Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

EMPIRICAL ANALYSIS OF THE PERFORMANCE OF VARIANCE ESTIMATORS IN SEQUENTIAL SINGLE–RUN RANKING & SELECTION: THE CASE OF TIME DILATION ALGORITHM

Giulia Pedrielli,
Yinchao Zhu,
Loo Hay Lee,

Department of Industrial & Systems Engineering
National University of Singapore
1 Engineering Drive 2
SINGAPORE, 117576

Haobin Li

IHPC
A*STAR
1 Fusionopolis Way
SINGAPORE 138632

ABSTRACT

Ranking and Selection has acquired an important role in the Simulation-Optimization field, where the different alternatives can be evaluated by discrete event simulation (DES). Black box approaches have dominated the literature by interpreting the DES as an oracle providing i.i.d. observations. Another relevant family of algorithms, instead, runs each simulator once and observes time series. This paper focuses on such a method, Time Dilation with Optimal Computing Budget Allocation (TD-OCBA), recently developed by the authors. One critical aspect of TD-OCBA is estimating the response given correlated observations. In this paper, we are specifically concerned with the estimator of the variance of the response which plays a crucial role in simulation budget allocation. We propose an empirical analysis over the performance impact on TD-OCBA of several variance estimators involved in resource allocation. Their performances are discussed in the typical probability of correct selection (PCS) framework.

1 INTRODUCTION

In this paper, we are interested in the sequential stochastic selection methods (S^3M) which optimize the Probability of Correct Selection (PCS). We assume that the number of candidate configurations is finite and the one with best expected performance is to be identified. S^3M has been widely applied as an optimization technique in presence of few number of discrete possible solutions in a stochastic environment.
In many applications, discrete event simulation (DES) models, as a major tool to provide prediction of system performance, are used as oracle in R&S techniques to evaluate the system behaviour. DES models are utilized mainly in two different ways: the output is taken either while the simulation is running a single long replication, or from multiple shorter replications once they are completed.
The issues posed by these two families of algorithms are very different. In a multiple replication case the length of the replication and the number of replications is a relevant decision, whereas in the single run case, we need to consider the correlation between observation. In fact, single–run simulation outputs a time series of realizations of performance indicator; while multiple replications provide a sequence of independent realizations of the output. As a result, the output analysis for these two families of algorithms is remarkably different (Schruben 1983).

S^3Ms algorithms based on independent multiple replications make use of simulation allocation rules in order to determine the number of simulations to allocate to each of the systems. Optimal Computing Budget Allocation is one of the most known procedures for budget allocation (Chen and Lee 2010). OCBA optimizes the probability of correct selection (i.e., the probability that the system estimated as the best

at the end of the procedure is the true best system) by allocating different number of replications to the candidate solutions. Similar to most of the developed S^3M procedures, OCBA treats the simulator as a black box not to be accessed while it is running.

Time Dilation (TD) falls in the family of single run S^3Ms. Specifically, multiple configurations are concurrently simulated in a unique run and time series observations are used to allocate computational resource to the different candidates (Schruben 1997). In its original version, TD does not have a fixed rule of budget allocation which is instead taylored to the specific applications. Realizing the potential of TD, authors recently developed TD-OCBA, which integrates TD and OCBA. As suggested in Pedrielli, Zhu, and Lee (2015), the output from a single–run simulation does not necessarily follow i.i.d. assumption required by OCBA. If a sample variance estimator is used for a non i.i.d. series, the bias will be significant and the OCBA rule will be severely impacted by the variance estimation. Hence, the variance estimator used in TD-OCBA becomes critical for the overall performance.

The estimation of the variance from simulation output has received a remarkable attention in the OR literature, for the purpose of point and confidence interval estimation. Goldsman and Nelson (2006) and the related literature provides estimators for correlated output such as non-overlapping/overlapping Batched Means, Standardized Time Series Area Estimators, and Cramer-von Mises estimators. While these estimators are asymptotically unbiased, we are interested in the behavior of the estimators given limited output length when used for optimization purposes and not only for estimation.

In order to investigate these properties, in this paper we focus on the study of the variance estimators and their influence on the TD-OCBA performance. Batch mean estimator, weighted area estimator and weighted Cramer-von Mises estimators have been studied and the experimental results for their performance in the typical PCS framework are shown and discussed.

The remainder of the paper is organized as follows: section 2 summarizes the main background of the paper. Section 3 details the problem of interest. Candidate estimators' performance are studied in section 4 for a theoretical case with known output distribution. Finally, section 5 closes the paper.

2 BACKGROUND

We would like to first introduce the notation used throughout the paper:

Table 1: Notations.

k	total number of systems
T	total budget, in number of replications(OCBA) or observations (TD-OCBA)
N_i	budget assigned to system i
v, v_i	time scale, time scale for system i
σ_i	standard deviation for system i
$\hat{\sigma}_i$	standard deviation estimation for system i
m	batch size
B_i	total number of batches for system i
$Y_{i,j}$	jth observation from system i
$\bar{Y}_i$	sample mean performance of system i, $\bar{Y}_i = \sum_{j=1}^{N_i} Y_{i,j}$
$\bar{Y}_{i,j}$	mean of jth batch of system i, $\frac{1}{m}\sum_{l=1}^{m} Y_{i,(j-1)m+l}$
$\bar{Y}_{i,j,l}$	mean of first l observation in jth batch of system i, $\frac{1}{l}\sum_{p=1}^{l} Y_{i,(j-1)m+p}$
b	index of best performing system
$\delta_{b,i}$	performance difference between system i and best performing system, $\bar{Y}_b - \bar{Y}_i$

Time Dilation grounds in the idea of redefining simulation experiments to include both the models and the systems being simulated. As a result of this, the simulation experimental unit of effort is rethought as real time, rather than discrete runs, batches, regenerative cycles, etc. (Schruben et al. 2003, Swisher

et al. 2000). This has far–reaching implications. In an integrated simulation experiment, each point in the experiment does not need to use the same units to measure the simulated time (say, hours or minutes), nor do the time scales need to be constant (Schruben 2010, Schruben 2013).
Hence, systems that are performing better can have their simulated time scales contracted. The result is that less simulation work is spent simulating the poorer performing systems; the better performing systems are simulated longer during the same interval of real time. This general idea has been called time dilation (Schruben 1997, Hyden et al. 2001, Hyden and Schruben 2000).
The clocks for design points that are performing relatively poorly can have their time scales dilated (increased), decreasing the relative numbers of events that are executed for that design point. More effort is spent simulating the winners than losers.

$$(1/\lambda_i) \cdot v_i\,[GTU] = 1/\lambda_i\,[LTU] \tag{1}$$

Equation (1), formalizes the time–scale update. If $1/\lambda_i$ represents the inter–arrival time and the global time unit (i.e., GTU) is $[hours]$, we might want to move the time unit of a good system to minutes by multiplying the inter–arrival by the conversion factor $v_i = \frac{[LTU]}{[GTU]}$, i.e., the time scale, with $v_i = 1/60$ in this example. By doing so, the simulator with associated time scale $v_i = 1/60$ will execute events 60-*times faster* than the candidates with associated time scale $v_j = 1$. This is the basic idea of time dilation and is the mechanism we exploit to differentiate computational budget among different solutions, instead of choosing the number of replications. Mapping time–dilated performance back to global simulated time is done simply by dividing each inter–event interval by the time–scale used simulating each point during that interval.

Sequential techniques that change the probability of where to run the next simulation appear to be well–suited for adaptation to the experimental unit of run time. In fact, the allocation ratio generated by a simulation allocation rule can be used to modify the time scale assigned to a specific simulation experiment. Optimal Computing Budget Allocation (OCBA) is one of the most successful techniques in the field of stochastic assignment of simulation budget (Chen and Lee 2010). OCBA formulates the allocation decision as a constrained maximization problem of the *Probability of Correct Selection* (PCS) subject to the total budget limitation:

$$\begin{aligned} &\max_{N_1,\ldots,N_k} PCS \\ &s.t.\ N_1 + N_2 + \ldots + N_k = T. \\ &N_i \in \mathbb{N}, i = 1,\ldots,k. \end{aligned} \tag{2}$$

where *PCS* is the probability that the *observed* best system is the system with best expected performance.

We can solve the aforementioned problem and derive the following allocation (Chen and Lee 2010):

$$\begin{aligned} &\frac{N_i}{N_j} = \left(\frac{\sigma_i/\delta_{b,i}}{\sigma_j/\delta_{b,j}}\right)^2, i,j \in 1,2,\ldots,k, and\ i \neq j \neq b, \\ &N_b = \sigma_b \sqrt{\sum_{i=1, i\neq b}^{k} \frac{N_i^2}{\sigma_i^2}} \end{aligned} \tag{3}$$

In order to fully utilize advantages from both TD and OCBA, in Pedrielli, Zhu, and Lee (2015), we employ the rule from OCBA to update the time scale in time-dilation. In order to do so, we need to compute the time scale ratio instead of reasoning in terms of number of replications as in equation (3). If we refer to α_i as the ratio of computational budget allocated to system i, i.e., $\alpha_i = \frac{N_i}{T}$, we can reformulate equation (3) as:

$$\frac{\alpha_i}{\alpha_j} = \left(\frac{\sigma_i/\delta_{b,i}}{\sigma_j/\delta_{b,j}}\right)^2, i,j \in 1,2,...,k, and\ i \neq j \neq b,$$
$$\alpha_b = \sigma_b \sqrt{\sum_{i=1, i\neq b}^{k} \frac{\alpha_i^2}{\sigma_i^2}}, \tag{4}$$
$$\sum_{p=1}^{k} \alpha_p = 1.$$

and we put forward a more general single–run ordinal optimization procedure, TD-OCBA, integrating TD and OCBA.

In order to implement TD-OCBA, each system is simulated according to a local clock and the overall simulation has a shared clock (global clock). Initial events for all systems are scheduled at the same time. And subsequent events will be scheduled to a common clock time stamp, by adding, timespan to wait for next event in local clock times the time scale, to the current common clock time stamp. The overall simulation always execute the event with earlier common clock time stamp. Update for the time scales happens firstly when each of the candidates has generated at least one batch of output data. Then the update will happen whenever there is a system that outputs m new observations. Finally the procedure will stop when total number of observations for all systems reaches T. The system with best observed performance will be selected. As previously highlighted, if a system is allocated a larger time scale v_i, the time between events will be increased (*dilated*) as well. As a result, being all systems subject to the same global time clock, larger time scale configurations will be simulating less events given a fixed simulation interval.

Algorithm 1: Time Dilated Optimization & OCBA (TD-OCBA).

Initialization: Set the total budget T (number of observations);
$l = 0, \nu_i \leftarrow \nu_0, d_i = 0$;
while $\sum_{i=1}^{k} N_i \leq T$ **do**
  for $i = 1,...,k$ *(parallel loop)* **do**
    Observe the sequence $\{Y_{i,d_i m+1}, Y_{i,d_i m+2}, ..., Y_{i,d_i m+m}\}$;
    Update: $\bar{Y}_i = \sum_{j=1}^{d_i m+m} \frac{Y_{i,j}}{d_i m+m}$;
    $\hat{\sigma}_i \leftarrow \sqrt{\text{variance estimation for system } i}$;
    $N_i \leftarrow N_i + m$;
    $d_i \leftarrow d_i + 1$;
  end
  Choose b s.t. $b \in \arg\min_{i=1,...,k} \bar{Y}_i$;
  $v_i \leftarrow \alpha_b/\alpha_i$, $i = 1,...,k$, determined in (4);
end

This procedure has been shown to have better performance than original TD and it can be applied to more general problems. At this point, we want to further develop the TD-OCBA procedure to improve the efficiency.

In order for the time scales to be correctly assigned, performance of the variance estimation is critical. For single–run simulations, the assumption of i.i.d. on the observations is not valid for most cases. Thus variance estimators that do not assume correlation structure are more suitable to be employed in the TD-OCBA procedure. To the best of authors' knowledge, Conway (1963) first concerned about variance estimation for simulation and since then it has been an interest field of statistical analysis. With techniques of batching and overlapping, several estimators have been developed in Schmeiser (1982), Goldsman et al.

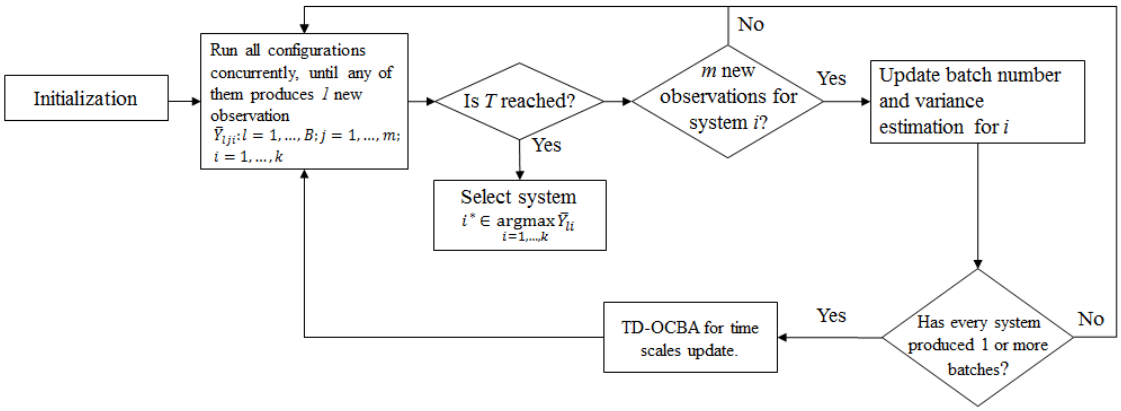

Figure 1: Flow chart for Algorithm 1.

(1990), Goldsman et al. (1999), Alexopoulos et al. (2007). While the asymptotic consistency of these estimators has been well studied, yet what affect the performance of TD-OCBA is the accuracy of the estimator. Intuitively, given same budget, TD-OCBA running with estimator of more accurate estimation should produce higher PCS. We are interested in this argument and thus this is the focus for this study.

3 VARIANCE ESTIMATION IN SINGLE RUN SIMULATION OPTIMIZATION

The interest of further studying the TD-OCBA lies in improving convergence rate of PCS. It is realized that this problem can be decomposed into 2 layers. The first layer problem is the allocation rule. We have not confirmed that use of original OCBA ratio guarantees optimized allocation for PCS. The optimization can be formulated as:

$$\begin{aligned} &\max_{\alpha_1,\ldots,\alpha_k} PCS \\ &s.t.\ \alpha_1 + \alpha_2 + \ldots + \alpha_k = 1. \\ &\alpha_i > 0, i = 1, \ldots, k. \end{aligned} \tag{5}$$

The second layer problem is concerned with the estimation for the variance associated to the simulation output since we do not know the correlation structure of the time series output generated by the simulator. Arguably, the more accurate the estimator is, the closer the time scale allocation is to the asymptotic optimal allocation.

While we need to use an asymptotically consistent indicator, we are also interested in the variance associated to this estimator. The first is a well studied problem, and several consistent estimators have been proposed. Alexopoulos et al. (2007) notes 3 aspects to assess the performance of variance estimator, namely: (1) bias, (2) variance, and (3) MSE (sum of variance and the square of bias).

The main control parameters to estimate the variance from a series of observations are:

- B indicating the number of batches;
- m indicating the batch size;
- ρ indicating the overlapping ratio, i.e., the portion of observations from the previous batch, which are *re*-used to form current batch;
- $\mathscr{E} = \{1, 2, \ldots, 5\}$ the estimator type, chosen among the available estimators and applied to all systems.

Now the second layer problem becomes, assigned the budget from level one, how to set the aforementioned parameters in order to minimize the estimated MSE associated with the output, namely, for each system

$i = 1, \ldots, k$:

$$P_i : \min_{B_i, m_i, \rho_i, \mathscr{E}_i} \hat{MSE}\left(\hat{\sigma}_i^2\right) \tag{6}$$

If the bias has been widely analyzed, the variance of such indicators has not been fully studied, however it impacts on the optimization performance. Indeed, given the limited budget of observations, TD-OCBA is sensitive to the accuracy of the estimators.

In this paper we investigate the effect of the bias and the variance of the estimators on the performance of TD-OCBA, which represent our second layer problem.

In this work. we choose estimators with known bias and variance are valuable for the assessment. In the search in literature, three families of estimators, namely batch mean estimator, weighted area estimator and CvM estimator and their overlapped form appear to be of particular interest.

3.1 Batch Mean Estimator

Non–overlapping Batch Mean (NBM) estimator was first discussed in Schmeiser (1982). Instead of using the standard variance estimator, the author groups observations into batches and estimates the variance of the sequence using the variances of the sample average computed for each batch. For each system configuration i, the estimator takes the following form:

$$\hat{\sigma}_i^2 = \frac{m}{B_i - 1} \sum_{k=1}^{B_i} (\bar{Y}_{i,k} - \frac{1}{B_i} \sum_{l=1}^{B_i} \bar{Y}_{i,l})^2. \tag{7}$$

This estimator represents the intuitive extension of the sample variance were the i.i.d. observations are replaced by the means of the different batches and the sample average is still the average of all the observations. We will refer to match mean estimator as $\mathscr{N}$.

3.2 Standardized Time Series

Schruben (1983) proposes the concept of *standardized time series*. For each system, the standardization involves converting the time series $Y_{i,1}, ..., Y_{i,n}$ into B batches with m outputs $(Y_{i,1}, ..., Y_{i,m})$,$(Y_{i,m+1}, ..., Y_{i,2m})$,..., $(Y_{i,(B_i-1)m+1}, ..., Y_{i,B_i m})$ and performing the following transformation:

$$T_{i,j}(t) \equiv \frac{\lfloor mt \rfloor (\bar{Y}_{i,j,\lfloor mt \rfloor} - \bar{Y}_{i,j,m})}{\sigma \sqrt{m}} \tag{8}$$

for $0 \leq t \leq 1$ and $j = 1, 2, ..., B_i$, where

$$\bar{Y}_{i,j,l} \equiv \frac{1}{l} \sum_{g=1}^{l} Y_{i,(j-1)m+g}$$

Based on this revolutionary idea, several estimators for the variance of time series observations have been proposed. We focus on two particular estimators: (1) the Weighted Area Estimator in section 3.3 and the Batched Cramer Von Mises Estimator presented in section 3.4.

3.3 Weighted Area Estimator

Goldsman et al. (1990) introduced the Batched area estimator (BAE) with the following form:

$$\hat{\sigma}_i^2 = \frac{1}{B_i} \sum_{j=1}^{B_i} \left\{ \frac{1}{m} \sum_{h=1}^{m} \left(f\left(\frac{h}{m}\right) \frac{h}{\sqrt{m}} \left(\frac{1}{h} \sum_{l=1}^{h} Y_{i,(j-1)m+l} - \frac{1}{m} \sum_{l=1}^{m} Y_{i,(j-1)m+l} \right) \right) \right\}^2. \tag{9}$$

where f is the weighting function that satisfies: $f(t)$ is continuous on interval $[0,1]$ and normalized so that $Var(\int_0^1 f(t)\sigma\mathscr{B}_0(t)dt) = 1$. $\mathscr{B}_0$ is a Brownian bridge process on $[0,1]$.

We will refer to BAE taking $f(t) = \sqrt{12}$ as $\mathscr{A}(f_0)$.

3.4 CvM Estimator

Batched CvM estimator(CvM) has been introduced in Goldsman, Kang, and Seila (1999). It takes the following form:

$$\hat{\sigma}_i^2 = \frac{1}{B_i}\sum_{j=1}^{B_i}\left\{\frac{1}{m}\sum_{h=1}^{m} g\left(\frac{h}{m}\right)\left(\frac{h}{\sqrt{m}}\left(\frac{1}{h}\sum_{l=1}^{h} Y_{i,(j-1)m+l} - \frac{1}{m}\sum_{l=1}^{m} Y_{i,(j-1)m+l}\right)\right)^2\right\}. \tag{10}$$

where g is the weighting function that satisfies: g has a continuous and bounded second derivative on $[0,1]$ and is normalized so that $E[\int_0^1 g(t)\sigma^2\mathscr{B}_0^2(t)dt] = \sigma^2$.

We will refer to CvM taking $g(t) = 6$ as $\mathscr{C}(g_0)$.

3.5 Overlapping Estimators

As the name suggests, these estimators involve overlapping of the batches. For non-overlapping estimators, batches are formed by each m new observations. In Alexopoulos et al. (2007), batches for overlapping estimators consist of $m-1$ observations from the most recent batch and only 1 observation is added to form a new batch.

In the TD-OCBA approach, such an update would require an excessive computational load. Therefore, we assume that the overlapping is determined by a ratio that the user (or the algorithm itself) can set statically or dynamically.

Hence, we introduce the overlapping ratio parameter o. Differently from the approach proposed in Alexopoulos et al. (2007), a new batch is formed from $\lfloor om \rfloor$ new outputs and $\lceil (1-o)m \rceil$ from the last batch. The overlapping ratio proposed in Alexopoulos et al. (2007) is $\frac{m-1}{m}$.

All the estimators presented in the previous sections can be applied to overlapped batches without any modification in the formulation.

While all the presented estimators are asymptotically unbiased, their bias terms and variance terms differ. Table 2 is taken from Alexopoulos et al. (2007). These estimators are chosen as they have exact approximate bias and variance.

Table 2: Approximate asymptotic bias and variance for estimators.

Nonoverlapping	(m/γ) Bias	(b/σ^4) Var	Overlapping ratio = $\frac{m-1}{m}$	(m/γ) Bias	(b/σ^4) Var
$\mathscr{N}$	1	2	$\mathscr{N}$	1	1.333
$\mathscr{A}(f_0)$	3	2	$\mathscr{A}(f_0)$	3	0.686
$\mathscr{C}(g_0)$	5	0.8	$\mathscr{C}(g_0)$	5	0.419

From Table 2, we can observe that for $\mathscr{N}$, $\mathscr{A}(f_0)$ and $\mathscr{C}(g_0)$, both approximated bias and variance are provided. Hence they are chosen to be incorporated in the TD-OCBA procedure to check the procedure performance (required budget) against (m/γ) Bias and (b/σ^4) Var. The batch size m is also varied as we are interested to see its effect on the budget.

4 NUMERICAL EXPERIMENTS & RESULTS

In this section, we use theoretical problems (i.e., problems for which we know the optimum as well as the true variance of the output for each candidate solution) to test the performance of the estimators. In particular, we generate two set of auto-regressive time series:

s1: $Y_{i,t+1} = c_i + 0.5Y_{i,t} + N(0,3)$, where $c_i = i$, $i = 1,...,10$.

s2: $Y_{i,t+1} = c_i + 0.5Y_{i,t} + N(0,6)$, where $c_i = i$, $i = 1,...,10$.

where the first group is considered to have small noise and second to have large noise.

The time series with smallest asymptotic mean is to be selected. The performance of an estimators is assessed by looking at the budget required to reach a fixed PCS level.
The statistic of the PCS is obtained out of 1000 macro-replications of the TD-OCBA algorithm. In the experiments, we observed that additional computational time required for the time scale update computation is negligible. This is particularly true when the simulation is significantly slow (i.e., the largest part of the computational time is required by the simulator itself). Hence, we can empirically conclude that the computation of estimator does not represent a computational burden in a realistic simulation optimization problem (realistic in terms of computational effort required by the simulation).
In the results, the budget reported is the average of the observations required for each of the macro-replications required to obtain a threshold PCS level. In the experiments we study the effect of the variance estimators and the batch size as well as the overlapping ratio (refer to Problem (6)).

Tables 3 and 4 show the results of the relative experiments.

Table 3: Budget for PCS to reach 0.95 for series 1.

	Nonoverlapping			Overlapping ratio = 0.2			Overlapping ratio = 0.5			Overlapping ratio = 0.8		
m	$\mathscr{N}$	$\mathscr{A}(f_0)$	$\mathscr{C}(g_0)$	$\mathscr{N}$	$\mathscr{A}(f_0)$	$\mathscr{C}(g_0)$	$\mathscr{N}$	$\mathscr{A}(f_0)$	$\mathscr{C}(g_0)$	$\mathscr{N}$	$\mathscr{A}(f_0)$	$\mathscr{C}(g_0)$
5	357.2	215.2	209.5	367.1	216.7	207.4	361.6	213.3	209.1	341.3	211.3	206.1*
10	239.4	229.1	212.9	230.0	226.7	212.3	234.1	224.5	211.2	221.5	216.8	209.0$^+$
20	278.6	296.3	278.9	277.2	292.5	278.4	281.5	291.7	275.7	281.6	284.4	274.2$^+$
50	505.5	505.4	504.9	505.2	505.8	507.0	506.1	506.7	505.0	505.8	504.7$^+$	506.7

Table 4: Budget for PCS to reach 0.90 for series 2.

	Nonoverlapping			Overlapping ratio = 0.2			Overlapping ratio = 0.5			Overlapping ratio = 0.8		
m	$\mathscr{N}$	$\mathscr{A}(f_0)$	$\mathscr{C}(g_0)$	$\mathscr{N}$	$\mathscr{A}(f_0)$	$\mathscr{C}(g_0)$	$\mathscr{N}$	$\mathscr{A}(f_0)$	$\mathscr{C}(g_0)$	$\mathscr{N}$	$\mathscr{A}(f_0)$	$\mathscr{C}(g_0)$
5	1670.5	783.4	590.1$^+$	2037.4	913.5	623.0	2983.2	985.0	664.1	2984.9	882.7	648.8
10	785.8	680.7	506.7	738.3	672.2	499.1	793.2	641.3	488.8*	836.3	593.9	490.8
20	587.6	631.2	513.0	560.8	633.2	505.0	595.6	602.3	501.7	603.9	565.4	498.0$^+$
50	690.6	743.5	695.7	677.3$^+$	740.0	688.7	688.8	731.7	686.6	700.8	722.1	677.8

In Table 3 and 4, highlighted cell contains the best performer of the 3 estimators for the same batch size and overlapping ratio. Non-overlapping can be inferred as having $o = 0$. $^+$ indicates the best performance in one row, thus for same batch size, comparing both overlapping and non-overlapping ratio cases. * denotes the overall best performance both in the row and table, i.e., comparing against all the parameters.
Similar experiments have been conducted for the original TD-OCBA in (Pedrielli et al. 2015) that uses the sample variance as estimator. In the original settings, the time scales are updated at *each* observation and no batching is adopted. In this setup, the original algorithm required an average budget of 248.2 observations to achieve a PCS of 0.95 for the case of time series 1, and an average budget of 881.3 to achieve a PCS level of 0.90 for the case of time series 2. It is noteworthy that, even though the original algorithm has an updating frequency about 1 order of magnitude larger than the batch-based algorithm, the original TD-OCBA with sample variance estimator is outperformed by TD-OCBA with $\mathscr{C}(g_0)$ when a proper batch size is chosen.

Further observations comparing the selected estimators can be drawn from the results.

The first is that the budget to reach fixed PCS is indeed related to the variance estimator. It is found that budget requirement varies by quite much using different estimator. From highlights and marks, it is noticed that $\mathscr{C}(g_0)$ performs better then $\mathscr{N}$ and $\mathscr{A}(f_0)$ in most cases.

The interesting part is that, according to the theoretical results presented in (Alexopoulos, Argon, Goldsman, Tokol, and Wilson 2007), $\mathscr{C}(g_0)$ has the largest Bias (m/γ) and the lowest variance ((b/σ^4)).

In other words we observe a positive relationship between the required budget and the variance of the indicator, whereas the bias does not seem to influence the performance.

By comparing the performance of overlapping and non-overlapping estimators, the overlapping version performs better most of the times. The overlapping coefficient impacts TD-OCBA in two ways: (1) the frequency at which the time scale is updated. From the description of the TD-OCBA procedure, it is known that time scales gets updated as soon as one new batch is formed, which results in the change in variance estimation. So for experiments using overlapping estimators, the time scale get updated more often; (2) overlapping estimators have lower associated variance than their non-overlapping counterpart and this may contribute to the difference in efficiency.

Lastly, the batch size is affecting the performance as well. The impact of the batch size can be well understood as it does not only affect the number of batches for each estimator, but it contributes to affect the frequency as which time scales are affected. Nevertheless, the budget to reach fixed PCS is not monotonic with the batch size. Though the estimator converges to true variance given infinitely large batch size, larger batch size results in less number of update of time scale for limited budget. In other words given finite budget, we have to trade off the batch estimator precision with the number of batch observations that can be obtained.

5 CONCLUTIONS & FUTURE WORK

Ranking and Selection has acquired an important role in the Simulation-Optimization field, where the different alternatives can be evaluated by discrete event simulation (DES). In fact, black box approaches have dominated the literature by interpreting the DES as an oracle providing i.i.d. observations.

Another relevant family of algorithms, instead, considers an oracle which produces time series observations as each simulator is ran only once (i.e., no multiple replicates are performed). In this paper, we focus on a specific single run sequential stochastic selection method (S^3M) recently developed by the authors, Time Dilation with Optimal Computing Budget Allocation (TD–OCBA).

As a single run approach, one of the most critical aspects of TD–OCBA is to deal with the estimate of the response given a series of correlated observations. In fact, in this paper, we derive a two-stage optimization structure for the class of single-run R&S algorithms. In particular, at a first stage we choose how to allocate the simulation budget to each candidate solution while at the second stage we want to minimize the MSE of the variance estimator playing with the estimator type the batch size and the overlapping ratio. Due to the difficulty in the estimation of the MSE, in this manuscript, we study the empirical performance of several variance estimators, different batch sizes as well as overlapping ratios in order to enhance the TD-OCBA procedure.

We propose an autoregressive process as case study and we observe a positive relationship between the asymptotic variance of the estimators and the required budget to reach fixed PCS values.

Also, generally, the overlapped estimators perform better than the non-overlapping counterpart. This is due to the fact that, given the same budget, more observations are available to the optimization procedure and this results in a positive effect in the search algorithm.

The required budget is, instead, non-monotonic in the batch size m. This is due to the fact that if, on the one hand larger batch size reduce the variance of the estimator, they also reduce the number of observations available in the scope of the search procedure, thus giving issues to the optimization.

Current work is focusing on the formulation of the second stage problem in order to enhance the TD-OCBA through an automated procedure that checks the MSE and returns a dynamic value for overlapping ratio and batch size.

Also, TD-OCBA opens to alternative first level problem formulation: since single-run procedures are easier to control in terms of running time, whereas the number of observations produces in a certain interval is typically a random variable. In such a setting, in addition to the probability of correct selection, we would like to formulate the problem in the framework of regret due to the fact that we do not know how many iterations the algorithm will produce in a specific run-time. As a result, we want to guarantee a good solution at any point in time.

REFERENCES

Alexopoulos, C., N. T. Argon, D. G. Goldsman, G. Tokol, and J. R. Wilson. 2007. "Overlapping Variance Estimators for Simulation". *Operations Research* 55 (6): 1090–1103.

Chen, C. H., and L. H. Lee. 2010. *Stochastic Simulation Optimization: An Optimal Computing Budget Allocation*. World scientific.

Conway, R. W. 1963. "Some Tactical Problems in Digital Simulation". *Management Science* 10 (1): 47–61.

Goldsman, D., K. Kang, and A. F. Seila. 1999. "Cramer-Von Mises Variance Estimators for Simulations". *Operations Research* 47 (2): 299–309.

Goldsman, D., M. Meketon, and L. Schruben. 1990. "Properties of Standardized Time Series Weighted Area Variance Estimators". *Management Science* 36 (5): 602–612.

Goldsman, D., and B. L. Nelson. 2006. "Chapter 15 Correlation-Based Methods for Output Analysis". In *Simulation*, Volume 13 of *Handbooks in Operations Research and Management Science*, 455 – 475. Elsevier.

Hyden, P., and L. W. Schruben. 2000. "Improved Decision Processes Through Simultaneous Simulation and Time Dilation". In *Proceedings of the 2000 Winter Simulation Conference*, edited by K. K. J. A. Joines, R. R. Burton and P. A. Fishwick, 743–748. Orlando, Florida: Institute of Electrical and Electronics Engineers, Inc.

Hyden, P., L. W. Schruben, and T. Roeder. 2001. "Resource Graphs for Modeling Large-scale, Highly Congested Systems". In *Proceedings of the 2001 Winter Simulation Conference*, edited by D. J. M. B. A. Peters, J. S. Smith and M. W. Rohrer, 523–529. Arlington, Virginia: Institute of Electrical and Electronics Engineers, Inc.

Pedrielli, G., Y. Zhu, and L. H. Lee. 2015. "Single–run Simulation Optimization Through Time Dilation and Optimal Computing Budget Allocation". In *Proceedings of the 10th Conference on Stochastic Models of Manufacturing and Service Operations*, 187–194.

Schmeiser, B. 1982. "Batch Size Effects in the Analysis of Simulation Output". *Operations Research* 30 (3): 556–568.

Schruben, L. W. 1983. "Confidence Interval Estimation Using Standardized Time Series". *Operations Research* 31 (6): 1090–1108.

Schruben, L. W. 1997. "Simulation Optimization Using Simultaneous Replications and Event Time Dilation". In *Proceedings of the 1997 Winter Simulation Conference*, edited by D. H. W. S. Andradbttir, K. J. Healy and B. L. Nelson, 177–180. Atlanta, Georgia: Institute of Electrical and Electronics Engineers, Inc.

Schruben, L. W. 2010. "Simulation Modeling for Analysis". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 20 (1): 2.

Schruben, L. W. 2013. "Simulation Modeling, Experimenting, Analysis, and Implementation". In *Proceedings of the 2013 Winter Simulation Conference*, edited by A. T. R. H. R. Pasupathy, S.-H. Kim and M. E. Kuhl, 678–690. Washington, D.C.: Institute of Electrical and Electronics Engineers, Inc.

Schruben, L. W., T. M. Roeder, W. K. Chan, P. Hyden, and M. Freimer. 2003. "Advanced Event Scheduling Methodology". In *Proceedings of the 2003 Winter Simulation Conference*, edited by D. F. S. Chick, P. J. Snchez and D. J. Morrice, 159–165. New Orleans, Louisiana: Institute of Electrical and Electronics Engineers, Inc.

Swisher, J. R., P. D. Hyden, S. H. Jacobson, and L. W. Schruben. 2000. "A Survey of Simulation Optimization Techniques and Procedures". In *Proceedings of the 2000 Winter Simulation Conference*, edited by K. K.

J. A. Joines, R. R. Burton and P. A. Fishwick, 119–128. Orlando, Florida: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

GIULIA PEDRIELLI is Research Fellow for the Department of Industrial & Systems Engineering at the National University of Singapore. Her research focuses on stochastic simulation-optimization in both single and multiple–objectives framework. She is developing her research in meta-model based simulation optimization and learning for simulation and simulation optimization. Her email address is giulia.pedrielli.85@gmail.com.

YINCHAO ZHU is a research engineer in National University of Singapore. He received his B.Eng. degree and B.Sci in 2012 from Engineering Science Program and Department of Mathematics at National University of Singapore. He is currently pursuing his PhD degree in the Department of Industrial and Systems Engineering. He has research interests in operations research and simulation optimization. His email address is yinchao.zhu@u.nus.edu.

LOO HAY LEE is Associate Professor and Deputy Head for the Department of Industrial and Systems Engineering, National University of Singapore. He received his B. S. (Electrical Engineering) degree from the National Taiwan University in 1992 and his S. M. and Ph. D. degrees in 1994 and 1997 from Harvard University. He is currently a senior member of IEEE, a committee member of ORSS, and a member of INFORMS. His research interests include production planning and control, logistics and vehicle routing, supply chain modeling, simulation-based optimization, and evolutionary computation. His email address is iseleelh@nus.edu.sg.

HAOBIN LI is Scientist for the Institute of High Performance Computing, A*STAR Singapore. He received his B.Eng. degree (1st Class Honors) in 2009 from the Department of Industrial and Systems Engineering at National University of Singapore, with minor in computer science; and Ph.D. degree from the same department in 2014. He has research interests in operations research, simulation optimization and designing high performance optimization tools with application on logistics and maritime studies. His email address is lihb@ihpc.a-star.edu.sg.