + b\right\} \\ &\ge 1 - \Pr\left\{[k] \notin I\right\} - \Pr\left\{\mu_{[k]} > \widehat{\mu}_{[k]} + b\right\}. \end{aligned} \tag{3.4}$$

Thus, if we can guarantee that

$$\Pr\left\{[k] \notin I\right\} \le \alpha_I \quad \text{and} \tag{3.5}$$

$$\Pr\left\{\mu_{[k]} > \widehat{\mu}_{[k]} + b\right\} \le \alpha_b' \tag{3.6}$$

where $\alpha_I + \alpha_b' = \alpha_b$, then the upper confidence limit will be valid as in Inequality (3.2).

Next consider the lower confidence limit, and notice that

$$\begin{aligned} \Pr\left\{\mu_{[k]} \ge \max_{i\in I} \widehat{\mu}_i - a\right\} &\ge \Pr\left\{\mu_{[k]} \ge \max_{i=1,2,\ldots,k} \widehat{\mu}_i - a\right\} \\ &= \Pr\left\{\widehat{\mu}_i \le \mu_{[k]} + a, i = 1, 2, \ldots, k\right\} \\ &\ge \Pr\left\{\widehat{\mu}_i \le \mu_i + a, i = 1, 2, \ldots, k\right\} \\ &\ge 1 - \sum_{i=1}^{k} \Pr\left\{\widehat{\mu}_i > \mu_i + a\right\}. \end{aligned} \tag{3.7}$$

Therefore, the lower confidence limit will be valid as in Inequality (3.1) if, for $i = 1, 2, \ldots, k$,

$$\Pr\{\widehat{\mu}_i > \mu_i + a\} \leq \alpha'_a = \alpha_a/k. \tag{3.8}$$

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 (3.6) and (3.8) 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 $x > 0$,

$$\Pr\{\widehat{\mu}_i - \mu_i > x\} \leq 1 - G_a(cx) \quad \text{and} \quad \Pr\{\widehat{\mu}_i - \mu_i < -x\} \leq 1 - G_b(cx), \tag{3.9}$$

where

$$a = \frac{1}{c} G_a^{-1}(1 - \alpha'_a), \tag{3.10}$$

$$b = \frac{1}{c} G_b^{-1}(1 - \alpha'_b), \quad \text{and} \tag{3.11}$$

$$c = \frac{1}{L}\left(G_a^{-1}(1 - \alpha'_a) + G_b^{-1}(1 - \alpha'_b)\right). \tag{3.12}$$

This determines the sampling scheme in such a way that it bounds the distribution of $\widehat{\mu}_i - \mu_i$ by a function that is free of dependence on i (see Sections 3.2.1 and 3.2.2 for examples).

Proposition 3.1.1. *Inequalities (3.1) and (3.2) hold if Inequalities (3.5) and (3.9) 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 (3.9) and Equation (3.10), $\Pr\{\widehat{\mu}_i - \mu_i > a\} \leq \alpha'_a$, while $\Pr\{\widehat{\mu}_i - \mu_i < -b\} \leq \alpha'_b$ by Inequality (3.9) and Equation (3.11). Thus, for all $i = 1, 2, \ldots, k$, Inequality (3.8) holds, which we already argued implies Inequality (3.1). Inequality (3.6) holds, and we have already argued that with Inequality (3.5) it implies Inequality (3.2). □

To show that a procedure delivers confidence limits with at least the coverage probabilities specified in Inequalities (3.1) and (3.2), we will verify that the screening procedure satisfies Inequality (3.5), 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 (3.9). These results provide a general framework for estimating $\mu_{[k]}$; the remainder of the chapter works out details for specific ways to form the subset I and the estimators $\widehat{\mu}_i$. 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 $\widehat{\mu}_i$ of the mean μ_i for each system $i \in I$.
4. Report the confidence interval $[\max_{i \in I} \widehat{\mu}_i - a, \max_{i \in I} \widehat{\mu}_i + b]$.

We obtain *efficient* procedures in two ways:

1. by reducing $|I|$, the number of means that we estimate, and
2. by employing efficient estimators $\widehat{\mu}_i$ of μ_i, so that the means we do estimate require as little computational effort as possible.

In Chapter 2, we reported on two-stage and multi-stage 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 chapter, 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.

3.2. 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.

Appendix A contains algorithms implementing procedures with various combinations of these features. Proofs of the procedures' validity appear in Appendix B.

3.2.1. 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 n_0 replications of X_i, which has the distribution of the discounted loss X under measure $\mathbf{P}_i$. The replications are used to estimate the variances $\sigma_i^2 := \mathrm{Var}[X_i]$. These are

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

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

$$N_i = \max \left\{ n_0, \left\lceil c^2 S_i^2 \right\rceil \right\} \tag{3.13}$$

are computed on the basis of the variance estimators S_i^2 and the scaling constant c as defined in Equation (3.12), 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 stage-1 sample averages

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

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

3.2.2. Controlled Control Variates

In this section, we present an extension of the Chen-Dudewicz 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. In addition, control variates in screening introduce technical complications: see 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 Appendix D. We introduce a q_i-dimensional vector C_i of control variates with known mean ξ_i. As 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, as the same financial variables are used in each case. In the basket put example, 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 (3.12) are given by $G_a(x) = G_b(x) = F_{t_{n_0-q-1}}(x) - \alpha_C$, so

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

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 (3.13) 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 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:

$$N_i = \max\left\{n_0, \left\lceil c^2\hat{\tau}_i^2 + \chi^2_{q_i,1-\alpha_C}\right\rceil\right\}. \tag{3.14}$$

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

3.2.3. Screening 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 $j \neq \ell$, X_{hj} and $X_{i\ell}$ 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 $I := \left\{ i | \forall h \neq i, \bar{X}_i \geq \bar{X}_h - W_{hi} \right\}$ where the threshold

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

The set I contains those systems which 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 decrease efficiency? The error spent on screening subtracts from the error that can be spent on estimating systems' means (Equations 3.5–3.6), 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. Therefore, 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

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

This follows from Equation (3.12) 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, the worst-case efficiency loss is only 7.7%. If $k = 64$, as in the basket put example, it is 8.8%. Even if $k = 2$, the worst-case efficiency loss is just 14%, although it is very unlikely the procedure will be used when the number of scenarios is so small. For these reasons, screening is very likely to improve efficiency, and even if it does not, it can not decrease efficiency by much.

As discussed in Chapter 2, the performance of the two-stage procedure depends significantly on the initial sample size n_0. When n_0 is small, increasing it tends to lead to improved screening, as 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 (3.13). It would be preferable to have a procedure that is less sensitive to n_0, and the multi-stage procedure described in the next section has this property.

3.2.4. Multi-Stage 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 $\bar{X}_i(\ell) := \sum_{j=1}^{N(\ell)} X_{ij}/N(\ell)$, where $N(\ell)$ is the total number of replications

sampled from each surviving system through screening stage ℓ. Appendix C.1 includes an explanation of why the sample size $N(\ell)$ is the same for each system still in contention.

There are three main aspects of the multi-stage procedure to resolve. We must specify

1. the screening stage sample sizes $N(\ell)$ for $\ell = 1, 2, \ldots, m-1$,
2. the screening thresholds $W_{hi}(\ell)$ for $\ell = 0, 1, \ldots, m-1$ and $h, i = 1, 2, \ldots, k$, and
3. the sample size N_i used in constructing the mean estimate $\widehat{\mu}_i$ 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 (3.5) and choose the final sample size so that Inequality (3.9) 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 $N(0), \ldots, N(m-1)$ 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 $N(\ell) = \lceil n_0 R^\ell \rceil$. 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 instance, sample sizes grew at a constant arithmetic 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 $\widehat{\mu}_i$, 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 (3.15) 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.

Next we consider the screening thresholds and error decomposition, given that sample sizes are fixed in advance. After each stage $\ell = 0, 1, \ldots, m-1$, screening takes place by

constructing

$$I(\ell+1) := \left\{ i \in I(\ell) \middle| \forall h \in I(\ell), \bar{X}_i(\ell) \geq \bar{X}_h(\ell) - W_{hi}(\ell) \right\}$$

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

$$W_{hi}(\ell) := t_{N(\ell)-1,1-\alpha_I/(m(k-1))} \frac{S_{hi}(\ell)}{\sqrt{N(\ell)}}, \tag{3.15}$$

where the stage-ℓ sample variance is

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

We use fully updated, cumulative sample variances to set the screening thresholds. Typically, multi-stage screening procedures for ranking and selection use only stage-0 sample variances to simplify inference. We find that it is valuable to use updated variance information and keep stage 0 very small, because a large fraction of systems were screened out at stage 0 in our examples.

After screening, we must choose a final sample size N_i for estimation of the mean μ_i by $\widehat{\mu}_i$. We cover the case with CV, which subsumes that without CV (Section 3.2.2). The scaling constant c comes from Equation (3.12) and $G_a(x) = G_b(x) = F_{t_{N(m-1)-q-1}}(x) - \alpha_C$. Equation (3.14) becomes

$$N_i = \max\{N(m-1), \lceil c^2 \hat{\tau}_i^2 (m-1) + \chi^2_{q_i,1-\alpha_C} \rceil\}, \tag{3.16}$$

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 $j = N(m-1)+1, N(m-1)+2, \ldots, N_i(m)$, and then the confidence limits are constructed around $\max_{i \in I(m)} \widehat{\mu}_i$, where each estimate $\widehat{\mu}_i$ is based on all replications $j = 1, 2, \ldots, N_i(m)$. That is, $\widehat{\mu}_i$ 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 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 (3.5) 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 instance, 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. The multi-stage procedure has an "$N(m-1)$ problem" in the same way. If we choose the maximum per-system screening budget $N(m-1)$ to be too small, not enough screening is done. If we choose $N(m-1)$ too large, then this multi-stage

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 (3.16) for any system that survives too long.

In the next section, an enhancement to the multi-stage 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. In Chapter 4 we present an adaptive multi-stage procedure that solves this problem.

3.2.5. 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 $I(\ell) = \{[k]\}$ often occurred for some screening stage $\ell < m-1$. 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

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

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 $M = \ell$ of stopping at an early stage ℓ is associated with low values of $S_{ih}^2(M)$ for all systems $h \neq i$, because when these sample variances are low, it helps system i to screen out all the others quickly. Low values of $S_{ih}^2(M)$ 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 $\hat{\tau}_i^2(\ell)$ 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

$$N_i = \max\{N(M-1), \lceil c^2\hat{\sigma}_i^2 + \chi^2_{q_i,1-\alpha_C}\rceil\}, \tag{3.17}$$

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 $M \geq \ell^*$. We only use variance information up through stage ℓ^* to determine the terminal sample size for estimation. That is, $\hat{\sigma}_i^2 = \hat{\tau}_i^2(\ell^*)$ is the sample residual variance of the regression of $X_{i1}, X_{i2}, \ldots, X_{iN(\ell^*)}$ on the control variates $C_{i1}, C_{i2}, \ldots, C_{iN(\ell^*)}$. Because $\hat{\tau}_i^2(\ell^*)$ is computed over a prespecified constant number $N(\ell^*)$ of replications, we can find associated chi-squared and t distributions. The scaling

constant c comes from Equation (3.12) with

$$G_a(x) = G_b(x) = F_{t_{N(\ell^*)-q-1}}(x) - \alpha_C. \tag{3.18}$$

3.2.6. Restarting

The critical values that determine the overall sample size for mean estimation depend upon 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. Therefore, when $K(M) := |I(M)|$, the number of systems remaining after screening ends at the random stage $M-1$, is small, it would 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 simulation stages.

After screening, we will obtain N_i new replications for each surviving system $i \in I(M)$ and form the estimators $\widehat{\mu}_i$ 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

$$N_i = \max\{n_i, \lceil c^2\hat{\sigma}_i^2 + \chi^2_{q_i,1-\alpha_C}\rceil\}. \tag{3.19}$$

The scaling constant c comes from Equation (3.12) with

$$G(x) = F_{t_{n-q-1}}(x) - \alpha_C, \tag{3.20}$$

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, since 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 $F_{N(M-1)-1,n_i-1}$. Assuming that n_i is large, the distribution of $(N(M-1)-1)\hat{\tau}_i^2(M-1)/\hat{\sigma}_i^2$ 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 (3.20) has many degrees of freedom and is thus approximately a normal distribution. This yields, from Equation (3.12), $c \approx (\Phi^{-1}(1-\alpha_a'') + \Phi^{-1}(1-\alpha_b''))/L$. Putting these approximations together, we set

$$n_i = \left(\frac{\Phi^{-1}(1-\alpha_a'') + \Phi^{-1}(1-\alpha_b'')}{L}\right)^2 \frac{(N(M-1)-1)\hat{\tau}_i^2(M-1)}{\chi^2_{N(M-1)-1,1-\epsilon}} + \chi^2_{q_i,1-\alpha_C}, \tag{3.21}$$

an approximate lower prediction limit for the desired size $c^2\hat{\sigma}_i^2 + \chi^2_{q_i,1-\alpha_C}$ in Equation (3.19).

3.3. Experimental Results

We now report selected results of computational experiments to test the efficiency and validity of the procedures developed in Section 3.2. We discuss the magnitude of the procedures' efficiency gains in Section 3.3.1, as well as the factors that contribute to them. This includes, in Section 3.3.2, an assessment of the extent to which efficiency depends on the choice of parameters such as sample sizes and error decomposition. Section 3.3.3 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 $N(m-1) = n_0 R^{m-1} = 30 \cdot 16384 = 491520$. 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 3.3 shows that this cost is not large relative to the simulation's total cost. For the multi-stage algorithms with early stopping, stopping is forbidden until after stage $\ell^* = 5$, yielding $N(\ell^*) = n_0 R^{\ell^*} = 30 \cdot 32 = 960$ replications to provide variance information for use in setting the final sample sizes. For the multi-stage algorithm with restarting, the significance level used in creating the prediction limit for the final sample size that underlies Equation (3.21) is $\epsilon = 1\%$.

The results presented in this section for a multi-stage procedure with early stopping are for a slight variant of the procedure described in Section 3.2.5, differing only as to the observations used in producing variance estimates for screening and terminal sample size computation (fewer are used). Both variants are valid, but the procedure described in Section 3.2.5 is simpler and would be expected to have slightly superior performance than the results presented in this section.

3.3.1. 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 Equation (3.13) or (3.14), 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

Table 3.1. 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

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 3.1 reports the efficiency of four procedures: the multi-stage procedure with restarting and CRN, the multi-stage procedure with early stopping and CRN, the multi-stage 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 3000 for the basket put example and 1000 for the options portfolio example. We chose these values to yield good performance for these examples, at this level of confidence and precision. Nonetheless, the two-stage procedure's performance is markedly inferior to that of the multi-stage procedure, primarily because the multi-stage 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 instance, it can not 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 variations on the multi-stage procedure that become possible without CRN, see 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 multi-stage procedure with CRN and restarting is within a few percent of this minimum size.

In summary, we recommend using a multi-stage 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. 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.

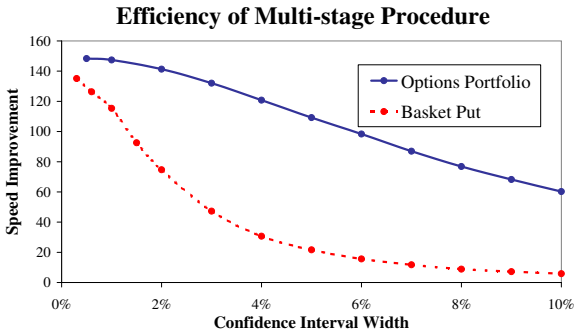

Figure 3.1. Effect of required precision on efficiency of the multi-stage procedure with early stopping and CRN relative to the standard procedure, at 99% confidence.

Figure 3.1 shows the effect of the confidence interval width L on the efficiency of the multi-stage 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.

In Table 3.1, we saw that the multi-stage 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 3.1, 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 multi-stage 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.

3.3.2. 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.

First we consider the effect of the sample sizes of the screening stages on the efficiency of the multi-stage procedure with restarting and CRN. The results are easier to interpret than for the multi-stage 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 $N(\ell) = \lceil n_0 R^\ell \rceil$, 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 systems. Figure 3.2 shows how these changes affect the efficiency of the multi-stage procedure with restarting and CRN.

The graphs in Figure 3.2 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

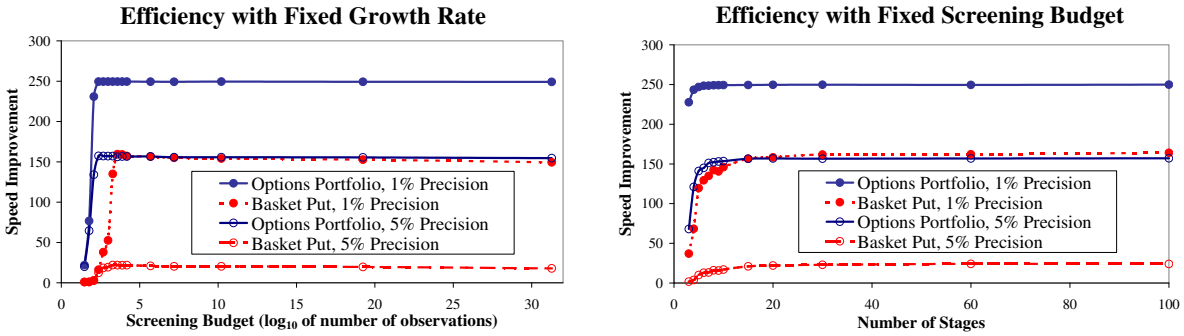

Figure 3.2. Effect of screening phase design on efficiency of the multi-stage procedure with restarting and CRN relative to the standard procedure, at 99% confidence.

$N(m-1)$ 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 3.2, 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 3.1, 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 Section 3.3.3 show that $n_0 = 30$ yielded adequate coverage for these examples.

Next we consider the effect of error allocation on the efficiency of the multi-stage 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 (3.5)). When using control variates, a portion α_C of both lower and upper error must be set aside for controlling them (Section 3.2.2). Figure 3.3 displays the effect of 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

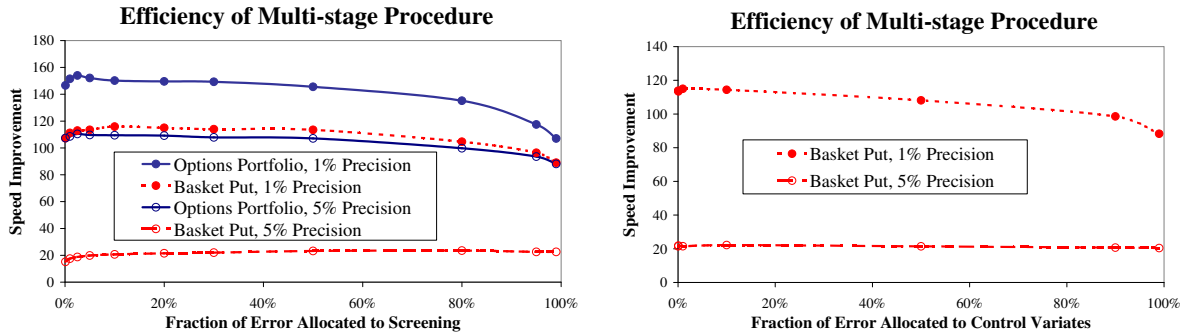

Figure 3.3. Effect of error allocation on efficiency of the multi-stage procedure with early stopping and CRN relative to the standard procedure, at 99% confidence.

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 insufficiency than excess. If α_I is too small, less screening takes place because the thresholds in Equation (3.15) 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 $m = 15$, $N(\ell^*) = 960$, 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 (3.14), 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 (Section 3.2.5) and restarting (Section 3.2.6). After some experimentation, we selected the first stage after which early stopping is allowed as $\ell^* = 5$. 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 multi-stage procedure with early stopping and without CRN on the basket put example, shown in Table 3.1. For the multi-stage 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 (3.21). A good value is $\epsilon = 1\%$.

3.3.3. Coverage

Our procedures come with coverage guarantees (3.1) and (3.2) for their confidence limits, but the guarantees are proved only for normally distributed data X_{ij}. 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 as in Chapter 2. Table 3.2 supports the conclusion that

Table 3.2. Error rates of multi-stage procedures with CRN at 95% confidence and 5% precision.

Error Prob.	Nominal	Estimate	With Early Stopping: Basket Put	With Early Stopping: Options Portfolio	With Restarting: Basket Put	With Restarting: 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 multi-stage 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 3.2 contain 5000 independent simulations. For each experiment, we report (in bold) the fraction of these 5000 simulations in which $\mu_{[k]} < \max_{i \in I} \widehat{\mu}_i - a$ as a point estimate of the lower error probability $\Pr\left\{\mu_{[k]} < \max_{i \in I} \widehat{\mu}_i - a\right\}$, and similarly for the upper error probability $\Pr\left\{\mu_{[k]} > \max_{i \in I} \widehat{\mu}_i + b\right\}$. 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 3.2 shows that the multi-stage 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 (3.8), even those which are screened out. This was the motivation for the procedure with restarting, which is indeed much less conservative.

3.4. Conclusions

In this chapter we proposed procedures for constructing a two-sided, fixed width confidence interval for the maximum or minimum 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 multi-stage 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.

CHAPTER 4

An Adaptive Procedure

A disadvantage of the procedures presented in Chapter 3 is that, in some cases, they might require some previous knowledge about the problem to be efficient. For example, having a large screening budget is usually good, as it allows the procedure to screen out most of the inferior systems. However, it might significantly decrease efficiency if more than one system has the maximum mean, or if some systems are nearly tied with the best. In such situations, screening might not be able to eliminate all systems but one. Even though the procedure with restarting is usually preferable over other alternatives, if screening is ineffective, restarting is wasteful of data. Before running the simulation, the user would have to decide whether or not to use restarting, and how much data to allocate to the screening stage. Making a good decision without substantial experience with simulation problems of the same form is difficult.

Without restarting, information generated during screening is reused during estimation of the confidence interval, so the amount of work done during screening is not very important. With restarting, information generated during screening is thrown away, so it is important to make sure that no excess work is done during screening. The advantage of restarting is that the new data are statistically independent of the screening exercise, so one may ignore the measures which were screened out, and design for the smaller problem.

In this chapter we develop an adaptive multi-stage procedure which combines good features of both approaches. The procedure is very efficient for all configurations, as it gains the benefits of restarting and of having a large budget to use for screening.

4.1. Adaptive Multi-stage Procedure

Our procedures produce a lower confidence limit that covers the coherent risk measure with probability at least $1-\alpha_a$, and an upper confidence limit that covers with probability at least $1-\alpha_b$. See Appendix B for a proof. The procedures spend some of this allowable error on screening (α_I), some on control variates (α_C), and the remainder on estimating the means of some systems. We use the control variate C_i for the output X_i of system i to improve estimation of the mean μ_i of system i.

The adaptive multi-stage procedure consists of two phases. Phase I ("pre-screening") consists of multi-stage screening whose purpose is, while controlling relative cost, to screen out as many inferior systems as possible, so that they do not contribute to the critical values that determine the overall sample size for mean estimation. No samples obtained during pre-screening are used during Phase II, which is an estimation procedure with additional multi-stage screening.

4.1.1. Phase I: Pre-screening

The sole purpose of the first phase is to reduce the number of systems and thus the natural bias of the estimation problem, making a fixed-width confidence interval attainable with fewer replications.

The maximal number of Phase I stages, m, is specified in advance. The first stage of Phase I is stage 0 and the first stage of Phase II is stage M, where the random variable $M \leq m$. The decision to proceed to Phase II is made randomly, on the basis of the simulated data, when the cost of continuing and doing one more stage of Phase I is greater than the estimated approximate savings due to further pre-screening. The growth rate R and the initial sample size n_0 are also specified in advance, so that the total sample size during stage ℓ is $N(\ell) = \lceil n_0 R^\ell \rceil$.

The initial sample size n_0 should be chosen so that sample averages are approximately normal. In most cases, $n_0 = 30$ is adequate. The procedure is most efficient if the growth factor R is between 1.2 and 2.0, while m is such that the total budget available for pre-screening is large. For example, if $R = 1.5$ and $m = 30$, the total budget available for Phase I is $\lceil n_0 R^{m-1} \rceil = 3,835,021$, which is large enough for most applications. We found that $R = 1.5$ and $m = 30$ worked well on all problems we consider. It was not possible to improve on the performance much by altering the parameters, as it was for the procedures presented in Chapter 3.

Let I be the set of systems that have not been screened out. Initially set $I \leftarrow \{1, \ldots, k\}$. Each stage $\ell = 0, \ldots, m-1$ of Phase I consists of the following steps:

(1) *Simulation.*

Simulate (X_{ij}, C_{ij}) for $j = N(\ell-1)+1, \ldots, N(\ell)$ and all $i \in I$.

(2) *Screening.*

For each $h, i \in I$ such that $h \neq i$, set

$$\begin{aligned} \bar{\bar{D}}_{hi} &\leftarrow \frac{1}{N(\ell)} \sum_{j=1}^{N(\ell)} (X_{hj} - X_{ij}), \\ S_{hi}^2 &\leftarrow \frac{1}{N(\ell)-1} \sum_{j=1}^{N(\ell)} (X_{hj} - X_{ij} - \bar{\bar{D}}_{hi})^2, \\ W_{hi} &\leftarrow \frac{t_{N(\ell)-1,1-\alpha_I/(2m(k-1))}}{\sqrt{N(\ell)}} S_{hi}, \end{aligned}$$

where $t_{\nu,r}$ is the r quantile of the t distribution with ν degrees of freedom.

Then set $I \leftarrow \left\{ i \in I | \forall h \in I, \bar{\bar{D}}_{hi} \geq -W_{hi} \right\}$.

(3) *Checking whether to proceed to Phase II.*

For each $i \in I$, compute the residual variance $\hat{\sigma}_i^2$ of regressing $X_{i,1}, \ldots, X_{i,N(\ell)}$ on $C_{i,1}, \ldots, C_{i,N(\ell)}$ and define

$$c_p := \frac{1}{L}(\Phi^{-1}(1 - \alpha_a/p + \alpha_C) + \Phi^{-1}(1 - \alpha_b + \alpha_I + \alpha_C)), \tag{4.1}$$

where Φ is the standard normal cumulative distribution function. If $\ell = m - 1$ or

$$|I| N(\ell)(R-1) > (c_{|I|}^2 - c_1^2) \max_{i \in I} \hat{\sigma}_i^2, \tag{4.2}$$

the procedure jumps to Phase II by setting $M \leftarrow \ell + 1$, which means that the next stage is the first stage of Phase II, and by setting $K \leftarrow |I|$, which is the number of systems left after pre-screening and which will be used for determining final sample sizes. Otherwise, set $\ell \leftarrow \ell + 1$ and return to Step 1.

Under the transition rule given by Inequality (4.2), pre-screening stops after stage $M-1$ when the cost of doing one more stage of pre-screening is greater than the approximate maximal savings due to continuation, computed under the assumption that after additional pre-screening there will be only one system left and it will have the largest variance.

4.1.2. Phase II: Screening and Estimation

Phase II begins by restarting, that is, throwing out all the data obtained in Phase I. The only effect of Phase I on Phase II is that Phase I determines the subset I of systems that Phase II handles. Phase II contains three parts.

First, in the initial stage M, the procedure determines the required total sample sizes N_i for each of the systems in I and the maximal necessary number P of subsequent screening stages. Second, in stages $M, \ldots, M+P-1$, the procedure does more screening. It maintains two sets of systems: the set I contains systems that have survived screening and from which the procedure has simulated as many samples as are required to construct the fixed-width confidence interval, while the set $\hat{I}$ contains systems that have survived screening so far, but which still require more sampling. Finally, once the required sample

size has been reached for all surviving systems, the procedure constructs a confidence interval.

Because M is the first stage after restarting, the procedure discards $\lceil n_0 R^{M-1} \rceil$ Phase I samples. To compensate for the discarded samples and keep the growth rate constant, during Phase II the procedure sets $N(\ell) \leftarrow \lceil n_0 R^{\ell-1}(R+1) \rceil$, $\ell \geq M$. This makes the total Phase II sample size grow at the rate R. It also makes the initial sample size of Phase II be $N(M) - N(M-1) \simeq n_0 R^M$, which is large enough to ensure high-quality variance estimates.

Initialize $\hat{I} \leftarrow I$ and then $I \leftarrow \emptyset$. Also initialize $N_i \leftarrow N(M)$ for all $i \in \hat{I}$. Each stage $\ell = M, \ldots, M+P$ consists of the following steps, except that only stage M contains Step 2, and Step 4 will not occur during stage $M+P$ because $\hat{I}$ will be empty then:

(1) *Simulation.*

Simulate (X_{ij}, C_{ij}) for $j = N(\ell-1)+1, \ldots, \min\{N_i, N(\ell)\}$ and all $i \in \hat{I}$.

Set $n \leftarrow N(\ell) - N(M-1)$.

(2) *Setting final sample sizes.*

If $\ell > M$, skip this step.

Set $\alpha_a'' \leftarrow \alpha_a/K - \alpha_C$ and $\alpha_b'' \leftarrow \alpha_b - \alpha_I - \alpha_C$, and set the scaling constant

$$c \leftarrow \frac{1}{L}(t_{n-q-1,1-\alpha_a''} + t_{n-q-1,1-\alpha_b''}), \tag{4.3}$$

where $q := \max_{i \in I} q_i$ and each q_i is the number of control variates in C_i.

For each $i \in \hat{I}$, compute the residual variance $\hat{\sigma}_i^2$ of regressing $X_{i,N(M-1)+1}, \ldots, X_{i,N(M)}$ on $C_{i,N(M-1)+1}, \ldots, C_{i,N(M)}$, and from it the total sample size

$$N_i \leftarrow \lceil c^2 \hat{\sigma}_i^2 + \chi^2_{q_i, 1-\alpha_C} \rceil + N(M-1), \tag{4.4}$$

where $\chi^2_{\nu,p}$ is the p quantile of the chi-squared distribution with ν degrees of freedom.

Set $P \leftarrow \lceil \log_R \max_{i \in I}(N_i / N(M)) \rceil$.

(3) *Updating I and $\hat{I}$.*

Add to I systems that have reached their required sample sizes and remove them from $\hat{I}$: set $I \leftarrow I \bigcup \left\{ i \in \hat{I} | N_i \leq N(\ell) \right\}$ and $\hat{I} \leftarrow \hat{I} \backslash I$.

(4) *Screening.*

For each $h, i \in \hat{I}$ such that $h \neq i$, set

$$\begin{aligned}
\bar{\bar{D}}_{hi} &\leftarrow \sum_{j=N(M-1)+1}^{N(\ell)} \frac{X_{hj} - X_{ij}}{n}, \\
S^2_{hi} &\leftarrow \sum_{j=N(M-1)+1}^{N(\ell)} \frac{(X_{hj} - X_{ij} - \bar{\bar{D}}_{hi})^2}{n-1}, \\
W_{hi} &\leftarrow \frac{1}{\sqrt{n}} t_{n-1, 1-\alpha_I/(2P(K-1))} S_{hi}.
\end{aligned}$$

Then set $\hat{I} \leftarrow \left\{ i \in \hat{I} | \forall h \in I, \bar{\bar{D}}_{hi} \geq -W_{hi} \right\}$.

(5) *Continue or compute confidence interval.*

If $\hat{I} \neq \emptyset$, set $\ell \leftarrow \ell + 1$ and return to Step 1.

Otherwise, for each $i \in I$, compute the estimate $\widehat{\mu}_i$ from the regression of $X_{i,N(M-1)+1}, \ldots, X_{i,N_i}$ on $C_{i,N(M-1)+1}, \ldots, C_{i,N_i}$. Set

$$\begin{aligned} a &\leftarrow \frac{1}{c} t_{N(M)-N(M-1)-q-1,1-\alpha''_a} \quad \text{and} \\ b &\leftarrow \frac{1}{c} t_{N(M)-N(M-1)-q-1,1-\alpha''_b}. \end{aligned}$$

The confidence interval is

$$(\max_{i \in I} \widehat{\mu}_i - a, \max_{i \in I} \widehat{\mu}_i + b).$$

4.1.3. Efficiency of the Rule for Restarting

The adaptive procedure offers two significant improvements over our previous procedures.

First, we do not need to specify a screening budget in advance. Choosing the screening budget too small or too big could have a very significant effect on the performance of our previous procedures, in some configurations making a simulation dozens of times slower: see Table 4.3 in Section 4.2. The adaptive procedure solves this problem by trying to screen out a system in Phase II only until its required sample size is reached. In effect, this allows the screening budget to be arbitrarily large, to vary by system, and to be determined adaptively by the required sample size.

Second, the adaptive procedure allows us to restart whatever the configuration of the means $\mu_1, \ldots, \mu_k$ may be. The effect of the decision whether or not to restart on performance is much less severe: as we will show below, usually we do not expect to save more than 40-80%. Restarting is usually beneficial because in a typical case there is only one best system. Having an adaptive pre-screening phase identifying a good time to restart allows us to achieve very good performance in a typical case, and reasonably good performance in all other cases.

How big are the benefits of pre-screening in a typical case? To answer this question let us first estimate the maximal possible savings due to restarting.

In the following analysis we make several simplifying assumptions. First, we assume that the estimate of the residual variance $\hat{\sigma}_i^2$ of system i is always approximately equal to the true residual variance σ_i^2. Second, we ignore the effect of the number of degrees of freedom on the sample sizes for estimation. Third, we assume that the effort required for screening out an inferior system is always the same, whether in Phase I, Phase II, or in an alternative procedure without pre-screening and restarting (such as the multi-stage procedure with early stopping in Chapter 3).

The total cost E of a simulation without pre-screening is the sum of the cost E_s of screening out inferior systems and the cost E_e of estimation of the surviving systems: $E = E_s + E_e$.

The total cost $\tilde{E}$ of a simulation with pre-screening is the sum of the pre-screening cost $\tilde{E}_p$, the cost $\tilde{E}_s$ of screening out inferior systems in Phase II, and the estimation cost $\tilde{E}_e$ of the surviving systems: $\tilde{E} = \tilde{E}_p + \tilde{E}_s + \tilde{E}_e$.

Under our assumptions, the sample size N_i in Equation (4.4) is approximately equal to $c^2\sigma_i^2$. Without pre-screening, the constant c in Equation (4.3) is approximately equal to c_k defined in Equation (4.1), where k is the initial number of systems. With pre-screening, c is approximately c_K, where K is the number of systems remaining after pre-screening. The smaller K, the bigger the benefit of pre-screening, because smaller c_K leads to smaller sample sizes for estimation.

We will assume that whether we simulate with pre-screening or not, the set I of the surviving systems is the same. This is generally so when pre-screening is stopped before the sample sizes for some systems exceed the sample sizes required for estimation, which is exactly the case when pre-screening could be beneficial.

A simulation without pre-screening costs $E = E_s + c_k^2 \sum_{i \in I} \sigma_i^2$, and a simulation with pre-screening costs $\tilde{E} = \tilde{E}_p + \tilde{E}_s + c_K^2 \sum_{i \in I} \sigma_i^2$. The latter is minimized when c_K^2 is as small as possible, which occurs when $K = 1$, i.e. there is only one system left after pre-screening. Also, under the assumptions we use in this section, the screening cost E_s is less than the total of the pre-screening and screening costs $\tilde{E}_p + \tilde{E}_s$, so the maximal efficiency improvement $E/\tilde{E}$ is achieved when the pre-screening and screening costs are negligible compared to estimation costs. This is a typical case in practice: pre-screening and screening are very fast compared to estimation, and they eliminate all but one system. Under

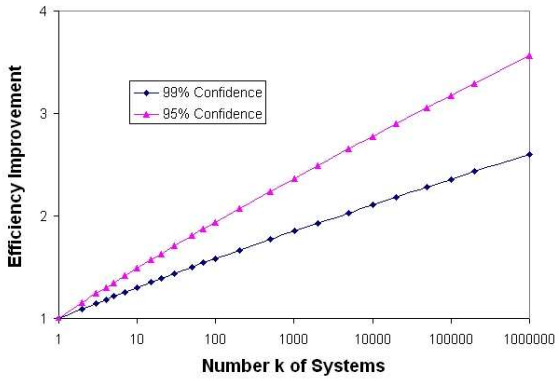

Figure 4.1. Maximal Efficiency Improvement Due to Restarting with $\alpha_a = 0.8\alpha$ and $\alpha_b = 0.2\alpha$

our assumptions, and if pre-screening and screening costs are negligible, the efficiency improvement due to restarting (i.e. due to having a pre-screening phase) is

$$\frac{E}{\tilde{E}} \approx \frac{c_k^2 \sum_{i \in I} \sigma_i^2}{c_K^2 \sum_{i \in I} \sigma_i^2} = \frac{c_k^2}{c_K^2} \leq \frac{c_k^2}{c_1^2}.$$

Figure 4.1 shows the maximal efficiency improvement c_k^2/c_1^2 as a function of the initial number of systems k. When the number of systems k is between 20 and 1000, the savings in a typical case are 40-80% at $1 - \alpha = 99\%$ confidence and 60-140% at $1 - \alpha = 95\%$ confidence.

Recall that the transition rule given by Inequality (4.2) chooses to restart when the cost of doing one more stage of pre-screening is greater than the approximate maximal savings due to continuation, computed under the assumption that after additional pre-screening there will be only one system left and it will have the largest variance. A typical case indeed has one clear best system, so the effort required for screening out

inferior systems is relatively small, the approximate maximal savings are relatively large, and pre-screening makes I a singleton.

How efficient is this transition rule in other situations? Let us consider a configuration when there are several systems which are tied for the best, while other systems are relatively easy to screen out. In this case we might worry that the cost of pre-screening could get too high before the adaptive procedure proceeds to Phase II. Is our transition rule still efficient?

Because now we are concerned that pre-screening may be too expensive, we assume that pre-screening lasts a long time and eliminates all inferior systems: the set $I(M)$ of systems used in Phase II equals I, the set of systems that survive screening and reach their required sample sizes, and the Phase II cost of screening $\tilde{E}_s = 0$. Again we assume that I is the same whether we use pre-screening or not: here we assume it contains only the systems that are tied. We now show how the transition rule in Inequality (4.2) provides a bound on $\tilde{E}_p - E_s$, the excess cost of pre-screening in the adaptive procedure over the cost of screening in a procedure without restarting. The effort required to screen out inferior systems is similar in either procedure, so $\tilde{E}_p - E_s \approx KN(M-1)$, the number of samples from the $K = |I|$ surviving systems that the adaptive procedure throws out by restarting.

Pre-screening stops after stage $\ell = M - 1$, the first time that the cost $(R-1)|I(\ell+1)|N(\ell)$ of the next stage exceeds $(c^2_{|I(\ell+1)|} - c_1^2)\max_{i \in I(\ell+1)} \hat{\sigma}_i^2(\ell)$. Under our present assumption that the residual variance estimates are approximately correct, this

yields the approximate upper bound

$$\begin{aligned} N(M-2) &\leq \frac{(c^2_{|I(M-1)|} - c_1^2)\max_{i\in I(M-1)}\sigma_i^2}{(R-1)|I(M-1)|} \\ &\leq \frac{(c_K^2 - c_1^2)\max_{i\in I}\sigma_i^2}{(R-1)K} \end{aligned}$$

because $I(M-1)$ contains $I(M) = I$ whose size is K, and c_p^2 defined in Equation (4.1) increases in p at a rate that is less than linear. Thus

$$\begin{aligned} KN(M-1) &\leq KRN(M-2) \\ &\leq \frac{R(c_K^2 - c_1^2)\max_{i\in I}\sigma_i^2}{R-1}. \end{aligned}$$

For $R = 1.5$, $R/(R-1) = 3$, and the relative efficiency improvement is

$$\begin{aligned} \frac{E}{\tilde{E}} &= \frac{E_s + c_k^2\sum_{i\in I}\sigma_i^2}{\tilde{E}_p + \tilde{E}_s + c_K^2\sum_{i\in I}\sigma_i^2} \\ &= \frac{E_s + c_k^2\sum_{i\in I}\sigma_i^2}{E_s + 3(c_K^2 - c_1^2)\max_{i\in I}\sigma_i^2 + c_K^2\sum_{i\in I}\sigma_i^2} \\ &\approx \frac{c_k^2\sum_{i\in I}\sigma_i^2}{3(c_K^2 - c_1^2)\max_{i\in I}\sigma_i^2 + c_K^2\sum_{i\in I}\sigma_i^2} \end{aligned}$$

approximately, if the cost E_s of screening is small. If the variances of the tied systems are approximately equal, this simplifies to

$$\frac{Kc_k^2}{3(c_K^2 - c_1^2) + Kc_K^2}.$$

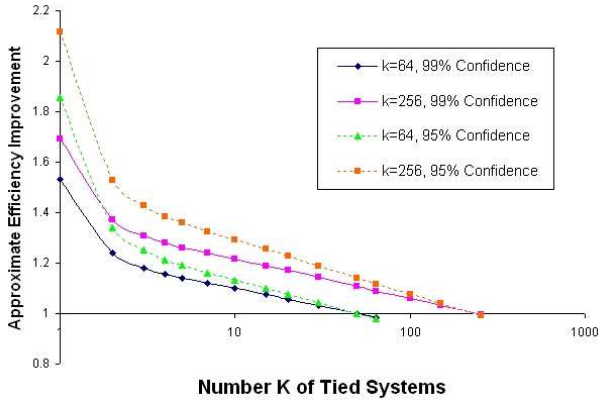

Figure 4.2. Effect of Ties on Approximate Efficiency Improvement Due to Restarting with $\alpha_a = 0.8\alpha$ and $\alpha_b = 0.2\alpha$

For $k = 256$ and $k = 64$ the efficiency improvements as a function of the number K of tied systems are shown in Figure 4.2. A value less than 1 represents a loss of efficiency. We see that even when some systems are tied, restarting with our transition rule can still produce substantial benefits. Even when all the systems are tied, the loss of efficiency is very slight.

The transition rule we presented is heuristic and is one of many similar rules that all work well. This rule is advantageous because of its simplicity and because it allows us to reap most of the benefits of restarting, without causing significant inefficiencies when restarting could be harmful. More efficient transition rules could be designed which take into account not only the sample variances of the systems, but also their sample means. However, such rules are complicated, and in most cases provide either small or no savings. Because the benefits seem insufficient to justify the additional complexity, we do not consider this approach here.

4.2. Performance of the Adaptive Multi-stage Procedure

In this section we continue to use the basket put and options portfolio examples to illustrate our procedure.

To test the adaptiveness of the procedure, in addition to the ordinary configuration with one best system, we also consider configurations "2 best" (obtained by adding a duplicate of the best system), "4 best" (by adding 3 duplicates), and "16 best" (by adding 15 duplicates), so that configuration "2 best" in the basket put example has $64 + 1 = 65$ systems in total, while configuration "16 best" has $64 + 15 = 79$ systems. This is not the same as in Figure 4.2, where the total number k of systems remains constant while the number K that are tied varies.

We split the $1 - \alpha = 1\%$ allowable error into components $\alpha_a = 0.8\%$ for the lower confidence limit and $\alpha_b = 0.2\%$ for the upper confidence limit. The error allocated to screening is $\alpha_I = 0.04\%$, and when using control variates, $\alpha_C = 0.002\%$ is allocated to controlling them. We choose initial sample size n_0 and the maximal number m of Phase I stages to be 30, and the growth factor R to be 1.5. We use CRN in all examples.

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, the largest mean.

We report efficiency as a speed improvement relative to the standard procedure, a modification of the two-stage procedure of Chen and Dudewicz (1976), as explained in

Table 4.1. Efficiency Relative to the Standard Procedure at 99% Confidence

Config.	Example					
	Options Portfolio			Basket Put		
	Precision					
	0.3%	1%	5%	0.3%	1%	5%
1 best	**252**	**244**	**154**	**208**	**158**	**22**
2 best	104	98	81	85	76	19
4 best	51	48	43	40	38	15
16 best	12	12	12	11	10	6.7

Chapter 3. That is, we report the ratio of the average number of samples required by the standard procedure to the average number of samples required by the adaptive multi-stage procedure. The results are summarized in Table 4.1. Recall that efficiency improvement can be larger than the number of systems k, which is 64 for the ordinary configuration of the basket put, and 256 for that of the options portfolio. The reason is that the improvement depends not only on k, but also on the size of the best system's standard deviation relative to the standard deviations of other systems.

Table 4.2 shows how much work the procedure does relative to the work required by the "clairvoyant" procedure, which knows in advance which systems are tied for the best, and applies the standard procedure to only these systems in isolation. That is, the clairvoyant procedure screens out all inferior systems by guessing right with no work.

Like the multi-stage procedure with restarting analyzed in Chapter 3, the adaptive procedure is less than 10% more expensive than estimating a single mean in the "1 best" configuration when a precise estimate is required. If there are ties the procedure first tries to break them, but when this becomes too expensive, proceeds to estimation: this is

Table 4.2. Sample Size Relative to the Clairvoyant Procedure at 99% Confidence

Config.	Example					
	Options Portfolio			Basket Put		
	Precision					
	0.3%	1%	5%	0.3%	1%	5%
1 best	**1.0**	**1.1**	**1.7**	**1.1**	**1.4**	**10**
2 best	1.2	1.2	1.5	1.2	1.3	5.4
4 best	1.1	1.2	1.3	1.2	1.2	3.2
16 best	1.1	1.1	1.1	1.1	1.1	1.7

its advantage over the multi-stage procedure with restarting. Table 4.2 demonstrates the robustness of the adaptive procedure's performance to configuration.

As we see from the last column of Table 4.2, in the configuration with no ties at 5% precision the adaptive procedure looks relatively inefficient compared to the clairvoyant procedure (10 times slower), but adding ties can make the adaptive procedure look more favorable. This is because 5% is a low precision, so the final sample size is not very large relative to the sample size required for screening. At 5% precision the clairvoyant procedure has a big advantage in screening perfectly for free.

Table 4.3 shows the efficiency improvement of the adaptive procedure relative to the most efficient procedure of Chapter 3: the multi-stage procedure with restarting. (In all cases reported in Table 4.3, the multi-stage procedure with early stopping was somewhat more expensive than the multi-stage procedure with restarting.) In some cases, the efficiency is slightly less than 1, i.e. the adaptive procedure required slightly more samples than the multi-stage procedure with restarting: the adaptive procedure does not always pick the best possible time to restart, but it picks a good time.

Table 4.3. Efficiency Relative to the Multi-Stage Procedure with Restarting at 99% Confidence

Configuration and Precision		Options Portfolio Number of screening stages m						Basket Put Number of screening stages m					
		5	10	15	20	25	30	5	10	15	20	25	30
1 best	0.3%	1.0	1.0	1.0	1.0	1.0	1.0	41	2.8	1.0	1.0	1.0	1.0
	1%	1.0	1.0	1.0	1.0	1.0	1.0	30	2.3	1.0	1.0	1.0	1.0
	5%	1.0	1.0	1.0	1.0	1.0	1.0	4.7	1.0	0.9	0.9	1.0	1.0
2 best	0.3%	0.9	0.9	0.9	0.9	1.0	1.8	17	1.6	0.9	1.0	1.6	6.1
	1%	0.9	0.9	0.9	1.0	2.0	10	16	1.5	1.0	1.8	7.8	54
	5%	0.9	0.9	1.2	4.0	25	205	4.6	1.0	1.6	6.5	44	328
4 best	0.3%	0.9	0.9	0.9	0.9	1.0	1.8	8.3	1.3	0.9	1.0	1.5	5.8
	1%	0.9	0.9	0.9	1.0	2.1	10	7.8	1.2	1.0	1.8	7.7	53
	5%	0.9	0.9	1.2	4.1	27	197	3.4	1.1	2.0	10	68	509
16 best	0.3%	0.9	0.9	0.9	0.9	1.0	1.7	3.1	1.0	1.0	1.0	1.6	5.7
	1%	0.9	0.9	1.0	1.1	2.0	9.2	3.0	1.1	1.1	1.8	7.7	52
	5%	0.9	1.0	1.2	4.3	27	201	2.1	1.1	2.7	15	110	833

The efficiency of both of the procedures depends heavily on the actual configuration of the means and the total screening budget of $n_0 R^{m-1}$ observations per system. We tested these procedures with $n_0 = 30$ and $R = 1.5$ while varying the maximal number of stages available for screening from 5 to 30, so that the total budget available for screening varied from 152 to $3,835,022$ observations per system. We set $R = 1.5$, not $R = 2$ as in Chapter 3, as this choice of the growth factor makes all procedures more efficient when there are ties.

The results in Table 4.3 illustrate the danger for our previous multistage procedures of choosing the budget for screening either too small or too large. What constitutes too small or too large depends on the actual configuration, whereas the adaptive procedure works well in all of them.

4.2.1. Similar Systems Nearly Tied for the Best

When simulating a coherent risk measure with common random numbers, several highly correlated systems may have nearly the largest mean. Such situations can occur when one or several factors that are usually important in computation of a risk measure turn out to be insignificant in a particular instance, or when parameters differ only slightly for some systems. For example, an equity derivative may have very similar values in generalized scenarios that differ only in interest rates. One might worry that simulation in this case is expensive and relatively inefficient, similar to what we see in Table 4.1.

However, even if the variances of the systems are large, the variances of the differences of the means of such systems will tend to be small. Unless some systems are identical, which is easy to recognize when simulating with common random numbers and in which case the duplicates should be taken out, small variances of the differences allow even very small differences in performance to be quickly detected, and even slightly inferior systems will be screened out relatively quickly.

For example, in the case of the basket put, the best system is the one that has pairwise correlation of 0.75 between the assets. Table 4.4 shows the effect of adding a system that has pairwise correlation of 0.74 between assets (configuration "2 similar"), adding 3 systems that have two out of three pairwise correlations of 0.74 and one pairwise correlation of 0.75 ("4 similar"), and adding 15 similar systems that have pairwise correlations of 0.75, 0.74 and 0.73 in various combinations ("16 similar"). In the case of the options portfolio, the best system (scenario) is the one that has the first and the fourth factors "up", while

the other two factors are unrestricted. We can add a similar system by assigning one of the "up" events a probability of $9/(10\sqrt{20})$ (in place of $1/\sqrt{20}$ in the best system) and the other a probability of $10/(9\sqrt{20})$. (configuration "2 similar"). In configuration "4 similar" we add two more systems by assigning one of the "up" events a probability of $99/(100\sqrt{20})$ and $98/(100\sqrt{20})$, while in configuration "16 similar" we add 12 more similar systems of this form.

From Table 4.4 we see that the increase in the average sample size due to adding similar systems is usually small. It is also not very sensitive to the similarity parameter, such as the pairwise correlation in the basket put example: in configuration "2 similar" it stays roughly the same whether we use correlation of 0.73 or 0.7499. Even though the two systems have almost exactly the same mean, the variance of the difference is so small that it is easily detected with common random numbers. The correlation between the best system and the similar system that we have added in configuration "2 similar" is 99.99% in the basket put example and 99.83% in the options portfolio example. Adding more such systems does not increase the sample size by much, as correlation is so high that the procedure will quickly screen out systems with smaller means. This increase is mostly due to the larger number of systems that need to be screened out, while the total sample size per system stays roughly the same.

This allows us to conclude that efficiency loss due to closeness of the best means should not in general be significant in financial applications, and that in most cases we will have a clear best.

Table 4.4. Increase in Average Sample Size Due to Adding Systems Similar to the Best at 99% Confidence

Config.	Example					
	Options Portfolio			Basket Put		
	Precision					
	0.3%	1%	5%	0.3%	1%	5%
2 similar	$< 1\%$	$< 1\%$	$< 1\%$	$< 1\%$	$< 1\%$	1%
4 similar	$< 1\%$	$< 1\%$	1%	$< 1\%$	2%	7%
16 similar	$< 1\%$	$< 1\%$	7%	1%	7%	28%

CHAPTER 5

Robustness of the Adaptive Procedure

5.1. Robustness to Non-normality

Under normal-theory assumptions, our procedures are exact, i.e. they deliver at least the nominal coverage probability. Although these assumptions are reasonable in many situations, they are usually not precisely correct. Our screening procedures use sample averages when the sample sizes are still small, and since the sample averages' distributions might be very far from normal, one might worry that screening errors might occur much more often than if distributions were normal.

It is comforting to know that the screening procedures are protected by the use of very conservative probability inequalities (such as the Bonferroni inequality) in their derivation. Error is allocated to pairwise comparisons between all systems during maximal possible number of stages, but many of these comparisons are never performed. Because of this, we can expect screening to be very robust to non-normality. In fact, in most of our experiments, all of which included 5,000 independent replications, screening errors never occurred.

Our estimation procedure will typically require large sample sizes. As we become more demanding, requiring a smaller confidence interval width or higher confidence, the final

Table 5.1. Effect of Strike Price and Initial Sample Size n_0 on Error Rates at 90% Confidence and 5% Precision in Basket Put Example

Strike Price (zero payoff probability)	n_0	Error Rate Upper (5% nominal)	Lower (5% nominal)	Screening (1% nominal)
	5	16%	4%	0%
$K = 85$	7	7%	2%	0%
($\approx 71\%$)	10	5%	1%	0%
	30	7%	6%	$\ll 1\%$
$K = 65$	50	5%	5%	0%
($\approx 92\%$)	100	4%	5%	0%
	30	42%	7%	3%
$K = 55$	100	7%	5%	0%
($\approx 98\%$)	300	4%	5%	0%

sample size becomes larger, making normality of mean estimators more plausible. For this reason moderate non-normality does not seem to be a problem for the final estimator.

However, if non-normality is extreme and the initial sample size is not adequate, the sample sizes after Phase I might be too small and the estimates of the variances which are used to compute final sample sizes could have a distribution that is far from (scaled) χ^2.

Let us consider the basket put example (see Table 5.1). In the ordinary configuration the strike price is 85 and the probability of a zero payoff is approximately 71%. If the probability of a zero payoff is 98% and n_0 is smaller than 200-300, estimates of the variances are so poor that coverage is inadequate. When the probability of a zero payoff is 92%, this can happen if n_0 is smaller than 50-100. When non-normality is not so extreme, such as in the case when the probability of a zero payoff is 90% or less, coverage is adequate as long as n_0 is larger than 10-20.

Table 5.2. Effect of Initial Sample Size n_0 on Error Rates at 90% Confidence and 5% Precision in Options Portfolio Example

n_0	Error Rate		
	Upper (5% nominal)	Lower (5% nominal)	Screening (1% nominal)
5	4%	5%	0%
10	4%	5%	0%
30	4%	5%	0%

In the options portfolio example non-normality is not very significant, so the coverage is adequate even when n_0 is very small: see Table 5.2.

The coverage is also adequate when distributions are heavy-tailed. For example, if in the basket put example logarithmic returns are not normal, but rather have the t distribution with 3 degrees of freedom, the coverage is adequate: see Table 5.3.

For our experiments in this section we chose relatively low 5% precision and 90% confidence. Because in this case the total sample sizes are smaller and therefore the sample averages are less normal, this should represent the hardest test for our procedure.

5.2. Empirical Analysis of Rare Errors

In this section we analyze the event of probability at most $1 - \alpha_a - \alpha_b$, in which the confidence interval does not contain the true value. Because screening is so conservative

Table 5.3. Error Rates with Log-t Returns in Basket Put Example at 90% Confidence and 5% Precision (n_0=30)

Error Rate		
Upper (5% nominal)	Lower (5% nominal)	Screening (1% nominal)
4%	5%	0%

and screening errors are so extremely rare, the error event consists primarily of estimation errors.

In Table 5.4 we present the relative root mean squared distances from the true largest mean to the nearest confidence limit: to the upper limit when the true value lies above the confidence interval and to the lower limit when the true value lies below the confidence interval, as a percentage of its width. These are conditional on the error event, i.e. they are distances given that the true value is above or below the confidence interval:

$$\frac{1}{L}\sqrt{\mathrm{E}\left[\left(\mu_k - (\max_{i\in I}\widehat{\mu}_i + b)\right)^2 \mid \mu_k > \max_{i\in I}\widehat{\mu}_i + b\right]}$$

for the upper distance, and

$$\frac{1}{L}\sqrt{\mathrm{E}\left[\left((\max_{i\in I}\widehat{\mu}_i - a) - \mu_k\right)^2 \mid \mu_k < \max_{i\in I}\widehat{\mu}_i - a\right]}$$

for the lower distance. If we were estimating the mean of just one system in isolation which has the same mean and variance as the best system and which is normally distributed, we would have relative root mean squared distance of approximately 17% for both the upper and the lower confidence limits at 90% confidence:

$$0.17 \approx \frac{1}{2z_{.95}}\sqrt{\int_{z_{.95}}^{\infty}(x - z_{.95})^2\frac{\phi(x)}{0.05}dx},$$

Table 5.4. Root Mean Squared Distance from True Value to Confidence Interval As Percentage of Its Width and Error Rates at 90% Confidence

	Options Portfolio Precision			Basket Put Precision		
	0.3%	1%	5%	0.3%	1%	5%
$n_0 = 30$						
Upper Distance	17%	17%	18%	16%	16%	13%
Lower Distance	18%	18%	18%	18%	16%	14%
Upper Error (5% nominal)	4%	4%	4%	4%	4%	5%
Lower Error (5% nominal)	5%	5%	5%	5%	5%	1%
$n_0 = 10$						
Upper Distance	17%	16%	17%	900%	92%	91%
Lower Distance	17%	17%	17%	55%	64%	21%
Upper Error (5% nominal)	4%	4%	4%	4%	4%	6%
Lower Error (5% nominal)	5%	5%	5%	5%	5%	1%

where ϕ is the pdf and $z_{.95} = \Phi^{-1}(0.95)$ is the 95%-quantile of the standard normal distribution. Table 5.4 shows that when non-normality of sample averages is not extreme the errors of the adaptive procedure on average are no more severe than the errors that happen when estimating a mean of a normal population.

However, when non-normality is extreme and n_0 has not been chosen adequately large, the estimation errors can be much more severe. For example, when using $n_0 = 10$ in the case of the basket put, we found that the coverage was adequate, but the root mean square distance from the upper confidence limit was approximately equal to the confidence interval width when using 1% precision, and it was about nine times that width when using 0.3% precision. (Recall that the confidence interval width is proportional to the precision.) Because non-coverage is a rare event, these large root mean squared distance estimates are not very precise, even though we used more than 5,000 replications to

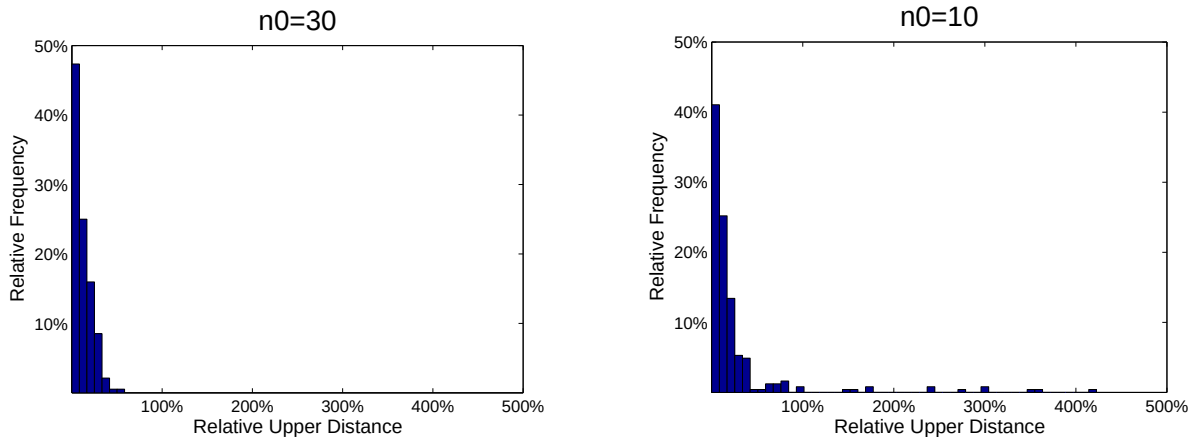

Figure 5.1. Distances from upper confidence limit at 90% confidence and 1% precision as percentages of confidence interval width, for $n_0 = 30$ and $n_0 = 10$.

estimate them. This indicates that when non-normality is extreme, the procedure might significantly under- or overestimate the risk measure: see Figure 5.1, representing the non-coverage events in a representative batch of 5,000 replications.

5.3. Conclusions

Unless non-normality is extreme and n_0 is too small, the procedure is very robust: the coverage is adequate, and even when the confidence interval does not contain the true value, the errors are usually not severe. In extreme cases we have to make sure that n_0 is large enough to get reliable variance estimates. Generally $n_0 = 30$ should be sufficient, but in some cases a preliminary assessment of normality of sample averages might be necessary in order to pick an appropriate initial sample size. The problem could also be fixed by importance sampling.

References

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

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

Boesel, J., Nelson, B. L., Kim, S., 2003. Using ranking and selection to 'clean up' after simulation optimization. Operations Research 51, 814–825.

Chen, H. J., Dudewicz, E. J., 1976. Procedures for fixed-width interval estimation of the largest normal mean. Journal of the American Statistical Association 71, 752–756.

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

Delbaen, F., 2002. Coherent risk measures on general probability spaces. In: Sandmann, K., Schönbucher, P. J. (Eds.), Advances in Finance and Stochastics: Essays in Honour of Dieter Sondermann. Springer-Verlag, New York, pp. 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.

Hochberg, Y., Tamhane, A. C., 1987. Multiple Comparison Procedures. John Wiley & Sons, New York.

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

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

Nelson, B. L., 1990. Control variate remedies. Operations Research 38, 974–992.

Nelson, B. L., Staum, J., 2006. Control variates for screening, selection, and estimation of the best. ACM Transactions on Modeling and Computer Simulation 16 (1), 1–24.

Nelson, B. L., Swann, J., Goldsman, D., Song, W., 2001a. Online companion for "Simple procedures for selecting the best simulated system when the number of alternatives is large". Available at `http://or.pubs.informs.org/Media/Nelson.pdf`.

Nelson, B. L., Swann, J., Goldsman, D., Song, W., 2001b. Simple procedures for selecting the best simulated system when the number of alternatives is large. Operations Research 49, 950–963.

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

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

Wilson, J. R., 2001. A multiplicative decomposition property of the screening-and-selection procedures of Nelson et al. Operations Research 49, 964–966.

APPENDIX A

Algorithms

In this appendix we specify the algorithms used in experiments. The algorithms are implementations of the procedures developed in Chapters 3 and 4. All algorithms are stated for the case where at most q control variates are used for any system, but this includes the case $q = 0$ where control variates are not used.

The algorithms are constructed for clarity rather than efficiency. They do not address computational issues such as how to update sample averages and variances, or the order in which to do the screening comparisons so as to reduce the number that actually have to be made.

A.1. The Standard Algorithm

This is a two-stage algorithm without screening. It is based on a procedure of Chen and Dudewicz (1976), but with ordinary sample means instead of generalized sample means, allowing for user-specified unequal error bounds associated with the lower and upper confidence limits, and using control variates.

(1) USER INPUT:

The user specifies the fixed confidence interval width $L > 0$ and the lower and upper error bounds α_a and α_b in $(0, 1/2)$.

(2) ALGORITHM PARAMETERS:

Choose the number of stage-0 replications $n_0 > q + 2$ and $\alpha_C < \min\{\alpha_a/k, \alpha_b\}$, the error component devoted to control variates.

(3) STAGE 0 SIMULATION:

Simulate (X_{ij}, C_{ij}) for all $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_0$.

(4) COMPUTE FINAL SAMPLE SIZES:

Set $\alpha_a'' \leftarrow \alpha_a/k - \alpha_C$, $\alpha_b'' \leftarrow \alpha_b - \alpha_C$, and the scaling constant

$$c \leftarrow \frac{1}{L}\left(t_{n_0-q-1,1-\alpha_a''} + t_{n_0-q-1,1-\alpha_b''}\right).$$

For each $i = 1, 2, \ldots, k$, compute the residual variance $\hat{\tau}_i^2$ of regressing $X_{i1}, X_{i2}, \ldots, X_{in_0}$ on $C_{i1}, C_{i2}, \ldots, C_{in_0}$, according to Appendix D, and from it the final sample size

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

(5) STAGE 1 SIMULATION:

Simulate (X_{ij}, C_{ij}) for all $i = 1, 2, \ldots, k$ and $j = n_0 + 1, n_0 + 2, \ldots, N_i$.

(6) COMPUTE CONFIDENCE INTERVAL:

For each $i = 1, \ldots, k$, compute the estimate $\widehat{\mu}_i$ from the regression of $X_{i1}, X_{i2}, \ldots, X_{iN_i}$ on $C_{i1}, C_{i2}, \ldots, C_{iN_i}$, according to Appendix D. Set

$$a \leftarrow \frac{1}{c}t_{n_0-q-1,1-\alpha_a''} \quad \text{and} \quad b \leftarrow \frac{1}{c}t_{n_0-q-1,1-\alpha_b''},$$

and the confidence interval is $[\max_{i=1,\ldots,k} \widehat{\mu}_i - a, \max_{i=1,\ldots,k} \widehat{\mu}_i + b]$.

A.2. A Two-Stage Algorithm with Screening

(1) USER INPUT:

The user specifies the fixed confidence interval width $L > 0$ and the lower and upper error bounds α_a and α_b in $(0, 1/2)$.

(2) ALGORITHM PARAMETERS:

Choose the number of stage-0 replications $n_0 > q + 2$, the error component $\alpha_I < \alpha_b$ devoted to screening, and $\alpha_C < \min\{\alpha_a/k, \alpha_b\}$, the error component devoted to control variates.

(3) STAGE 0 SIMULATION:

Simulate (X_{ij}, C_{ij}) for all $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_0$.

For each $h, i = 1, 2, \ldots, k$ such that $h \neq i$, set

$$\begin{aligned} \bar{D}_{hi} &\leftarrow \frac{1}{n_0}\sum_{j=1}^{n_0}(X_{hj} - X_{ij}), \\ S^2_{hi} &\leftarrow \frac{1}{n_0 - 1}\sum_{j=1}^{n_0}(X_{hj} - X_{ij} - \bar{D}_{hi})^2, \quad \text{and} \\ W_{hi} &\leftarrow t_{n_0-1,1-\alpha_I/(k-1)}\frac{S_{hi}}{\sqrt{n_0}}. \end{aligned}$$

Set $I \leftarrow \left\{h = 1, 2, \ldots, k | \forall i \in I(\ell), \bar{D}_{hi} \geq -W_{hi}\right\}$.

(4) COMPUTE FINAL SAMPLE SIZES:

Set $\alpha_a'' \leftarrow \alpha_a/k - \alpha_C$, $\alpha_b'' \leftarrow \alpha_b - \alpha_I - \alpha_C$, and the scaling constant

$$c \leftarrow \frac{1}{L}\left(t_{n_0-q-1,1-\alpha_a''} + t_{n_0-q-1,1-\alpha_b''}\right),$$

where $q := \max_{i\in I} q_i$ and q_i is the number of control variates in C_i.

For each $i \in I$, compute the residual variance $\hat{\tau}_i^2$ of regressing $X_{i1}, X_{i2}, \ldots, X_{i,n_0}$ on $C_{i1}, C_{i2}, \ldots, C_{i,n_0}$, according to Appendix D, and from it the final sample size

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

(5) FINAL STAGE SIMULATION:

Simulate (X_{ij}, C_{ij}) for all $i \in I$ and $j = n_0+1, n_0+2, \ldots, N_i$.

(6) COMPUTE CONFIDENCE INTERVAL:

For each $i \in I$, compute the estimate $\widehat{\mu}_i$ from the regression of $X_{i1}, X_{i2}, \ldots, X_{iN_i}$ on $C_{i1}, C_{i2}, \ldots, C_{iN_i}$, according to Appendix D. Set

$$a \leftarrow \frac{1}{c}t_{n_0-q-1,1-\alpha_a''} \quad \text{and} \quad b \leftarrow \frac{1}{c}t_{n_0-q-1,1-\alpha_b''},$$

and the confidence interval is $[\max_{i\in I}\widehat{\mu}_i - a, \max_{i\in I}\widehat{\mu}_i + b]$.

A.3. A Multi-Stage Algorithm with Early Stopping

(1) USER INPUT:

The user specifies the fixed confidence interval width $L > 0$ and the lower and upper error bounds α_a and α_b in $(0, 1/2)$.

(2) ALGORITHM PARAMETERS: Choose

(a) the number of stage-0 replications $n_0 = N(0) > q + 2$,

(b) the maximum number m of screening stages,

(c) the number $\ell^* \in \{1, 2, \ldots, m-1\}$ of screening stages at which early stopping is not allowed,

(d) the factor $R > 1$ by which the sample size grows at each screening stage,

(e) the error component $\alpha_I < \alpha_b$ devoted to screening, and

(f) the error component $\alpha_C < \min\{\alpha_a/k, \alpha_b - \alpha_I\}$ devoted to control variates.

(3) INITIALIZATION:

Set $\ell \leftarrow 0$, $I(0) \leftarrow \{1, 2, \ldots, k\}$, and $N(-1) \leftarrow 0$.

(4) SCREENING STAGE SIMULATION:

Simulate (X_{ij}, C_{ij}) for all $i \in I(\ell)$ and $j = N(\ell - 1) + 1, N(\ell - 1) + 2, \ldots, N(\ell)$.

For each $h, i \in I(\ell)$ such that $h \neq i$, set

$$\begin{aligned}
\bar{\bar{D}}_{hi} &\leftarrow \frac{1}{N(\ell)} \sum_{j=1}^{N(\ell)} (X_{hj} - X_{ij}), \\
S_{hi}^2 &\leftarrow \frac{1}{N(\ell)-1} \sum_{j=1}^{N(\ell)} (X_{hj} - X_{ij} - \bar{\bar{D}}_{hi})^2, \quad \text{and} \\
W_{hi} &\leftarrow t_{N(\ell)-1,1-\alpha_I/(m(k-1))} \frac{S_{hi}}{\sqrt{N(\ell)}}.
\end{aligned}$$

Set $I(\ell+1) \leftarrow \left\{ h \in I(\ell) | \forall i \in I(\ell), \bar{\bar{D}}_{hi} \geq -W_{hi} \right\}$.

(5) PROCEED TO NEXT STAGE:

Increment $\ell \leftarrow \ell + 1$.

If $\ell \leq \ell^*$, or if $\ell < m$ and $|I(\ell)| > 1$, set $N(\ell) \leftarrow \lceil n_0 R^\ell \rceil$ and return to Step 4.

Otherwise, set $M \leftarrow \ell$.

(6) COMPUTE FINAL SAMPLE SIZES:

Set $\alpha_a'' \leftarrow \alpha_a/k - \alpha_C$, $\alpha_b'' \leftarrow \alpha_b - \alpha_I - \alpha_C$, and the scaling constant

$$c \leftarrow \frac{1}{L} \left(t_{N(\ell^*)-q-1,1-\alpha_a''} + t_{N(\ell^*)-q-1,1-\alpha_b''} \right).$$

For each $i \in I(M)$, compute the residual variance $\hat{\sigma}_i^2$ of regressing X_{i1}, X_{i2}, ..., $X_{iN(\ell^*)}$ on C_{i1}, C_{i2}, ..., $C_{iN(\ell^*)}$, according to Appendix D, and from it the final sample size

$$N_i \leftarrow \max \left\{ N(M-1), \left\lceil c^2 \hat{\sigma}_i^2 + \chi^2_{q_i,1-\alpha_C} \right\rceil \right\}.$$

(7) FINAL STAGE SIMULATION:

Simulate (X_{ij}, C_{ij}) for all $i \in I(M)$ and $j = N(M-1)+1, \ldots, N_i$.

(8) COMPUTE CONFIDENCE INTERVAL:

For each $i \in I(M)$, compute the estimate $\widehat{\mu}_i$ from the regression of $X_{i1}, X_{i2}, \ldots, X_{iN_i}$ on $C_{i1}, C_{i2}, \ldots, C_{iN_i}$, according to Appendix D. Set

$$a \leftarrow \frac{1}{c} t_{N(\ell^*)-q-1,1-\alpha_a''} \quad \text{and} \quad b \leftarrow \frac{1}{c} t_{N(\ell^*)-q-1,1-\alpha_b''},$$

and the confidence interval is $[\max_{i \in I(M)} \widehat{\mu}_i - a, \max_{i \in I(M)} \widehat{\mu}_i + b]$.

A.4. A Multi-Stage Algorithm with Restarting

(1) USER INPUT:

The user specifies the fixed confidence interval width $L > 0$ and the lower and upper error bounds α_a and α_b in $(0, 1/2)$.

(2) ALGORITHM PARAMETERS: Choose

(a) the number of stage-0 replications $n_0 = N(0) > q + 2$,

(b) the maximum number m of screening stages,

(c) the factor $R > 1$ by which the sample size grows at each screening stage,

(d) the error component $\alpha_I < \alpha_b$ devoted to screening,

(e) the error component $\alpha_C < \min\{\alpha_a/k, \alpha_b - \alpha_I\}$ devoted to control variates, and

(f) the prediction confidence level $0 < \epsilon < 1/2$ for use in choosing the number of replications in the first stage of the restarted procedure.

(3) INITIALIZATION:

Set $\ell \leftarrow 0$, $I(0) \leftarrow \{1, 2, \ldots, k\}$, and $N(-1) \leftarrow 0$.

(4) SCREENING STAGE SIMULATION:

Simulate (X_{ij}, C_{ij}) for all $i \in I(\ell)$ and $j = N(\ell-1)+1, N(\ell-1)+2, \ldots, N(\ell)$.

For each $h, i \in I(\ell)$ such that $h \neq i$, set

$$\begin{aligned} \bar{\bar{D}}_{hi} &\leftarrow \frac{1}{N(\ell)} \sum_{j=1}^{N(\ell)} (X_{hj} - X_{ij}), \\ S_{hi}^2 &\leftarrow \frac{1}{N(\ell)-1} \sum_{j=1}^{N(\ell)} (X_{hj} - X_{ij} - \bar{\bar{D}}_{hi})^2, \quad \text{and} \\ W_{hi} &\leftarrow t_{N(\ell)-1,1-\alpha_I/(m(k-1))} \frac{S_{hi}}{\sqrt{N(\ell)}}. \end{aligned}$$

Set $I(\ell+1) \leftarrow \left\{ h \in I(\ell) | \forall i \in I(\ell), \bar{\bar{D}}_{hi} \geq -W_{hi} \right\}$.

(5) PROCEED TO NEXT STAGE:

Increment $\ell \leftarrow \ell + 1$.

If $\ell < m$ and $|I(\ell)| > 1$, set $N(\ell) \leftarrow \lceil n_0 R^\ell \rceil$ and return to Step 4.

Otherwise, set $M \leftarrow \ell$.

(6) FIRST STAGE OF MEAN ESTIMATION:

Set $\alpha_a'' \leftarrow \alpha_a/|I(M)| - \alpha_C$ and $\alpha_b'' \leftarrow \alpha_b - \alpha_I - \alpha_C$.

For each $i \in I(M)$, compute the residual variance $\hat{\tau}_i^2$ of regressing $X_{i1}, X_{i2}, \ldots, X_{iN(M-1)}$

on $C_{i1}, C_{i2}, \ldots, C_{iN(M-1)}$, according to Appendix D, and set

$$n_i \leftarrow \max\left\{q+3, \left\lceil \left(\frac{\Phi^{-1}(1-\alpha_a'') + \Phi^{-1}(1-\alpha_b'')}{L}\right)^2 \frac{(N(M-1)-1)\hat{\tau}_i^2}{\chi^2_{N(M-1)-1,1-\epsilon}} + \chi^2_{q_i,1-\alpha_C} \right\rceil\right\}.$$

Set $n \leftarrow \min_{i\in I(M)} n_i$ and the scaling constant

$$c \leftarrow \frac{1}{L}\left(t_{n-q-1,1-\alpha_a''} + t_{n-q-1,1-\alpha_b''}\right).$$

Simulate (X_{ij}, C_{ij}) for all $i \in I(M)$ and $j = N(M-1)+1, \ldots, N(M-1)+n_i$. For each $i \in I(M)$, compute the residual variance $\hat{\sigma}_i^2$ of regressing $X_{i,N(M-1)+1}$, $X_{i,N(M-1)+2}$, …, $X_{i,N(M-1)+n_i}$ on $C_{i,N(M-1)+1}$, $C_{i,N(M-1)+2}$, …, $C_{i,N(M-1)+n_i}$, according to Appendix D, and set the final sample size

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

(7) SECOND STAGE OF MEAN ESTIMATION:

Simulate (X_{ij}, C_{ij}) for all $i \in I(M)$ and $j = N(M-1)+n_i+1, \ldots, N(M-1)+N_i$.

(8) COMPUTE CONFIDENCE INTERVAL:

For each $i \in I(M)$, compute the estimate $\widehat{\mu}_i$ from the regression of $X_{i,N(M-1)+1}$, $X_{i,N(M-1)+2}$, …, $X_{i,N(M-1)+N_i}$ on $C_{i,N(M-1)+1}$, $C_{i,N(M-1)+2}$,…, $C_{i,N(M-1)+N_i}$, according to Appendix D. Set

$$a \leftarrow \frac{1}{c}t_{n-q-1,1-\alpha_a''} \quad \text{and} \quad b \leftarrow \frac{1}{c}t_{n-q-1,1-\alpha_b''},$$

and the confidence interval is $[\max_{i\in I(M)} \widehat{\mu}_i - a, \max_{i\in I(M)} \widehat{\mu}_i + b]$.

A.5. An Adaptive Multi-Stage Algorithm

This algorithm implements the adaptive multi-stage procedure presented in Chapter 4.

(1) USER INPUT:

The user specifies the fixed confidence interval width L and the lower and upper error bounds α_a and α_b.

(2) ALGORITHM PARAMETERS: Choose

(a) the number of stage-0 replications $n_0 = N(0)$,

(b) the maximum number m of pre-screening stages,

(c) the factor R by which the sample size grows at each stage,

(d) the error component $\alpha_I < \alpha_b$ devoted to screening,

(e) the error component $\alpha_C < \min\{\alpha_a/k, \alpha_b - \alpha_I\}$ devoted to the dispersion of the control variates' sample average from its mean.

(3) INITIALIZATION:

Set $\ell \leftarrow 0$, $M \leftarrow m$, $P \leftarrow m$, $K \leftarrow k$, $Q \leftarrow 0$, $\hat{I} \leftarrow \{1, \ldots, k\}$, $N(-1) \leftarrow 0$, and $N_i \leftarrow \lceil n_0 R^m \rceil, i \in \hat{I}$.

(4) SIMULATION:

Simulate (X_{ij}, C_{ij}) for $j = N(\ell - 1) + 1, \ldots, \min\{N_i, N(\ell)\}$ for all $i \in \hat{I}$.

Set $F \leftarrow N(\ell) - Q$.

(5) SETTING FINAL SAMPLE SIZES:

If $\ell \neq M$, skip this step.

Set $\alpha_a'' \leftarrow \alpha_a/K - \alpha_C$ and $\alpha_b'' \leftarrow \alpha_b - \alpha_I - \alpha_C$, and set the scaling constant $c \leftarrow \frac{1}{L}\left(t_{F-q-1,1-\alpha_a''} + t_{F-q-1,1-\alpha_b''}\right)$, where $q := \max_{i\in I} q_i$ and q_i is the number of control variates in C_i.

For each $i \in \hat{I}$, compute the residual variance $\hat{\sigma}_i^2$ of regressing $X_{i,N(M-1)+1}, \ldots, X_{i,N(M)}$ on $C_{i,N(M-1)+1}, \ldots, C_{i,N(M)}$, and from it the sample size

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

Set $P \leftarrow \lceil \log_R \max_{i\in\hat{I}}(N_i/N(M)) \rceil$.

(6) UPDATING I AND $\hat{I}$:

If $\ell \leq M$, skip this step.

Set $I \leftarrow I \bigcup \left\{ i \in \hat{I} | N_i < N(\ell) \right\}$ and $\hat{I} \leftarrow \hat{I} \backslash I$.

(7) SCREENING:

For each $h, i \in \hat{I}$ such that $h \neq i$,

- Set $\bar{\bar{D}}_{hi} \leftarrow \sum_{j=Q+1}^{N(\ell)} (X_{hj} - X_{ij})/(N(\ell) - Q)$.
- Set $S_{hi}^2 \leftarrow \sum_{j=Q+1}^{N(\ell)} (X_{hj} - X_{ij} - \bar{\bar{D}}_{hi})^2/(F-1)$.
- Set $W_{hi} \leftarrow t_{F-1,1-\alpha_I/(2(K-1)P)} S_{hi}/\sqrt{N(\ell) - Q}$.

Set $\hat{I} \leftarrow \left\{ i \in \hat{I} | \forall h \in \hat{I}, \bar{\bar{D}}_{hi} \geq -W_{hi} \right\}$.

(8) CHECKING WHETHER TO PROCEED TO PHASE II:

If $\ell \geq M$, skip this step.

For each $i \in \hat{I}$, compute the residual variance $\hat{\sigma}_i^2$ of regressing $X_{i,1}, \ldots, X_{i,N(\ell)}$ on $C_{i,1}, \ldots, C_{i,N(\ell)}$ and define

$$c_p := \frac{1}{L}(\Phi^{-1}(1 - \alpha_a/p - \alpha_C) + \Phi^{-1}(1 - \alpha_b - \alpha_I - \alpha_C)).$$

If $\ell = m - 1$ or

$$|\hat{I}|N(\ell)(R-1) > (c_{|\hat{I}|}^2 - c_1^2) \max_{i \in \hat{I}} \hat{\sigma}_i^2,$$

then set $M \leftarrow \ell + 1$, $Q \leftarrow N(M-1)$, $K \leftarrow |\hat{I}|$, and $I \leftarrow \emptyset$.

(9) CONTINUE OR COMPUTE CONFIDENCE INTERVAL:

If $\hat{I} \neq \emptyset$:

- Increment $\ell \leftarrow \ell + 1$.
- Set $N(\ell) \leftarrow \lceil n_0 R^\ell \rceil$ if $\ell < M$, or $N(\ell) \leftarrow \lceil n_0 R^{\ell-1}(1+R) \rceil$ if $\ell \geq M$.
- Return to step (4).

Otherwise, for each $i \in I$, compute the estimate $\widehat{\mu}_i$ from the regression of $X_{i,Q+1}, \ldots, X_{i,N_i}$ on $C_{i,Q+1}, \ldots, C_{i,N_i}$. Set

$$a \leftarrow \frac{1}{c} t_{F-q-1,1-\alpha_a''} \quad \text{and} \quad b \leftarrow \frac{1}{c} t_{F-q-1,1-\alpha_b''},$$

and the confidence interval is $(\max_{i \in I} \widehat{\mu}_i - a, \max_{i \in I} \widehat{\mu}_i + b)$.

APPENDIX B

Proofs

B.1. Validity of the Basic Procedures with Screening

We present a proof for the two-stage algorithm. A generalization shows that the error bounds (2.2) and (2.3) hold for the multi-stage algorithm too.

B.1.1. Lower Confidence Limit

The basis for bounding

$$p := \Pr\left[\mu_{[k]} \geq \max_{i \in I} \bar{X}_i - a\right] \geq 1 - \alpha$$

is the motivating observation that $\max_{i \in I} \bar{X}_i \leq \max_{i=1,\ldots,k} \bar{X}_i$. Even for a system $j \notin I$, i.e. which has been screened out, $\bar{X}_j$ is defined on the probability space, although we do not simulate it. So we have

$$\begin{aligned} p &\geq \Pr\left[\mu_{[k]} \geq \max_{i=1,\ldots,k} \bar{X}_i - a\right] \\ &= \Pr\left[\forall i = 1, \ldots, k,\ \bar{X}_i \leq \mu_{[k]} + a\right] \\ &\geq \Pr\left[\forall i = 1, \ldots, k,\ \bar{X}_i \leq \mu_i + a\right] \end{aligned}$$

because $\mu_i \leq \mu_{[k]}$. Using independence,

$$p \geq \prod_{i=1}^{k} \Pr\left[\bar{X}_i \leq \mu_i + a\right] = \prod_{i=1}^{k} \Pr\left[\frac{\bar{X}_i - \mu_i}{S_i/\sqrt{N_i}} \leq \frac{a\sqrt{N_i}}{S_i}\right].$$

From (2.4) and (2.5), $a = bF_\nu^{-1}((1-\alpha)^{1/k})/F_\nu^{-1}(1-\beta_1)$, while from (2.6), $b\sqrt{N_i}/S_i \geq F_\nu^{-1}(1-\beta_1)$. Therefore $a\sqrt{N_i}/S_i \geq F_\nu^{-1}((1-\alpha)^{1/k})$ so

$$p \geq \prod_{i=1}^{k} F_\nu\left(F_\nu^{-1}((1-\alpha)^{1/k})\right) = 1 - \alpha.$$

B.1.2. Upper Confidence Limit

The probability of interest is

$$\begin{aligned} q &:= \Pr\left[\mu_{[k]} \leq \max_{i \in I} \bar{X}_i + b\right] \\ &\geq \Pr\left[\mu_{[k]} \leq \bar{X}_{[k]} + b, [k] \in I\right] \\ &= \Pr\left[\mu_{[k]} \leq \bar{X}_{[k]} + b,\right. \\ &\qquad \left.\forall j \neq k\ \bar{X}_{[k]} \geq \bar{X}_{[j]} - W_{[k],[j]}\right]. \end{aligned} \tag{B.1}$$

Define

$$Z_k := \frac{\bar{X}_{[k]} - \mu_{[k]}}{\sigma_{[k]}/\sqrt{N_{[k]}}}$$

and for $j \neq k$,

$$Z_j := \frac{(\bar{X}_{[k]} - \bar{X}_{[j]}) - (\mu_{[k]} - \mu_{[j]})}{\sqrt{(\sigma_{[k]}^2 + \sigma_{[j]}^2)/n_0}}.$$

The probability (B.1) can be rewritten as $\Pr[\cap_{i=1}^{k} E_i]$ where the event E_k is that $-Z_k \leq b\sqrt{N_{[k]}}/\sigma_{[k]}$ and for $j \neq k$, the event E_j is that $-Z_j \leq (W_{[k],[j]}+\mu_{[k]}-\mu_{[j]})/\sqrt{(\sigma^2_{[k]}+\sigma^2_{[j]})/n_0}$.

Now we need to condition on the first-stage sample variances, because they appear in the event E_j for $j \neq k$ through $W_{[k],[j]}$, and also determine the sample sizes $N_{[k]}$, which is present in E_k. Let $\mathcal{F}$ represent the information in $(S_1^2, \ldots, S_k^2)$. The conditional distribution of each Z_i is normal with mean 0. Their joint conditional distribution is such that each $\mathrm{Cov}[Z_i, Z_j|\mathcal{F}] > 0$. By Slepian's inequality (Hochberg and Tamhane 1987, Thm. A2.2.1),

$$\Pr[\cap_{i=1}^{k} E_i|\mathcal{F}] \geq \Pr[E_k|\mathcal{F}] \prod_{j \neq k} \Pr[E_j|\mathcal{F}].$$

Taking expectations,

$$\begin{aligned} \Pr[\cap_{i=1}^{k} E_i] &\geq \mathrm{E}\left[\Pr[E_k|\mathcal{F}] \prod_{j \neq k} \Pr[E_j|\mathcal{F}]\right] \\ &\geq \Pr[E_k] \prod_{j \neq k} \Pr[E_j] \end{aligned}$$

where the second line follows from Kimball's inequality (Hochberg and Tamhane 1987, Thm. A2.2.6).

The appendix of Nelson et al. (2001) shows that the product over $j \neq k$ is greater than or equal to $1 - \beta_0$. This relates to the probability of correct screening:

$$\Pr[[k] \in I] \geq \prod_{j \neq k} \Pr[E_j] \geq 1 - \beta_0.$$

The first factor

$$\begin{aligned}\Pr[E_k] &= \Pr\left[-\frac{\bar{X}_{[k]}-\mu_{[k]}}{\sigma_{[k]}/\sqrt{N_{[k]}}} \le \frac{b\sqrt{N_{[k]}}}{\sigma_{[k]}}\right] \\ &= \Pr\left[\mu_{[k]} \le \bar{X}_{[k]} + b\right] \\ &= \Pr\left[-\frac{\bar{X}_{[k]}-\mu_{[k]}}{S_{[k]}/\sqrt{N_{[k]}}} \le \frac{b\sqrt{N_{[k]}}}{S_{[k]}}\right],\end{aligned}$$

and by (2.6), $b\sqrt{N_{[k]}}/S_{[k]} \ge F_\nu^{-1}(1-\beta_1)$, so this probability is at least $F_\nu\left(F_\nu^{-1}(1-\beta_1)\right) = 1-\beta_1$. This relates to the probability of coverage without screening:

$$\begin{aligned}\Pr\left[\mu_{[k]} \le \max_{i=1,\ldots,k} \bar{X}_i + b\right] &\ge \Pr\left[\mu_{[k]} \le \bar{X}_{[k]} + b\right] \\ &= \Pr[E_k] \ge 1-\beta_1.\end{aligned}$$

Putting all the pieces together, and using $1-\beta = (1-\beta_0)(1-\beta_1)$, we conclude that (2.3) holds.

B.2. Validity of the Procedures with CRN, CV and Dynamic Stopping

The proofs rely on Proposition 3.1.1. We show that Inequalities (3.5) and (3.9) hold.

Proposition B.2.1. *If for each $i = 1, 2, \ldots, k$, the observations $X_{i1}, X_{i2}, \ldots$ are independent and identically distributed (i.i.d.) normal random variables, then the standard procedure (Algorithm A.1) without control variates makes Inequalities (3.1) and (3.2) hold.*

Proof. This procedure has no screening, so $\alpha_I = 0$, $I = \{1, 2, \ldots, k\}$, and Inequality (3.5) holds trivially.

Let G_a and G_b be the cumulative distribution function $F_{t_{n_0-1}}$ of the t distribution with $n_0 - 1$ degrees of freedom. Because the error probability bounds α'_a and α'_b are both in $(0, 1/2)$, while $F_{t_{n_0-1}}(0) = 1/2$ and $\lim_{x\to\infty} F_{t_{n_0-1}}(x) = 1$, $G_a(0) < 1 - \alpha'_a < \lim_{x\to\infty} G_a(x)$ and $G_b(0) < 1 - \alpha'_b < \lim_{x\to\infty} G_b(x)$.

In the absence of control variates, $\widehat{\mu}_i = \sum_{j=1}^{N_i} X_{ij}/N_i$. The distribution of $(\widehat{\mu}_i - \mu_i)/(S_i/\sqrt{N_i})$ is t with $n_0 - 1$ degrees of freedom (Hochberg and Tamhane, 1987, Thm. 2.1). By Equation (14), $cS_i/\sqrt{N_i} \le 1$. Thus, for $x \ge 0$,

$$\Pr\{\widehat{\mu}_i - \mu_i \le x\} \ge \Pr\left\{\widehat{\mu}_i - \mu_i \le \frac{xcS_i}{\sqrt{N_i}}\right\} = \Pr\left\{\frac{\widehat{\mu}_i - \mu_i}{S_i/\sqrt{N_i}} \le xc\right\} = F_{t_{n_0-1}}(xc).$$

Similar reasoning provides the other half of Inequality (3.9). □

When we employ control variates, the terminal sample size in our procedures is of the form

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

However, this formula is a convenient approximation for the exact required sample size

$$\min_{n\ge n_0}\left\{n : \left(\frac{n-q}{q}\right)\left(\frac{n}{c^2\hat{\tau}^2} - 1\right) \ge \mathcal{F}_{1-\alpha_C,q,n-q}\right\},$$

where $\mathcal{F}_{1-\alpha_C,q,n-q}$ is the $1-\alpha_C$ quantile of the F distribution with $(q, n-q)$ degrees of freedom (Nelson and Staum, 2006). Although the proofs that follow refer to algorithms incorporating the approximation, they depend on having the exact required sample size.

Proposition B.2.2. *If for each $i = 1, 2, \ldots, k$, $X_{ij} = \mu_i + (C_{ij} - \xi_i)'\beta_i + \eta_{ij}$, where the residuals $\{\eta_{ij}, j = 1, 2, \ldots\}$ and controls $\{C_{ij}, j = 1, 2, \ldots\}$ are independent sets of i.i.d. normal random variables, β_i is an unknown constant vector, $E[C_{i1}] = \xi_i$, and $E[\eta_{i1}] = 0$, then the standard procedure (Algorithm A.1) and the two-stage procedure with screening (Algorithm A.2) make Inequalities (3.1) and (3.2) hold.*

Proof. Inequality (3.9) follows from Prop. 4 of Nelson and Staum (2006), using $G_a(x) = G_b(x) = F_{t_{n_0-q-1}}(x) - \alpha_C$, where q is the number of controls.

The error probability bounds α'_a and α'_b are both in $(0, 1/2)$, while $\alpha_C < \min\{\alpha'_a, \alpha'_b\}$. From $G_a(0) = G_b(0) = 1/2 - \alpha_C$ and $\lim_{x\to\infty} G_a(x) = \lim_{x\to\infty} G_b(x) = 1 - \alpha_C$, it follows that $G_a(0) < 1 - \alpha'_a < \lim_{x\to\infty} G_a(x)$ and $G_b(0) < 1 - \alpha'_b < \lim_{x\to\infty} G_b(x)$.

If there is no screening (Algorithm A.1), Inequality (3.5) holds trivially. If there is screening (Algorithm A.2), Inequality (3.5) follows from reasoning along the lines of the appendix of Nelson et al. (2001b): first, by construction of I, the probability of correct screening $\Pr\{[k] \in I\} = \Pr\{\forall i = 1, 2, \ldots, k,\ \bar{D}_{[k]i} \geq -W_{[k]i}\}$. Next, define $\sigma^2_{hi} := \mathrm{Var}[X_h - X_i]$ and $Z_i := (\bar{D}_{[k]i} - (\mu_{[k]} - \mu_i))/(\sigma_{[k]i}/\sqrt{n_0})$, which is standard normal.

By symmetry of the standard normal distribution,

$$\begin{aligned}\Pr\{[k] \in I\} &= \Pr\left\{\forall i = 1, \ldots, k,\ Z_i \leq \frac{W_{[k]i} + (\mu_{[k]} - \mu_i)}{\sigma_{[k]i}/\sqrt{n_0}}\right\} \\ &\geq \Pr\left\{\forall i = 1, \ldots, k,\ Z_i \leq t_{n_0-1,1-\alpha_I/(k-1)} \frac{S_{[k]i}}{\sigma_{[k]i}}\right\},\end{aligned}$$

by definition of $W_{[k]i}$ and using $\mu_{[k]} - \mu_i \geq 0$. Applying the Bonferroni inequality, the probability of correct screening is at least

$$1 - \sum_{i=1}^{k} \Pr\left\{\ Z_i > t_{n_0-1,1-\alpha_I/(k-1)} \frac{S_{[k]i}}{\sigma_{[k]i}}\right\}.$$

The term for $i = [k]$ is zero because $Z_{[k]} = 0$, while the other $k - 1$ terms are $\alpha_I/(k-1)$ because Z_i and $(n_0-1)(S_{[k]i}/\sigma_{[k]i})^2$ are independent and their distributions are respectively standard normal and chi-squared with $n_0 - 1$ degrees of freedom. Consequently, $\Pr[[k] \in I] \geq 1 - \alpha_I$. □

Proposition B.2.3. *If for each $i = 1, 2, \ldots, k$, $X_{ij} = \mu_i + (C_{ij} - \xi_i)'\beta_i + \eta_{ij}$, where the residuals $\{\eta_{ij}, j = 1, 2, \ldots\}$ and controls $\{C_{ij}, j = 1, 2, \ldots\}$ are independent sets of i.i.d. normal random variables, β_i is an unknown constant vector, $E[C_{i1}] = \xi_i$, and $E[\eta_{i1}] = 0$, then the multi-stage procedure with early stopping (Algorithm A.3) makes Inequalities (3.1) and (3.2) hold.*

Proof. Inequality (3.5) follows from a screening error decomposition via the Bonferroni inequality:

$$\Pr\{[k] \notin I(m)\} \le \sum_{\ell=0}^{m-1} \sum_{i \neq [k]} \Pr\left\{\bar{X}_{[k]}(\ell) < \bar{X}_i(\ell) - W_{[k]i}(\ell)\right\} \le \sum_{\ell=0}^{m-1} \sum_{i \neq [k]} \frac{\alpha_I}{m(k-1)} = \alpha_I.$$

The univariate inference $\Pr\left\{\bar{X}_{[k]}(\ell) < \bar{X}_i(\ell) - W_{[k]i}(\ell)\right\} \le \alpha_I/(m(k-1))$ is the same as in the proof of Prop. B.2.2 because the sample sizes $N(\ell)$ are constants.

Inequality (3.9) holds with $G_a(x) = G_b(x) = F_{t_{N(\ell^*)-q-1}}(x) - \alpha_C$ by Prop. 4 of Nelson and Staum (2006), which applies because there is a residual variance estimator (called $\hat{\sigma}_i^2$ here and $\hat{\tau}_i^2(n_0)$ there) formed from a regression using an initial sample of a fixed number of observations (called $N(\ell^*)$ here and n_0 there), and the final sample size N_i is set in the same way as a function of the residual variance estimator. □

Proposition B.2.4. *If for each $i = 1, 2, \ldots, k$, $X_{ij} = \mu_i + (C_{ij} - \xi_i)'\beta_i + \eta_{ij}$, where the residuals $\{\eta_{ij}, j = 1, 2, \ldots\}$ and controls $\{C_{ij}, j = 1, 2, \ldots\}$ are independent sets of i.i.d. normal random variables, β_i is an unknown constant vector, $E[C_{i1}] = \xi_i$, and $E[\eta_{i1}] = 0$, then the multi-stage procedure with restarting (Algorithm A.4) makes Inequalities (3.1) and (3.2) hold.*

Proof. Steps 6–8 of Algorithm A.4 are simply the standard algorithm (Algorithm A.1) applied with unequal initial sample sizes n_i and a set $I(M)$ of systems both of which are determined by Steps 1–5 of Algorithm A.4. We can view this as a randomly generated simulation problem, where restarting makes the random variates used in Steps 6–8 independent

of the mechanism in Steps 1–5 that randomly generates the problem. We compensate for the unequal sample sizes by using $n := \min_{i \in I(M)} n_i$ in setting the degrees of freedom while computing the scaling constant c. Decreasing the degrees of freedom increases the final sample size and thus also increases the probability that the confidence interval contains the largest mean $\mu_{[k]}$. Applying Prop. 4 of Nelson and Staum (2006) to the randomly generated problem shows that Inequality (3.9) holds with $G_a(x) = G_b(x) = F_{t_{n-q-1}}(x) - \alpha_C$ for each $i \in I(M)$. Because there is no screening in Steps 6–8, $I = I(M)$ and Prop. 3.1.1 implies

$$\Pr\left\{\max_{i\in I}\mu_i \geq \max_{i\in I}\widehat{\mu}_i - a\right\} \geq 1-\alpha_a \quad \text{and} \quad \Pr\left\{\max_{i\in I}\mu_i \leq \max_{i\in I}\widehat{\mu}_i + b\right\} \geq 1-\alpha_b+\alpha_I.$$

The reason that the upper bound for the probability of a violation of the upper confidence limit is $\alpha_b - \alpha_I$ is that Step 6 of Algorithm A.4 sets $\alpha_b'' \leftarrow \alpha_b - \alpha_I - \alpha_C$ while the corresponding Step 4 of Algorithm A.1 sets $\alpha_b'' \leftarrow \alpha_b - \alpha_C$ because no screening takes place in the standard algorithm.

Consider the lower confidence limit and notice that $\mu_{[k]} := \max_{i=1,2,\ldots,k} \mu_i \geq \max_{i\in I} \mu_i$, whatever the subset $I \subseteq \{1, 2, \ldots, k\}$ generated by Steps 1–5 of Algorithm A.4 may be. Consequently,

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

which verifies Inequality (3.1). Next consider the upper confidence limit and notice that if $[k] \in I$, then $\mu_{[k]} = \max_{i \in I} \mu_i$. Consequently,

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

From the result of Prop. 3.1, we found $\Pr\left\{\max_{i \in I} \mu_i > \max_{i \in I} \widehat{\mu}_i + b\right\} \leq \alpha_b - \alpha_I$. Because Steps 3–5 of Algorithms A.3 and A.4, which perform screening, are the same, the proof of Prop. B.2.3 applies here and shows that Inequality (3.5) holds: $\Pr\{[k] \notin I\} \leq \alpha_I$. The result is $\Pr\left\{\mu_{[k]} \leq \max_{i \in I} \widehat{\mu}_i + b\right\} \geq 1 - \alpha_I - (\alpha_b - \alpha_I) = 1 - \alpha_b$, which verifies Inequality (3.2). □

B.3. Validity of the Adaptive Procedure

Proposition B.3.1. *If for each $i = 1, 2, \ldots, k$, $X_{ij} = \mu_i + (C_{ij} - \xi_i)'\beta_i + \eta_{ij}$, where the residuals $\{\eta_{ij}, j = 1, 2, \ldots\}$ and controls $\{C_{ij}, j = 1, 2, \ldots\}$ are independent sets of i.i.d. normal random variables, β_i is an unknown constant vector, $E[C_{i1}] = \xi_i$, and $E[\eta_{i1}] = 0$, then the adaptive multi-stage procedure makes Inequalities (3.1) and (3.2) hold.*

Proof. While I is the set of systems that survives screening after Phase II and $[k]$ is the best system, let $I(M)$ be the set of systems that survives screening in Phase I, and let

$[k]_M$ be the best system in $I(M)$. We decompose the screening error α_I in the following way: allocate $\alpha_I/2$ to Phase I and $\alpha_I/2$ to Phase II.

Phase I has at most m stages and there are at most k systems during any stage, so there are at most $m(k-1)$ comparisons with system $[k]$ during screening in Phase I. Therefore during Phase I we use screening thresholds

$$W_{hi}(\ell) = \frac{S_{hi}(\ell) t_{N(\ell)-1,1-\alpha_I/(2m(k-1))}}{\sqrt{N(\ell)}}$$

at stage ℓ for differences of sample averages of observations generated during stages 1 to ℓ. Phase II has at most P screening stages and there are at most K systems during any stage, so there are at most $P(K-1)$ comparisons with system $[k]_M$ during screening. Therefore during Phase II we use thresholds

$$W_{hi}(\ell) = \frac{S_{hi}(\ell) t_{N(\ell)-N(M-1)-1,1-\alpha_I/(2P(K-1))}}{\sqrt{N(\ell)-N(M-1)}}$$

at stage ℓ for differences of sample averages generated during stages M to ℓ. By the Bonferroni inequality, $\Pr[[k] \notin I(M)] \leq \alpha_I/2$ and $\Pr[[k]_M \notin I] \leq \alpha_I/2$.

Applying Prop. 4 of Nelson and Staum (2006) to the randomly generated problem of estimating the value of the best system in $I(M)$ shows that Inequality (3.9) holds with $G_a(x) = G_b(x) = F_{t_{N(M)-N(M-1)-q-1}}(x) - \alpha_C$ for each $i \in I(M)$. Proposition 3.1.1 then

implies that

$$\Pr\left\{\max_{i\in I(M)}\mu_i \geq \max_{i\in I}\widehat{\mu}_i - a\right\} \geq 1-\alpha_a \quad \text{and} \quad \Pr\left\{\max_{i\in I(M)}\mu_i \leq \max_{i\in I}\widehat{\mu}_i + b\right\} \geq 1-\alpha_b+\alpha_I/2.$$

Consider the lower confidence limit and notice that $\mu_{[k]} = \max_{i=1,2,\ldots,k}\mu_i \geq \max_{i\in I(M)}\mu_i$, whatever the subset $I(M) \subseteq \{1,2,\ldots,k\}$ generated after Phase I may be. Consequently,

$$\Pr\left\{\mu_{[k]} \geq \max_{i\in I}\widehat{\mu}_i - a\right\} \geq \Pr\left\{\max_{i\in I(M)}\mu_i \geq \max_{i\in I}\widehat{\mu}_i - a\right\} \geq 1-\alpha_a,$$

which verifies Inequality (3.1). Next consider the upper confidence limit and notice that if $[k] \in I(M)$, then $\mu_{[k]} = \max_{i\in I(M)}\mu_i$. Consequently,

$$\begin{aligned}
\Pr\left\{\mu_{[k]} \leq \max_{i\in I}\widehat{\mu}_i + b\right\} &\geq \Pr\left\{[k]\in I(M), \mu_{[k]} \leq \max_{i\in I}\widehat{\mu}_i + b\right\} \\
&= \Pr\left\{[k]\in I(M), \max_{i\in I(M)}\mu_i \leq \max_{i\in I}\widehat{\mu}_i + b\right\} \\
&\geq 1 - \Pr\{[k]\notin I(M)\} - \Pr\left\{\max_{i\in I(M)}\mu_i > \max_{i\in I}\widehat{\mu}_i + b\right\} \\
&\geq 1-\alpha_I/2-(\alpha_b-\alpha_I/2) = 1-\alpha_b,
\end{aligned}$$

which verifies Inequality (3.2).

□

APPENDIX C

Variants

This appendix contains remarks about possible variants of the procedures discussed in the text.

C.1. Common Random Numbers

C.1.1. Grouping

Common random numbers are intended to induce positive correlation between systems, reducing the variances of the differences of their sample means, and thus facilitating screening. However, common random numbers may instead induce negative correlation between some pairs of systems, which inflates the variance of the difference of their sample means. If this were known in advance, it would be possible to divide the systems into groups such that no group contains a pair of systems with negative correlation under common random numbers. Then one would give each group its own set of common random numbers, independent of those belonging to all other groups. This approach ensures that all systems have nonnegative correlation, so that common random numbers cannot hurt screening. Moreover, this approach delivers a multiplicative error decomposition, as explained in Section C.3.1. However, we found that this was not helpful for the examples we considered. To screen out an inferior system i quickly requires that there be *some*

superior system h such that the expectation of the difference $\bar{X}_h - \bar{X}_i$ is large relative to its standard deviation. We found that typically a system has negative correlation only with a few of the superior systems, not all of them, and that the negative correlations are small in magnitude. Consequently, negative correlations have a very small effect on screening. The multiplicative error decomposition discussed in Section C.3.1 also has only a very small effect on simulation efficiency. Thus, grouping systems to avoid negative correlation has only very slight benefits. These benefits are less important than the drawback that some pairs of systems with positive correlation are split between different groups, because one member of the pair has negative correlation with a third system, and thus the benefits of common random numbers for this pair are lost. In conclusion, we recommend not dividing systems into groups that are simulated independently.

C.2. Multi-stage Procedures without CRN

The sample size during screening should be the same for all systems when using CRN. Suppose instead that screening featured comparisons of averages over samples of unequal size, $\sum_{j=1}^{n_h} X_{hj}/n_h$ and $\sum_{j=1}^{n_i} X_{ij}/n_i$, where $n_h < n_i$. The variance of the difference between these averages is $\sigma_i^2/n_i - 2\sigma_{hi}^2/n_i + \sigma_h^2/n_h$, where $\sigma_{hi}^2 = \mathrm{Cov}(X_{hj}, X_{ij})$. Using only n_h replications to form both sample averages, $\mathrm{Var}[\sum_{j=1}^{n_h}(X_{ij} - X_{hj})/n_h] = (\sigma_i^2 - 2\sigma_{hi}^2 + \sigma_h^2)/n_h$. The change in variance due to using the extra replications X_{ij} for $j = n_h + 1, n_h + 2, \ldots, n_i$ is $(1/n_i - 1/n_h)(\sigma_i^2 - 2\sigma_{hi}^2) = (1/n_h - 1/n_i)(2\rho_{hi}\sigma_h - \sigma_i)\sigma_i$, where ρ_{hi} is the correlation that common random numbers induce between X_h and X_i. When

$\rho_{hi} > \sigma_i/(2\sigma_h)$, this change is positive, meaning that the inclusion of extra replications of X_i actually *increases* the variance of the difference used in screening, making screening less effective. Thus, when common random numbers are effective in inducing high correlation, the use of unequal sample sizes during screening is a mistake.

In the absence of common random numbers, it would be possible to allow different systems to have different sample sizes during screening, and to replace sample variances of differences S^2_{hi} with sums of sample variances S^2_h and S^2_i. In Chapter 2 we describe a scheme for choosing different sample sizes during screening. However, the presence of unequal sample sizes in screening complicates matters. The screening threshold

$$W_{hi} = t_{n_0-1,1-\alpha_I/(m(k-1))}\sqrt{\frac{S^2_h}{N_h(\ell)} + \frac{S^2_i}{N_i(\ell)}}$$

in Chapter 2 can only be proved to deliver $\Pr\{[k] \notin I\} \leq 2\alpha_I$: see the appendix of Nelson et al. (2001b). However, $\Pr\{[k] \notin I\} \leq \alpha_I$ holds in limiting cases and held reliably in extensive simulation experiments (Nelson et al., 2001a). This issue does not affect our procedures with common random numbers.

C.3. Error Spending

C.3.1. Multiplicative Decomposition

In Chapter 2, we used a multiplicative decomposition $1 - \alpha_b = (1 - \alpha_I)(1 - \alpha'_b)$. This is frequently possible in settings such as Inequality (3.4); see also Wilson (2001). However,

we found that multiplicative decomposition provided negligible efficiency gains over additive decomposition. Furthermore, in the presence of common random numbers, discussed in Section 4.3, it is easier to establish coverage bounds given an additive decomposition.

In the case of independent sampling of the X_i, or by means of Slepian's inequality (Hochberg and Tamhane, 1987, Thm. A2.2.1) in the case when common random numbers induce nonnegative correlation among all systems ($\text{Corr}[X_i, X_j] \geq 0$ for all i, j), one may use a multiplicative decomposition in Inequality (3.7) instead of an additive decomposition. That is, instead of α_a/k in Inequality (3.8), we would have $1 - (1 - \alpha_a)^{1/k}$. The result is a reduction in the required sample sizes to attain the fixed confidence interval width, but we found that this effect was negligible in practice.

C.3.2. Unequal Allocations

In Inequality (3.7), we could allocate error unequally across systems as long as the individual error probabilities sum to α_a. If we could guess in advance some information about the systems, we might allocate less error to those systems that are more likely to be screened out or have lower variances.

When systems are simulated independently, it is possible to give unequal allocations of error in constructing the various thresholds $W_{hi}(\ell)$. While it would require good advance guesses about the problem's structure to motivate unequal allocations across systems, the m screening stages are different because some come before others, and the earlier ones have higher variances associated with the sample averages. Therefore, it might make sense

to allocate more error to earlier stages so as to screen out systems more quickly at first, but we do not explore this possibility here.

APPENDIX D

Control Variate Estimators

This appendix provides definitions and notation for our use of control variates; it is based on Nelson and Staum (2006). We assume that X_{ij}, the jth output from the simulation of system i, can be represented as

$$X_{ij} = \mu_i + (C_{ij} - \xi_i)'\beta_i + \eta_{ij}.$$

For each system $i = 1, 2, \ldots, k$ and any sample size n, $\{\eta_{ij}, j = 1, 2, \ldots, n\}$ are i.i.d. $\mathcal{N}(0, \tau_i^2)$ random variables. The $q_i \times 1$ vector C_{ij} is called the *control variate*; for fixed i and $j = 1, 2, \ldots, n$ the control variates are also i.i.d., are independent of η_{ij}, and have known expected value ξ_i. The multiplier β_i is a $q_i \times 1$ vector of unknown constants that captures the relationship between the output X_{ij} and the control C_{ij}, while η_{ij} represents that part of the variability in X_{ij} that is not explained by the controls. We define the CV estimator; the development is based on Nelson (1990).

Let

$$X_i(n) = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{in} \end{pmatrix} \quad \text{and} \quad C_i(n) = \begin{pmatrix} C_{i1} \\ C_{i2} \\ \vdots \\ C_{in} \end{pmatrix}$$

be vectors of the output and controls across all n replications from system i. Define the sample mean of the outputs and controls as $\bar{X}_i(n) := \sum_{j=1}^{n} X_{ij}/n$ and $\bar{C}_i(n) := \sum_{j=1}^{n} C_{ij}/n$. In this appendix, for clarity we append "(n)" to represent quantities defined across n replications.

To define the CV point estimator, let

$$L_i'(n) := \left[(C_{i1} - \bar{C}_i(n)), (C_{i2} - \bar{C}_i(n)), \ldots, (C_{in} - \bar{C}_i(n))\right].$$

Then the CV estimator of μ_i is

$$\begin{aligned} \widehat{\mu}_i(n) &= \left[\frac{1}{n}\mathbf{1}_{n\times 1}' - \left(\bar{C}_i(n) - \xi_i\right)' \left(L_i'(n)L_i(n)\right)^{-1} L_i'(n)\right] X_i(n) \\ &= \bar{X}_i(n) - \left(\bar{C}_i(n) - \xi_i\right)' \widehat{\beta}_i, \end{aligned}$$

where $\mathbf{1}_{n\times 1}$ is a column n-vector whose entries all equal one, and $\widehat{\beta}_i$, defined by the equations immediately above, is the usual least-squares regression slope coefficient (Nelson,

1990). Also define

$$\widehat{\tau}_i^2(n) \;:=\; \frac{1}{n-q_i-1}\sum_{j=1}^{n}\left[X_{ij}-\widehat{\mu}_i(n)-(C_{ij}-\xi_i)'\widehat{\beta}_i(n)\right]^2$$

as the residual variance estimator.

It is shown in Nelson and Staum (2006) that if the assumptions made in this appendix hold and C_{ij} has a multivariate normal distribution, then $\widehat{\mu}_i(N_i)-\mu_i$ satisfies Inequality (10) with $G_a(x)=G_b(x)=F_{t_{n_0-q-1}}(x)-\alpha_C$.