*Proceedings of the 2004 Winter Simulation Conference*
*R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.*

# OPTIMAL COMPUTING BUDGET ALLOCATION UNDER CORRELATED SAMPLING

Michael C. Fu

The Robert H. Smith School of Business
University of Maryland
College Park, MD 20742, U.S.A.

Jian-Qiang Hu

Department of Manufacturing Engineering
Boston University
Brookline, MA 02446, U.S.A.

Chun-Hung Chen

Dept Systems Engineering & Operations Research
George Mason University
Fairfax, VA 22030, U.S.A.

Xiaoping Xiong

The Robert H. Smith School of Business
University of Maryland
College Park, MD 20742, U.S.A.

## ABSTRACT

We consider the optimal computing budget allocation (OCBA) problem where the simulated designs are correlated. The exact optimal allocation is presented for two designs, and an approximation proposed for more than two designs. Numerical experiments for several examples compare the approximation with some other allocation procedures. Results on convergence rates and the non-Gaussian setting are also discussed.

## 1 INTRODUCTION

The optimal computing budget allocation (OCBA) problem, introduced in Chen et al. (1997, 2000), is to allocate a fixed simulation budget among a finite number of designs, in order to maximize the probability of correct selection. Previous OCBA work has addressed only the setting where the different designs are sampled independently. This work considers the correlated case, providing an alternative approximate solution to the one in Fu et al. (2004) and presenting a brief summary of recent results in Xiong and Fu (2004). These results are useful in at least two ways: specifying efficient allocations based on estimates of the means, variances, and correlations; and estimating the gain in computational efficiency by inducing correlation in the experiments. The exact optimal allocation is derived for two designs. For the general case (more than two designs), we present an approximation whose solution matches the independent case of Chen et al. (2000) when the correlation is taken to zero. In that work, the optimal allocation depends on the individual design variances and pairwise mean differences with the best. The allocations derived here include an additional dependence to be expected: the correlations with the best design.

## 2 PROBLEM SETTING

The objective is to find an allocation of the total number of simulation replications that maximizes the probability of correct selection, denoted by PCS, where "correct selection" will be defined as selecting the design with largest mean. Letting $\mu_i$ denote the mean for design $i$, we assume without loss of generality that design 1 is the best, i.e., $\mu_1 > \mu_i \; \forall i > 1$. Denoting $N_i$ as the # simulation replications allocated to design $i$ and $\bar{J}_i$ as the sample average for design $i$ over $N_i$ replications, the problem is to select $N_1, N_2, , \ldots, N_k$ in order to maximize $P(\bar{J}_1 - \bar{J}_i > 0, \; i = 2, \ldots, k)$, subject to the budget constraint $N_1 + N_2 + \ldots + N_k = T$, where $T$ is the total computing budget (# simulation replications).

We further assume that the samples are jointly normally distributed, with paired covariance and correlations given by $C_{ij}$ (equal to variance $\sigma_i^2$ for $j = i$) and $\rho_{ij} = C_{ij}/(\sigma_i \sigma_j)$, respectively. Then, $\{\bar{J}_i, i = 1, \ldots, k\}$ and $\{\bar{J}_1 - \bar{J}_i, i = 2, \ldots, k\}$ are also multivariate normal with respective covariance matrices $\Theta = \left[\frac{C_{ij}}{\max(N_i, N_j)}\right]_{k \times k}$ and $\Lambda = [\lambda_{ij}]_{(k-1)\times(k-1)} = X^T \Theta X$, where $X$ is a $k \times (k-1)$ matrix with 1's in the first row, -1's in the diagonal starting on the first entry of the second row, and 0's otherwise. For example, for $k = 2$, we have

$$\lambda_{11} = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2} - \frac{2C_{12}}{\max(N_1, N_2)}. \qquad (1)$$

## 3 SOLUTION FOR TWO DESIGNS

For two designs, the problem reduces to minimizing $\lambda_{11}$ given by (1), i.e., the following optimization problem:

$$\min_{N_1,N_2} \left\{ \frac{\sigma_1/\sigma_2}{N_1} + \frac{\sigma_2/\sigma_1}{N_2} - \frac{2\rho_{12}}{\max(N_1, N_2)} \right\} \tag{2}$$

subject to $N_1 + N_2 = T$.

Simplifying notation, we drop the subscript on $\rho_{12}$ and define $c = \sigma_1/\sigma_2$, assuming without loss of generality that $c \geq 1$. Then (2) can be rewritten as

$$\min_{N_1,N_2} \left\{ \frac{c}{N_1} + \frac{1/c}{N_2} - \frac{2\rho}{\max(N_1, N_2)} \right\}. \tag{3}$$

When $\rho = 0$, it is known that the optimal solution is

$$\frac{N_1}{N_2} = \frac{\sigma_1}{\sigma_2},$$

i.e., the simulation allocation depends only on the standard deviations when there is no correlation in the sampling.

To handle the difficulty introduced by the $\max(N_1, N_2)$ in (3), the analysis is divided into two cases: (a) $N_1 \geq N_2$, and (b) $N_1 \leq N_2$. Furthermore, we now treat the minimization of (3) as a continuous variable optimization problem.

For case (a) $N_1 \geq N_2$, the function to be minimized in (3) becomes the following:

$$h_a(N_1) = \frac{c - 2\rho}{N_1} + \frac{1/c}{T - N_1}. \tag{4}$$

Differentiating and setting to zero, one obtains:

$$\frac{N_1}{N_2} = \sqrt{c(c - 2\rho)}. \tag{5}$$

The ratio given by (5) can provide a further indication of the effect of correlation on the relative allocation. In the case of independent simulations (i.e., $\rho = 0$), $N_1/N_2 = c$, where $c$ is the ratio of the standard deviations.

Substituting $N_2 = T - N_1$, we obtain the solution

$$N_1^a = T \frac{\sqrt{c(c - 2\rho)}}{1 + \sqrt{c(c - 2\rho)}}. \tag{6}$$

However, two additional conditions need to be satisfied in order for this to be valid: $c - 2\rho \geq 0$, and $c(c - 2\rho) \geq 1$, since $N_1 \geq N_2$ was assumed. The second condition reduces to $c - 1/c \geq 2\rho$, which implies the first condition, since $c > 0$. If this is not satisfied, then (4) is monotonically increasing in $N_1$, and the solution occurs at the boundary $N_1 = N_2$.

For case (b) $N_1 \leq N_2$, we have objective function

$$h_b(N_1) = \frac{c}{N_1} + \frac{1/c - 2\rho}{T - N_1}.$$

Similar to case (a), the first-order condition leads to

$$\frac{N_2}{N_1} = \sqrt{\frac{1}{c}\left(\frac{1}{c} - 2\rho\right)}, \tag{7}$$

and we can show that if $1/c - c \geq 2\rho$, then $h_b(N_1)$ is minimized at

$$N_1^b = T \frac{1}{1 + \sqrt{\frac{1}{c}(\frac{1}{c} - 2\rho)}}; \tag{8}$$

otherwise, $h_b(N_1)$ is monotonically decreasing in $N_1$, and the solution occurs at the boundary $N_1 = N_2$.

Combining cases (a) and (b), we have the overall solution based on three regions for $2\rho$: (Recall that we assumed $c \geq 1$ without loss of generality, so that $N_1$ corresponds to the system with the higher variance and $c - 1/c \geq 1/c - c$.)

(i) If $2\rho \geq c - 1/c$, then the optimum allocation is $N_1 = N_2 = T/2$;

(ii) If $1/c - c \leq 2\rho \leq c - 1/c$, then the optimum allocation is given by (6), which has $N_1 \geq N_2$. Note that taking $2\rho = c - 1/c$ in (6) also gives $N_1 = N_2 = T/2$, so there is continuity with the previous case;

(iii) If $2\rho \leq 1/c - c$, then the optimum allocation is given by either (6) or (8), depending on which of $h_a(N_1^a)$ and $h_b(N_1^b)$ is smaller.

Thus, to complete the allocation specification, we need to compare $h_a(N_1^a)$ and $h_b(N_1^b)$ in the region $2\rho \leq 1/c - c$:

$$\begin{aligned} h_a(N_1^a) &= \frac{1}{cT}\left(1 + \sqrt{c(c - 2\rho)}\right)^2 \\ &= \frac{1}{cT}\left(1 + c^2 - 2c\rho + 2c\sqrt{1 - 2\rho/c}\right), \\ h_b(N_1^b) &= \frac{c}{T}\left(1 + \sqrt{\frac{1}{c}(\frac{1}{c} - 2\rho)}\right)^2 \\ &= \frac{1}{cT}\left(c^2 + 1 - 2c\rho + 2c\sqrt{1 - 2\rho c}\right). \end{aligned}$$

Since $c \geq 1$ and $2\rho \leq 1/c - c \leq 0$, we have $\sqrt{1 - 2\rho/c} \leq \sqrt{1 - 2\rho c}$, so that $h_a(N_1^a) < h_b(N_1^b)$ if $c > 1$ and equality occurring when $c = 1$. Therefore, the optimum allocation is given by (6) in the region $2\rho \leq 1/c - c$, which is the same as in the region $1/c - c \leq 2\rho \leq c - 1/c$.

Summarizing mathematically, we can write

$$N_1^* = \begin{cases} N_1^a \geq N_2^* & 2\rho \leq c - 1/c, \\ T/2 = N_2^* & 2\rho \geq c - 1/c, \end{cases} \tag{9}$$

and $N_2^* = T - N_1^*$. More discussion of the solution and a simplified analysis is contained in Fu et al. (2004).

## 4 APPROXIMATE SOLUTION FOR THREE DESIGNS

We now consider the case of three designs, i.e., $k = 3$. For this case, we want to maximize

$$f(N_1, N_2, N_3) = \int_{u_{11}x_1 \geq \mu_2 - \mu_1} \int_{u_{22}x_2 \geq \mu_3 - \mu_1 - u_{12}x_1} e^{-\frac{x_1^2 + x_2^2}{2}} dx_2 dx_1, \tag{10}$$

s.t. $N_1 + N_2 + N_3 = T$, where $u_{11} = \sqrt{\lambda_{11}}$, $u_{12} = \lambda_{12}/\sqrt{\lambda_{11}}$, $u_{22} = \sqrt{(\lambda_{11}\lambda_{22} - \lambda_{12}^2)/\lambda_{11}}$, and $N_1, N_2, N_3$ enter via the following equations:

$$\lambda_{11} = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2} - \frac{2C_{12}}{\max(N_1, N_2)}, \tag{11}$$

$$\lambda_{12} = \frac{\sigma_1^2}{N_1} - \frac{C_{12}}{\max(N_1, N_2)} - \frac{C_{13}}{\max(N_1, N_3)} \tag{12}$$

$$+ \frac{C_{23}}{\max(N_2, N_3)}, \tag{13}$$

$$\lambda_{22} = \frac{\sigma_1^2}{N_1} + \frac{\sigma_3^2}{N_3} - \frac{2C_{13}}{\max(N_1, N_3)}. \tag{14}$$

In general, this maximization problem is analytically intractable. Therefore, we propose an approximate problem, where again we treat the variables as continuous. First, we note that $\exp(-(x_1^2 + x_2^2)/2)$ decreases exponentially with respect to $x_1^2 + x_2^2$, so as a surrogate for maximizing $f(N_1, N_2, N_3)$, which involves integration over an infinite region

$$\{(x_1, x_2) | u_{11}x_1 \geq \mu_2 - \mu_1, u_{12}x_1 + u_{22}x_2 \geq \mu_3 - \mu_1\},$$

we instead maximize the size of the circle centered at the origin that is contained in this region, which is equivalent to maximizing the circle's radius given by $\min(d_2, d_3)$, where

$$d_2 = \frac{\mu_1 - \mu_2}{u_{11}} = \frac{\mu_1 - \mu_2}{\sqrt{\lambda_{11}}} \tag{15}$$

$$d_3 = \frac{\mu_1 - \mu_3}{\sqrt{u_{12}^2 + u_{22}^2}} = \frac{\mu_1 - \mu_3}{\sqrt{\lambda_{22}}} \tag{16}$$

are the distances from the origin to the lines $u_{11}x_1 = \mu_2 - \mu_1$ and $u_{12}x_1 + u_{22}x_2 = \mu_3 - \mu_1$, respectively, i.e.,

$$\arg \max_{N_1, N_2, N_3} \int_{u_{11}x_1 \geq \mu_2 - \mu_1} \int_{u_{22}x_2 \geq \mu_3 - \mu_1 - u_{12}x_1} e^{-\frac{x_1^2 + x_2^2}{2}} dx_2 dx_1$$

$$\approx \arg \max_{N_1, N_2, N_3} \min(d_2, d_3) \equiv (N_1', N_2', N_3').$$

The error of this approximate solution decreases exponentially with $T$ (see Fu et al. 2004 for details).

Next, we discuss how to maximize $\min(d_2, d_3)$. Based on (11), (14), (15), and (16), we have

**Lemma 1** *$d_i$ is monotone with respect to $N_i$ ($i = 2, 3$); in particular, increasing if $C_{1i} < \frac{1}{2}\sigma_i^2$ or $C_{1i} > \frac{1}{2}\sigma_i^2$, $N_i < N_1$; else decreasing (if $C_{1i} > \frac{1}{2}\sigma_i^2$, $N_i > N_1$).*

This gives the following result:

**Lemma 2** *$\min(d_2, d_3)$ is maximized at $d_2 = d_3$.*

Based on Lemma 2, we set $d_2 = d_3$, which leads to

$$\frac{\lambda_{11}}{\lambda_{22}} = \left(\frac{\mu_1 - \mu_2}{\mu_1 - \mu_3}\right)^2. \tag{17}$$

This approach can be further extended to the case of more than three designs, which we do in the next section, where an iterative solution procedure is proposed. First, however, we consider a special case where we can obtain an explicit analytical solution.

### 4.1 A Special Case for Three Designs

We consider a special case where designs 2 and 3 are symmetric, i.e., $\mu_2 = \mu_3$, $\sigma_2 = \sigma_3$, and $\sigma_{12} = \sigma_{13}$, so that $N_2 = N_3$, $\lambda_{11} = \lambda_{22}$, and $d_2 = d_3$. Then the original problem of maximizing $\min(d_2, d_3)$ is reduced to minimizing $\lambda_{11} = \lambda_{22}$, i.e.,

$$\min_{N_1, N_2} \left\{ \frac{\sigma_1/\sigma_2}{N_1} + \frac{\sigma_2/\sigma_1}{N_2} - \frac{2\rho_{12}}{\max(N_1, N_2)} \right\} \tag{18}$$

s.t. $N_1 + 2N_2 = T$, which is similar to the case of two designs. Again, define $c = \sigma_1/\sigma_2$, but instead of assuming that $c \geq 1$, we need to consider two regions: $c \geq 1/\sqrt{2}$ and $c \leq 1/\sqrt{2}$.

We again divide the analysis into two cases: (a) $N_1 \geq N_2 (= N_3)$, and (b) $N_1 \leq N_2 = (N_3)$, and treat the minimization problem as a continuous variable optimization problem. For case (a) $N_1 \geq N_2$, the function to be minimized in (18) becomes:

$$g_a(N_1) = \frac{c - 2\rho_{12}}{N_1} + \frac{2/c}{T - N_1}, \tag{19}$$

and the analogous solution is given by

$$\frac{N_1}{N_2} = \sqrt{2c(c-2\rho_{12})}. \tag{20}$$

Comparing Equation (20) with Equation (5), we can see that the relative weight for allocating the computing budget to the best design becomes greater as we increase the number of designs. This partially explains the observations made in Chen et al. (2000) that $N_1$ is much larger than $N_i (i = 2, \ldots, k)$ when the number of designs $k$ is large.

Substituting $2N_2 = T - N_1$ leads to the solution analogous to (6):

$$N_1^a = T\frac{\sqrt{(c-2\rho_{12})c/2}}{1+\sqrt{(c-2\rho_{12})c/2}}. \tag{21}$$

However, this solution is only valid under $N_1 \geq N_2$, which from (20) implies that $2\rho_{12} \leq c - 1/(2c)$. When $2\rho_{12} > c - 1/(2c)$, $g_a(N_1)$ is minimized at the boundary $N_1 = N_2$.

For case (b) $N_2 \geq N_1$, the objective function is

$$g_b(N_1) = \frac{c}{N_1} + \frac{2(1/c - 2\rho_{12})}{T - N_1}, \tag{22}$$

which is minimized at

$$\frac{N_2}{N_1} = \sqrt{\frac{1}{2c}\left(\frac{1}{c} - 2\rho_{12}\right)},$$

giving the solution

$$N_1^b = T\frac{1}{1+\sqrt{\frac{2}{c}\left(\frac{1}{c} - 2\rho_{12}\right)}}, \tag{23}$$

if $2\rho_{12} \leq 1/c - 2c$; otherwise, $g_b(N_1)$ is minimized at the boundary $N_1 = N_2$.

When $c \geq 1/\sqrt{2}$, we have $c - 1/(2c) \geq 0 \geq 1/c - 2c$. Therefore, combining cases (a) and (b), we have the overall solution based on three regions for $2\rho_{12}$:

(i) If $2\rho_{12} \geq c - 1/(2c)$, then the optimum allocation is $N_1 = N_2 = N_3 = T/3$;

(ii) If $1/c - 2c \leq 2\rho_{12} \leq c - 1/(2c)$, then the optimum allocation is given by (21), which has $N_1 \geq N_2$;

(iii) If $2\rho_{12} \leq 1/c - 2c$, then the optimum allocation is given by either (21) or (23), depending on which of $g_a(N_1^a)$ and $g_b(N_1^b)$ is smaller.

Finally, we compare $g_a(N_1^a)$ and $g_b(N_1^b)$ in the region $2\rho_{12} \leq 1/c - 2c$:

$$\begin{aligned} g_a(N_1^a) &= \frac{2}{cT}\left(1+\sqrt{\frac{c(c-2\rho_{12})}{2}}\right)^2, \\ g_b(N_1^b) &= \frac{c}{T}\left(1+\sqrt{\frac{2}{c}(\frac{1}{c}-2\rho_{12})}\right)^2. \end{aligned}$$

Hence

$$\begin{aligned} & g_a(N_1^a) - g_b(N_1^b) \\ &= \frac{2}{T}\left(\rho_{12} + \sqrt{2(1-2\rho_{12}/c)} - \sqrt{2(1-2c\rho_{12})}\right) \\ &= \frac{2}{T}\left(\rho_{12} - \frac{4\rho_{12}(1-c^2)/c}{\sqrt{2(1-2\rho_{12}/c)}+\sqrt{2(1-2c\rho_{12})}}\right) \\ &= \frac{2\rho_{12}}{T}\left(1 - \frac{4(1-c^2)/c}{\sqrt{2(1-2\rho_{12}/c)}+\sqrt{2(1-2c\rho_{12})}}\right). \end{aligned}$$

Since $c \geq 1/\sqrt{2}$ and $2\rho_{12} \leq 1/c - 2c$ $(\leq 0)$, we have

$$1 - \frac{4(1-c^2)/c}{\sqrt{2(1-2\rho_{12}/c)}+\sqrt{2(1-2c\rho_{12})}}$$

$$\geq 1 - \frac{2\sqrt{2}}{\sqrt{2(1-2\rho_{12}/c)}+\sqrt{2(1-2c\rho_{12})}} \geq 0.$$

Therefore, $g_a(N_1^a) \leq g_b(N_1^b)$ for $2\rho_{12} \leq 1/c - 2c$, which leads to

$$N_1^* = \begin{cases} T\frac{\sqrt{(c-2\rho_{12})c/2}}{1+\sqrt{(c-2\rho_{12})c/2}} \ (\geq N_2) & 2\rho_{12} \leq c - 1/(2c); \\ T/3 \ (= N_2) & 2\rho_{12} \geq c - 1/(2c) \geq 0. \end{cases} \tag{24}$$

When $c \leq 1/\sqrt{2}$, we have $1/c - 2c \geq 0 \geq c - 1/(2c)$, and the three regions for $\rho_{12}$ are:

(i) $2\rho_{12} \geq 1/c - 2c$: in this region the optimum allocation is $N_1 = N_2 = N_3 = T/3$;

(ii) $c - 1/(2c) \leq 2\rho_{12} \leq 1/c - 2c$: in this region the optimum allocation is given by (23), which has $N_2 \geq N_1$;

(iii) $2\rho_{12} \leq c - 1/(2c)$: in this region the optimum allocation is given by either (21) or (23), depending on which of $g_a(N_1^a)$ and $g_b(N_1^b)$ is smaller.

Unfortunately, in this case, which of $g_a(N_1^a)$ and $g_b(N_1^b)$ is smaller depends on the actual values of $c$ $(\leq 1/\sqrt{2})$ and

$\rho_{12}$, so we write:

$$N_1^* = \begin{cases} T\frac{1}{1+\sqrt{\frac{2}{c}(\frac{1}{c}-2\rho_{12})}} \quad (\le N_2) \\ \qquad \text{if } c - 1/(2c) \le 2\rho_{12} \le 1/c - 2c; \\ \arg\min\left\{ g_a\left(T\frac{\sqrt{(c-2\rho_{12})c/2}}{1+\sqrt{(c-2\rho_{12})c/2}}\right), \right. \\ \qquad \left. g_b\left(T\frac{1}{1+\sqrt{\frac{2}{c}(\frac{1}{c}-2\rho_{12})}}\right)\right\} \\ \qquad \text{if } 2\rho_{12} \le c - 1/(2c) \le 0; \\ T/3 \quad (= N_2) \qquad \text{if } 2\rho_{12} \ge 1/c - 2c \ge 0. \end{cases} \tag{25}$$

Combining (24) and (25), the solution is as follows:

$$N_1^* = \begin{cases} T\frac{\sqrt{(c-2\rho_{12})c/2}}{1+\sqrt{(c-2\rho_{12})c/2}} \quad (\ge N_2) \\ \qquad c \ge 1/\sqrt{2}, 2\rho_{12} \le c - 1/(2c); \\ T\frac{1}{1+\sqrt{\frac{2}{c}(\frac{1}{c}-2\rho_{12})}} \quad (\le N_2) \\ \qquad c \le 1/\sqrt{2},\ c - 1/(2c) \le 2\rho_{12} \le 1/c - 2c; \\ \arg\min\left\{ g_a\left(T\frac{\sqrt{(c-2\rho_{12})c/2}}{1+\sqrt{(c-2\rho_{12})c/2}}\right), \right. \\ \qquad \left. g_b\left(T\frac{1}{1+\sqrt{\frac{2}{c}(\frac{1}{c}-2\rho_{12})}}\right)\right\} \\ \qquad c \le 1/\sqrt{2}, 2\rho_{12} \le 1/c - 2c \le 0; \\ T/3 \quad (= N_2) \qquad \text{otherwise.} \end{cases} \tag{26}$$

**Remark**. Consistent with the results obtained for the two-design case in the previous section, when the correlation $\rho_{12}$ is positive and high enough — i.e., the last situation in (26) where $2\rho_{12} \ge c - 1/(2c) \ge 0$ for $c \ge 1/\sqrt{2}$ or $2\rho_{12} \ge 1/c - 2c \ge 0$ for $c \le 1/\sqrt{2}$ — the equal allocation ($N_1 = N_2 = N_3 = T/3$) is again optimal.

## 5 MORE THAN THREE DESIGNS

The approach used for the three-design case that led to Equation (17) can also be applied to the general case of $k$ designs ($k \ge 3$). Let

$$d_{i+1} = \frac{\mu_1 - \mu_{i+1}}{\sqrt{\lambda_{ii}}},$$

which is the distance from the origin to hyperplane $\sum_{m=1}^{i} u_{m,i} x_m = \mu_{i+1} - \mu_1$ ($i = 1, \ldots, k-1$). It can be similarly shown that $\min(d_2, \cdots, d_k)$ is maximized when

$$d_i = d_j \quad \forall i, j > 1,$$

which leads to the constraint (for any $j = 1, \ldots, k-1$)

$$\frac{\lambda_{jj}}{\lambda_{ii}} = \left(\frac{\mu_1 - \mu_{j+1}}{\mu_1 - \mu_{i+1}}\right)^2, \quad \text{for } i = 1, \ldots, k-1, i \ne j, \tag{27}$$

where

$$\lambda_{ii} = \frac{\sigma_1^2}{N_1} + \frac{\sigma_{i+1}^2}{N_{i+1}} - \frac{2C_{1,i+1}}{\max(N_1, N_{i+1})},\ i = 1, \ldots, k-1. \tag{28}$$

Equation (27) is central to the optimal allocation of the computing budget, relating the factors that have the most significant impact on the computing budget allocations: i) the variances for each design; ii) the differences in means between each of the designs and the best design; and iii) the correlations between each of the designs and the best design. The relationships implied by equation (27) among the first two factors also appear to be consistent with our intuition from the independent case. For example, as the difference between $\mu_1$ and $\mu_{i+1}$ increases, the allocation to design $i+1$ (i.e., $N_{i+1}$) decreases. Intuitively speaking, this means that design $i+1$ is much worse than design 1 and so less attention should be paid to this inferior design. On the other hand, if the variance for design $i+1$ (i.e., $\sigma_{i+1}^2$) increases, $N_{i+1}$ should increase due to the high uncertainty for design $i+1$.

Recall that we wish to maximize $d_j$, which is equivalent to minimizing $\lambda_{jj}$ subject to (27) and $\sum_{i=1}^{k} N_i = T$.

To simplify notation, we define for $N_1 \ge N_i$,

$$\tilde{C}_{1i} = \sigma_1^2 - 2C_{1i}, \quad \tilde{C}_{ii} = \sigma_i^2, \tag{29}$$

and for $N_1 < N_i$,

$$\tilde{C}_{1i} = \sigma_1^2, \quad \tilde{C}_{ii} = \sigma_i^2 - 2C_{1i}. \tag{30}$$

This allows us a more compact expression for $\lambda_{ii}$:

$$\lambda_{ii} = \frac{\tilde{C}_{1,i+1}}{N_1} + \frac{\tilde{C}_{i+1,i+1}}{N_{i+1}}. \tag{31}$$

Without loss of generality, we fix $j = 1$ in (27), i.e., we wish to minimize $\lambda_{11}$. Defining

$$\beta_{i+1} \equiv \frac{\lambda_{11}}{\lambda_{ii}} = \left(\frac{\mu_1 - \mu_2}{\mu_1 - \mu_{i+1}}\right)^2, \tag{32}$$

we define a new (Lagrangian) objective function that incorporates the constraints:

$$L(N_1, N_2, \ldots, N_k, \delta, \delta_3, \ldots, \delta_k) = \lambda_{11} + \sum_{i=3}^{k} \delta_i(\lambda_{11} - \beta_i \lambda_{i-1,i-1}) + \delta(\sum_{i=1}^{k} N_i - T), \tag{33}$$

where $\delta, \delta_3, \ldots, \delta_k$ are Lagrange multipliers. Then

$$L(N_1, N_2, \ldots, N_k, \delta, \delta_3, \ldots, \delta_k) = \left(\frac{\tilde{C}_{12}}{N_1} + \frac{\tilde{C}_{22}}{N_2}\right)\left(1 + \sum_{i=3}^{k} \delta_i\right) - \sum_{i=3}^{k} \delta_i \beta_i \left(\frac{\tilde{C}_{1i}}{N_1} + \frac{\tilde{C}_{ii}}{N_i}\right) + \delta\left(\sum_{i=1}^{k} N_i - T\right), \quad (34)$$

and again treating the variables as continuous,

$$\frac{\partial L}{\partial N_1} = -\frac{\tilde{C}_{12}}{N_1^2}\left(1 + \sum_{i=3}^{k} \delta_i\right) + \sum_{i=3}^{k} \delta_i \beta_i \frac{\tilde{C}_{1i}}{N_1^2} + \delta, \quad (35)$$

$$\frac{\partial L}{\partial N_2} = -\frac{\tilde{C}_{22}}{N_2^2}\left(1 + \sum_{i=3}^{k} \delta_i\right) + \delta, \quad (36)$$

$$\frac{\partial L}{\partial N_i} = \delta_i \beta_i \frac{\tilde{C}_{ii}}{N_i^2} + \delta, \qquad \text{for } i \geq 3. \quad (37)$$

Strictly speaking, due to (29) and (30), the Kuhn-Tucker conditions require $(N_i - N_1)\partial L/\partial N_i = 0$, but as an easier approximation, we simply solve $\partial L/\partial N_i = 0$, $i = 1, 2, \ldots, k$, which does not necessarily hold when the optimal values of $N_i$ and $N_1$ are equal. Under this condition, (35)-(37) lead to

$$N_1 = \sqrt{\sum_{i=2}^{k} \frac{\tilde{C}_{1i}}{\tilde{C}_{ii}} N_i^2}. \quad (38)$$

For independent systems (i.e., $C_{1i} = 0 \Longrightarrow \tilde{C}_{1i} = \sigma_1^2$), (38) reduces to the formula derived in Chen et al. (2000). It also agrees with the $k = 2$ solution obtained in (5) and (7). In our setting, we can solve for each $N_i$ using (31) and (32), which we combine as

$$\left(\frac{\mu_1 - \mu_2}{\mu_1 - \mu_i}\right)^2 = \beta_i = \frac{\frac{\tilde{C}_{12}}{N_1} + \frac{\tilde{C}_{22}}{N_2}}{\frac{\tilde{C}_{1i}}{N_1} + \frac{\tilde{C}_{ii}}{N_i}}, \quad i = 2, ..., k. \quad (39)$$

In principle, these equations can be solved to determine $\{N_i\}$. However, since they depend on the values of the means and variances/covariances, we propose the following two-stage algorithm, where the initial stage is used to estimate the means and variances/covariances, and the second stage calculates $\{N_i\}$ based on these estimates to allocate the bulk of the computational budget (assuming $kn_0 \ll T$).

In order to implement the CBA algorithm, the step marked (*) requires further details. It was purposely left unspecified, because it is a numerical estimation step that can be done in several ways. Here, we describe one possible implementation, which was used in our numerical examples reported in the next section.

Note that for any $\{N_i\}$ satisfying (38) and (39), $\{MN_i\}$ also satisfies these equations (where $M > 0$ is any constant). Thus, one can find any solution $\{N_i\}$ satisfying (38) and (39) first, and then normalize according to $\sum_i N_i = T$. In particular, arbitrarily choose a value for $N_1$ (e.g., $N_1 = 1$), and then solve for $N_i$ in terms of $N_2$ using (39). Substituting into (38), one can then solve for $N_2$ and then have the rest of the $N_i$. Doing this leads to

$$N_i = \frac{\beta_i \tilde{C}_{ii}}{\tilde{C}_{12} + \tilde{C}_{22}/N_2 - \beta_i \tilde{C}_{1i}}, \quad i > 2, \quad (40)$$

$$1 = \sum_{i=2}^{k} \frac{\beta_i^2 \tilde{C}_{ii} \tilde{C}_{1i}}{\left(\tilde{C}_{12} + \tilde{C}_{22}/N_2 - \beta_i \tilde{C}_{1i}\right)^2}, \quad (41)$$

the latter to be solved for $N_2$.

In actual implementation, the means and variances/covariances are estimated using the corresponding sample quantities. Let $*$ denote the index of the design with current best sample mean, and $**$ denote the index of the design with the current 2nd best sample mean. Set $N_* = 1$ and solve for $N_{**}$ using Equation (41), with the variable $N_2$ replaced by $N_{**}$, and then apply (40) to find $N_i$ for the remaining $i \neq *, **$, where in both Equations (40) and (41), $\{\beta_i\}$ and $\{\tilde{C}_{ij}\}$ will be estimated by the respective sample quantities (denoted sample means by $\{\bar{X}_i\}$ and latter estimates by $\{\widehat{\beta}_i\}$ and $\{\widehat{\tilde{C}_{ij}}\}$), e.g.,

$$\widehat{\beta}_i = \left(\frac{\bar{X}_* - \bar{X}_{**}}{\bar{X}_* - \bar{X}_i}\right)^2, \quad i \neq *, \quad \text{so that } \widehat{\beta}_{**} = 1. \quad (42)$$

Once (40) and (41) are successfully solved, the $\{N_i\}$ are renormalized and appropriately rounded to integers so that they sum to the total computation budget $T$. There is one problem that can be encountered for high enough positive correlation relative to the variances of the designs (this is even true theoretically, not just empirically), which is the situation in which all of the coefficients of $N_i^2, i > 1$ ($i \neq *$ in the samples) are negative in (38). This corresponds to the last condition in (9) and (26) in the respective two-design case and three-design special symmetric case. As a result, if this occurs, we simply chose the equal allocation.

In finding the solution, the most time-consuming step is to find $\{N_{**}\}$ which satisfies (41). Many numerical methods are available for this purpose. One has to calculate the summation term in (41) a few times for different trial values of $\{N_{**}\}$ in order to find the solution. The computational burden, however, of this procedure is still negligible compared with carrying out a typical discrete-event simulation.

**Correlated Budget Allocation (CBA) Algorithm**

- **Inputs:** $k$ (# designs), $T$ (total simulation budget), and $n_0$ (initial sample size for each design).
- **Step 1.** Estimate $\{\mu_i\}$ and $\{C_{ij}\}$, based on $n_0$ replications of each design.
- **Step 2.** Use estimates of $\{\mu_i\}$ and $\{C_{ij}\}$, including the indices for the best and 2nd best designs based on the sample means, in (29), (30), (38), and (39) to determine new $\{N_i\}$. (*)
- **Step 3.** Perform $(N_i - n_0)^+$ additional replications of design $i$, $i = 1, ..., k$.
- **Return** design with largest (overall) sample mean.

For $k = 3$, we can work out a solution without having to resort to this iterative procedure.

$$1 = \frac{\beta_2^2 \tilde{C}_{22}\tilde{C}_{12}}{\left(\tilde{C}_{12} + \tilde{C}_{22}/N_2 - \beta_2\tilde{C}_{12}\right)^2} + \frac{\beta_3^2 \tilde{C}_{33}\tilde{C}_{13}}{\left(\tilde{C}_{12} + \tilde{C}_{22}/N_2 - \beta_3\tilde{C}_{13}\right)^2},$$

$$\beta_2 = 1, \beta_3 = (\mu_1 - \mu_2/\mu_1 - \mu_3)^2 .$$

Let

$$x = \tilde{C}_{22}/N_2, \quad \alpha_i = \beta_i^2 \tilde{C}_{ii}\tilde{C}_{1i},\ i = 1, 2, \quad \gamma = \tilde{C}_{12} - \beta_3\tilde{C}_{13}.$$

$$1 = \frac{\tilde{C}_{22}\tilde{C}_{12}}{x^2} + \frac{\beta_3^2 \tilde{C}_{33}\tilde{C}_{13}}{\left(\tilde{C}_{12} + x - \beta_3\tilde{C}_{13}\right)^2},$$

$$x^2 = \alpha_2 + \frac{\alpha_3 x^2}{(\gamma + x)^2} \Longrightarrow (x^2 - \alpha_2)(\gamma + x)^2 = \alpha_3 x^2,$$

yielding the quartic equation

$$x^4 + 2\gamma x^3 + (\gamma^2 - \alpha_2^2 - \alpha_3)x^2 - 2\gamma\alpha_2 x - \alpha_2\gamma^2 = 0,$$

which has a known "closed-form" solution.

We now consider a special case $N_1 >> N_i (i = 2, \ldots, k)$. Based on (27), we have

$$\frac{N_i}{N_j} \approx \frac{\sigma_i^2(\mu_1 - \mu_j)^2}{\sigma_j^2(\mu_1 - \mu_i)^2} \quad \forall i, j > 1, \tag{43}$$

which is the same as the result obtained in Chen et al. (2000) for independent systems.

## 6 NUMERICAL EXAMPLES

We consider a few numerical examples to compare the performance of the CBA algorithm proposed in the last section with the OCBA algorithm (Chen et al. 2000) — which henceforth we will call the *independent* budget allocation (IBA) algorithm, since it is derived based on independent sampling, and to contrast it with the correlated version proposed here — and with the simple baseline benchmark alternative of equal budget allocation (EBA), i.e., $N_1 = N_2 = \cdots = N_k = T/k$. For each example, the probability of correct selection, denoted by PCS, is estimated using 100,000 independent "macro" replications of each budget allocation algorithm, leading to approximately three decimal places of precision. All examples had 10 designs with total computing budget of $T = 500$ and initial allocation of $n_0 = 10$ (100 total samples), leaving 400 samples to be allocated in the second stage of the IBA and CBA algorithms. Another evaluation of the computational savings was obtained by increasing $T$ successively for EBA in order to estimate the total sampling budget required to achieve the same level of PCS as CBA.

### 6.1 Equal Variance Example

This is adopted from Example 1 in Chen et al. (2000): $\tilde{J}_{im} \sim N(10 - i, 6^2), i = 1, 2, \cdots, 10$, first is best design. Four levels of correlation considered: -0.2, 0.0, 0.2, 0.9. The estimated PCS is given by the following (total sampling budget required by EBA to reach CBA-achieved PCS shown in parentheses):

| correlation | EBA | CBA | IBA |
|---|---|---|---|
| -20% | 0.758 | 0.817 (760) | 0.811 |
| 0% | 0.778 | 0.858 (870) | 0.858 |
| 20% | 0.808 | 0.893 (910) | 0.889 |
| 90% | 0.996 | 0.996 (500) | 0.964 |

In all three procedures, positive correlation leads to improvement in simulation efficiency, whereas negative correlation has negative impact on the simulation efficiency. Except at the very highest correlation level, IBA outperforms EBA, and CBA performs the best, providing a savings of over 50% in the total computational budget over EBA to achieve the same level of PCS. With very high positive correlation, however, the allocation of CBA coincides with EBA, which indeed outperforms the IBA allocation.

### 6.2 Unequal Variances Example

Same as previous, except the means and variances were (i.i.d.) randomly generated from $U(0, 10)$ and $U(24, 48)$ distributions, respectively:

| $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $\mu_i$ | 8.34 | 0.98 | 6.49 | 0.10 | 8.03 |
| $\sigma_i^2$ | 34.35 | 44.34 | 28.12 | 35.93 | 44.49 |
| $i$ | 6 | 7 | 8 | 9 | 10 |
| $\mu_i$ | 6.21 | 9.78 | 9.10 | 1.32 | 3.27 |
| $\sigma_i^2$ | 43.39 | 39.72 | 24.31 | 42.10 | 24.42 |

Thus, in this example, the best design is design 7, and the results are as follows:

| correlation | EBA | CBA | IBA |
|---|---|---|---|
| -20% | 0.653 | 0.697 (670) | 0.690 |
| 0% | 0.667 | 0.713 (670) | 0.713 |
| 20% | 0.691 | 0.771 (790) | 0.741 |
| 90% | 0.957 | 0.957 (500) | 0.827 |

Again, CBA outperforms both EBA and IBA.

### 6.3 Single-Server FCFS $U/U/1$ Queue

The interarrival times are $U(4, 14)$ and service times $U(1, 15.5 + 0.5i)$, $i = 1, .., 10$, with design performance $(\mu_i)$ as the expected negative (to make it a maximization problem) average system time over the first 15 served customers, so design 1 is the best. We compare two cases: one with independent sampling (indicated by "IND") and one with correlated sampling (indicated by "CRN"), using the same arrival process but independent service times. We also consider a sequential version of CBA, denoted by "seqCBA" in the results below, that starts with $n_0$ samples for each design as before, but then instead of allocating *all* of the remaining samples (400 for this particular example) at once, the computing budget is distributed *incrementally* by an amount $\Delta$ ($\Delta = 20$ used in this particular example) in each iteration until the total computing budget $T$ is exhausted. The two-stage procedure is a special case with $\Delta = T - n_0 k$. The results are as follows:

| | EBA | CBA | IBA | seqCBA |
|---|---|---|---|---|
| IND | 0.828 | 0.932 (1140) | 0.932 | 0.967 (1720) |
| CRN | 0.857 | 0.943 (1020) | 0.935 | 0.977 (1680) |

Again, CBA is more efficient than the other two compared algorithms, with the relative performance similar to that of the $\rho = 0$ and $\rho = 0.2$ cases in the previous two examples, although the reduction in computational burden is even more significant, as EBA requires more than twice the budget as CBA for the same PCS level. The sequential version further improves the efficiency significantly, as now EBA requires more than triple the budget as seqCBA for the same PCS level. Similar results were observed in Chen, He, and Yücesan (2003) and Fu et al. (2004).

## 7 RECENT EXTENSIONS

We summarize two recent significant results regarding the non-Gaussian setting and the rate of convergence of the allocation algorithm, the details of which are described in Xiong and Fu (2004). Under the setting where the samples come from a non-Gaussian distribution with well-defined moment generating function and such that the copula between any two sampling distributions is linear (which includes the independent case), we can establish a procedure similar to the that in Fu et al. (2004) to locate an approximate solution. The second result addresses the issue of convergence rate. Let $\{\widetilde{N}_i, i = 1, \cdots, k\}$ and $\{N_i, i = 1, \cdots, k\}$ be the approximate solution and the exact solution, respectively. Under the framework of Gaussian distributions and some further mild assumptions, we can show $\max_i |\widetilde{N}_i - N_i| < C$ for some constant $C$ that is *independent* of the total computing budget $T$.

## 8 ACKNOWLEDGMENTS

Michael C. Fu was supported in part by NSF under Grants DMI-9988867 and DMI-0323220, and by AFOSR under Grants F496200110161 and FA95500410210. Jian-Qiang Hu was supported in part by NSF under Grant EEC-0088073. Chun-Hung Chen was supported in part by NSF under Grants DMI-0002900, DMI-0049062 and IIS-0325074, by NASA Ames Research Center under Grants NAG-2-1565 and NAG-2-1643, by FAA under Grant 00-G-016, and by George Mason University Research Foundation. Xiaoping Xiong was supported in part by NSF under Grant DMI-0323220.

## REFERENCES

Chen, C.H., D. He, E. Yücesan. 2003. Better-than-optimal simulation runs allocation. In *Proceedings of the 2003 Winter Simulation Conference*, ed. P. J. Sanchez, S. Chick, D. Ferrin, and D. J. Morrice, 490-495. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

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

Chen, H.C., C.H. Chen, L. Dai, E. Yücesan. 1997. New development of optimal computing budget allocation for discrete event simulation. In *Proceedings of the 1997 Winter Simulation Conference*, ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson, 334-341. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Fu, M.C., J.Q. Hu, C.H. Chen, X. Xiong. 2004. Simulation allocation for determining the best design in the presence of correlated sampling. *INFORMS Journal on Computing*, submitted.

Xiong, X. and M.C. Fu. 2004. Optimal computing budget allocation: rate of convergence and non-Gaussian results. in preparation.

## AUTHOR BIOGRAPHIES

**MICHAEL C. FU** is a Professor in the Robert H. Smith School of Business, with a joint appointment in the Institute for Systems Research and an affiliate appointment in the Department of Electrical and Computer Engineering, all at the University of Maryland. He received degrees in mathematics and EE/CS from MIT, and an M.S. and Ph.D. in applied mathematics from Harvard University. His research interests include simulation and applied probability modeling, particularly with applications towards manufacturing systems, inventory control, and financial engineering. He currently serves as Simulation Area Editor of *Operations Research*. He is co-author (with J.Q. Hu) of the book, *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*, which received the INFORMS College on Simulation Outstanding Publication Award in 1998. His e-mail address is <mfu@rhsmith.umd.edu>.

**JIAN-QIANG HU** an Associate Professor with the Department of Manufacturing Engineering, Boston University. He received his B.S. in applied mathematics, from Fudan University, China, and his M.S. and Ph.D. degrees in applied mathematics, from Harvard University. His research interests include traffic engineering and QoS for communication networks, design of optical networks, queueing theory, stochastic discrete event systems, and Monte Carlo simulation. He is a co-author of the book, *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*, which won the 1998 Outstanding Simulation Publication Award from the INFORMS College on Simulation. He is an associate editor of *Operation Research*, and was a past associate editor of *IIE Transaction on Design and Manufacturing* (1998-2001). He is a member of INFORMS and IEEE. His email address is <hqiang@bu.edu>.

**CHUN-HUNG CHEN** is an Associate Professor of Systems Engineering at George Mason University. He received his Ph.D. from Harvard University in 1994. His research interests are mainly in the development of very efficient methodology for simulation and optimization, and its application to engineering design and air traffic management. He served as the Co-Editor of the Proceedings of the 2002 Winter Simulation Conference. Dr. Chen won the 1994 Harvard University Eliahu I. Jury Award for the Best Thesis in the field of Control. He is one of the recipients of the 1992 MasPar Parallel Computer Challenge Award. He is a member of INFORMS and a senior member of IEEE. His email address is <cchen9@gmu.edu>.

**XIAOPING XIONG** is a Ph.D. candidate in the Decision & Information Technology Department at the Robert H. Smith School of Business, University of Maryland, College Park. He has a master's degree in Management Science from Peking University. His research interests include options pricing, simulation, and stochastic approximation algorithms. In 2003, he was awarded the Frank T. Paine Award for Academic Excellence, and in 2004, he received the Abraham Golub Memorial Dissertation Proposal Prize. His email address is <xxiong@rhsmith.umd.edu>.