Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

POMDP-BASED RANKING AND SELECTION

Ruihan Zhou
Yijie Peng

Guanghua School of Management
Peking University
5 Yiheyuan Road
Beijing 100871, P. R. CHINA

ABSTRACT

In this paper, we formulate the ranking and selection (R&S) problem as a stochastic control problem under the Bayesian framework. We propose to use particle filter to approximate the posterior distribution of states under the general Bayesian framework. The learning and decision are treated under the umbrella of a partially observable Markov decision process and a rollout policy based on Monte Carlo simulation is proposed. This policy can use one or more classic R&S approaches as base policies to efficiently learn the value function by rolling out simulation trajectories. We present numerical examples to demonstrate the effectiveness of the rollout policy and the performance of our policy is significantly improved relatively to the base policies.

1 INTRODUCTION

For decision-making problems in the real world, managers usually need to compare several alternatives based on a certain measure for selecting the one with the best performance. Specifically, in a random environment, there are several alternatives with unknown average performance, and it is necessary to learn the alternative with the best performance by random sampling. Therefore, we need to design a sampling policy to allocate simulation budget dynamically and select the optimal alternative based on the entire sample information (Hong et al. 2021). In simulation, the aforementioned problems are often referred to as Ranking and Selection (R&S).

A well-studied paradigm in R&S is the framework of indifference zone (IZ), which could be traced back to the study of Bechhofer (1954) and Rinott (1978). The recent advances can be found in Kim and Nelson (2001), Frazier et al. (2007), Luo et al. (2015) and Ni et al. (2015). The sampling allocation algorithm in IZ framework ensures that the probability of correction selection (PCS) reaches a certain level. In IZ framework, the guarantee of PCS is primary and the (statistical) efficiency of finding the best solution is secondary. Due to the need to guarantee PCS level in a worst-case configuration, the sampling algorithm in the IZ framework tends to allocate more simulation replications than required to guarantee PCS level.

Therefore, many recent studies focus on improving the efficiency of finding the optimal alternative. Representative policies include: optimal computing budget allocation (OCBA) (Chen et al. 2006; Chen et al. 2000), expected value of information (EVI) (Chick et al. 2010; Inoue and Koichiro 2001), knowledge gradient (KG) (Frazier et al. 2007; Gupta and Miescke 1996), and expected improvement (EI) (Donald et al. 1998; Ryzhov 2016), where the sampling allocation decisions are derived under the framework of static optimization or one-step forward greedy optimization. Recently, Peng et al. (2016) and Peng et al. (2018) formulate the dynamic sampling decisions in R&S using dynamic programming as a stochastic control problem (SCP), whose optimal solution satisfies the Bellman equation. Peng et al. (2018) prove that under the common conditions of R&S, sampling allocation decision does not affect the Bayesian posterior distribution based on sample observation information. Thus, the SCP is a Markov decision process (MDP).

Many random optimization problems can be modeled as partially observable Markov decision process (POMDP). However, due to curse-of-dimensionality, value iteration, policy iteration and other classical solving policies, are often intractable to solve POMDP problems. Bertsekas and Castanon (1998) propose a rollout scheme through Monte Carlo simulation. Given sufficient simulation budget, Monte Carlo sampling based on any base policy can ensure that new policy obtained by rollout would not be worse than the original base policy. The rollout policy has been used in some papers related to dynamic programming applications. Tesauro and Galperin (1996) also apply the policy in the context of computer backgammon based on simulation. The terminology "rollout" is introduced by Tesauro as a synonym for repeated play of a given backgammon position to estimate the expected score by Monte Carlo simulation.

Any POMDP can be converted into an equivalent MDP by merging information states, which is often called information-state MDP (ISMDP). Since Peng et al. (2018) model the R&S problem as an MDP, R&S can also be described by POMDP and derive its corresponding rollout policy. In this paper, we study the simulation budget allocation policy in R&S problem from the perspective of POMDP, where the state variables in POMDP follow a fixed distribution. We introduce a rollout policy that can effectively integrate existing R&S simulation sampling policies. Specifically, we take several popular sampling policies as the base policies and conduct numerical experiments under different parameter distributions. The results show that the performance of rollout policies is much better than their base policies. In order to improve applicability, we also propose to use particle filter to update the posterior distribution of parameter when the parameter distribution has no conjugate priors. To the best of our knowledge, this is the first work to study the rollout policy of R&S problem to enhance the performance of base policies.

The rest of this paper is organized as follows. In Section 2, we define the R&S problem, show how R&S can be expressed as a SCP, and give the related Bellman optimality equation. In Section 3, We focus on the principle and properties of the rollout policy. The numerical experimental results are presented in Section 4. In Section 5, the conclusion is given and the prospect is proposed.

2 R&S UNDER BAYESIAN FRAMEWORK

2.1 Problem Formulation

Suppose there are N alternatives with unknown mean μ_i, $i = 1,\ldots,N$. Our goal is to select the alternative with the highest mean as the best alternative, i.e.,

$$\langle 1 \rangle \triangleq \arg\max_{i=1,\ldots,N} \mu_i,$$

where μ_i is obtained by sampling estimation. Let $X_{i,t}$ be the t-th sampling of alternative i. Assume $X_t \triangleq (X_{1,t},\ldots,X_{N,t})$, $t \in \mathbb{Z}^+$, is a joint sampling distribution and follows independent identical distribution (i.i.d.), i.e., $X_t \sim Q(\cdot;\theta)$, whose probability density function is $q(\cdot;\theta)$, where $\theta \in \Theta$ contains all unknown parameters in the parameter family. The marginal distribution of alternative i is $Q_i(\cdot;\theta_i)$, the density is $q_i(\cdot;\theta_i)$, θ_i contains all the unknown the parameters in the marginal distribution, $\mu_i \in \theta_i$, and $(\theta_1,\ldots,\theta_N) \in \theta$. In addition, we assume that the unknown parameters follow a prior distribution, i.e., $\theta \sim F(\cdot;\zeta_0)$, where ζ_0 contains all hyperparameters of the parameter family of the prior distribution.

The dynamic decision in R&S problem can be expressed as an allocation and selection (A&S) policy (Peng et al. 2016). The allocation policy $\mathscr{A}_t(\cdot)$ represents the allocation of the t-th sampling budget to an alternative according to a certain standard, while the selection policy $\mathscr{S}(\cdot)$ represents the selection of the best alternative after exhausting all the sampling budget.

Allocation policies are a sequence of mappings, $\mathscr{A}_t(\cdot) = (A_1(\cdot),\ldots,A_t(\cdot))$, where $A_t(\varepsilon^a_{t-1}) \in \{1,\ldots,N\}$, which allocates the t-th sampling budget to an alternative based on the information set ε^a_{t-1} collected through all the previous steps. The information at step t is given by the formula

$$\varepsilon^a_t \triangleq \{\mathscr{A}_t(\varepsilon^a_{t-1});\varepsilon_t\},$$

where ε_t contains all sample information and prior information ζ_0. Define

$$A_{i,t}(\varepsilon_{t-1}^a) \triangleq \mathbf{1}\{A_t(\varepsilon_{t-1}^a) = i\}.$$

Peng et al. (2018) describe the information collection process in R&S problem that follows the sampling allocation policy. Taking the allocation of four samples among the three alternatives as an example, given the prior information ζ_0, the collected information set ε_4^a is determined by the following two tables. The allocation decision expressed in right table determines the (bold) observable elements in left table.

$$\begin{array}{ccc|ccc}
X_{1,1} & \mathbf{X_{2,1}} & X_{3,1} & A_{1,1}(\zeta_0)=0 & A_{2,1}(\zeta_0)=1 & A_{3,1}(\zeta_0)=0 \\
\mathbf{X_{1,2}} & X_{2,2} & X_{3,2} & A_{1,2}(\varepsilon_1^a)=1 & A_{2,2}(\varepsilon_1^a)=0 & A_{3,2}(\varepsilon_1^a)=0 \\
\mathbf{X_{1,3}} & X_{2,3} & X_{3,3} & A_{1,3}(\varepsilon_2^a)=1 & A_{2,3}(\varepsilon_2^a)=0 & A_{3,3}(\varepsilon_2^a)=0 \\
X_{1,4} & X_{2,4} & \mathbf{X_{3,4}} & A_{1,4}(\varepsilon_3^a)=0 & A_{2,4}(\varepsilon_3^a)=0 & A_{3,4}(\varepsilon_3^a)=1
\end{array} \tag{1}$$

Meanwhile, Peng et al. (2018) show that sampling decision-making and information flow have an interactive relationship. In (1), the interaction between sampling allocation decision and information flow is shown as follows:

$$\zeta_0 \;\rightarrow\; \varepsilon_1^a = \{A_1(\zeta_0)=2;\varepsilon_1\} \;\rightarrow\; \ldots \;\rightarrow\; \varepsilon_4^a = \{A_1(\zeta_0)=2,\ldots,A_4(\varepsilon_3^a)=3;\varepsilon_4\}. \tag{2}$$

As t increases, sampling decisions and information sets are nested within each other. We reorganize our sample observations and put them together in chronological a order, i.e., $\overline{X}_i^{(t)} \triangleq (\overline{X}_{i,1},\ldots,\overline{X}_{i,t_i})$, where $t_i \triangleq \sum_{l=1}^{t} A_{i,l}(\varepsilon_{l-1}^a)$, $i=1,\ldots,N$. Although t_i is also a mapping from the information set, we do not express the dependency on the information set in the notation for simplicity. For example, (3) shows how to reorganize sample observations in (1). We can get

$$\begin{array}{ccc}
\overline{X}_{1,1}=X_{1,2} & \overline{X}_{2,1}=X_{2,1} & \overline{X}_{3,1}=X_{3,4} \\
\overline{X}_{1,2}=X_{1,3} & &
\end{array} \tag{3}$$

The selection is a mapping $\mathscr{S}(\varepsilon_T^a) \in \{1,\ldots,N\}$, which makes the final selection at step T and selects the best solution according to the information collected through the allocation. The reward for a given final choice is a function of θ, i.e., $V(\theta;i)|_{i=\mathscr{S}}$. In R&S, the two most commonly used reward functions are

$$V_P(\theta;i) \triangleq \mathbf{1}\{i=\langle 1\rangle\},$$

$$V_E(\theta;i) \triangleq \mu_i - \mu_{\langle 1\rangle},$$

where subscripts P and E represent PCS and expected opportunity cost (EOC), respectively. $\langle i\rangle$, $i=1,\ldots,N$, is the indices of alternatives in a descending order , which satisfies $\mu_{\langle 1\rangle} > \ldots > \mu_{\langle N\rangle}$. $V_P=1$ if the selected alternative is the true best; otherwise, $V_P=0$. V_E is the difference between the mean of the chosen alternative and the mean of the true best, which measures the EOC of the selection decision. It should be noted that due to the uncertainty of the parameter θ, the values of the final reward V_P and V_E are unknown and need to be quantified through the prior distribution of the parameters in the Bayesian framework.

2.2 R&S as Stochastic Control

Peng et al. (2018) represent the dynamic decision in R&S as an SCP. In the Bayesian framework, the expected values of A&S policy $(\mathscr{A},\mathscr{S})$ can be defined recursively in SCP as

$$V_T(\varepsilon_T^a;\mathscr{A},\mathscr{S}) \triangleq \mathbb{E}[V(\theta;i)|\varepsilon_T^a]|_{i=\mathscr{S}(\varepsilon_T^a)} = \int_{\theta\in\Theta} V(\theta;i)F(d\theta|\varepsilon_T^a)|_{i=\mathscr{S}(\varepsilon_T^a)},$$

where $\mathscr{A} \triangleq \mathscr{A}_T$, $F(\cdot|\varepsilon_t^a)$ is the posterior distribution of θ based on information set ε_t^a. The $d\cdot$ in $d\theta$ represents the Lebesgue measure for continuous distribution and the counting measure for discrete distribution.

$$\begin{aligned} V_t(\varepsilon_t^a;\mathscr{A},\mathscr{S}) &\triangleq \mathbb{E}[V_{t+1}(\varepsilon_t^a \cup \{X_{i,t+1}\};\mathscr{A},\mathscr{S})|\varepsilon_t^a]|_{i=A_{t+1}(\varepsilon_t^a)} \\ &= \int_{\mathscr{X}_i} V_{t+1}(\varepsilon_t^a \cup \{x_{i,t+1}\};\mathscr{A},\mathscr{S}) Q_i(dx_{i,t+1}|\varepsilon_t^a)|_{i=A_{t+1}(\varepsilon_t^a)}, \end{aligned}$$

$\mathscr{X}_i$ is the support set of $X_{i,t+1}$ distribution, and $Q_i(\cdot|\varepsilon_t^a)$ is the predictive distribution of $X_{i,t+1}$ based on information set ε_t^a. Posterior and predictive distributions can be expressed by using Bayesian rules as follows:

$$F(d\theta|\varepsilon_t^a) = \frac{L(\varepsilon_t^a;\theta)F(d\theta;\zeta_0)}{\int_{\theta\in\Theta} L(\varepsilon_t^a;\theta)F(d\theta;\zeta_0)}, \tag{4}$$

$$Q_i(dx_{i,t+1}|\varepsilon_t^a) = \frac{\int_{\theta\in\Theta} Q_i(dx_{i,t+1};\theta_i)L(\varepsilon_t^a;\theta)F(d\theta;\zeta_0)}{\int_{\theta\in\Theta} L(\varepsilon_t^a;\theta)F(d\theta;\zeta_0)}, \tag{5}$$

where $L(\cdot)$ is the likelihood of the sample. The posterior and predictive distributions of normal sampling distributions are discussed in the next section. Through the formula of SCP, we define an optimal A&S policy as

$$(\mathscr{A}^*,\mathscr{S}^*) \triangleq \sup_{\mathscr{A},\mathscr{S}} V_0(\zeta_0;\mathscr{A},\mathscr{S}). \tag{6}$$

We give the Bellman equation of SCP (6). As from the analysis above, the sampling decision and information sets are nested. To avoid tracking the history of the whole sampling allocation decisions, Theorem 1 given by Peng et al. (2018) shows that the posterior distribution and predictive distribution of step t are determined by ε_t.Under the canonical assumptions in R&S, sampling distribution policy will not affect the Bayesian structure, i.e., $(X_{1,t},\ldots,X_{N,t})$, $t\in\mathbb{Z}^+$, is independent, and dependencies between the sampling distributions Q of different alternatives are allowed. Thus, if we define ε_t as the state of step t, then SCP (6) satisfies the optimality equation of MDP.

Next we give a formal description of the R&S problem as an MDP. The MDP has state ε_t, action A_{t+1}, $0\le t<T$, and $\mathscr{S}$, $t=T$. When $0\le t<T$, there is no reward; when $t=T$, the reward is $V_T(\varepsilon_T;\mathscr{S})$. The information set changes as follows at $0\le t<T$,

$$\{\zeta_0,\overline{X}_1^{(t)},\ldots,\overline{X}_N^{(t)}\} \to \{\zeta_0,\overline{X}_1^{(t)},\overline{X}_i^{(t)},X_{i,t+1},\ldots,\overline{X}_N^{(t)}\}|_{i=A_{t+1}} \tag{7}$$

where $X_{i,t+1}\sim Q_i(\cdot|\varepsilon_t)$, $i=A_{t+1}$.We can recursively calculate the optimal A&S policy $(\mathscr{A}^*,\mathscr{S}^*)$ for SCP (6) by using the following Bellman equation:

$$V_T(\varepsilon_T) \triangleq V_T(\varepsilon_T;i)|_{i=\mathscr{S}^*(\varepsilon_T)}, \tag{8}$$

$$\begin{aligned} V_T(\varepsilon_T;i) &\triangleq \mathbb{E}[V(\theta;i)|\varepsilon_T], \\ \mathscr{S}^*(\varepsilon_T) &= \arg\max_{i=1,\ldots,N} V_T(\varepsilon_T;i), \end{aligned}$$

for $0\le t<T$, we have

$$V_t(\varepsilon_t) \triangleq V_t(\varepsilon_t;i)|_{i=A_{t+1}^*(\varepsilon_t)}, \tag{9}$$

$$\begin{aligned} V_t(\varepsilon_t;i) &\triangleq \mathbb{E}[V_{t+1}(\varepsilon_t;X_{i,t+1})|\varepsilon_t], \\ A_{t+1}^*(\varepsilon_t) &= \arg\max_{i=1,\ldots,N} V_t(\varepsilon_t;i). \end{aligned}$$

For an MDP, the equivalence between SCP's optimal policy (6) and the optimal policy (8), (9) can be established directly by induction.

2.3 Update of Posterior Distribution

Under the Bayesian framework, the problem of state estimation is to recursively update the belief θ_t of the current state based on existing information set ε_t. This belief is described by the distribution $f(\cdot|\varepsilon_t)$, which needs to be calculated recursively by predicting and updating steps. The predicting step is to calculate the posterior probability density of current state based on the existing prior information given by $f(\cdot|\varepsilon_{t-1})$. The updating step is to calculate the posterior probability density by using the latest measurement values to modify the prior probability density.

The dimension of the MDP state space mentioned above grows as the steps grow. Under conjugacy, the information set ε_t can be completely determined by the posterior hyperparameter, i.e., $\varepsilon_t = \zeta_t$. Thus, the dimension of the state space is the dimension of the hyperparameter, which is fixed at any step. The following provides a concrete form of conjugation for independent normal distributions with known variances.

Under the Bayesian framework, the conjugate prior of normal distribution $N(\mu_i, \sigma_i^2)$ with unknown mean and known variance is a normal distribution $N(\mu_i^{(0)}, (\sigma_i^{(0)})^2)$, and the posterior distribution of μ_i is $N(\mu_i^{(t)}, (\sigma_i^{(t)})^2)$, where

$$\mu_i^{(t)} = (\sigma_i^{(t)})^2 \left(\frac{\mu_i^{(0)}}{(\sigma_i^{(0)})^2} + \frac{t_i m_i^{(t)}}{\sigma_i^2}\right), \quad m_i^{(t)} \triangleq \frac{\sum_{l=1}^{t_i} \overline{X}_{i,l}}{t_i},$$

$$(\sigma_i^{(t)})^2 = \left(\frac{1}{(\sigma_i^{(0)})^2} + \frac{t_i}{\sigma_i^2}\right)^{-1}.$$

The predictive distribution of $X_{i,t+1}$ is $N(\mu_i^{(t)}, \sigma_i^2 + (\sigma_i^{(t)})^2)$. If $(\sigma_i^{(0)})^2 \to \infty$, then $\mu_i^{(t)} = m_i^{(t)}$. In this case, the prior is a non-informative prior. For a normal distribution with unknown variance, there is a normal conjugate prior.

Next we discuss a feasible way to update a posterior distribution for parameters without conjugate priors. Monte Carlo simulation could be a computationally feasible choice. Due to Markovian property, it is natural to assume that the θ_t at the current time t is only related to θ_{t-1} at the previous time. Also, suppose that the information set ε_t measured at time t is only related to θ_{t-1}. We can generate a large number of sample paths (particles), and use particle filter to iteratively update the posterior measure (Doucet 2001). R&S only requires a special case of particle filter where the particles do not mutate to update the posterior distribution as follows (Peng et al. 2018):

$$\hat{F}(\cdot|\varepsilon_t) = \frac{1}{K}\sum_{i=1}^{K} \mathbf{1}_{\theta_{i,j}^{(t)}}(\cdot) \to \hat{F}(\cdot|\varepsilon_{t+1}) = \frac{1}{K}\sum_{i=1}^{K} \mathbf{1}_{\theta_{i,j}^{(t+1)}}(\cdot),$$

where K is the number of particles, $\theta_{i,j}^{(t)}$ is the j-th particle for θ_i at step t, $j = 1, \ldots, K$, and $\mathbf{1}(\cdot)$ is the delta-measure with mass on x. The particles $\theta_{i,j}^{(t+1)}$ are resampled from $\theta_{i,j}^{(t)}$ with weights

$$w_{i,j} = \frac{q_i(X_{i,t}; \theta_{i,j}^{(t)})}{\sum_{l=1}^{K} q_i(X_{i,t}; \theta_{i,l}^{(t)})}, j = 1, \ldots, K,$$

where $q_i(\cdot)$ is the density for the sampling distribution of the i-th alternative, $i = 1, \ldots, N$. The pseudo-code of the particle filter is shown in Algorithm 1.

3 ROLLOUT POLICY

The idea of rollout policy with a single base policy is: given a fixed budget, perform K rollouts with the corresponding base policy π at the i-th decision node (i.e., selecting the i-th alternative) at step t (i.e.,

Algorithm 1: Particle Filter

Input: $F(\cdot|\varepsilon_t)$, $\varepsilon_t^{(i)}$, N, K
Output: the estimate distribution of the state t: $\hat{F}(\cdot|\varepsilon_{t+1})$

for *i=1 to N* **do**
  particle set initialization: generate sample particle set $\{\theta_{i,j}^{(t)}\}$ by prior F;
  for *j=1 to K* **do**
    calculate and normalize particle weights $w_{i,j} = \frac{q_i(X_{i,t};\theta_{i,j}^{(t)})}{\sum_{l=1}^{K} q_i(X_{i,t};\theta_{i,l}^{(t)})}$;
    resample the particle set$\{\theta_{i,j}^{(t)}, w_{i,j}\}$ to get the new particle set $\{\theta_{i,j}^{(t+1)}, \frac{1}{K}\}$;
  end
  $\hat{F}(\cdot|\varepsilon_{t+1}) = \frac{1}{K}\sum_{i=1}^{K} \mathbf{1}_{\theta_{i,j}^{(t+1)}}(\cdot)$.
end

allocating the $t+1$-th sampling budget), and estimate the action value obtained by action $A_{t+1}^{(i)}$. Budget t is allocated to the alternative with the highest current action value.

Next, we account for the rollout policy in detail and start with the basic notation. $\varepsilon_{t+1}^{(i)}$ represents the updated state after selecting the i-th alternative through action $A_{t+1}^{(i)}$ in ε_t. See see (3) (7) for details on how to update the state. A_{t+1} represents a set of possible actions for allocating $t+1$-th simulation replication, $A_{t+1} = \{A_{t+1}^{(1)}, \ldots, A_{t+1}^{(N)}\}$, and the action that allocates budget t to alternative i is $A_{t+1}^{(i)} = i$, $i = 1, \ldots, N$.

Monte Carlo simulation is used to generate K sample trajectories of dimension $T-t \times N$ at step t, where the remaining budget is $T-t$. These simulations entail the generation of K trajectories by the base policy π, resulting in getting the reward $r_{t,k}^{(i)}$. Note that the allocation process in the R&S problem does not generate rewards,which are only collected at the end of the process through numerous repetitions of rollouts to estimate the current PCS by mean reward. $r_{t,1}^{(i)}, \ldots, r_{t,K}^{(i)}$ constitute independent and identically distributed random variables following the Bernoulli distribution with a distribution parameter of $\mathrm{PCS}^{\pi}(\varepsilon_{t+1}^{(i)})$. $\mathrm{PCS}^{\pi}(\varepsilon_{t+1}^{(i)})$ represents the PCS value in $\varepsilon_{t+1}^{(i)}$ when using policy π to complete the remaining allocation process. $r_{t,k}^{(i)}$ takes value 1 if the optimal alternative is selected and 0 otherwise, with probabilities of $\mathrm{PCS}^{\pi}(\varepsilon_{t+1}^{(i)})$ and $1-\mathrm{PCS}^{\pi}(\varepsilon_{t+1}^{(i)})$, respectively. We define $V^{\pi}(\varepsilon_t, A_{t+1}^{(i)})$ as the theoretical action value of taking action $A_{t+1}^{(i)}$ in ε_t using base policy π, where

$$V^{\pi}(\varepsilon_t, A_{t+1}^{(i)}) = \mathrm{PCS}^{\pi}(\varepsilon_{t+1}^{(i)}).$$

$V_t^{(i)}(\varepsilon_{t+1}^{(i)})$ is an approximation of $V^{\pi}(\varepsilon_t, A_{t+1}^{(i)})$ estimated by rollout, i.e.,

$$V_t^{(i)}(\varepsilon_{t+1}^{(i)}) = \frac{1}{K}\sum_{k=1}^{K} r_{t,k}^{(i)},$$

and when $K \to \infty$, $V_t^{(i)}(\varepsilon_{t+1}^{(i)}) \to \mathrm{PCS}^{\pi}(\varepsilon_{t+1}^{(i)})$. The accuracy of reward value estimation improves with the increase of simulation rollout number K. According to the law of large numbers, the average of $r_{t,k}^{(i)}$ is an unbiased estimate of V^{π}. Theorem 1 in (Bertsekas and Castanon 1998) justifies that the rollout policy is guaranteed to perform at least as well as its base policy if the base policy is consistent. Many common base policies of the greedy type in R&S are consistent. Therefore, in the R&S problems discussed in this paper, the rollout policy is guaranteed to not be worse than the given base policy π when the action value estimate is accurate enough. Not only does the "accuracy" here mean that the estimate is close to the theoretical

value $\text{PCS}^{\pi}(\varepsilon_{t+1}^{(i)})$ but also it refers to the accurate sequence of $V_t^{(i)}(\varepsilon_{t+1}^{(i)})$, i.e., if $\text{PCS}^{\pi}(\varepsilon_{t+1}^{(i)}) > \text{PCS}^{\pi}(\varepsilon_{t+1}^{(j)})$, then $V_t^{(i)}(\varepsilon_{t+1}^{(i)}) > V_t^{(j)}(\varepsilon_{t+1}^{(j)})$. The above situation does not always happen when the number of rollouts is finite, however, we still can guarantee the policy improvement with a certain probability in the current state.

After obtaining the action value, the action taken is:

$$A_{t+1} = \arg\max_{i=1,\ldots,N} V_t^{(i)}(\varepsilon_{t+1}^{(i)}).$$

Algorithm 2 gives the pseudo-code for rollout of a single base policy.

Algorithm 2: Single Base Policy Rollout Iteration

Input: ζ_0, N, T, F_0
Output: S^*

$\varepsilon_0 = \zeta_0$;
for *t=1 to T-1* **do**
  draw a set of simulation sample space from belief F_t;

  for *i=1 to N* **do**
    $A_{t+1}^{(i)} \leftarrow$ take action on the previous state ε_t;
    $\varepsilon_{t+1}^{(i)} \leftarrow$ update the state based on $A_{t+1}^{(i)}$;

    $F_{t+1}^{(i)} \leftarrow$ update the parameter posterior distribution;

    $V_{t+1}^{(i)}(\varepsilon_{t+1}^{(i)}) \leftarrow$ estimate the value by K rollouts using base policy π;

  end
  $A_{t+1}^*(\varepsilon_t) \leftarrow \arg\max_{i=1,\ldots,N}(V_{t+1}^{(i)}(\varepsilon_{t+1}^{(i)}))$;

  $\varepsilon_{t+1} \leftarrow$ update the state based on A_{t+1}^*;

  $F_{t+1} \leftarrow$ update the parameter posterior distribution;

end
$S^* \leftarrow$ select the alternative with the largest parameter posterior mean.

We can also use the parallel rollout policy (Chang et al. 2004). Given a base policy set $\Pi = \{\pi_1, \ldots, \pi_H\}$, calculate the reward value according to each single policy rollout, the reward value of alternative i is the maximum among H single policy reward values, and the budget is finally allocated to the alternative with the largest current reward value.The parallel rollout policy is guaranteed to be at least as good as the best base policy in Π. The formula of reward value is

$$V_{t,\pi_h}^{(i)}(\varepsilon_{t+1}^{(i)}) = \frac{1}{K}\sum_{k=1}^{K} r_{t,k,\pi_h}^{(i)},$$

and when $K \to \infty$, $V_{t,\pi_h}^{(i)}(\varepsilon_{t+1}^{(i)}) \to V^{\pi_h}(\varepsilon_t, A_{t+1}^{(i)}) = \text{PCS}^{\pi_h}(\varepsilon_{t+1}^{(i)})$. The optimal value function is

$$\tilde{V}_t^*(\varepsilon_{t+1}^{(i)}) = \max_{\pi_h \in \Pi}(V_{t,\pi_h}^{(i)}(\varepsilon_{t+1}^{(i)})),$$

which in turn leads to the following parallel rollout policy:

$$\tilde{A}^*_{t+1}(\varepsilon_t) = \arg\max_{i=1,\ldots,N} (\tilde{V}^*_t(\varepsilon_t^{(i)})).$$

This approach addresses the problem of base policy selection when it is impossible to determine which allocation policy is more effective in the current environment.

4 SIMULATION EXPERIMENTS

In this section, we test the rollout procedure with different base policies by comparing it with existing classical sampling approaches. The following five sampling allocation policies are used for comparison: KG with uninformative prior (Frazier et al. 2007); AOAP with uninformative prior (Peng et al. 2018); EI with uninformative prior (Donald et al. 1998); EA that sequentially allocates the first to the last alternative in a cyclical manner; PTV, for which the number of allocated simulation replications to each alternative is proportional to its sample variance, implemented sequentially by the "most starving" rule.

In order to better understand the numerical experiment, in conjunction with the description of the R&S problem in Section 2, we once again clarify how is synthetic data generated (to mimic a real world use case) and what are algorithm settings. Under the assumption of a normal sampling distribution, i.i.d. samples from each alternative i are generated by $X_i \sim N(\mu_i, \sigma_i^2)$, where μ_i is a ground truth mean performance randomly generated at the start of each experiment and σ_i is the true standard deviation. We conduct numerical experiments under the assumption that μ_i follows normal, beta, gamma, and beta + binomial distributions, respectively. The selection policy is fixed as

$$\hat{S}(\varepsilon_t) = \langle 1 \rangle_t,$$

which selects the alternative with the largest posterior mean. The numerical performance of each sampling procedure is measured by the PCS following the sampling procedure to allocate a fixed amount of simulation budget, i.e.,

$$\mathrm{PCS_t} \triangleq \mathbb{E}\left[\mathbf{1}\{\hat{S}(\varepsilon_t) = \langle 1 \rangle\}\right.$$

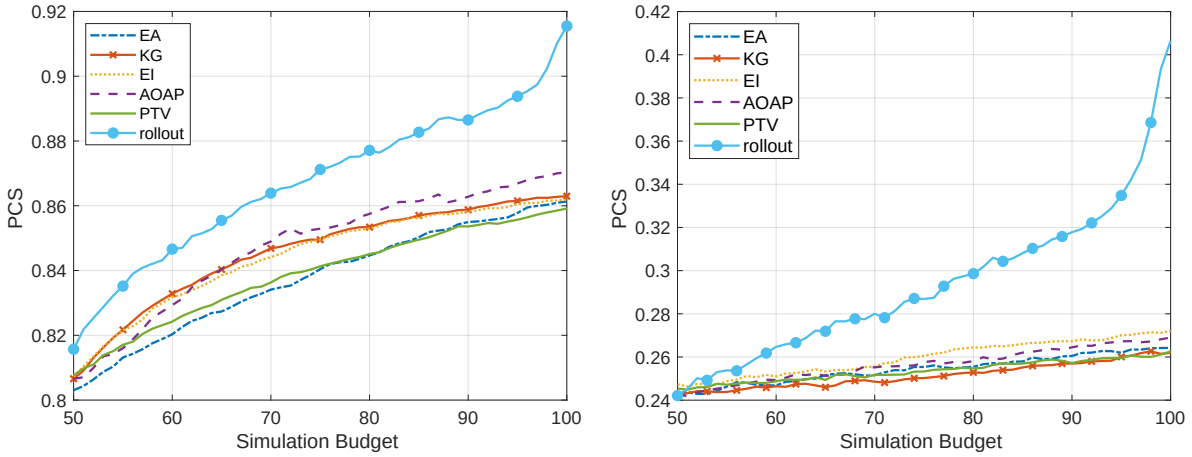
Numerical experiments show that the proposed policy greatly improves the performance of the base policy in various scenarios. All statistics in this section are estimated from 10^4 independent macro-simulations.

4.1 Normal Distribution with Conjugate Prior

To compare rollout policy whose base policy is EA and AOAP, respectively, with KG, AOAP, EA and PTV in the normal sampling distributions, we use two typical scenarios with high-confidence and low-confidence (Peng et al. 2018). The priors for the unknown means of the normal sampling distributions are assumed to be the normal conjugate priors introduced in Section II, i,e, $\mu_i \sim N(\mu_i^{(0)}, (\sigma_i^{(0)})^2)$. We control the confidence of the scenario by adjusting the mean $\mu_i^{(0)}$ and standard deviation $\sigma_i^{(0)}$ of the true mean μ_i.

A High-Confidence Scenario In the first example, the number of alternatives is $N = 5$ and the total simulation budget is $T = 100$. The high-confidence scenario prior hyper-parameters are given by $\mu_i^{(0)} = 0$, $(\sigma_i^{(0)})^2 = 1$, the true standard deviation $\sigma_i = 1$, $i = 1, \ldots, 5$.

High-confidence scenarios are qualitatively described by three characteristics: large differences between the performance of different scenarios, small variances, and large simulation budget. The standard deviation of the prior distribution controls the dispersion of the true mean following the prior distribution. This example can be categorized as a high-confidence scenario. The first 50 sampling budgets are allocated equally to each alternative. and the remaining simulation replications are allocated according to different sampling allocation procedures.The number of rollouts per step is $K = 50$. Figure 1 (a) shows the trajectory of PCS when the simulation budget increases from 50 to 100.



(a) high-confidence scenario

(b) low-confidence scenario

Figure 1: PCS of the six policies for normal distribution scenarios.

As can be seen from Figure 1 (a), the rollout policies significantly outperforms other policies with the increase of simulation budget, and the rollout policies lead to 4% increase in PCS compared with the classic R&S policies at the end. EA and AOAP are almost indistinguishable, whereas KG is slightly better than other classic R&S policies.

A Low-Confidence Scenario In this example, the number of alternatives is $N = 5$ and the total simulation budget is $T = 100$. The low-confidence scenario prior hyper-parameters are $\mu_i^{(0)} = 0$, $(\sigma_1^{(0)})^2 = 0.002$, $i = 1$, $(\sigma_i^{(0)})^2 = 0.001$, $i = 2,\ldots,5$, the standard deviation $\sigma_i = 1$, $i = 1,\ldots,5$.

Low-confidence scenarios are qualitatively described by three characteristics: the differences between the means of competing designs are small, the variances are large, and small simulation budget. The first 50 initial simulation replications are allocated equally to five alternatives. The number of rollouts per step is $K = 50$.

As can be seen from Figure 1 (b), PCS of four base sequential policies slowly increase with the increase of simulation budget, AOAP and KG have a slight advantage over the other two classic R&S policies. PCS of the rollout policies grows rapidly, significantly faster than PCS obtained by other policies and lead to 14%-20% increase in PCS at the end. Moreover, it is observed that AOAP outperforms EA, and the rollout policy with AOAP as the base policy also performs significantly better than the one with EA as the base policy, which corroborates our theoretical result that the rollout policy performance depends on the base policy performance.

4.2 General Bayesian Distribution

We use sampling importance resampling filter (SIR), a popular particle filter algorithm which combines sequential importance sampling and resampling to update the posterior distribution of parameters. How to choose the number of sample particles N_P is an important problem in the trade-off between compuyation time and accuracy. We conduct experiment with $N_P = 50, 100, 500$ in the above two normal distribution scenarios of Section 4.1, and select Root Mean Squared Error (RMSE) to measure the estimation accuracy. RMSE between policies with conjugate prior and SIR in Table 1 and CPU time in Table 2 indicates that, in general, more particles increase the estimation accuracy and computation time. In experiments, $N_P = 50$ is adopted for SIR.

Under the general Bayesian framework, we assume that the sampling distributions are normal with unknown mean and known variance, and give numerical examples to test the proposed rollout policy (the base policy is EA), EI, EA and PTV in different distribution scenarios. The priors for the unknown means

Table 1: RMSE of each policy.

High	Policy	EA-SIR50	EA-SIR100	EA-SIR500	KG-SIR50	KG-SIR100	KG-SIR500
	RMSE	0.010	0.009	0.004	0.010	0.009	0.005
Low	Policy	EA-SIR50	EA-SIR100	EA-SIR500	KG-SIR50	KG-SIR100	KG-SIR500
	RMSE	0.011	0.009	0.005	0.010	0.006	0.003

Table 2: CPU time of each policy.

High	Policy	EA-SIR50	EA-SIR100	EA-SIR500	KG-SIR50	KG-SIR100	KG-SIR500
	time(s)	0.0216	0.0235	0.1003	0.0243	0.0371	0.1847
Low	Policy	EA-SIR50	EA-SIR100	EA-SIR500	KG-SIR50	KG-SIR100	KG-SIR500
	time(s)	0.0092	0.0214	0.1938	0.0145	0.0281	0.2146

are assumed to be beta distribution $B(\alpha,\beta)$, gamma distribution $G(\varphi,\gamma)$ and a Normal-Binomial distribution $N(\mu^*,(\sigma^*)^2)+B(N,p)$ respectively. In the following three examples, the number of alternatives is $N=5$ and the total simulated budget is $T=100$. The true standard deviation is $\sigma_i=1$, $i=1,\ldots,5$. The first 50 simulation replications are equally allocated to each alternative. The number of rollouts per step is $K=50$. The number of sampling particles in particle filter is $N_P=50$.

Beta Distribution The Beta distribution scenario hyper-parameters in the prior are given by $\alpha=1$, $\beta=3$, i.e., $\mu_i\sim B(1,3)$. From Figure 2 (a), we can see the PCSs of EI, EA and PTV increase with the simulation budget. The PCS growth rate of the rollout policy is obviously faster, and the PCS of the proposed rollout is significant larger than the PCSs obtained by the other policies.

Gamma Distribution The Gamma distribution scenario hyperparameters in the prior are given by $\varphi=2$, $\gamma=1$, i.e., $\mu_i\sim G(2,1)$.

As can be seen from Figure 2 (b), the three sequential policies have comparable performance, the PCSs of EI, EA and PTV increase with the simulation budget, and EI has a slight advantage over EA and PTV. The numerical results show that rollout performance is close to EI before the simulated budget reaches 60, and PCS of the rollout policy increases much faster than PCS of EI after that.

Normal-Binomial Distribution The Normal-Binomial distribution scenario hyper-parameters in the prior are given by $\mu_i^*=0$, $(\sigma_i^*)^2=0.001$, $i=1,\ldots,5$, $N=5$, and $p=0.5$, i.e., $\mu_i\sim N(0,0.001)+B(5,0.5)$.

From Figure 2 (c), we can know there is little difference between the three sampling policies in this scenario, PTV and EA have a slight advantage, EI has the worst performance, and the rollout policy again shows a tremendous advantage in this scenario.

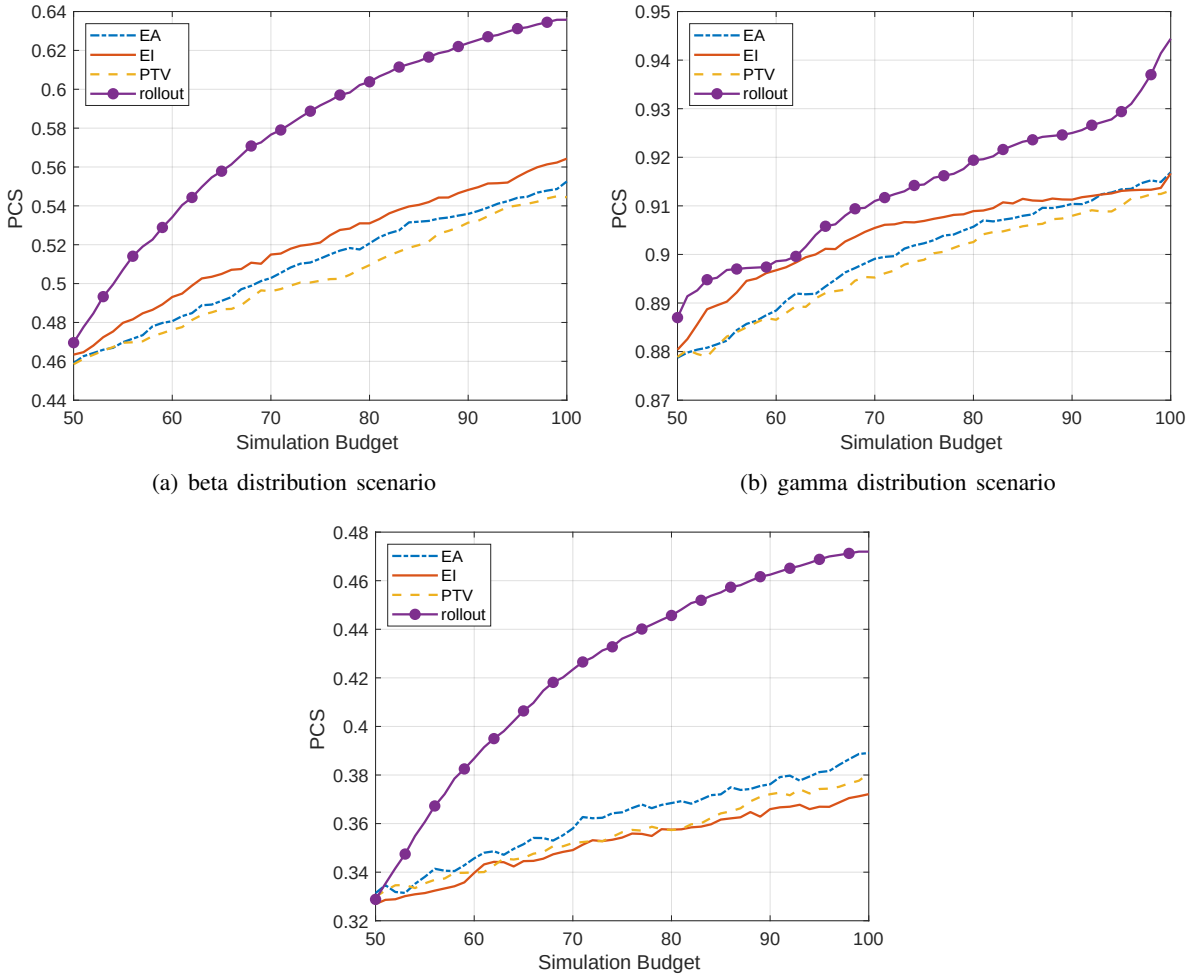
Through the above numerical experiments, we can see that rollout can greatly improve the performance of base policies in all scenarios where the performance of the classic R&S policies is poor. And in scenarios where the policies perform well, rollout also achieves an improvement.

5 CONCLUSION

In this paper, we study the R&S problem under a POMDP paradigm, and we propose a rollout policy to solve POMDP through online Monte Carlo simulation. By combining with existing allocation policies, the proposed rollout policy effectively improves the PCS of the existing allocation policies. The rollout policy is consistent and its performance is guaranteed to be at least as good as the base policy with a certain probability.

The policy proposed in this paper has a wide range of applicability. Updating parameter posterior distribution by particle filter makes it possible to solve general R&S problems. Several numerical results show that this policy can perform well in various scenarios of R&S problem.

The limitation of the rollout policy lies in the high consumption of computing resources for large-scale R&S problems. In the future, deep neural network combined with rollout can be adopted to build an AI framework to solve this problem. The idea resembles AlphaGo, which integrates MCTS and deep neural



(a) beta distribution scenario

(b) gamma distribution scenario

(c) normal-binomial distribution scenario

Figure 2: PCS of the four policies for general Bayesian distribution scenario.

network trained offline by rollout. In addition, extending this framework to the R&S problem under the non-normal sampling assumption is also a possible direction.

REFERENCES

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.

Bertsekas, D., and D. Castanon. 1998. "Rollout Algorithms for Stochastic Scheduling Problems". *Journal of Heuristics* 5:89–108.

Chang, H. S., R. Givan, and E. Chong. 2004. "Parallel Rollout for Online Solution of Partially Observable Markov Decision Processes". *Discrete Event Dynamic Systems* 14(3):309–341.

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

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.

Donald, R. J., S. Matthias, and J. W. William. 1998. "Efficient Global Optimization of Expensive Black-Box Functions". *Journal of Global Optimization* 13:455–492.

Doucet, A. 2001. *Sequential Monte Carlo Methods*. Wiley Online Library.

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.

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* 8:321–343.

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

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.

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* 63:1177–1194.

Ni, E. C., D. F. Ciocan, S. G. Henderson, and S. R. Hunter. 2015. "Efficient Ranking and Selection in Parallel Computing Environments". *Operations Research* 65(3):821–836.

Peng, Y., C. H. Chen, E. Chong, and M. C. Fu. 2018. "A Review of Static and Dynamic Optimization for Ranking and Selection". In *Proceedings of the 2018 Winter Simulation Conference*, 1909–1920. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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.

Rinott, Y. 1978. "On two-stage selection procedures and related probability-inequalities". *Communications in Statistics-theory and Methods* 7:799–811.

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

Tesauro, G., and G. R. Galperin. 1996. "On-line Policy Improvement using Monte-Carlo Search". In *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*, 1068–1074.

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.