Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

FIXED-PRECISION RANKING AND SELECTION AS MARKOV DECISION PROCESS

Ruihan Zhou[1], and Yijie Peng[1]

[1]Wuhan Institute of Artificial Intelligence & Guanghua School of Management, Peking University, Beijing, CHINA

ABSTRACT

In this study, we conceptualize the fixed-precision ranking and selection (R&S) problem as a stochastic control problem and subsequently model it using Markov Decision Process (MDP). Our approach aims to study the fixed-precision paradigm of R&S within a stochastic dynamic programming framework. To address the fixed-precision R&S challenge, we employ AlphaRank, an innovative artificial intelligence technique for tackling fixed-budget R&S problems. This procedure intelligently handles learning and decision-making through deep reinforcement learning, thereby addressing the R&S problem where the mean differences between various alternatives tend to approach zero. We use a numerical example to illustrate the efficacy of AlphaRank in solving fixed-precision R&S problems. Notably, this method mitigates, to a certain extent, the issues faced by traditional fixed-precision programs, which often require excessive sampling to reach a specified accuracy level.

1 INTRODUCTION

In decision-making scenarios, a common challenge involves evaluating a set of alternatives based on specific performance metrics. This study focuses on such comparative analyses with the objective of identifying the best alternative, defined by having either the largest (or smallest) mean performance. This task is particularly complex in stochastic environments where the mean performances of these alternatives are unknown and must be estimated through statistical sampling from stochastic systems. Consequently, a systematic selection procedure is essential to ascertain the required number of samples from each alternative and subsequently determine the best alternative based on the collected sample information. These kinds of decision-making problems are frequently referred to as ranking and selection (R&S).

R&S problems dated from the 1950s, primarily in the fields of agriculture and clinical studies (Bechhofer 1954). During this era, it is customary to test for the homogeneity among multiple alternatives, such as comparing grain yields or the efficacy of different drug treatments. A typical inquiry might involve assessing whether various grains yielded identical average outputs or if multiple drug therapies achieve the same average effectiveness. When the homogeneity of their means is rejected statistically, a pressing question emerges: which alternative is the best? This problem is initially put forth by Paulson (1949), sparking the early advances of R&S procedures.

R&S procedures can be classified into frequentist and Bayesian procedures, based on the probability models used for analyzing collected samples, as demonstrated in works like Chick (2006), Kim and Nelson (2006), and Branke et al. (2008). In contrast, this study adopts a different perspective, in line with Gabillon et al. (2012), Hunter and Nelson (2017), and Hong et al. (2021), by categorizing R&S problems into fixed-precision and fixed-budget procedures, which are differentiated by their distinct objectives. Particularly, fixed-precision procedures are designed to offer a statistical guarantee that the selected alternative is the best, or at least nearly so. Conversely, fixed-budget procedures focus on the strategic allocation of a given sampling budget in either optimal or approximately optimal manners. To delineate these two categories, Hong et al. (2021) illustrate that they essentially follow two different formulations, i.e., the hypothesis-testing and dynamic-programming formulations, respectively. This perspective has been widely recognized

and adopted in the literature, leading to the development of new procedures within these two formulations, as exemplified by the works of Batur and Choobineh (2012) and Peng et al. (2018).

Fixed-precision procedures are developed to provide a statistical guarantee that the optimal (or at least approximately optimal) alternative will be selected. These procedures aim to achieve a predefined selection probability of $1-\alpha$ for the best alternative. The challenge in achieving this objective of these procedures critically depends on the closeness of the mean values of the alternatives in competition. To solve this issue, Bechhofer (1954) and Rinott (1978) propose an indifference-zone (IZ) framework. Recent advances in the well-established IZ framework include Kim and Nelson (2001), Frazier et al. (2007), Ni et al. (2017), Luo et al. (2015), Hong et al. (2021), Zhong and Hong (2021) and many others. In this framework, parameter δ is introduced to define the smallest discernible difference. Under this paradigm, an IZ sampling procedure can choose either the best alternative or an alternative whose performance is within δ range with a probability of $1-\alpha$. Additionally, Fan et al. (2016) introduce an IZ-free procedure, enhancing the identification of alternatives across any range of mean differences. In fixed-precision R&S, the primary focus is on ensuring the statistical probability level selecting the best alternative, even though this often leads to allocating more simulation observations than necessary.

In the realm of fixed-budget procedures, the primary objective is to judiciously allocate simulation observations to maximize the efficacy in pinpointing the best alternative. Key methodologies in this arena include Optimal Computing Budget Allocation (OCBA) (Chen et al. 2006; Chen et al. 2000) Expected Value of Information (EVI) (Chick et al. 2010; Inoue 2001), Knowledge Gradient (KG) (Frazier et al. 2007; Gupta and Miescke 1996), and Expected Improvement (EI) (Jones et al. 1998; Ryzhov 2016). Peng et al. (2016) and Peng et al. (2018) introduce a stochastic dynamic programming (SDP) framework to capture sequential sampling and formulate the sampling-allocation problem as a Markov decision process (MDP). Traditional fixed-budget R&S research has not directly addressed the original SDP problem. Instead, the focus has been on static allocation approximations or one-step look-ahead approximations. Zhou and Peng (2023) introduce a Monte Carlo-based rollout method for learning and decision-making in the fixed-budget R&S problem. To overcome challenges related to computational complexity and efficiency, Zhou et al. (2024) develop AlphaRank, an AI-driven solution employing deep reinforcement learning (DRL). This involves obtaining a neural network (NN) through offline pre-training, which is then used for online allocation decisions. Significantly, the rollout policy is incorporated as a means of policy improvement during the pre-training phase.

Different from these works, the goal of this study is to focus on the fixed-precision R&S problem, and explain how the existing AlphaRank procedures fit in the framework. AlphaRank procedure is a natural fit for solving the fixed-budget R&S problem, utilizing approximate dynamic programming (ADP) methods within a SDP framework. Subsequently, we model the fixed-precision R&S problem as an infinite-horizon MDP with a defined stopping time, ensuring the probability of correct selection (PCS) of the best alternative up to a pre-specified probability level $1-\alpha$. The PCS can be approximated either by the estimated value function derived from a rolling horizon rollout policy or through a well-trained NN in AlphaRank. This method significantly curtails the volume of simulation observations required to attain a certain PCS level. In a sense, AlphaRank, when applied to the fixed-precision R&S problem, also can be regarded as an IZ-free procedure, accommodating scenarios where the pairwise mean differences between alternatives might be arbitrarily close to zero.

This paper only focuses on selecting the best mean. However, it is worth noting that certain related challenges may also fall under the umbrella of R&S problems. These encompass various combinations of objectives and performance metrics used for comparisons. Examples include ranking all alternatives, selecting the top-m alternatives, or identifying a subset that contains the best. Performance metrics might range from quantiles to proportions. Interested readers may refer to comprehensive reviews in Kim and Nelson (2006).

The rest of this paper is organized as follows. Section 2 describes the MDP modeling for the fixed-precision R&S problem. Section 3 offers a concise overview of the rollout technique, followed by

a comprehensive description of the AlphaRank procedure, including its detailed setup and pre-training process. Section 4 presents the results of the numerical experiments utilizing AlphaRank. Section 5 provides conclusions.

2 MDP MODELING OF FIXED-PRECISION R&S

In this section, we commence with the formulation of the fixed-precision R&S problem, and then we proceed to model the problem as a MDP with a defined stopping time.

2.1 Formulation

We capture the statistics of a set of alternatives indexed by $i = 1, 2, \ldots, N$, and observe a sequence of data $X_{i,1}, X_{i,2}, \ldots, X_{i,t}$, where $X_{i,t}$ is the observation at time step t for alternative i. Note that the t-th index of the observations of alternative i may be different from the t-th index of steps, because the i-th alternative is not necessarily allocated at each step. We assume that the observations follow a distribution F_s with mean μ_i^{true} and are independent and identically distributed (i.i.d.) for each alternative, i.e., $X_{i,t} \sim F_s$.

We use Assumption 1 to describe the structure of the R&S problem in the case of normal sampling distribution, which is one of the most common assumptions in R&S research.

Assumption 1 Suppose that the samples follow a normal distribution, i.e., $X_{i,t} \sim N(\mu_i^{true}, (\sigma_i^{true})^2)$, where parameter μ_i^{true} is unknown and $(\sigma_i^{true})^2$ is known.

Note that the specific case we present under the assumption of a normal sampling distribution in R&S is primarily for better understanding. The MDP modeling of the R&S problem introduced in Section 2, and the solution method proposed in Section 3, are not confined to this assumption but can be applied under other distributional assumptions in both requentist and Bayesian framework such as Peng et al. (2018), Chen and Ryzhov (2019) and Gao et al. (2017).

We assume that there is only one best alternative, although the following paradigm can be generalized to problems with multiple best alternatives. Our objective is to find the best alternative defined by $\arg\max_{i=1,\ldots,N} \mu_i^{true}$, where μ_i^{true} is estimated in different ways under different assumptions after allocating t simulation observations. To describe the precision of selection, one common way is to use the probability that the selected alternative is the true best, i.e., PCS. Let s_t be the current state after spending t sampling budget, which contains complete environment information, including the sample information. Let the selection made after allocating t simulation observations be $\widehat{S}_t$. The PCS given the current state s_t is

$$\mathrm{PCS}(s_t) = \Pr\left(\widehat{S}_t = \arg\max_{i=1,\ldots,N} \mu_i^{true} \,\middle|\, s_t \right).$$

For fixed-precision R&S problem, under a fixed precision $1-\alpha$ $(0 < \alpha < 1 - 1/N)$, the objective is to deliver a PCS guarantee as

$$\mathrm{PCS}(s_\tau) \geq 1 - \alpha,$$

and then τ is the number of simulation observations at the end of the experiment.

2.2 MDP Modeling

In the fixed-precision R&S problem, we guarantee a probability of $1-\alpha$. Formally, the problem can be stated as follows. After obtaining t samples, we can derive statistical characteristics of the alternatives based on the sample information. For example, under Assumption 1, at step t, the statistics ε_t of the alternatives can be simply defined as

$$\varepsilon_t = \{\bar{X}_t, \bar{\sigma}_t^2\},$$

where $\bar{X}_t = \{\bar{X}_{1,t}, \ldots, \bar{X}_{N,t}\}$ and $\bar{\sigma}_t^2 = \{\bar{\sigma}_{1,t}^2, \ldots, \bar{\sigma}_{N,t}^2\}$. $\bar{X}_{i,t}$ and $\bar{\sigma}_{i,t}^2$ are the sample mean and sample variance of alternative i, respectively. Under different assumptions, other statistical characteristics can be collected

according to their settings. For instance, in scenarios where there is assumed correlation among alternatives, the correlation between pairs of samples could be considered (Zhang and Peng 2024). Similarly, under assumptions where the unknown true mean has a conjugate prior, characteristics like the prior mean and prior variance of the parameters can be taken into account (Zhou et al. 2024).

The number of simulation observations that have been allocated is t, and p_t is the current precision. Therefore, the state space s_t is

$$s_t = \{\varepsilon_t, t, p_t\}.$$

The state after allocating alternative i can be defined as

$$s_{t+1}^{(i)} = \{\varepsilon_{t+1}^{(i)}, t+1, p_{t+1}^{(i)}\},$$

where $\varepsilon_{t+1}^{(i)}$ represents the updated statistics after the $(t+1)$-th simulation observation has been allocated to alternative i. After each sampling, the state transitions to a new state based on the outcome of the observation and the remaining simulation budget. In the current discussion, the state transition mechanism is to update $\bar{X}_{t+1}$, $\bar{\sigma}_{t+1}^2$ according to the latest allocated sample. Let $T_{i,t}$ represent the number of simulation observations allocated to alternative i with a total of t allocated simulation observations, $\sum_{i=1,\ldots,N} T_{i,t} = t$. If the $t+1$-th simulation observation is allocated to alternative i, then $T_{i,t+1} = T_{i,t} + 1$ and the observation is $X_{i,t+1}$. Specifically, the sample mean $\bar{X}_{i,t+1}$ and sample variance $\bar{\sigma}_{i,t+1}^2$ of alternative i are updated as follows:

$$\bar{X}_{i,t+1} = \frac{T_{i,t} \cdot \bar{X}_{i,t} + X_{i,t+1}}{T_{i,t+1}},\ \bar{\sigma}_{i,t+1}^2 = \frac{T_{i,t}}{T_{i,t+1}} \cdot \left(\bar{\sigma}_{i,t}^2 + \frac{(\bar{X}_{i,t} - X_{i,t+1})^2}{T_{i,t+1}} \right).$$

Parameters of other alternatives except for alternative i being allocated will not be updated at the $t+1$-th step. In the absence of precision updating the explicit function expression, we can obtain the updated $p_{t+1}^{(i)}$ by approximation. The rollout technique described in Section 3 is a suitable technique for estimating PCS, which uses action value $Q_t^{(i)}$ to approximate $p_{t+1}^{(i)}$. Further, we can also approximate PCS with NNs, for example, using the output V_t of AlphaRank's value NN in Section 3 to approximate $p_{t+1}^{(i)}$.

The MDP modeling of fixed-precision R&S problem can be characterized as follows. The continuation value in the Bellman equation takes into account that the agent may decide to stop the process at any time, and the expected value of PCS related to continuing versus stopping is considered in the optimal policy. In this problem, the stopping time is $\tau \doteq \min\{t \in \mathbb{Z} : p_t \geq 1-\alpha\}$, and the equation is expressed as follows. For $0 \leq p_t < 1-\alpha$, we have

$$a_{t+1}^* = \arg\max_{i=1,\ldots,N} Q(s_t, i),$$

$$V_t(s_t) = Q\left(s_t, a_{t+1}^*\right) = \mathbb{E}\left[Q\left(s_t, a_{t+2}^*\right) \middle| s_t, a_{t+1}^*\right] = \mathbb{E}\left[V_{t+1}(s_{t+1}) \middle| s_t, a_{t+1}^*\right],$$

where a_{t+1}^* is the optimal allocation policy at $t+1$-th step, $Q(s_t, i)$ is the state-action-value function, which represents the reward value of choosing alternative i, and s_{t+1} is determined by s_t and newly allocated observation $X_{a_{t+1}^*, t+1}$. In principle, the action values $Q(s_t, i)$ can be calculated using classic SDP techniques, such as value iteration or policy iteration.

To identify the best alternative from the set of N alternatives based on the observations, the alternative that maximizes the posterior PCS, based on the information of all allocated simulation observations, is selected, given the current state. At the time when PCS condition $p_t \geq 1-\alpha$ is satisfied for the first time, i.e., $t = \tau$, the optimal selection policy at the state s_τ of step τ is (Peng et al. 2016)

$$S_\tau^* = \arg\max_{i=1,\ldots,N} \Pr\left(\mu_i^{true} \geq \mu_j^{true} \middle| s_\tau\right), \tag{1}$$

and the value function after allocating τ simulation observations is

$$V_\tau(s_\tau) = Q(s_\tau, S_\tau^*) = \mathbb{E}[V(s_\tau, S_\tau^*) | s_\tau],$$

where

$$V(s_\tau, S^*_\tau) = \mathbf{1}\{\widehat{S}_\tau = \arg\max_{i=1,\ldots,N} \mu_i^{true}\}.$$

After reaching the stopping time, the simulation allocation procedure terminates.

Under most circumstances, except for some specific and stringent assumptions, directly computing equation (1) presents a considerable challenge. To circumvent this, it is common to utilize approximate optimal selection policies. These policies may include selecting the largest sample mean under frequentist assumptions or opting for the largest posterior mean of the true mean under Bayesian assumptions. In this study, we chose to implement the policies of selecting the largest sample mean as our approximate optimal selection, i.e.,

$$\widehat{S}_\tau = \arg\max_{i=1,\ldots,N} \bar{X}_{i,\tau}.$$

3 METHODOLOGY

In this section, we first introduce the AlphaRank procedure proposed by Zhou et al. (2024) to solve fixed-budget R&S problems, because it is the building block of our procedure. We then develop a new fixed-precision procedure whose pre-training process has some modifications compared to the original fixed-budget version to ensure that the procedure can meet the given statistical guarantees.

3.1 AlphaRank Procedure

This subsection will address three core aspects: the essence of AlphaRank, the source of its capability to enhance the performance of base policies, and its applicability in resolving fixed-precision R&S problems.

AlphaRank trains the NN models to study the behavior of the rollout policy introduced by Zhou et al. (2024). This rollout policy represents an effective online sampling policy based on Monte-Carlo simulation, although it encounters some challenges related to computational efficiency, which is crafted to facilitate learning and decision-making in the fixed-budget R&S problem with a budget of T. In this context, the policy considers the final PCS achieved by selecting an alternative at the current step t, and subsequently allocating the remaining H observations based on a base policy. This resultant PCS is treated as the action-value function for the chosen alternative at step t. To solve the computational issue of rollout policy, AlphaRank procedure involves the pre-training of a series of NN models with high precision. This pre-training is conducted offline, utilizing a predetermined prior distribution. Once trained, these NN models can be directly applied for making allocation decisions in practical scenarios.

Specifically, each NN model is intricately designed to output the estimated action value for every alternative. This estimation is based on input data that reflects the current state, such as the collected sample information and the remaining budget. Specifically, in a fixed-budget R&S scenario with a budget of T, the input of the NN, denoted as $input_t$ at state t, encompasses the statistical data of the alternatives along with the number of steps to explore forward in the rollout, H, i.e.,

$$input_t = \{\varepsilon_t, H\}.$$

For example, in the normal case, the input could be $input_t = \{\bar{X}_t, \bar{\sigma}_t^2, H\}$, which has 2 sample statistics including sample mean, sample variance, parameter prior mean and variance, and each statistic is an N-dimensional vector. The actual number of steps to explore forward H in the $input_t$ can be adjusted according to the actual situation. For example, H can represent the remaining budget when the budget T for the problem is small, i.e., $H = T - t$. When a rollout does not encompass the full set of remaining $T - t$ steps, H can be a fixed constant.

The output of NN $output_t$ is an action value vector

$$output_t = V_t = \left(V_t^{(1)}, \ldots, V_t^{(N)}\right),$$

which evaluates the expected PCS for performing each action at the current state t, where $V_t^{(i)}$ represents the action value of selecting alternative i at step t.

Next, we specifically explain how the rollout policy estimates the potential PCS of each action as its action value. Let $a_{t+1}^{(i)}$ be the possible action that allocates the $(t+1)$-th simulation observation to alternative i, there is a theoretical action value of taking action $a_{t+1}^{(i)}$ in s_t which is represented by $Q(s_t, a_{t+1}^{(i)})$. Let $s_{t+1}^{(i)}$ be the updated state after selecting the i-th alternative through action $a_{t+1}^{(i)}$ at current state s_t. Given a base policy π, $Q(s_t, a_{t+1}^{(i)})$ represents the PCS value $\mathrm{PCS}^{\pi}(s_{t+1}^{(i)})$ that can be obtained when the remaining simulation observations are allocated through the base policy π after action $a_{t+1}^{(i)}$ is selected at s_t. Since it is difficult to have an explicit form to calculate $Q(s_t, a_{t+1}^{(i)})$ precisely, we approximate this value by Monte Carlo simulation. The rollouts entail the generation of K trajectories by the base policy π, which means that starting from $s_{t+1}^{(i)}$, the remaining H observations are allocated according to π. Then we can get the rewards $r_{t,k}^{(i)}$. In the k-th rollout, $r_{t,k}^{(i)} = 1$ when the selection is correct, and $r_{t,k}^{(i)} = 0$ otherwise, with probabilities $\mathrm{PCS}^{\pi}(s_{t+1}^{(i)})$ and $1 - \mathrm{PCS}^{\pi}(s_{t+1}^{(i)})$, respectively. Therefore, $Q_t^{(i)}(s_{t+1}^{(i)})$ can be calculated by

$$Q_t^{(i)}\left(s_{t+1}^{(i)}\right) = \frac{1}{K}\sum_{k=1}^{K} r_{t,k}^{(i)} = \frac{1}{K} R_{t,K}^{(i)}.$$

As $K \to \infty$, $Q_t^{(i)}(s_{t+1}^{(i)}) \to Q(s_t, a_{t+1}^{(i)}) = \mathrm{PCS}^{\pi}(s_{t+1}^{(i)})$. Upon estimating the action value, the $(t+1)$-th simulation observation is then strategically allocated to the alternative that exhibits the highest calculated action value at that point, i.e., the sample-allocation action of the rollout policy is

$$a_{t+1}^{roll} = \arg\max_{i=1,\ldots,N} Q_t^{(i)}\left(s_{t+1}^{(i)}\right).$$

The selection after all T simulation observations have been allocated is

$$S_T^{roll}\left(s_T^{(a_T^{roll})}\right) = \arg\max_{i=1,\ldots,N} \Pr\left(\mu_i^{true} \geq \mu_j^{true} \middle| s_T\right).$$

For simplicity, as discussed in Section 2, we can also approximate the selection of an alternative. The diagram illustrating the rollout process is presented in Figure 1. This figure depicts the decision-making process and the state transitions occurring within the rollout, using a scenario with two alternatives to showcase how each allocation decision impacts the state of the alternatives. In Figure 1, different types of circles represent distinct states: black circles denote alternatives that have yet to receive any simulation allocations; hollow circles indicate alternatives currently being allocated simulations; and gray circles symbolize the simulation observations, which are essentially samples drawn from the updated prior distribution.

Zhou et al. (2024) prove that the rollout policy is statistically guaranteed to perform at least as well as its base policy within a certain probability threshold. Consequently, leveraging an offline-trained NN to learn the value functions estimated by the rollout policy and then utilizing this NN to address the SDP challenge online can obviate the need for an actual rollout. This approach significantly accelerates the process while aiming to preserve decision quality.

Furthermore, during each round of pre-training iterations, the NN from the preceding training round is adopted as the base policy for the rollout in the current round. This strategy guarantees that the NN, which assimilates the behavior of the current round's rollout policy, exhibits superior performance compared to the NN from the previous round. Detailed training procedures will be further expounded in Section 3.2.

A vital aspect to note is that the observations utilized in the rollout are not drawn from the simulation models but are instead sampled from the current updated prior. This makes them computationally more accessible and permits their acquisition offline, where the simulation models might not be available. This factor plays a crucial role in the pre-training process, enabling a more efficient and feasible approach.

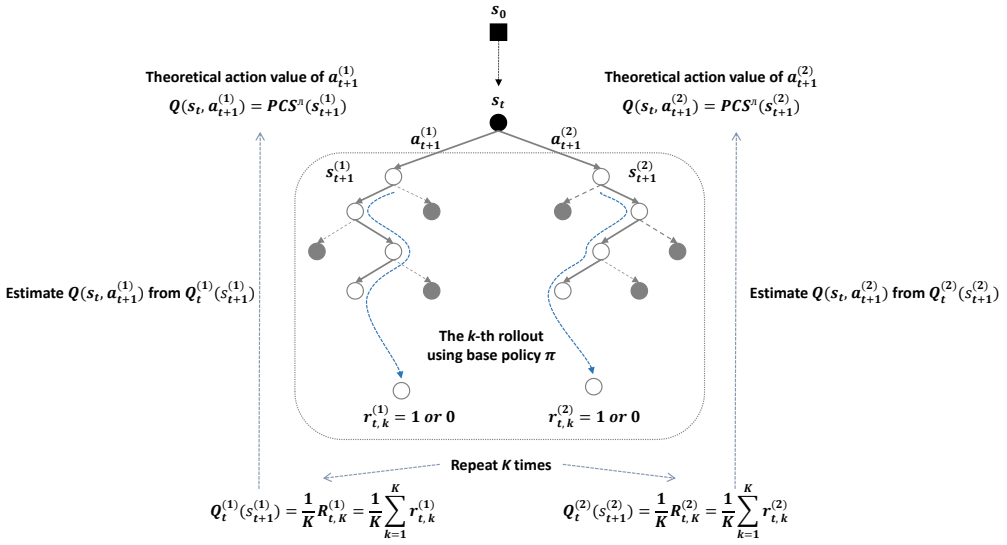

Figure 1: The decisions in the rollout process.

In Section 2, we discuss how, when the fixed-precision R&S problem is formulated as a MDP with a stopping time, its state is characterized by the current sample statistical characteristics and the precision of PCS under that state. For the fixed-budget R&S problem, the trained NN takes in these current sample statistical characteristics and the number of step look forward (remaining budget). Its output is the estimated PCS achieved after the remaining budget is allocated. This setup enables us to apply the NN for tackling fixed-precision R&S problems. In particular, since fixed-precision R&S problem is regarded as a fixed-budget problem with an infinite budget, when utilizing a well-trained NN, we input the current sample statistical characteristics along with a predetermined small numerical value, H, representing the number of step look forward in the rollout. We use the output value $V_t^{(i)}$ to approximate the PCS value obtained after selecting alternative i and then allocating H additional simulation observations. Sampling is ceased and an additional H simulation observations are allocated only when the output estimated value meets the stopping criteria, i.e.,

$$\min_{i=1,\ldots,N} V_t^{(i)} \geq 1-\alpha, \tag{2}$$

and the total budget expended is thus $t+H$. The reason for choosing the lower bound rather than the upper bound of the estimated value in (2) to meet the given precision level is that, despite the accuracy of the NN estimation, there will always be some error compared to the actual PCS. This approach better ensures that the actual use of the NN achieves the specified precision. For computational complexity analysis regarding rollout policy, NN training, and direct use of NN as allocation policy, see Zhou et al. (2024).

3.2 Pre-training Process of AlphaRank

In a manner akin to the techniques used in training image classifiers, we pre-train a NN using a dataset that is generated from prior information. This process significantly enhances the NN's capability to adapt to the dynamics of sampling in R&S problems. The NN is subsequently deployed for online sampling in R&S scenarios. The input data for the NN, reflecting current states like statistics of alternatives and the remaining budget, and the output targets of this process are the estimated action values generated through a rollout policy. The NN is trained based on this rollout policy, which undergoes iterative refinement. In each successive training round, it is used as the base policy in the rollout of next round. The efficiency of the NN's allocation is assessed by applying the currently trained NN directly as an allocation policy. This involves evaluating whether the resulting PCS satisfies our predefined stopping criteria. This continuous

improvement loop ensures that the NN becomes progressively adept at making allocation decisions in R&S problems.

The NN is trained to optimize the loss between NN predicted value V_t and rollout action value Q_t, where $Q_t = (Q_t^{(1)}, \ldots, Q_t^{(N)})$. We utilize the cross-entropy with a regularization term as the loss function for training NN:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\left[V_t^{(i)} \log Q_t^{(i)} + (1 - V_t^{(i)}) \log(1 - Q_t^{(i)})\right] + c||\cdot||^2.$$

The training and evaluation process of AlphaRank is depicted as follows.

- **Training data generation** Utilizing a rollout policy, the training data for our study are generated through a process akin to self-play. Within this rollout, the NN assumes the role of the base policy. In the initial training round, the NN, being untrained, lacks the proficiency to guide the allocation of simulation observations effectively. Therefore, in these early rounds, classic fixed-budget R&S sampling procedures such as EA and OCBA can be employed as the base policy for the rollout. Since in the fixed-precision problem, the number of steps to rollout forward is a fixed value H, the generation of the training dataset requires specifying a dataset size M. Consequently, in each round, a total of M simulation observations are allocated, generating M pieces of data. Each data contains its state vector and the estimated PCS value from rolling out H steps forward in the corresponding state.
- **NN training** The dataset amassed through data generation process is instrumental in training a new NN. This NN is specifically designed to learn the behavior of the rollout policy. Once trained, it is then employed to direct the rollout in a self-play mode during the subsequent round of training. This cyclical approach allows the NN to progressively refine its understanding and implementation of the rollout policy, thereby enhancing its decision-making capabilities in successive iterations.
- **NN evaluation** A single iteration of training data generation and NN training is referred to as a round. After several such iterations, the new NN is used directly as a policy to guide simulation resource allocation, as well as a policy used for PCS estimation at each step. For the fixed-precision R&S problem, the primary evaluation criterion is whether the PCS reaches the desired precision level of $1-\alpha$. The secondary consideration is the reduction of the simulation budget. This contrasts with the fixed-budget R&S problem, where the sampling process ceases once all simulation observations are allocated, the sampling process in the fixed-precision R&S problem continues if $p_{t+H} \approx \min_{i=1,\ldots,N} V_t^{(i)} < 1-\alpha$, with H simulation observations being allocated until $\min_{i=1,\ldots,N} V_t^{(i)} \geq 1-\alpha$. If the PCS yielded by the new NN surpasses that of the previous NN, the training parameters are updated to reflect this improved performance. If not, the existing NN is maintained. This method ensures a continuous enhancement in the precision and efficiency of the simulation resource allocation.
- **Stopping rule** For the fixed-precision R&S problem, the training process continues until the PCS exceeds $1-\alpha$, and no further improvement is observed in the NN evaluation, which signifies that the sampling policy, as dictated by the NN, has attained the specified level of precision.

The pseudo-code of the pre-training process is shown in Algorithm 1 and the pipeline of the training is depicted in Figure 2.

4 NUMERICAL EXPERIMENTAL RESULTS

In this section, we conduct a comparative analysis of AlphaRank and two traditional IZ procedures within the scope of fixed-precision R&S problems. The performance metrics presented in this section are estimated from 10^5 independent macro-simulations.

Algorithm 1: Pre-training

Input: number of alternatives N, times of rollout K, number of forward steps in rollout H, size of the training dataset M

while *the PCS in evaluation does not satisfy the stopping rule* **do**
  for *t=1 to M* **do**
    Calculate the current statistics of alternatives ε_t according to a_t^{roll}.
    for *k=1 to K* **do**
      for *i=1 to N* **do**
        Rollout H steps forward with NN as the base policy and get the reward $r_{t,k}^{(i)}$.
      end
    end
    Calculate the value function $Q_t^{(i)} = \frac{1}{K}\sum_{k=1}^{K} r_{t,k}^{(i)}$.
    Collect the training data $\{\varepsilon_t, Q_t\}$.
    The sampling action is $a_{t+1}^{roll} = \arg\max_{i=1,\ldots,N} Q_t^{(i)}$.
  end
  NN training: update parameters of NN by minimizing *Loss* with the Adam optimizer.
  NN evaluation: NN is used as allocation policy in simulation and then the corresponding PCS is obtained.
end

Output: the trained NN

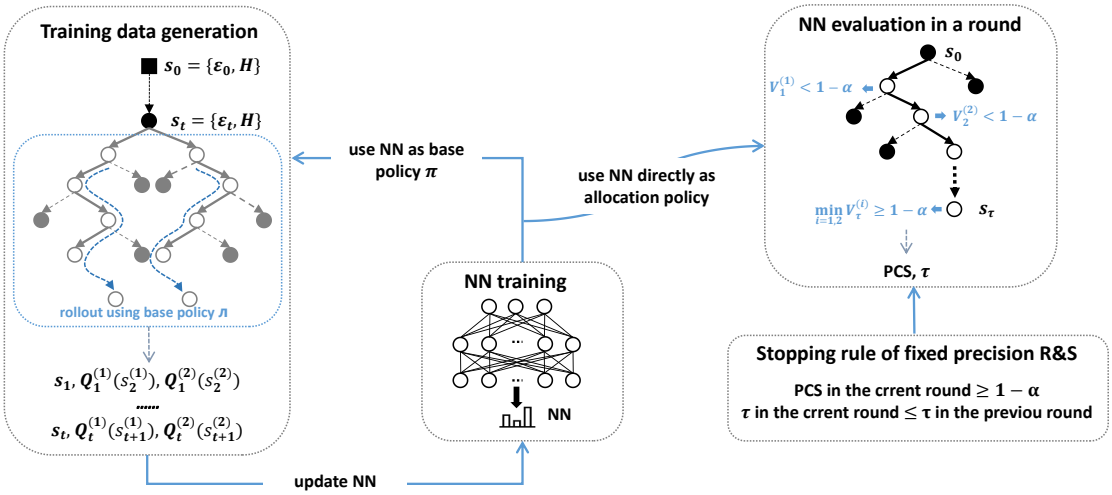

Figure 2: NN training and evaluating architecture with the number of alternatives=2.

4.1 Benchmark

We first provide a brief introduction to the concept of the IZ framework and two benchmark procedures. In R&S problems, we may not be able to select the best with the desired precision, when the means are sufficiently close to each other. To overcome this obstacle, Bechhofer (1954) introduced a so-called IZ parameter $\delta > 0$, which refers to the smallest mean difference worth detecting. Given the IZ, the R&S problems are modified to select the best alternative, when all the inferior alternatives are outside the IZ of the best, i.e., $\mu_{[N]}^{true} - \delta > \mu_{[N-1]}^{true}$, where $\mu_{[N]}^{true} > \mu_{[N-1]}^{true} > \cdots > \mu_{[1]}^{true}$. Therefore, if $\mu_{[N]}^{true} - \mu_{[i]}^{true} \leq \delta$,

$i = 1, \ldots, N-1$, we believe that there is no difference in choosing $\mu_{[i]}^{true}, \ldots, \mu_{[N]}^{true}$ as the best alternative. Considering the fixed-precision constraint, most of the existing R&S procedures are designed under the IZ formulation. These procedures are often called IZ procedures. In this experiment, we compare AlphaRank with two classical IZ procedures, i.e., Rinott (Rinott 1978) and Kim and Nelson's (Kim and Nelson 2001) procedures. Next, we provide a brief description of the two procedures.

Rinott procedure First, generate T_0 samples for each alternative i and calculate the sample variance $\bar{\sigma}_{i,T_0}^2$. Second, the total sample size T_i allocated to alternative i is set to be positively proportional to its sample variance, i.e.,

$$T_i = \max\left\{T_0, \left\lceil \frac{h_R^2 \bar{\sigma}_{i,T_0}^2}{\delta^2} \right\rceil\right\}.$$

Generate $T_i - T_0$ samples from alternative i and calculate the sample mean $\bar{X}_{i,T_i}$. Finally, select the alternative with the largest sample mean as the best. In our experiment, to avoid the complexity in calculating h_R, we use the variation of Rinott procedure proposed by Clark and Yang (1986), which adopts Bonferroni's inequality and sets it approximately as the $1-\alpha/(N-1)$ quantile of a t-distribution with $T_0 - 1$ degrees of freedom.

Kim and Nelson's (KN) procedure The primary aim of such a sequential procedure is to quickly identify and eliminate those alternatives that appear notably subpar, thereby optimizing the overall computational effort needed to find the best. It also uses an additional initial stage of sampling to estimate the unknown variances. Once these variances are estimated, the procedure transitions to the screening of alternatives. For each pair of alternatives j and i, it constructs the partial-sum process for their mean difference, represented as $\{t(\bar{X}_{j,t} - \bar{X}_{i,t}) : n = 1, 2, \ldots\}$. At each step t, KN procedure evaluates whether this partial-sum process moves beyond a pre-defined triangular region, and decisions are made based on this assessment. Finally, the only alternative remaining within the specified region is selected as the best. The details of these formulas are not covered in this brief overview; interested readers can refer to Kim and Nelson (2001) for comprehensive information.

4.2 Experiment

The experiment is conducted with $X_{i,t} \sim N(\mu_i^{true}, (\sigma_i^{true})^2)$, $i = 1, 2, \ldots, 5$. We use $\mu_i^{true} \sim N(\mu_i, \sigma_i^2)$ to generate μ_i^{true}. Different examples are set by varying the hyper-parameters μ_i, σ_i^2 and $(\sigma_i^{true})^2$. For Rinott and KN procedures, we set the IZ parameter $\delta = 0.05$. The first 50 simulation observations are allocated equally to each alternative for estimating the sample means and variances, i.e., $T_0 = 10$. The settings and results of other hyper-parameters are presented in Table 1.

Table 1: Average performance of Rinott procedure, KN procedure, and AlphaRank in Experiment 4.

Parameter Settings				Rinott		KN		AlphaRank	
μ	σ_i^2	σ^{true}	$1-\alpha$	budget	PCS	budget	PCS	budget	PCS
0	0.01	1	0.5	537	0.736	188	0.591	58	0.4912
0	1	1	0.9	2797	1	366	0.916	102	0.9056
0	1	1	0.95	3570	1	561	0.997	124	0.9483

The results in Table 1 indicate that the number of samples consumed by Rinott and KN is much higher than that required by AlphaRank, e.g., about 2879% and 452% of AlphaRank's consumption, respectively, at 95% confidence, and is highly sensitive to the pre-specified target. The PCS of AlphaRank is in close agreement with the target $1-\alpha$. To increase the probability guarantee level from 90% to 95%, the Rinott procedure and KN procedure require 773 and 195 more simulation observations, respectively, and their

actual PCSs reach 1 approximately, whereas AlphaRank requires only 22 more simulation observations and its actual PCS is very close to 95%.

5 CONCLUSION REMARK

In this study, we formulate the fixed-precision R&S problem under the umbrella of MDP. We utilize AlphaRank, a cutting-edge AI approach that uses DRL and rollout techniques to effectively tackle this problem. A series of extensive numerical experiments demonstrate AlphaRank's efficacy, overcoming a common limitation in previous methodologies where the number of simulation observations often surpasses what is necessary to guarantee a satisfactory level of PCS.

The application of AI methods to R&S problems creates numerous research opportunities. For example, Zhou et al. (2024) propose the DCR framework, merging the concepts of "divide-and-conquer" and "recursion", using small, well-trained NN models to solve large-scale R&S problems effectively by leveraging a parallel computational platform. Future work includes adapting this method to large-scale fixed-precision scenarios. It is also noteworthy to mention that our discussions have so far been predicated on the assumption of normality. In future developments, AlphaRank's functionality can be expanded to include a broader spectrum of distribution assumptions, significantly widening its range of application.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 72325007, 72250065, and 72022001.

REFERENCES

Batur, D. and F. Choobineh. 2012. "Stochastic Dominance Based Comparison for System Selection". *European Journal of Operational Research* 220(3):661–672 https://doi.org/10.1016/j.ejor.2012.02.018.

Bechhofer, R. E. 1954. "A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances". *Annals of Mathematical Statistics* 25:16–39.

Branke, J., C. Schmidt, and S. E. Chick. 2008. "Selecting a Selection Procedure". *Management Science* 53(12):1916–1932.

Chen, C. H., D. He, and M. Fu. 2006. "Efficient Dynamic Simulation Allocation in Ordinal Optimization". *IEEE Transactions on Automatic Control* 51(12):2005–2009.

Chen, C. H., J. Lin, and E. Chick. 2000. "Simulation Budget Allocation for Further Enhancing the Efficiency of Ordinal Optimization". *Discrete Event Dynamic Systems* 10:251–270.

Chen, Y. and I. O. Ryzhov. 2019. "Balancing Optimal Large Deviations in Ranking and Selection". In *2019 Winter Simulation Conference (WSC)*, 3368–3379 https://doi.org/10.1109/WSC40007.2019.9004810.

Chick, S. E. 2006. "Chapter 9 Subjective Probability and Bayesian Methodology". In *Handbooks in Operations Research and Management Science*, edited by S. G. Henderson and B. L. Nelson, Volume 13, 225–257. San Diego: Elsevier https://doi.org/10.1016/S0927-0507(06)13009-1.

Chick, S. E., J. Branke, and C. Schmidt. 2010. "Sequential Sampling to Myopically Maximize the Expected Value of Information". *INFORMS Journal on Computing* 22(1):71–80.

Clark, G. M. and W.-n. Yang. 1986. "A Bonferroni Selection Procedure When Using Commom Random Numbers with Unknown Variances". In *1986 Winter Simulation Conference (WSC)*, 313–315 https://doi.org/10.1109/WSC60868.2023.10407663.

Fan, W., L. J. Hong, and B. L. Nelson. 2016. "Indifference-Zone-Free Selection of the Best". *Operations Research* 64(6):1499–1514.

Frazier, P. I., W. B. Powell, and S. Dayanik. 2007. "A Knowledge-Gradient Policy for Sequential Information Collection". *SIAM Journal on Control & Optimization* 47(5):2410–2439.

Gabillon, V., M. Ghavamzadeh, and A. Lazaric. 2012. "Best Arm Identification: a Unified Approach to Fixed Budget and Fixed Confidence". In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 3212–3220. New York: Curran Associates Inc.

Gao, S., W. Chen, and L. Shi. 2017. "A New Budget Allocation Framework for the Expected Opportunity Cost". *Operations Research* 65(3):787–803.

Gupta, S. S. and K. J. Miescke. 1996. "Bayesian Look Ahead One-Stage Sampling Allocations for Selection of the Best Population". *Journal of Statistical Planning & Inference* 54(2):229–244.

Hong, L. J., W. Fan, and J. Luo. 2021. "Review on Ranking and Selection: A New Perspective". *Frontiers of Engineering Management* 8(5):321–343.

Hunter, S. R. and B. L. Nelson. 2017. "Parallel Ranking and Selection". In *Advances in Modeling and Simulation: Seminal Research from 50 Years of Winter Simulation Conferences*, edited by A. Tolk, J. Fowler, G. Shao, and E. Yücesan, 249–275. Berlin: Springer https://doi.org/10.1007/978-3-319-64182-9_12.

Inoue, C. K. 2001. "New Two-Stage and Sequential Procedures for Selecting the Best Simulated System". *Operations Research* 49(5):732–743.

Jones, D. R., M. Schonlau, and W. J. Welch. 1998. "Efficient Global Optimization of Expensive Black-Box Functions". *Journal of Global Optimization* 13:455–492.

Kim, S. H. and B. L. Nelson. 2001. "A Fully Sequential Procedure for Indifference-Zone Selection in Simulation". *ACM Transactions on Modeling & Computer Simulation* 11(3):251–273.

Kim, S.-H. and B. L. Nelson. 2006. "Chapter 17 Selecting the Best System". In *Handbooks in Operations Research and Management Science*, edited by S. G. Henderson and B. L. Nelson, Volume 13, 501–534. San Diego: Elsevier https://doi.org/10.1016/S0927-0507(06)13017-0.

Luo, J., L. J. Hong, B. L. Nelson, and Y. Wu. 2015. "Fully Sequential Procedures for Large-Scale Ranking-and-Selection Problems in Parallel Computing Environments". *Operations Research* 63(5):1177–1194.

Ni, E. C., D. F. Ciocan, S. G. Henderson, and S. R. Hunter. 2017. "Efficient Ranking and Selection in Parallel Computing Environments". *Operations Research* 65(3):821–836.

Paulson, E. 1949. "A Multiple Decision Procedure for Certain Problems in the Analysis of Variance". *Annals of Mathematical Statistics* 20(1):95–98.

Peng, Y., C. H. Chen, M. C. Fu, and J. Q. Hu. 2016. "Dynamic Sampling Allocation and Design Selection". *INFORMS Journal on Computing* 28(2):195–208.

Peng, Y., E. K. P. Chong, C.-H. Chen, and M. C. Fu. 2018. "Ranking and Selection as Stochastic Control". *IEEE Transactions on Automatic Control* 63(8):2359–2373 https://doi.org/10.1109/TAC.2018.2797188.

Rinott, Y. 1978. "On Two-Stage Selection Procedures and Related Probability-Inequalities". *Communications in Statistics-theory and Methods* 7(8):799–811.

Ryzhov, I. O. 2016. "On the Convergence Rates of Expected Improvement Methods". *Operations Research* 64(6):1515–1528.

Zhang, Z. and Y. Peng. 2024. "Sample-Efficient Clustering and Conquer Procedures for Parallel Large-Scale Ranking and Selection". *arXiv preprint* arXiv.2402.02196.

Zhong, Y. and L. J. Hong. 2021. "Knockout-Tournament Procedures for Large-Scale Ranking and Selection in Parallel Computing Environments". *Operations Research* 70:432–453.

Zhou, R., L. J. Hong, and Y. Peng. 2024. "AlphaRank: An Artificial Intelligence Approach for Ranking and Selection Problems". *arXiv preprint* arXiv.2402.00907.

Zhou, R. and Y. Peng. 2023. "POMDP-Based Ranking and Selection". In *2023 Winter Simulation Conference (WSC)*, 3388–3399 https://doi.org/10.1109/WSC60868.2023.10407663.

AUTHOR BIOGRAPHIES

RUIHAN ZHOU is a Ph.D. candidate in the Department of Management Science and Information Systems in Guanghua School of Management at Peking University, Beijing, China. Her research interests include simulation optimization and artificial intelligence. Her email address is rhzhou@stu.pku.edu.cn.

YIJIE PENG is an Associate Professor in Guanghua School of Management at Peking University. His research interests include stochastic modeling and analysis, simulation optimization, machine learning, data analytics, and healthcare. He is a member of INFORMS and IEEE, and serves as an Associate Editor of the Asia-Pacific Journal of Operational Research and the Conference Editorial Board of the IEEE Control Systems Society. His email address is pengyijie@pku.edu.cn.