THREE-STAGE SELECTION APPROACH WITH THE INITIAL SIMULATION SAMPLE SIZE

Mohammad H. Almomani¹§, Adam Baharum², Rosmanjawati Abdul Rahman³
School of Mathematical Sciences
Universiti Sains Malaysia
Penang, 11800, MALAYSIA

Abstract: The article studies the effect of the initial simulation sample size on the performance of Three-Stage selection approach. The aim of the study was to see the effects of the initial simulation sample size on the selection approach in selecting the best simulated designs when the number of alternatives system is large. In Three-Stage selection approach, the ordinal optimization technique is used in the first stage to select a subset that overlap with the set of the best m% designs with high probability, and then subset selection procedure is used to get a smaller subset that contains the best among the subset that is selected by the first stage. Finally, the indifference-zone procedure is used to select the best design among the survivors in the second stage. In fulfilling the aim of the study, the Three-Stage approach was implemented on M/M/1 queuing system under some parameter settings, with a different choice of the initial simulation sample sizes. The results show that the choice of the initial sample size does affect the performance of Three-Stage selection approach.

AMS Subject Classification: 62F07, 90C06, 49K30, 49K35
Key Words: ranking and selection, ordinal optimization, subset selection, indifference-zone, simulation optimization, initial simulation sample size

Received: June 17, 2011 © 2011 Academic Publications, Ltd.
§Correspondence author
1. Introduction

We are considering the problem of selecting the best design from a finite and large set of alternatives, where the expected value of each alternative can be inferred by simulation. Statistical selection procedures are used to identify the best of a finite set of simulation alternative designs, where the best design defined as the design that has the minimum (maximum) expected value of performance measures. In this paper, we study the effect of the initial simulation sample size on the performance of one of the selection approach that proposed by Almomani and Alrefaei [1], this approach known as a Three-Stage selection approach.

Three-Stage selection approach as present in Almomani and Alrefaei [1], Almomani et al [2] and Almomani et al [3] identify the best design by using three stages. At first stage, the Ordinal Optimization (OO) procedure is used to select a subset that overlap with the set of the best $m\%$ designs with high probability. At second stage, Subset Selection (SS) procedure is used to get a smaller subset that contains the best among the subset that was selected at the first stage before. Finally, the Indifference-Zone (IZ) procedure is used to select the best design among the subset from the second stage.

In many selection approaches, sample size in the first stage $t_1$ plays an important role to the performance of these approaches. In fact, the initial sample size $t_1$ cannot be too small since we might get poor estimates for the sample mean and variances. On the other hand, $t_1$ cannot be too large, because in the first stage there exist many noncritical designs and by giving a large number of samples will result in losing a large number of samples and also wasting computation time. However, Chen et al [4] and Chen et al [5] suggested that $t_1$ should be between 10 and 20 as a good choice for the initial simulation sample size. Unfortunately, there is no clear formula to calculate an appropriate value of the initial simulation sample size $t_1$ for the selection approaches, when the number of alternatives is large. To know the effects of the initial simulation sample size $t_1$ on the performance of Three-Stage approach, we are applied it on M/M/1 queuing design with a different choice of the initial simulation sample size $t_1$.

The paper is organized as follows; In Section 2, we present the problem statement. In Section 3, we review the OO, SS and IZ procedures. We present Three-Stage selection approach in Section 4, followed by a numerical illustration in Section 5. Finally, concluding remarks are presented in Section 6.
2. Problem Statement

We consider the following optimization problem

$$\min_{\theta \in \Theta} f(\theta)$$

where $f(\theta) = E[L(\theta, Y)]$, with $f$ is the expected performance measure of some complex stochastic design, $\Theta$ is arbitrary feasible solution set, that finite and has no structure, $\theta$ is a vector representing the system design parameters, $Y$ represents all the random effect of the design and $L$ is a deterministic function that depends on $\theta$ and $Y$.

The goal of the selection procedure is to identify the best $n$ simulated designs, where the best in this paper is defined as the design with the smallest mean, which is unknown and to be inferred from simulation. Suppose that there are $n$ designs. Let $Y_{ij}$ represent the $j^{th}$ output from design $i$, and let $Y_i = \{Y_{ij}, j = 1, 2, \ldots\}$ denote the output sequence from design $i$. We assume that $Y_{ij}$ are independent and identically distributed (i.i.d.) normal with unknown means $\mu_i = E(Y_{ij})$ and variances $\sigma^2_i = Var(Y_{ij})$. In addition, we assume that the $Y_1, Y_2, \ldots, Y_n$ are mutually independent. The normality assumption is not a problem here because simulation outputs are obtained from an average performance or batch means, so by using the Central Limit Theorem the normality assumption will hold. In practice the $\sigma^2_i$ are unknown, so we estimate it using sample variances $S^2_i$. Throughout this paper we assume that a smallest mean is better, so if the ordered $\mu_i$-values are denoted by $\mu[1] \leq \mu[2] \leq \ldots \leq \mu[n]$, then the design having mean $\mu[1]$ is referred to as the best design. The correct selection occurs when the design selected by the selection procedure is the same as the true best design.

3. Background

3.1. Ordinal Optimization (OO)

OO concentrates on isolating a subset of good designs with high probability and reduces the required simulation time dramatically for discrete event simulation. The OO procedure has been proposed by Ho et al [6]. Since then, it has emerged as an efficient technique for simulation and optimization. In this procedure, the aim is to find a good enough solution, rather than to estimate accurately the performance value of these designs.
If we simulate the design to estimate the expected performance measure, the confidence interval of this estimate cannot be improved faster than $1/\sqrt{n}$ where $n$ is the number of replications used, see Chen et al [4]. This rate may be good for many problems, when there is a small number of alternatives but it is not good enough for a complex simulation with a larger number of alternatives. The reason is that, since each sample requires one simulation run, we need a large number of samples when we are dealing with a huge number of alternative designs in the solution set which is very hard and may be impossible. In this case, one could compromise the objective to get a good enough solution rather than doing extensive simulation, which is impossible in many cases. However, in many real world applications, using a simpler model to get an accurate estimate is somehow impossible.

Suppose that the correct selection is to select a subset $G$ of $g$ designs from the feasible solution set $\Theta$ that contains at least one of the top $m\%$ best designs. Since we assume that $\Theta$ is very huge then the probability of correct selection is given by $P(CS) \approx (1 - (1 - \frac{m}{100})^g)$. Now, suppose that the correct selection is to select a subset $G$ of $g$ designs that contains at least $r$ of the best $s$ designs. Let $S$ be the subset that contains the actual best $s$ designs, then the probability of correct selection can be obtained using the hypergeometric distribution as, $P(CS) = P(\left| G \cap S \right| \geq r) = \sum_{i=r}^{g} \binom{g}{i} \binom{n-s}{g-i} / \binom{n}{g}$. However, we assume that the number of alternatives is very large then $P(CS)$ can be approximated by the binomial random variable. Therefore, $P(CS) \approx \sum_{i=r}^{g} \binom{g}{i} \left( \frac{m}{100} \right)^i \left( 1 - \frac{m}{100} \right)^{g-i}$, where we assume that $s/n \times 100\% = m\%$. It is also clear that the value $P(CS)$ increases when the sample size $g$ increases. More details of OO procedure can be found in Deng and Ho [7], Lee et al [8], Li et al [9], Zhao et al [10] and Ho et al [11].

3.2. Subset Selection (SS)

SS procedure screens out the feasible solution set and eliminates non-competitive designs and constructs a subset that contains the best design with high probability. We can use this procedure when the number of alternatives is relatively large to select (a random-size) subset that contains the actual best design. We require that $P(CS) \geq 1 - \alpha$, where the correct selection is selecting a subset that contains the actual best design, and $1 - \alpha$ is a predetermined probability. The SS procedure dating back to Gupta [12], who presented a single stage procedure for producing a subset containing the best design with a specified probability. Extensions of this work which is relevant to the simulation setting
include Sullivan and Wilson [13] who derived a two stage SS procedure that determines a subset of maximum size \( m \) that, with a specified probability will contain designs that are all within a pre-specified amount of the optimum.

### 3.3. Indifference-Zone (IZ)

The main aim of IZ procedure is selecting the best design among \( n \) designs when \( n \leq 20 \). Suppose we have \( n \) alternative designs that are normally distributed with unknown means \( \mu_1, \mu_2, \ldots, \mu_n \) and suppose that these means are ordered as \( \mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[n]} \). We seek to locate the design that has the best minimum mean \( \mu_{[1]} \). The IZ is defined to be the interval \([\mu_{[1]}, \mu_{[1]} + \delta]\), where \( \delta \) is a predetermined small positive real number. We are interested in selecting an alternative \( k \) such that \( \mu_k \in [\mu_{[1]}, \mu_{[1]} + \delta] \).

Let correct selection here is selecting an alternative whose mean belongs to the indifference zone. We would like the correct selection to take place with high probability, say with a probability not smaller than \( 1 - \alpha \) where \( 1/n \leq 1 - \alpha \leq 1 \). Care must be taken in specifying the value of \( \delta \), because the design requires to implicitly determine the common single-stage sample size required for the \( n \) competing designs. If \( \delta \) is too small, then the number of observations is expected to be large. Also when the \( 1 - \alpha \) is large then the number of observations is expected to be large.

The IZ procedure consists of two stages. In the first stage, all designs are sampled using \( t_1 \) simulation runs to get an initial estimate of expected performance measure and their variances. Next, depending on the information obtained in the first stage, we compute how many more samples are needed in the second stage for each design to guarantee that \( P(CS) \geq 1 - \alpha \). Rinott [14] has presented a procedure that is applicable when the data are normally distributed and all designs are simulated independently of each others. This procedure consists of two stages for the case when variances are completely unknown. Tamhance [15] has presented a simple procedure that is valid when variances may not be equal.

There are several combined selection procedures that presented to solve the selection problem when the number of alternatives large. Nelson et al [16] proposed a two-stage subset selection procedure. The first stage is to reduce the number of competitive designs. These designs are carried out to the second stage in which involved with the IZ procedure using the information gathered from the first stage. Alrefaei and Almomani [17] proposed two sequential algorithms for selecting a subset of \( k \) designs that is contained in the set of the top \( s \) designs. Another comprehensive review of SS and IZ procedures can be
found in Bechhofer et al [18], Goldsman and Nelson [19], Kim and Nelson [20] and Kim and Nelson [21].

4. Three-Stage Approach

This approach consists three stages; in the first stage, a subset $G$ is selected randomly from the feasible solution set $\Theta$. Let $g$ denotes the size of the subset $G$, where $g$ is a relatively small number, but the probability that $G$ contains one of the best $m\%$ alternatives is high. In the second stage, the SS procedure is applied on the subset $G$ to select a subset $I$ that contains the best design with high probability, where $|I| \leq 20$. Note that the size of the subset $G$ is small so we can use the SS procedure. Finally, the IZ procedure is applied on the set $I$ to select the best design. Note that the size of set $I$ is less than or equal 20, so that we can use the IZ procedure. The algorithm of the Three-Stage approach as proposed in Almomani and Alrefaei [1] is described as follows:

Step 0: Determine the size of the set $G$ that will be selected in the first stage that satisfies $P(G$ contains at least one of the best $m\%$ designs $) \geq 1 - \alpha_1$, $|G| = g$. Determine the number of samples $t_1 \geq 2$ to be observed for each design in the selected subset $G$, the indifference zone $\delta$. $\alpha_2$ and $\alpha_3$ are the significance level for the second and the third phase, respectively. Let $t = t_{1-\alpha/2, t_1-1}$ be the upper critical value of $t$-distribution with $t_1 - 1$ degree of freedom.

Step 1: Select a subset $G$ of size $g$ randomly from $\Theta$.

Step 2: For each design $\theta \in \Theta$, take a random sample of length $t_1$ observations, $Y_i(\theta) \ (1 \leq i \leq t_1)$.

Step 3: Calculate the first stage sample mean and variances $\bar{Y}^{(1)}(\theta)$ and $s^2(\theta)$ as, $\bar{Y}^{(1)}(\theta) = \frac{\sum_{i=1}^{t_1} Y_i(\theta)}{t_1}$ and $s^2(\theta) = \frac{\sum_{i=1}^{t_1} (Y_i(\theta) - \bar{Y}^{(1)})^2}{t_1-1}$.

Step 4: Let $W_{ij} = t \left( \frac{s_i^2}{t_1} + \frac{s_j^2}{t_1} \right)^{1/2}$ be the pooled variance for all $i \neq j$.

Step 5: Set $I = \{i : 1 \leq i \leq g \text{ and } \bar{Y}_i^{(1)} \leq \bar{Y}_j^{(1)} - [W_{ij} - \delta], \forall i \neq j\}$, where $[x]^− = x$ if $x < 0$ and $[x]^− = 0$ otherwise.

Step 6: If $I$ contains a single index, then this design is the desired design. Otherwise, for all $i \in I$, compute the second stage sample size $T_i =$
max\{t_1, [(h_{si}/\delta)^2]\}, where h = h(1 - \alpha/2, t_1, |I|) be the Rinott’s constant and can be obtained from tables of Wilcox [22].

**Step 7:** For each design \(i \in I\), take additional \(T_i - t_1\) random samples, \(Y(i)\) and compute the overall sample means \(\bar{Y}^{(2)}(i) = \frac{\sum_{j=1}^{T_i} Y_j(i)}{T_i}\) where \(Y_j(i)\) (\(1 \leq i \leq g, 1 \leq j \leq T_i\)) is the observed samples in the second phase for each design.

**Step 8:** Select as best the design \(i \in I\) with the smallest \(\bar{Y}^{(2)}(i)\).

**Remarks:**

- The initial simulation sample size \(t_1\) is the number of observations that are taken in the first stage (OO procedure) in order to get an initial estimate of mean and variance for each design. Note that, if \(t_1\) is too small, we might get a poor estimate of \(\sigma_i^2 (s_i^2)\). In particular, it could be that \(s_i^2\) is much greater than \(\sigma_i^2\), leading to an unnecessarily large value of \(T_i\).

- The Rinott’s constant \(h = h(1 - \alpha/2, t_1, |I|)\) is determined by the desired confidence level \((1 - \alpha/2)\), the initial simulation sample size \(t_1\), and the number of designs in the set \(I\) \(|I|\). From tables of Wilcox [22] we note that, the constant \(h\) increases in \(|I|\), and decreases in \(\alpha\) and \(t_1\). The experiment design factor that is under controlled is \(t_1\).

- Nelson et al [16] have shown that with probability at least \(1 - (\alpha_2 + \alpha_3)\) this approach selects the best design in the subset \(G\). Therefore, if \(G\) contains at least one of the top \(m\%\) designs, then this approach selects a good design with probability \(1 - (\alpha_2 + \alpha_3)\). From the OO, the selected set \(G\) from Step 1, contains at least one of the best \(m\%\) designs with probability \(\approx 1 - (1 - \frac{m}{100})^g = 1 - \alpha_1\), where \(g = |G|\). Therefore, \(P(\text{the selected design in Step 8 is in the top } m\% \text{ designs}) \geq 1 - (\alpha_1 + \alpha_2 + \alpha_3)\).

5. Numerical Example

We consider optimizing M/M/1 queuing designs where the inter arrival times and the service times are exponentially distributed and the design has one server. Suppose we have \(n\) M/M/1 queuing designs and we want to select one of the best \(m\%\) queuing designs that has the minimum average waiting time per customer, given that the arrival rate \(\lambda\) is fixed and the service rate \(\mu \in [a, b]\). We
implement the Three-Stage selection approach for solving this problem under some parameter settings.

Suppose we have 1000 queuing designs and we want to select one of the best 5% designs. Since the number of alternatives is large, we cannot use the ranking and selection (SS and IZ) procedures directly, so we use the OO procedure to choose a small subset $G$ that contains at least one from the top $m\%$ designs with high probability, then we use the SS procedure and the IZ procedure to select the best among the selected subset $G$.

In this example, we assume that $\lambda = 1$ and $\mu \in \Theta = \{4.0, 4.001, \ldots, 5.0\}$, which means we have 1000 different M/M/1 designs. In the first stage, we select a subset of 100 designs from $\Theta$, then the probability that at least one of the observed top 100 designs is one of the actual top 50 designs is approximately $1 - (1 - \frac{50}{1000})^{100} \approx 0.99$. Then we use the SS procedure to select a subset that contains the best designs among the 100 designs with probability greater than or equal 0.995, then use the IZ procedure to select the best design among the survivals of the SS with probability greater than or equal 0.995. Therefore, the probability of correct selection (selecting a design that belongs to the best 5% designs) is greater than or equal 0.98.

Table 1 contains the results of this experiment, when 10 replications are used for selecting one of the top 5% designs, $|I|$ is the number of designs survived in stage 2 (the subset selection), $T$ is the total samples used in each run, $t_1$ is the initial simulation sample size used to estimate the objective function values for all members of $G$, and “Best” means the index of the design that has been selected by Three-Stage algorithm as the best design. It is clear that this algorithm gets the correct selection always. Note that the analytical probability of correct selection is $P(CS) \geq 0.98$.

In the next experiment, we select a subset $G$ that contains 50 designs in the first stage by OO, the probability that at least one of the observed top 50 designs is one of the actual top 50 designs is approximately $1 - (1 - \frac{50}{1000})^{50} \approx 0.92$. Then, we use the SS procedure to select a subset that contains the best design from these 50 designs with probability greater than or equal 0.995, we use the IZ procedure to select the best design among the survivals of the SS with probability greater than or equal 0.995. Then the probability of correct selection is greater than or equal 0.91. Table 2 contains a result of the 10 replications of this experiment.

These two experiments are repeated 100 replications. Note that, in the first experiment we use $t_1 = 50$ to guarantee that $|I| \leq 20$, but for the second experiment we use $t_1 = 20$ guarantees that $|I| \leq 20$ since $G$ is smaller in the second experiment. Table 3 contains the probability of correct selection. It is
clear that the probability of correct selection is very high in both experiments.

<table>
<thead>
<tr>
<th>g</th>
<th>$\bar{T}$</th>
<th>Three-Stage $P(CS)$</th>
<th>Analytical $P(CS)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>191086</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>50</td>
<td>22436</td>
<td>0.96</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Table 3: The numerical illustration for n=1000, m%=5%

To see the effect of $t_1$ on the Three-Stage selection approach we use $t_1 = 10$, 

Table 1: The numerical illustration for n=1000, g=100, m%=5%

Table 2: The numerical illustration for n=1000, g=50, m%=5%
Table 4 and Table 5 contain the results of 10 replications when \( g = 100 \) and 50 and we want to select one of the best 5% designs from \( \Theta \), and Table 6 contains the results of 10 replications when \( g = 100 \) and we want to select one of the best 1% designs from \( \Theta \), respectively.

| Replications | \(|I|\) | \(t_1\) | \(T = \sum T_i\) | Best |
|--------------|--------|--------|-----------------|------|
| 1            | 72     | 10     | 60980           | 988  |
| 2            | 64     | 10     | 66890           | 989  |
| 3            | 72     | 10     | 62420           | 990  |
| 4            | 69     | 10     | 64470           | 987  |
| 5            | 58     | 10     | 57770           | 994  |
| 6            | 54     | 10     | 62480           | 955  |
| 7            | 73     | 10     | 63260           | 964  |
| 8            | 72     | 10     | 56340           | 989  |
| 9            | 74     | 10     | 58350           | 984  |
| 10           | 55     | 10     | 59350           | 969  |

Table 4: The numerical illustration for \( n=1000, g=100, m\%=5\% \)

| Replications | \(|I|\) | \(t_1\) | \(T = \sum T_i\) | Best |
|--------------|--------|--------|-----------------|------|
| 1            | 26     | 10     | 28920           | 969  |
| 2            | 23     | 10     | 34040           | 422  |
| 3            | 28     | 10     | 31470           | 995  |
| 4            | 35     | 10     | 36370           | 999  |
| 5            | 32     | 10     | 38080           | 990  |
| 6            | 29     | 10     | 32710           | 984  |
| 7            | 26     | 10     | 31740           | 958  |
| 8            | 32     | 10     | 35790           | 995  |
| 9            | 36     | 10     | 32550           | 994  |
| 10           | 26     | 10     | 30810           | 58   |

Table 5: The numerical illustration for \( n=1000, g=50, m\%=5\% \)

When we fixed \( t_1 = 10 \) in Tables 4, 5 and 6, we note that the size of the set \( I \) is large (i.e. \(|I| \geq 20\)) so we cannot apply the IZ procedure in the set \( I \), since we cannot find the constant \( h \) when \(|I| \) is large. We approximate the constant \( h \) and we apply Three-Stage algorithm, we found that in one replication the
Replications | $|I|$ | $t_1$ | $T = \sum T_i$ | Best
--- | --- | --- | --- | ---
1 | 72 | 10 | 63450 | 990
2 | 77 | 10 | 59320 | 989
3 | 49 | 10 | 66740 | 987
4 | 72 | 10 | 60390 | 981
5 | 70 | 10 | 65110 | 982
6 | 69 | 10 | 57320 | 980
7 | 58 | 10 | 62610 | 984
8 | 65 | 10 | 60040 | 984
9 | 75 | 10 | 61310 | 967
10 | 58 | 10 | 58530 | 993

Table 6: The numerical illustration for $n=1000$, $g=100$, $m\%=1\%$

selected design is far away from the best design as in Table 5, and we found that the $P(CS)$ is very different from the analytical $P(CS)$. For example, in the experiment that is described in Table 6 we find that $P(CS) = 0.21$ where the analytical $P(CS) = 0.62$. So we need to determine the number of replications $t_1$ to guarantee that $|I| \leq 20$ and then we can use IZ procedure on the set $I$.

6. Conclusion

In this paper, we discuss the effect of the initial simulation sample size $t_1$ on the performance of Three-Stage selection approach that is used to selecting a good simulated design, when the number of alternatives is large. In the first stage, we use the OO procedure to select a relatively small subset $G$ where the intersection with the set that contains the actual best $m\%$ alternatives is high. Then SS procedure are used to select a smaller subset $I$ that contains the best alternative of $G$ where $|I| < 20$. These are followed by the IZ procedure to select the best alternative of $I$ which highly results in selecting an alternative that belongs to the top $m\%$ alternatives. We apply Three-Stage selection approach in M/M/1 queuing design under some parameters setting, and we test different values for $t_1$, to study the effect of $t_1$ on Three-Stage approach. We note that Three-Stage approach is affected by $t_1$ in different rooms. The size of set $I$ is large so it is not adequate to apply the IZ procedure in the set $I$. We approximate the constant $h$ and we apply Three-Stage algorithm, we found that the value of $P(CS)$ is very different from the analytical $P(CS)$. In a future work a “zeroth
stage" of sampling should be added to the Three-Stage selection approach in order to determine an adequate choice for the value of the initial simulation sample size $t_1$ for each design.

Acknowledgments

The researchers would like to thank School of Mathematical Sciences, USM and the USM fellowship scheme for the financial support.

References


