Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

A PRELIMINARY STUDY OF OPTIMAL SPLITTING FOR RARE-EVENT SIMULATION

John F. Shortle
Chun-Hung Chen

Department of Systems Engineering and Operations Research
4400 University Dr., MS 4A6
Fairfax, V.A. 22030, U.S.A.

ABSTRACT

Efficiency is a big concern when using simulation to estimate rare-event probabilities, since a huge number of simulation replications may be needed in order to obtain a reasonable estimate of such a probability. Furthermore, when multiple designs must be compared, and each design requires simulation of a rare event, then the total number of samples across all designs can be prohibitively high. This paper presents a new approach to enhance the efficiency for rare-event simulation. Our approach is developed by integrating the notions of level splitting and optimal computing budget allocation. The goal is to determine the optimal numbers of simulation runs across designs and across a number of splitting levels so that the variance of the rare-event estimator is minimized.

1 INTRODUCTION

Simulation is a powerful tool that can be used to analyze a wide variety of systems. In principle, given an accurate model and ample computer time, simulation can provide accurate answers to a number of questions. However, in the context of rare events, a key limitation is the computation time needed to obtain a reasonable estimate of the rare-event probability. For example, consider a rare event that occurs with probability 10^{-9}. If we simulate the system 10^9 times, then we see, on average, one occurrence of this rare event. Even if we can simulate 10,000 runs per second, we need about 1 day just to observe one event. Many more simulations are needed in order to obtain a reasonable confidence interval. Further, when multiple designs must be compared, and each design requires simulation of a rare event, then the total number of simulation replications across all designs can be very high. The objective of this paper is to improve the simulation efficiency to determine the best of multiple designs where rare events are of concern.

There are two main approaches that have been used in the literature to improve the efficiency of rare-event simulations: importance sampling and splitting. The idea of importance sampling (Glynn 1994, Heidelberger 1993) is to change the sample space so that rare events are more probable. The main challenge is that this usually requires specific knowledge about the problem, so solutions tend to be highly sensitive to the assumptions of the model. Another approach is splitting (Glasserman et al. 1999, L'Ecuyer, Demers, and Tuffin 2006). The basic idea of splitting is to create separate copies of the simulation whenever the simulation gets closer to the rare event of interest. Effectively, this multiplies runs that are more likely to reach the rare event and kills runs that are not promising, thus improving the likelihood of observing the rare event. It has been shown that splitting has the potential to significantly reduce the variance for rare-event problems.

In the comparison of multiple designs, the standard Monte Carlo approach is to simply simulate each design for an equal amount of time or an equal number of replications. Then, the simulated metrics are compared to choose the best design among all alternatives. As one might imagine, there are better ways to do this. For example, after an initial simulation period, one may see that some designs are performing poorly, while others are performing well. One can then adjust the computing allocation to simulate the promising designs more frequently and the other designs less frequently, if at all. Chen et al. (1997), Chen et al. (2000), and Chen et al. (2008) presented an Optimal Computing Budget Allocation (OCBA) approach and demonstrated that the intelligent use of the simulated outputs from different alternatives can dramatically improve simulation efficiency. Extensions of the OCBA approach include Lee et al. (2004), who consider multiple objective functions, Trailovic and Pao (2004), who consider the objective of minimizing variance, and Fu et al. (2007), who consider correlated sampling. In addition, Hyden and Schruben (2000), Chick and Inoue (2001), Kim and Nelson (2006), and Branke, Chick, and Schmidt (2007) also demonstrated that the simulation efficiency can be significantly improved by utilizing more simulation information.

978-1-4244-2708-6/08/$25.00

In this paper, we present a new optimal splitting technique for rare-event simulation. The key idea is to integrate the notions of OCBA into splitting methods to optimally allocate the limited computing budget so that the overall efficiency is maximized. For simulation analysis of a single system, we want to determine the optimal numbers of simulation runs among a number of splitting levels so that the variance of the rare-event estimator is minimized. For a comparison of multiple designs, we want to determine the optimal numbers of simulation runs for each splitting level in all designs so that the overall simulation efficiency is maximized.

The paper is organized as follows. In the next section, we introduce the idea of the splitting technique for rare-event simulation. Section 3 formulates the problem of our proposed optimal splitting technique. Some cases are studied in Section 4. Section 5 concludes the paper.

2 LEVEL SPLITTING

Our proposed method is based on a promising variance reduction technique, called multilevel splitting (Garvels 2000). The basic idea of this method is to consider the rare event as the intersection of a nested sequence of events. The probability of the rare event is the product of conditional probabilities, each of which can be estimated more accurately than the rare event itself, for a given simulation effort. This transforms the rare-event problem into a set of much easier problems. It has been shown that the efficiency gain can be orders of magnitude in using splitting techniques versus standard Monte Carlo simulation (L'Ecuyer, Demers, and Tuffin 2006).

Figure 1 shows the basic idea of level splitting. In the figure, the y-axis measures the proximity of the system to the rare event set. The process is assumed to start at state 0, and the rare-event set is defined by a threshold level L. For example, the y-axis could denote the length of a queue, and L could denote the maximum queue size before buffer overflow. The interval $[0,L]$ is partitioned into m stages by choosing levels $0 = L_0 < L_1 < \cdots < L_m = L$. Whenever the simulation crosses a level, it is "split" into separate simulation runs. These runs are independently simulated starting from the splitting point. In this way, more computer time is spent on runs that are closer to the rare event. Once the simulation is complete, runs that hit the rare event are appropriately normalized so that an unbiased estimate is given for the rare-event probability. This is described more precisely in a moment.

There are many different ways of implementing the splitting idea (L'Ecuyer, Demers, and Tuffin 2006). In this paper, we consider one type of level splitting, called *fixed-effort* splitting. Let $X = \{X_t, t \geq 0\}$ be a stochastic process with state space state χ. We assume that the stochastic process is a Markov process. Let $h : \chi \to \mathbb{R}$ be a map of

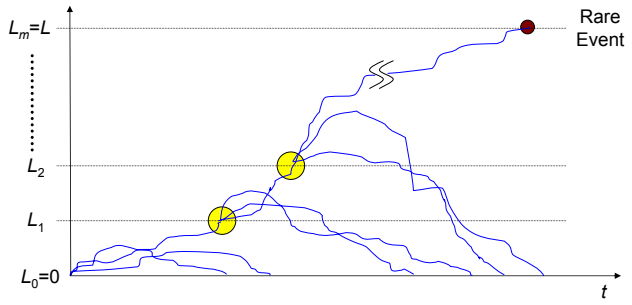

Figure 1: The basic concept of level splitting.

the state space to the "level" of the process. We assume that $h(X_0) = 0$. The rare event R is defined as the set of states whose level is larger than some constant $L > 0$. That is, $R \equiv \{x \in \chi : h(x) \geq L\}$. $h(x)$ is called the *importance function*. Let T_R be the first time the process X enters the rare event set R, and let T_S be the first time the process returns to the starting point (level 0) after leaving it. The probability we wish to estimate is

$$\gamma \equiv \Pr\{T_R < T_S\}.$$

For example, if R denotes a buffer overflow in a queue and X_0 denotes an empty system, then $\{T_R < T_S\}$ is the event that a buffer overflow occurs before the queue returns to an empty state (starting from an empty state).

Let $T_i \equiv \inf\{t > 0 : h(X_t) \geq L_i\}$, for $i = 1,2,\ldots,m$, be the time for the process to first reach level i (starting from level 0). Let $D_i \equiv \{T_i < T_S\}$ be the event that the process reaches level i before returning to level 0. Thus, the rare event probability is

$$\Pr\{T_R < T_S\} = \Pr\{T_m < T_S\} = \Pr\{D_m\}.$$

Let $p_i \equiv \Pr\{D_i | D_{i-1}\}$, for $i = 2,\ldots,m$, be the probability that the process reaches level i (before returning to level 0) given that the process has reached level $i-1$ (before returning to level 0). Also, let $p_1 \equiv \Pr\{D_1\}$. Since $D_m \subset D_{m-1} \subset \cdots \subset D_1$, we have

$$\gamma = \Pr\{D_m\} = \Pr\{D_1\}\Pr\{D_2|D_1\}\cdots\Pr\{D_m|D_{m-1}\} = p_i.$$

Note that p_i is a conditional probability. The idea of multilevel splitting is to estimate each probability p_i separately, by starting a large number of simulations in states at level L_{i-1} conditional on the event D_{i-1}.

To implement the splitting technique, we start N_1 independent simulation runs from the initial state X_0 and simulate each of them until time $\min(T_1, T_S)$. Let Q_1 be the number of those runs that reach level 1 before returning to the starting state. Then $\hat{p}_1 \equiv Q_1/N_1$ is an unbiased estima-

tor of p_1. Upon hitting the threshold L_1, the end states of the Q_1 simulations are collected into a set denoted A_1, which become the starting states for the next stage of simulation.

At stage i (simulation from level L_{i-1} to level L_i) for $i \geq 2$, draw N_i starting states at random, with replacement, from the set A_{i-1}. This is equivalent to taking N_i samples from the empirical distribution of states at level L_{i-1}. With these N_i starting points, each simulation progresses independently until time $\min(T_i, T_S)$. This is called "splitting" of the sample path. Let Q_i be the number of runs that reach level i before the returning to the initial state (level 0). Then $\hat{p}_i \equiv Q_i/N_i$ is an unbiased estimator of p_i, which is a binomial random variable with parameters N_i and p_i. Upon hitting a threshold L_i, the end states of the Q_i simulations are collected into the set A_i for the next stage of simulation. This procedure repeats at every level, for $i = 2, \ldots, m$. After that, an unbiased estimator for the rare-event probability γ is

$$\hat{\gamma} = \hat{p}_1 \hat{p}_2 \cdots \hat{p}_m.$$

L'Ecuyer, Demers, and Tuffin (2006) showed that the variance of the estimation can be dramatically reduced as compared with standard Monte Carlo simulation.

3 THE OPTIMAL SPLITTING PROBLEM

We present a novel idea to reduce the total simulation time in the rare-event design problem. Under this setting, there are k alternate designs and rare-event simulations must be conducted for all k designs. Thus, a large number of simulation replications (or runs) for each design must be conducted to insure that the estimation variance for each design is sufficiently low so that the best design can be correctly identified.

Existing splitting techniques can improve simulation efficiency for each design locally, but do not use the information of relative probability/mean estimations among different designs. Our new approach intends to reduce estimation variance by looking at both the local information of different splitting levels and the overall topology of the design space. In particular, we want to allocate simulation resources to different levels in different alternative designs to maximize the quality of decision making, defined here as the probability of selecting the best design. To do so, we integrate the notion of Optimal Computing Budget Allocation (OCBA).

Unlike the OCBA setting where each simulation replication is a complete Monte Carlo sample (from level L_0 to L_m), in this paper, we apply the splitting technique to conduct the simulation where a simulation replication involves multi-level (or multi-stage) simulations. For notation, let N_{ij} be the number of simulation runs for design i during stage j (i.e., the number of simulation paths starting at level L_{j-1} for design i), and let p_{ij} be the probability of reaching level L_j

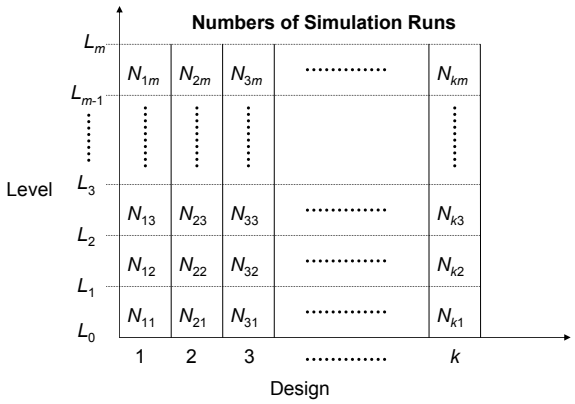

Figure 2: Decision variables in the simulation problem.

(before returning to level 0) conditional on starting from level L_{j-1}, for design i. After some initial simulation runs for stage j in design i are conducted, we can estimate p_{ij} and associated confidence intervals. As N_{ij} increases, the estimation of p_{ij} becomes better. In this setting, we want to intelligently choose N_{ij} for all i and j so that the overall simulation efficiency is maximized. Figure 2 depicts the concept by showing the decision variables N_{ij}.

Let b_{ij} be the one-run average simulation cost for design i during stage j. Then the total simulation cost is approximately

$$\sum_{i=1}^{k} \sum_{j=1}^{m} b_{ij} N_{ij}.$$

Let $\Pr\{\text{CS}\}$ be the probability of correctly selecting the best design based on simulation output. $\Pr\{\text{CS}\}$ increases as more simulation runs are performed. Specifically, we maximize the probability of correct selection $\Pr\{\text{CS}\}$ with a constraint on the total computation time:

$$\max_{N_{ij}} \Pr\{\text{CS}\} \quad \text{such that} \quad \sum_{i=1}^{k} \sum_{j=1}^{m} b_{ij} N_{ij} \leq T. \qquad (1)$$

4 CASE STUDIES

In the special case that $m = 1$, there are no intermediate levels, so no level-splitting techniques are used; this corresponds to a single row in Figure 2. In this case, the problem simplifies to the OCBA problem considered in Chen et al. (2000). The solution has shown the possibility to enhance simulation efficiency by an order of magnitude.

In a similar manner, in the special case that $k = 1$, the problem simplifies to an optimal level-splitting allocation for a single design; this corresponds to a single column in Figure 2. Chen and Shortle (2008) offer an asymptotic optimal solution for (1) of this special case as follows.

The N_{ij} are chosen to satisfy the following constraint

$$N_{11}\sqrt{b_{11}p_{11}} = N_{12}\sqrt{b_{12}p_{12}} = \cdots = N_{1m}\sqrt{b_{1m}p_{1m}}. \quad (2)$$

If we assume that the b_{ij}'s are equal, then this says that more simulation effort is spent on levels with smaller probabilities. That is, as p_{1j} decreases, N_{1j} increases.

The general solution to (1) remains ongoing research. In this section, we consider a specific smaller problem where $k = 2$ and $m = 2$. In other words, there are two designs and two splitting levels. The solution to this problem offers some insights for solutions to the general problem (1). For simplicity, we assume that the simulation cost for each level is approximately the same. That is, the b_{ij}'s are equal and can be ignored in the formulation.

To choose the better of two designs, we need to estimate $\gamma_1 - \gamma_2$, where γ_i is the rare-event probability estimator for design i. To maximize the probability of correct selection $\Pr\{\text{CS}\}$, we minimize

$$\begin{aligned}\text{Var}[\hat{\gamma}_1 - \hat{\gamma}_2] &= \text{Var}[\hat{\gamma}_1] + \text{Var}[\hat{\gamma}_2] \\ &= \text{Var}[\hat{p}_{11}\hat{p}_{12}] + \text{Var}[\hat{p}_{21}\hat{p}_{22}],\end{aligned}$$

where $\hat{p}_{ij}$ denotes the estimator for the conditional probability for design i from level $j-1$ to level j. Thus the problem is

$$\begin{aligned}\min_{N_{11},N_{12},N_{21},N_{22}} \ & \text{Var}[\hat{p}_{11}\hat{p}_{12}] + \text{Var}[\hat{p}_{21}\hat{p}_{22}] \\ \text{s.t.} \quad & N_{11} + N_{12} + N_{21} + N_{22} = T. \end{aligned} \quad (3)$$

Assuming that N_{ij} is a fixed number, then $N_{ij}\hat{p}_{ij}$ is a binomial random variable with parameters (p_{ij}, N_{ij}). Thus

$$\begin{aligned}\text{Var}[\hat{p}_{i1}\hat{p}_{i2}] &= \text{E}[\hat{p}_{i1}^2\hat{p}_{i2}^2] - \text{E}[\hat{p}_{i1}]^2\text{E}[\hat{p}_{i2}]^2 \\ &= \left(p_{i1}^2 + \frac{p_{i1}(1-p_{i1})}{N_{i1}}\right)\left(p_{i2}^2 + \frac{p_{i2}(1-p_{i2})}{N_{i2}}\right) \\ &\qquad - p_{i1}^2 p_{i2}^2 \\ &\approx \frac{p_{i1}p_{i2}}{N_{i1}N_{i2}}\left[p_{i2}N_{i2} + p_{i1}N_{i1} + 1\right],\end{aligned}$$

where the last approximation follows since $1 - p_{ij} \approx 1$. In general, we must have $N_{ij}p_{ij} > 1$ so that there is at least one hit to the next level on average (when simulating from level $j-1$ to j). We further consider an asymptotic condition where the N_{ij}'s are large so that we can write

$$\begin{aligned}\text{Var}[\hat{p}_{i1}\hat{p}_{i2}] &\approx \frac{p_{i1}p_{i2}}{N_{i1}N_{i2}}\left[p_{i2}N_{i2} + p_{i1}N_{i1}\right] \\ &= \gamma_i^2\left(\frac{1}{p_{i1}N_{i1}} + \frac{1}{p_{i2}N_{i2}}\right).\end{aligned} \quad (4)$$

With (4), the optimal splitting problem in (3) can be written as

$$\begin{aligned}\min_{N_{11},N_{12},N_{21},N_{22}} \ & \left(\frac{\gamma_1^2/p_{11}}{N_{11}} + \frac{\gamma_1^2/p_{12}}{N_{12}} + \frac{\gamma_2^2/p_{21}}{N_{21}} + \frac{\gamma_2^2/p_{22}}{N_{22}}\right) \\ \text{s.t.} \quad & N_{11} + N_{12} + N_{21} + N_{22} = T. \end{aligned} \quad (5)$$

By finding the stationary point of the Lagrangian relaxation of (5), we obtain an optimal solution to the problem for the computing budget allocation between designs 1 and 2, and their corresponding stages 1 and 2 as follows:

$$\frac{N_{11}}{\gamma_1/\sqrt{p_{11}}} = \frac{N_{12}}{\gamma_1/\sqrt{p_{12}}} = \frac{N_{21}}{\gamma_2/\sqrt{p_{21}}} = \frac{N_{22}}{\gamma_2/\sqrt{p_{21}}}. \quad (6)$$

The literature typically assumes equal probabilities among the different stages (e.g., $p_{11} = p_{12}$ and $p_{21} = p_{22}$) implying equal runs among the stages ($N_{11} = N_{12}$ and $N_{21} = N_{22}$). Equation (6) shows the specific optimal allocation when the probabilities are different. It is intuitive that N_{ij} decreases as p_{ij} increases. That is, fewer simulations are required in a stage that is less rare. However, it is possibly counter-intuitive that N_{ij} increases as γ_i increases. That is, more simulations are required for the design that is less rare. To get a sense for the improvement that is possible, we further study the following cases.

4.1 Case 1: Standard Simulation

We allocate an equal number of runs to designs 1 and 2 (i.e., $N_{11} = N_{21} = T/2$). In other words, simulations are run starting at level 0 (the standard starting location). No splitting is done. In standard simulation, there is no explicit control over the number of runs from level 1 to level 2. Thus, the number of times we simulate from level 1 to level 2 (N_{12} corresponding to design 1 and N_{22} corresponding to design 2) are random variables with expectations $\text{E}[N_{12}] = p_{11}N_{11}$ and $\text{E}[N_{22}] = p_{21}N_{21}$. In other words, N_{11} and N_{21} are fixed numbers and are considered to be decision variables, while N_{12} and N_{22} are random variables. Since we are concerned with rare-event simulation, the second level simulation numbers N_{12} and N_{22} are very small, so the total budget allocation is approximately T (though slightly higher):

$$N_{11} + N_{12} + N_{21} + N_{22} \approx T.$$

Since we are not using level-splitting, we do not use (6). Instead, we can directly compute $\text{Var}[\hat{\gamma}_1 - \hat{\gamma}_2] = \text{Var}[\hat{\gamma}_1] +$

$\text{Var}[\hat{\gamma}_2]$:

$$\begin{aligned}\text{Var}[\hat{\gamma}_1]+\text{Var}[\hat{\gamma}_2] &= \frac{\gamma_1(1-\gamma_1)}{N_{11}}+\frac{\gamma_2(1-\gamma_2)}{N_{21}} \\ &\approx \frac{2\gamma_1}{T}+\frac{2\gamma_2}{T}=\frac{2(\gamma_1+\gamma_2)}{T}.\end{aligned}$$

4.2 Case 2: Level Splitting Without OCBA

As in Case 1, we allocate an equal number of runs to designs 1 and 2 (i.e., $N_{11}=N_{21}$). But this time, we control the number of runs from level 1 to level 2. To get a sense for the possible speedup, we suppose that probabilities among different levels are equal within a given design. That is, $p_{11}=p_{12}=\sqrt{\gamma_1}$ and $p_{21}=p_{22}=\sqrt{\gamma_2}$. In this case, there is no effort to control the distribution of runs among different designs, only the number of runs among different levels within a design. Standard results from level splitting [or (2)] implies that performance is optimized when $N_{11}=N_{12}$ and $N_{21}=N_{22}$. Thus $N_{11}=N_{12}=N_{21}=N_{22}=T/4$.

$$\begin{aligned}&\text{Var}[\hat{p}_{11}\hat{p}_{12}]+\text{Var}[\hat{p}_{21}\hat{p}_{22}] \\ &\approx \gamma_1^2\left(\frac{1}{p_{11}N_{11}}+\frac{1}{p_{12}N_{12}}\right)+\gamma_2^2\left(\frac{1}{p_{21}N_{21}}+\frac{1}{p_{22}N_{22}}\right) \\ &= \gamma_1^2\left(\frac{4}{\sqrt{\gamma_1}T}+\frac{4}{\sqrt{\gamma_1}T}\right)+\gamma_2^2\left(\frac{4}{\sqrt{\gamma_2}T}+\frac{4}{\sqrt{\gamma_2}T}\right) \\ &= \frac{8\gamma_1^{3/2}}{T}+\frac{8\gamma_2^{3/2}}{T}\end{aligned}$$

Compared with Case 1 (standard simulation), the variance contribution due to each design is reduced here by a factor of $4\sqrt{\gamma_i}$, which is generally much less than 1 for small γ_i.

4.3 Case 3: OCBA Without Level Splitting

The optimal allocation for two designs is $N_{11}/N_{21}=\sigma_1/\sigma_2$, where σ_1 and σ_2 are the standard deviations of the outcomes of single simulations of designs 1 and 2 (e.g., Chen 2002). That is, $\sigma_1=\sqrt{\gamma_1(1-\gamma_1)}\approx\sqrt{\gamma_1}$ and $\sigma_2=\sqrt{\gamma_2(1-\gamma_2)}\approx\sqrt{\gamma_2}$. Thus, $N_{11}/N_{21}\approx\sqrt{\gamma_1/\gamma_2}$. As in Case 1, the number of times we simulate from level 1 to level 2, N_{12} and N_{22}, are random variables, with small expectations. Thus, we roughly have $N_{11}+N_{21}\approx T$. Combining these two results gives and

$$N_{11}=\frac{\sqrt{\gamma_1}}{\sqrt{\gamma_1}+\sqrt{\gamma_2}}T \quad \text{and} \quad N_{21}=\frac{\sqrt{\gamma_2}}{\sqrt{\gamma_1}+\sqrt{\gamma_2}}T.$$

So

$$\begin{aligned}\text{Var}[\hat{\gamma}_1]+\text{Var}[\hat{\gamma}_2] &= \frac{\gamma_1(1-\gamma_1)}{N_{11}}+\frac{\gamma_2(1-\gamma_2)}{N_{21}} \\ &\approx \frac{\gamma_1+\sqrt{\gamma_1\gamma_2}}{T}+\frac{\gamma_2+\sqrt{\gamma_1\gamma_2}}{T} \\ &= \frac{\gamma_1+\gamma_2+2\sqrt{\gamma_1\gamma_2}}{T} \\ &= \frac{\left(\sqrt{\gamma_1}+\sqrt{\gamma_2}\right)^2}{T}\end{aligned}$$

We make the following remarks:

1. The variances of the terms from the individual designs are equal;
2. Suppose we let $\gamma_2=c\gamma_1$, so that c represents the ratio of the two rare-event probabilities. Then, the ratio of the variances for Case 3 (OCBA) and Case 1 (standard simulation) reduces to

$$\frac{(1+c)\gamma_1+2\gamma_1\sqrt{c}}{2(1+c)\gamma_1}=\frac{(1+\sqrt{c})^2}{2(1+c)}. \qquad (7)$$

   (a) If $c=1$ (i.e., $\gamma_1=\gamma_2$), then (7) reduces to 1. In other words, when the rare events of the two designs are equal, there is no improvement using OCBA.

   (b) If $c\to 0$ or $c\to\infty$, then (7) goes to $1/2$. In other words, the simulation efficiency of OCBA improves as the rare-event probabilities become more unequal, up to a maximum improvement of a factor of $1/2$.

4.4 Case 4: OCBA and Level Splitting

From (6),

$$\frac{N_{11}}{\gamma_1^{3/4}}=\frac{N_{12}}{\gamma_1^{3/4}}=\frac{N_{21}}{\gamma_2^{3/4}}=\frac{N_{22}}{\gamma_2^{3/4}}.$$

This implies that $N_{11}=N_{12}$ and $N_{21}=N_{22}$ (i.e., equal runs among the two levels, for a fixed design) and also that

$$N_{11}=N_{12}=\frac{\gamma_1^{3/4}}{2(\gamma_1^{3/4}+\gamma_2^{3/4})}T$$

and

$$N_{21}=N_{22}=\frac{\gamma_2^{3/4}}{2(\gamma_1^{3/4}+\gamma_2^{3/4})}T.$$

From (4),

$$\begin{aligned}
&\mathrm{Var}[\hat{p}_{11}\hat{p}_{12}] + \mathrm{Var}[\hat{p}_{21}\hat{p}_{22}] \\
&\quad \approx \gamma_1^2 \left(\frac{1}{p_{11}N_{11}} + \frac{1}{p_{12}N_{12}} \right) \\
&\qquad + \gamma_2^2 \left(\frac{1}{p_{21}N_{21}} + \frac{1}{p_{22}N_{22}} \right) \\
&\quad = \left(\frac{\gamma_1^{3/2}}{N_{11}} + \frac{\gamma_1^{3/2}}{N_{12}} \right) + \left(\frac{\gamma_2^{3/2}}{N_{21}} + \frac{\gamma_2^{3/2}}{N_{22}} \right) \\
&\quad = \frac{4\gamma_1^{3/4}(\gamma_1^{3/4} + \gamma_2^{3/4})}{T} + \frac{4\gamma_2^{3/4}(\gamma_1^{3/4} + \gamma_2^{3/4})}{T}.
\end{aligned}$$

Compared with standard simulation (Case 1), the ratio of the resulting variances is

$$\frac{\frac{4\gamma_1^{3/4}(\gamma_1^{3/4}+\gamma_2^{3/4})}{T} + \frac{4\gamma_2^{3/4}(\gamma_1^{3/4}+\gamma_2^{3/4})}{T}}{\frac{2(\gamma_1+\gamma_2)}{T}} = 2\frac{(\gamma_1^{3/4} + \gamma_2^{3/4})^2}{(\gamma_1+\gamma_2)}$$

If $\gamma_2 = c\gamma_1$, this ratio reduces to

$$\frac{2\sqrt{\gamma_1}(1+c^{3/4})^2}{(1+c)}$$

1. If $c = 1$ (that is, $\gamma_1 = \gamma_2$) then the ratio is $4\sqrt{\gamma_1}$, which is the previous level-splitting result without OCBA.
2. If $c \to 0$ or $c \to \infty$, then the ratio goes to $2\sqrt{\gamma_1}$ or $2\sqrt{\gamma_2}$, respectively, which is the best-case improvement and is the product of the level-splitting result and the best-case OCBA improvement. However, the overall ratio is not the same as the product of the two ratios from Case 2 and Case 3. If we assume that $\gamma_1 > \gamma_2$ (or $c < 1$), which yields no loss in generality, then the best-case improvement is $2\sqrt{\gamma_1}$.

In this example where $k = 2$, the objective function in (1) decomposes into separable problems. So, we can first determine the optimal allocation within a single design and then determine the optimal allocation between designs. As a result, the overall improvement in simulation efficiency can be seen in some sense as the product of the two effects. However, the general problem is much more complex when $k > 2$. In this case, the problem does not separate, so the improvement is expected to be greater than the product of the two effects.

5 CONCLUSIONS

In this paper, we presented a new idea of optimal splitting for rare-event simulation and decision making among multiple alternatives. The key idea was to integrate the notions of optimal budget allocation and level-splitting methods to optimally allocate the budget over a fixed computing resource. We presented a formulation of the optimal computing budget problem and provided an approximate solution in the case of two designs and two levels. The maximum improvement factor for this example was $2\sqrt{\gamma_1}$, where γ_1 was the probability of the rare event for design 1, and we have assumed without loss of generality that $\gamma_1 > \gamma_2$. This reduction factor can be viewed as the product of the reduction factor for a 2-stage level-splitting approach with equally spaced levels ($4\sqrt{\gamma_1}$) and the maximum reduction factor for a 2-design budget allocation ($1/2$). The solution we gave to this problem depended on the separability of the problem. For larger problems, there is no separability, so finding the optimal solution becomes more challenging.

REFERENCES

Branke, J., S. E. Chick, and C. Schmidt. 2007. Selecting a selection procedure. *Management Science* 53 (12): 1916–1932.

Chen, C. H. 2002. Very efficient simulation for engineering design problems with uncertainty. In *Modeling and Simulation-Based Life Cycle Engineering*, ed. K. Chong, S. Saigal, and S. Thynell, 291–302. London: Spon Press.

Chen, C. H., D. He, M. Fu, and L. H. Lee. 2008. Efficient simulation budget allocation for selecting an optimal subset. To appear in *INFORMS Journal on Computing*.

Chen, C. H., J. Lin, E. Yücesan, and S. E. Chick. 2000. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Journal of Discrete Event Dynamic Systems: Theory and Applications* 10:251–270.

Chen, C. H., and J. Shortle. 2008. Optimal splitting for rare-event simulation. working paper, George Mason University.

Chen, H. C., C. H. Chen, L. Dai, and E. Yücesan. 1997. New development of optimal computing budget allocation for discrete event simulation. In *Proceedings of the 1997 Winter Simulation Conference*, 334–341. Piscataway, NJ: IEEE.

Chick, S., and K. Inoue. 2001. New two-stage and sequential procedures for selecting the best simulated system. *Operations Research* 49:16091624.

Fu, M. C., J. Q. Hu, C. H. Chen, and X. Xiong. 2007. Simulation allocation for determining the best design in the presence of correlated sampling. *INFORMS Journal on Computing* 19:101–111.

Garvels, M. 2000. *The splitting method in rare event simulation*. Ph. D. thesis, University of Twente, The Netherlands.

Glasserman, P., P. Heidelberger, P. Shahabuddin, and T. Zajic. 1999. Multilevel splitting for estimating rare event probabilities. *Operations Research* 47 (4): 585–600.

Glynn, P. W. 1994. Efficiency improvement technique. *Annals of Operations Research* 53 (1): 175–197.

Heidelberger, P. 1993. Fast simulation of rare events in queueing and reliability models. In *Performance Evaluation of Computer and Communication Systems*, ed. L. Donatiello and R. Nelson, 165–202. Springer Verlag.

Hyden, P., and L. Schruben. 2000. Improved decision processes through simultaneous simulation and time dilation. In *Proceedings of the Simulation Conference*, ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson, 743–748. Piscataway, NJ: IEEE.

Kim, S.-H., and B. L. Nelson. 2006. Selecting the best system. In *Handbook in Operations Research and Management Science: Simulation*, ed. S. G. Henderson and B. L. Nelson. Elsevier.

L'Ecuyer, P., V. Demers, and B. Tuffin. 2006. Splitting for rare-event simulation. In *Proceedings of 2006 Winter Simulation Conference*, ed. L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, 137–148. Piscataway, NJ: IEEE.

Lee, L. H., E. P. Chew, S. Y. Teng, and D. Goldsman. 2004. Optimal computing budget allocation for multi-objective simulation models. In *Proceedings of the 2004 Winter Simulation Conference*, ed. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 586–594. Piscataway, NJ: IEEE.

Trailovic, L., and L. Y. Pao. 2004. Computing budget allocation for efficient ranking and selection of variances with application to target tracking algorithms. *IEEE Transactions on Automatic Control* 49:58–67.

AUTHOR BIOGRAPHIES

JOHN F. SHORTLE is an Associate Professor of Systems Engineering and Operations Research at George Mason University. He is a member of the Center for Air Transportation Systems Research at GMU and a member of the Center for Network-Based Systems, a collaborative initiative between Noblis and GMU. His research interests include simulation and queueing applications in air transportation and telecommunications. He served as a co-editor for the 2007 Winter Simulation Conference Proceedings. Previously, he worked in the telecommunications industry at US WEST Advanced Technologies. He received a B.S. in mathematics from Harvey Mudd College in 1992 and a Ph.D. and M.S. in operations research at UC Berkeley in 1996. In 2000, he received the Daniel H. Wagner Award for Excellence in Operations Research Practice. His e-mail address is <jshortle@gmu.edu>.

CHUN-HUNG CHEN is a Professor of Systems Engineering and Operations Research at George Mason University. He received his Ph.D. from Harvard University in 1994. His research interests are mainly in development of very efficient methodologies for simulation and optimization, and its application to engineering design and air traffic management. Dr. Chen has served as Co-Editor of the Proceedings of the 2002 Winter Simulation Conference, Program Co-Chair for 2007 INFORMS Simulation Society Workshop, and the Coordinator of Contributed Papers for 2008 INFORMS Annual Meeting. He is currently an associate editor of IEEE Transactions on Automatic Control, International Journal of Simulation and Process Modeling, and the Book Series on System Engineering and Operations Research for WSPC. His email address is <cchen9@gmu.edu>.