Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

THOMPSON SAMPLING MEETS RANKING AND SELECTION

Yijie Peng
Gongbo Zhang

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

ABSTRACT

Ranking and selection has been actively studied in simulation. We briefly review the ranking and selection problem and some existing sampling procedures. Thompson sampling is originally proposed for the multi-armed bandits problem, whereas its variant top-two Thompson sampling can perform better in the ranking and selection problem. We compare the top-two Thompson sampling comprehensively with some popular sampling procedures from both theoretical and numerical perspectives.

1 INTRODUCTION

Ranking and selection (R&S) has been actively studied in simulation for decades (Bechhofer 1954; Bechhofer et al. 1995; Kim 2013). The goal of a classic R&S problem is to select the best alternative with the smallest (or largest) mean from a finite number of alternatives with unknown means $\mu_1, \cdots, \mu_k$. In the context of simulation, each alternative is typically a complex stochastic model (e.g., queueing network), so the mean of the random output of the model (e.g., average system time) usually does not have an analytical form but can be estimated by simulation. Correlations between the outputs of different stochastic models could be introduced by common random numbers (Chick and Inoue 2001a). For example, random numbers can be shared for generating arrival processes of different queueing models so that positive correlations between simulation outputs of different models are introduced.

The unknown means of alternatives can be estimated by independent random samples (simulation replications) following a joint distribution, i.e., $(X_{1,\ell}, \cdots, X_{k,\ell}) \sim Q(\cdot;\theta)$, $\ell \in \mathbb{Z}^+$, where θ contains all unknown parameters of the sampling distribution. Obviously, $\mu_1, \cdots, \mu_k$ are a part of θ. The most common assumption in the literature of R&S is that the samples of different alternatives, i.e., $X_{i,\ell}, X_{j,\ell}$, $i \neq j$, follow independent normal distributions. In practice, the output of a stochastic model is often an average of many random variables (e.g., average system time of 100 customers) so that the normal assumption of samples is justified by certain central limit theorem. If the samples do not follow normal distribution, then a macro-replication that is an average of a batch of simulation replications is used as a sample in experiment, which is approximately normally distributed. By law of large number, sample means converge to means almost surely (a.s.), i.e., $\sum_{\ell=1}^{n} X_{i,\ell}/n \to \mu_i$ a.s., as $n \to \infty$, $i = 1, \cdots, k$. Thus as simulation budget goes to infinity, the best alternative could be eventually selected. However, simulation replications are typically expensive so that simulation budget is often very limited in practice.

Figure 1 presents 99% corresponding confidence intervals for the means of five alternatives. In this example, it is highly unlikely for alternatives 1, 4, and 5 to be the best alternative, so it appears to be unreasonable to spend a significant amount of simulation replications on learning the means of alternatives 1, 4, and 5 more accurately rather than focusing on learning the means of the other two more promising alternatives to distinguish which one is actually the best. This example motivates the study of a central

issue in the research of R&S, i.e., how to allocate simulation replications intelligently to improve sampling efficiency. Intuitively, allocating more simulation replications to the alternatives with smaller means and higher variances seems to be reasonable for efficiently selecting the best alternative. This intuition is summarized as "mean-variance" or "exploitation-exploration" tradeoff in the literature.

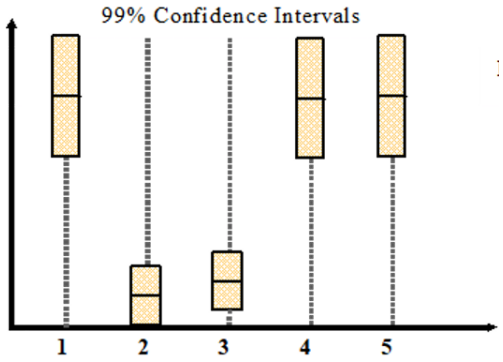


Figure 1: 99% confidence intervals for the means of five alternatives.

To formalize the study on the issue, we need to introduce a metric to measure sampling efficiency. In R&S, a popular metric is the probability of correct selection (PCS), i.e., $PCS(\theta) := P\big(\mu_{\mathscr{S}} < \mu_i,\ i \neq \mathscr{S}|\theta\big) = \mathbb{E}\big[\mathscr{L}_{0-1}(\mathscr{E}_T)|\theta\big]$, where $\mathscr{S}$ is the selection decision made after allocating a total of T simulation replications, $\mathscr{E}_t$ is the information set obtained after allocating t simulation replications, and $\mathscr{L}_{0-1}(\mathscr{E}_t) := \mathbf{1}\{\mathscr{S} = i^*\}$ is the zero-one loss after allocating t simulation replications with $i^* := \arg\max_{i=1,\cdots,k} \mu_i$; another popular metric is expected opportunity cost (EOC), i.e., $EOC(\theta) := \mathbb{E}\big[\mathscr{L}_l(\mathscr{E}_T)|\theta\big]$, where $\mathscr{L}_l(\mathscr{E}_t) := \mu_{\mathscr{S}} - \mu_{i^*}$ is the zero-one loss after allocating t simulation replications (Branke et al. 2007). Notice that for the metric commonly used in R&S, the loss only occurs at the end after allocating all T simulation replications.

Multi-armed bandits (MAB) problem has even longer history and richer literature than R&S (Robbins 1952; Lai 1987; Russo et al. 2018). The name of MAB comes from a motivating story in which a gambler enters a casino and sits down at a slot machine with multiple arms (alternatives) that can be pulled (sampled). Each time when he/she pulls an arm, the slot machine outputs a random reward (simulation replication) independent of the past. The goal of a classic MAB problem is the same as that of R&S, i.e., selecting the best arm from a finite number of arms with unknown expected rewards (means). A major difference than R&S lies in that the metric for measuring sampling efficiency is usually an expected total regrets accumulated each time when pulling an arm, i.e., $\mathbb{E}\big[\mathscr{L}_{cr}(\mathscr{E}_T)|\theta\big]$, where $\mathscr{L}_{cr}(\mathscr{E}_T) := \sum_{t=1}^{T}\big(\mu_{A_t} - \mu_{i^*}\big)$ with A_t being the decision for allocating the t-th sample based on available information. Interestingly, a stream of researches recently gaining steam in MAB study a best arm identification (BAI) or pure exploration problem, which uses same metric (e.g., PCS) as in R&S.

A well balance of exploitation and exploration is also the key for intelligently sampling each alternative in the MAB problem. ε-Greedy, upper confidence bound (UCB) (Auer et al. 2002), and Thompson sampling (TS) (Thompson 1933) are popular sampling procedures in the MAB domain, which can achieve certain asymptotically optimal growth rate of the expected total regrets. TS appears to be superior because it does not require to well tune some hyper-parameter and typically leads to better finite-sample performance than the other two procedures. However, the performance of TS is notoriously poor for the R&S or BAI problem. Recently, Russo (2020) proposes a top-two Thompson sampling (TTTS) procedure which performs well for the R&S problem. Compared to TS, TTTS tends to explore more when allocating samples.

Both TS and TTTS are Bayesian sequential sampling procedures. Peng et al. (2018a) show that the sequential sampling decisions in R&S can be formulated as Markov decision process (MDP) under a Bayesian framework, and rigorously establish the Bellman equation for MDP. In principle, stochastic dynamic programming can be used to compute the optimal sampling policy, but in general this is computationally

intractable due to curse-of-dimensionality. This problem is clearly stated in both Peng et al. (2018a) and Russo (2020), which take different routes to circumvent the computational difficulty.

In R&S, numerous existing sampling procedures can be basically categorized into two branches developed under two distinctive philosophies. An earlier branch is studied under an indifference zone (IZ) paradigm (Goldsman and Nelson 1998; Kim and Nelson 2006), which tries to guarantee a predetermined PCS level for parameters satisfying an IZ assumption, i.e., $\min_{\theta\in\Xi} PCS(\theta) \geq \alpha$ with $\Xi := \{\theta \in \Theta:\ \mu_{i^*} + \delta < \mu_i,\ i \neq i^*\}$. Since the sampling procedures derived in the IZ paradigm need to guarantee a PCS level even for the worse case which is called the slippage configuration (Branke et al. 2007), they usually require to allocate more simulation replications than necessary for guaranteeing a PCS level in practice. A recent theoretical breakthrough has been made by Fan et al. (2016) to remove the IZ assumption for guaranteeing PCS, but the drawback of conservativeness still remains. Application backgrounds particularly suitable for methods in this branch include drug test, where guaranteeing a high probability that a lifesaving drug is most effective is valuable. However, those methods may not be best suited for applications like designing a complex manufacturing system, where the possible number of alternative designs could be huge, running a simulation model would be expensive, and a managerial decision might need to be made fast adaptive to varying market conditions. The famous Go-playing artificial intelligence (AI) algorithm AlphaGo is build upon a Monte Carlo tree search (MCTS) backbone (Silver et al. 2016), where there are a scale of 10^{170} possible states, which are more than the number of atoms in the universe. For estimating the state-action value function by Monte Carlo simulations following procedures of node selection and rollout, squandering more simulation source than necessary in order to guarantee PCS would make the algorithm fail in such as a hard problem. In R&S, the other relatively new branch is to maximize a performance metric given a simulation budget constraint (Chen and Lee 2011; Powell and Ryzhov 2012). The methods in this branch do not generally guarantee PCS but can typically lead to better performance given a fixed simulation budget than those in the earlier branch. In the literature of R&S, the first branch is referred to as the frequentist or fixed-precision branch, and the second branch is called the Bayesian or fixed-budget branch. Recently, Hong et al. (2021) view the methods in the first branch from a hypothesis testing perspective, whereas they offer a dynamic programming perspective for the second branch.

TS and TTTS can be well fitted into the Bayesian branch, and we briefly introduce the methods in this branch. Optimal computing budget allocation (OCBA) in Chen et al. (2000) is developed by solving a static optimization problem, i.e., $\max_{n_1+\cdots+n_k=T} PCS(\theta)$ with n_i being the number of simulation replications allocated to alternative i. With several approximations, an asymptotically optimal sampling ratio with an analytical formula can be obtained given unknown parameter θ. Then an estimate for θ is plugged into the OCBA formula using initial samples equally allocated to each alternative in the first stage so that the remaining samples can be allocated according to the sampling ratio suggested by OCBA for enhancing sampling efficiency. Here the sampling efficiency should not be defined as the PCS given a specific parameter θ. For a given parameter, we can directly select the best alternative by sorting means which are a part of θ without running simulation. Therefore, a reasonable performance metric could be an integrated PCS (IPCS) over a weighting prior measure $F(\cdot)$ on parameter space Θ, i.e., $IPCS := \int_{\theta\in\Theta} PCS(\theta)F(d\theta)$. Under a Bayesian framework, a dynamic allocation and selection (A&S) policy to optimize IPCS is defined in Peng et al. (2016), and a Bellman equation for solving the optimal A&S policy is rigorously established in Peng et al. (2018a). Well-known Bayesian sequential sampling procedures include expected value of improvement (EVI) in Chick and Inoue (2001b), knowledge gradient (KG) in Frazier et al. (2008), and expected improvement (EI) in Ryzhov (2016). KG and EI sequentially allocate a sample to optimize myopic surrogate criteria. Asymptotically optimal allocation procedure (AOAP) proposed in Peng et al. (2018a) sequentially allocates a sample to optimize an approximation of the value function one-step look ahead in the MDP. The asymptotic sampling ratio following AOAP is proved to optimize the large deviations rate of PCS defined in Glynn and Juneja (2004). TTTS is also proved to achieve asymptotically optimality which is defined differently than that in Glynn and Juneja (2004).

In this work, we will first introduce TTTS and compare its theoretical properties with some existing sampling procedures in the Bayesian branch of R&S. Then comprehensive numerical comparisons between TTTS and various R&S sampling procedures will be conducted in many experiments, which appears to have never been done in the literature. Based on numerical results, we will provide guidelines for practitioners on how to choose suitable sampling procedures in different applications.

The rest of the paper is organized as follows. In Section 2, we view TTTS from a stochastic dynamic programming perspective in the Bayesian framework of R&S and discuss theoretical properties of TTTS. Section 3 presents numerical experiments. Conclusions are given in the last section.

2 THOMPSON SAMPLING IN BAYESIAN FRAMEWORK OF R&S

We first introduce a Bayesian framework for the R&S problem. Suppose θ follows a prior distribution $F(\cdot;\zeta_0)$ that reflects our prior knowledge on the unknown parameter, where ζ_0 in the prior distribution is called a hyper-parameter. Let $X_i^{(t)} := (X_{i,1},\cdots,X_{i,t_i})$ and $\mathscr{E}_t := \{\zeta_0, X_1^{(t)},\cdots,X_k^{(t)}\}$, where t_i is the number of simulation replications received by alternative i after allocating t simulation replications. The posterior distributions can be calculated using Bayes rule. In the case where the prior distribution is a conjugate prior of the sampling distribution, the posterior distribution lies in the same parametric family of the prior distribution, i.e., $F(\cdot;\zeta_t)$ where ζ_t is the posterior hyper-parameter. Under the conjugate prior, the information set $\mathscr{E}_t$ can be completely determined by the posterior hyper-parameters, i.e., $\mathscr{E}_t = \zeta_t$. The conjugate prior for the normal distribution $N(\mu_i,\sigma_i^2)$ with unknown mean and known variance is a normal distribution $N(\mu_i^{(0)},(\sigma_i^{(0)})^2)$. The posterior distribution of μ_i is $N(\mu_i^{(t)},(\sigma_i^{(t)})^2)$, where $\mu_i^{(t)} = (\sigma_i^{(t)})^2\big(\mu_i^{(0)}/(\sigma_i^{(0)})^2 + t_i\bar{X}_i^{(t)}/\sigma_i^2\big)$ and $(\sigma_i^{(t)})^2 = \big(1/(\sigma_i^{(0)})^2 + t_i/\sigma_i^2\big)^{-1}$ with $\bar{X}_i^{(t)} := \sum_{\ell=1}^{t_i} X_{i,\ell}/t_i$. If $\sigma_i^{(0)} \to \infty$, $\mu_i^{(t)} = \bar{X}_i^{(t)}$, and the prior is the uninformative prior in this case. For general sampling and prior distributions, algorithms for updating posterior distribution efficiently can be found in Russo et al. (2018).

The dynamic decisions in the R&S problem can be formulated as an A&S policy (Peng et al. 2016). The allocation policy is a sequence of mappings $\mathscr{A}_t(\cdot) = (A_1(\cdot),\cdots,A_t(\cdot))$ that sequentially allocates each sample to an alternative based on collected information, and the selection policy $\mathscr{S}(\cdot)$ makes a final decision to select the best alternative after exhausting all simulation replications. For a sequential allocation policy, $t_i = \sum_{\ell=1}^{t} \mathbf{1}\{A_\ell(\mathscr{E}_{\ell-1}) = i\}$. For simplicity, the selection decision is usually fixed as picking the alternative with the smallest sample or posterior mean, and more attention has been focused on sampling decision in the literature. Exceptions on discussing the optimal selection policy include Peng et al. (2016) and Eckman and Henderson (2022). Peng et al. (2018a) shows that the sampling allocation policy would not affect the Bayesian structure under a canonical assumption in R&S, and the optimal A&S policy $(\mathscr{A}_T^*,\mathscr{S}^*)$ satisfies the following Bellman equation: $V_T(\mathscr{E}_T) := V_T(\mathscr{E}_T;i)\big|_{i=\mathscr{S}^*(\mathscr{E}_T)}$, where $V_T(\mathscr{E}_T;i) := \mathbb{E}[\mathbf{1}\{i=i^*\}|\mathscr{E}_T]$ and $\mathscr{S}^*(\mathscr{E}_T) = \arg\max_{i=1,\cdots,k} V_T(\mathscr{E}_T;i)$, and for $0 \le t < T$, $V_t(\mathscr{E}_t) := V_t(\mathscr{E}_t;i)\big|_{i=A_{t+1}^*(\mathscr{E}_t)}$, where $V_t(\mathscr{E}_t;i) := \mathbb{E}[V_{t+1}(\mathscr{E}_t,X_{i,t+1})|\mathscr{E}_t]$ and $A_{t+1}^*(\mathscr{E}_t) = \arg\max_{i=1,\cdots,k} V_t(\mathscr{E}_t;i)$. Both Peng et al. (2018a) and Russo (2020) realize computational intractability of solving this Bellman equation except for some toy examples.

Russo (2020) proposes "simple and intuitive rules for adaptively allocating measurement effort" and then analyzes the asymptotic properties of the proposed method. Based on available information $\mathscr{E}_{t-1}$, TS allocates the t-th simulation replication randomly to alternative i with the posterior probability that alternative i is the best, i.e., $\alpha_{i,t} := P(\mu_i > \mu_j,\ j \ne i|\mathscr{E}_t)$, $i = 1,\cdots,k$. This can be easily implemented by allocating the t-th simulation replication to alternative $I_t = \arg\max_{i=1,\cdots,k} \tilde{\mu}_i$, where $\tilde{\mu}_i$, $i = 1,\cdots,k$, are sampled from posterior distribution $\tilde{\theta} \sim F(\cdot|\mathscr{E}_t)$. Assuming independent normal sampling distributions with known variances and conjugate priors, $\tilde{\mu}_i \sim N(\mu_i^{(t)},(\sigma_i^{(t)})^2)$, $i = 1,\cdots,k$. Numerical results show that TS performs poorly in R&S, because TS tends to overly prefer the alternative with the largest posterior probability in allocating samples. Notice that following a policy sampling all alternatives infinitely often, $\alpha_{i^*,t} \to 1$ as $t \to \infty$. To encourage more exploration, TTTS proposed in Russo (2020) allocates the t-th simulation replication to alternative I_t with a given probability β, and with probability $1-\beta$, it samples an

alternative different than I_t. To generate this alternative, TTTS continues sampling $J_t = \arg\max_{i=1,\cdots,k} \tilde{\mu}_i$ until $J_t \neq I_t$, and then allocate the t-th simulation replication to alternative J_t. TTTS samples alternative i with probability $\psi_{i,t} := \alpha_{i,t}\beta + \alpha_{i,t}(1-\beta)\sum_{j\neq i} \frac{\alpha_{j,t}}{1-\alpha_{j,t}}$, where the first part in the summation is the probability that alternative i is chosen as I_t and the second part is the probability that it is chosen as J_t. Following a policy sampling all alternatives infinitely often, $\psi_{i^*,t} \to \beta$ as $t \to \infty$. In implementation, the loop for computing J_t could be very long when $\max_{i=1,\cdots,k} \alpha_{i,t}$ is close to one. For computational efficiency, we may need to end the loop when the number of iterations reaches certain cutoff value.

The optimal policy in MDP can be proved to be a deterministic mapping from the state space to decision space (Bertsekas 1995). However, a randomized policy may have computational benefits. TTTS is a randomized policy that makes sampling decisions based on posterior distributions adaptive to available sample information, i.e., $A_t \sim P(\cdot|\mathscr{E}_{t-1})$. Under certain regularity conditions, Russo (2020) proves that the posterior probability of incorrect selection cannot converge to zero at a rate faster than $e^{-t\Gamma^*}$ following any adaptive allocation policy, i.e., $-\lim_{t\to\infty} \frac{1}{t}\log F(\Theta_{i^*}^c|\mathscr{E}_t) = -\lim_{t\to\infty} \frac{1}{t}\log\left(\sum_{i\neq i^*} \alpha_{i,t}\right) \leq \Gamma^*$, where $\Theta_i := \{\theta \in \Theta:\ \mu_i > \mu_j,\ j \neq i\}$, and $\Gamma^* := \max_{\psi} \min_{\theta' \in \Theta_{i^*}^c} \sum_{i=1}^k \psi_i d(\theta_i'||\theta_i)$ with $\psi = (\psi_1, \cdots, \psi_k)$ being a probability distribution and $d(\theta_i'||\theta_i)$ being the Kullback-Leibler divergence between the marginal sampling distribution $Q_i(\cdot;\theta_i')$ and $Q_i(\cdot;\theta_i)$ of the i-th alternative; following TTTS with parameter β, $-\lim_{t\to\infty} \frac{1}{t}\log F(\Theta_{i^*}^c|\mathscr{E}_t) = \Gamma_\beta^*$, where $\Gamma_\beta^* := \max_{\psi:\psi_{i^*}=\beta} \min_{\theta' \in \Theta_{i^*}^c} \sum_{i=1}^k \psi_i d(\theta_i'||\theta_i)$, $\lim_{t\to\infty} \bar{\psi}_t = \psi^\beta$, where $\bar{\psi}_t = (\bar{\psi}_{1,t}, \cdots, \bar{\psi}_{k,t})$ with $\bar{\psi}_{i,t} := \sum_{\ell=1}^t \psi_{i,\ell}/t$ and $\psi^\beta := \arg\max_{\psi:\psi_{i^*}=\beta} \min_{\theta' \in \Theta_{i^*}^c} \sum_{i=1}^k \psi_i d(\theta_i'||\theta_i)$; moreover, $\Gamma^* = \Gamma_{\beta^*}^*$ with $\beta^* := \arg\max_\beta \Gamma_\beta^*$. If β_t is adaptive to $\mathscr{E}_{t-1}$ and $\beta_t \to \beta^*$, then TTTS using β_t as the probability for choosing I_t and J_t is asymptotically optimal in the sense that the posterior probability of incorrect selection converges in the fastest rate. This kind of β_t is generally difficult to compute, but Russo (2020) shows that simply choosing $\beta = 1/2$ would usually lead to good performance and $\Gamma^* \leq 2\Gamma_{1/2}$.

For normal sampling distribution, $\Gamma^* = \max_\psi \min_{i\neq i^*} \frac{(\mu_{i^*}-\mu_i)^2}{2(\sigma_{i^*}/\psi_{i^*}+\sigma_i/\psi_i)}$, which coincides with the optimal decreasing rate of the probability of incorrect selection defined by Glynn and Juneja (2004), i.e., $-\min_\psi \lim_{t\to\infty} \frac{1}{t}\log P\left(\bar{X}_{i^*}^{(t)} < \bar{X}_i^{(t)},\ i \neq i^*|\theta\right)$ with $t_i = t\psi_i,\ i = 1, \cdots, k$. Although asymptotically optimal sampling ratio can be defined for general sampling distribution in Glynn and Juneja (2004), most sampling procedures in R&S are derived for normal sampling distribution. An exception is balancing optimal large deviations rate (BOLD) in Chen and Ryzhov (2019), which is a sequential sampling procedure developed for general sampling distribution. KG, EI, and AOAP are sequential Bayesian sampling procedures for normal sampling distribution, and their sampling policies are deterministic mappings A_t adaptive to $\mathscr{E}_{t-1}$, which have analytical forms. Following AOAP, the sampling ratio $(t_1, \cdots, t_k)/t$ converges to the optimal probability ψ^* that leads to the optimal rate Γ^* of normal sampling distribution as $t \to \infty$ (Peng et al. 2018a), but the sampling ratios of KG and EI are not asymptotically optimal (Peng et al. 2018).

3 NUMERICAL EXPERIMENTS

In this section, we conduct numerical experiments to test the performance of different sampling procedures under normal, Bernoulli and exponential sampling distributions, respectively. In all numerical examples, statistical efficiency of the sampling procedures is measured by the IPCS estimated by 10^5 independent macro runs. The IPCS is presented as a function of simulation budget.

3.1 Normal Sampling Distribution

For normal sampling distribution, we implement ten sampling procedures: equal allocation (EA), which equally allocates simulation budget to estimate the performance of each alternative; "most starving" sequential OCBA in Chen and Lee (2011), denoted as OCBA for simplicity; EI; KG; BOLD; AOAP; UCB, which sequentially allocates each simulation budget to alternative $\arg\max_{i=1,\cdots,k}\left(\bar{X}_i^{(t)} + \sqrt{2\ln t/t_i}\right)$; TTTS; IZ procedure in Rinott (1978), which takes n_0 simulation replications of each alternative and calculates sample

variances $\bar{\sigma}_i^2$ at first stage, and then allocates $\max\left\{\left\lceil h^2\bar{\sigma}_i^2/\delta^2\right\rceil - n_0, 0\right\}$ additional simulation replications to alternative i at second stage, where h and δ are IZ parameters. We implement TTTS with cutoff values 10 and 100, denoted as TTTS 10 and TTTS 100, respectively.

In all numerical examples, the number of initial simulation replications for parameter estimation is chosen to be $n_0 = 10$. For IZ, let $\delta = 0.2$ and constant h is computed with 95% PCS target. We conduct eight numerical experiments under normal sampling distributions. For each macro experiment in examples N.1-N.4, simulation replications are generated independently from normal distribution $N\left(\mu_i, \sigma_i^2\right)$. For each macro experiment in examples N.5-N.8, μ_i, $i = 1, \cdots, k$, are generated from normal conjugate priors $N(\mu_i^{(0)}, (\sigma_i^{(0)})^2)$, and simulation replications are generated independently from normal distribution $N\left(\mu_i, \sigma_i^2\right)$. The parameter settings are summarized in Table 1.

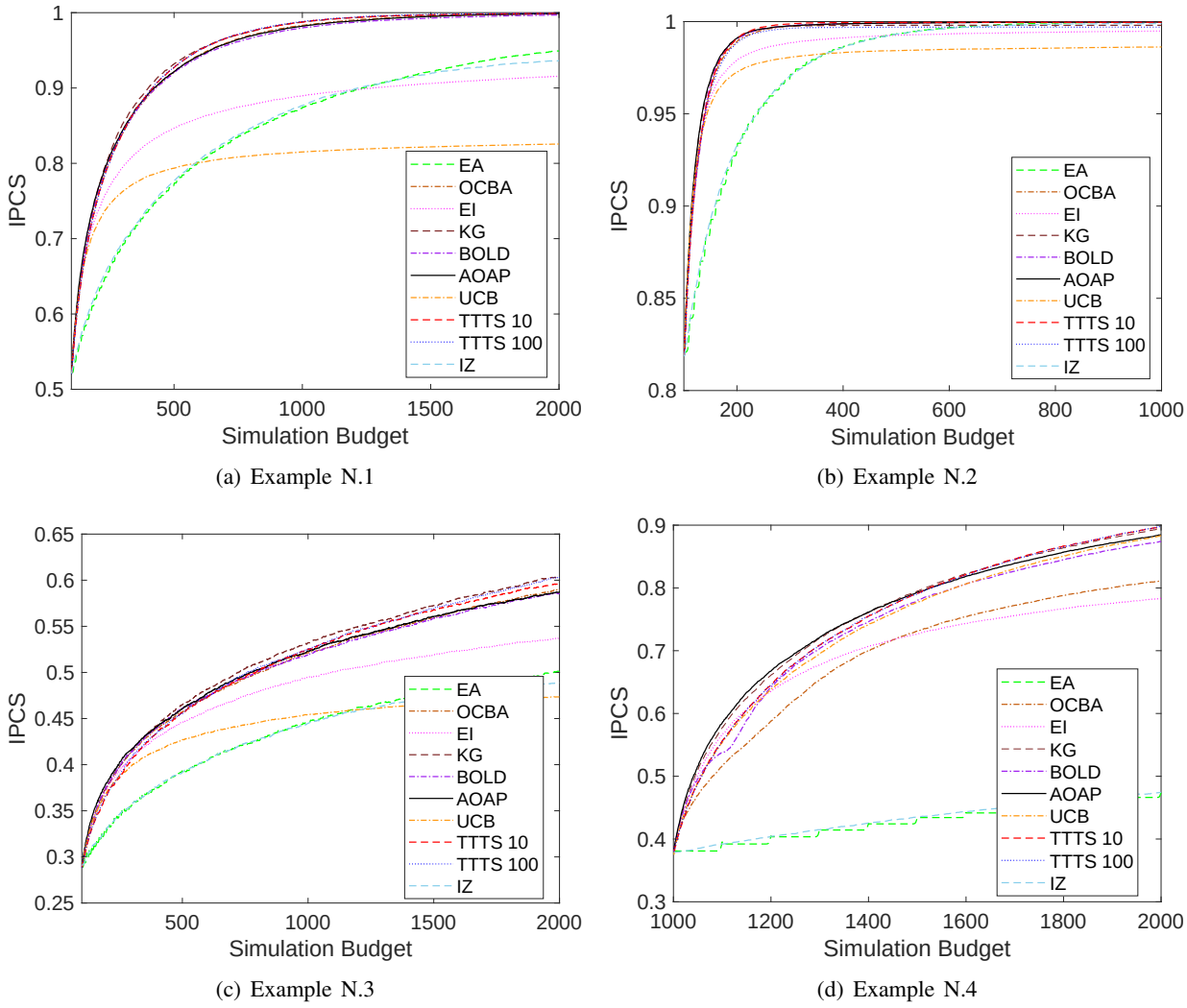
Table 1: Parameter settings of the numerical examples in Section 3.1.

Examples	k	μ_i	σ_i	T	Examples	k	$\mu_i^{(0)}$	$\sigma_i^{(0)}$	σ_i	T
N.1	10	$i-1$	6	2000	N.5	5	0	0.1	$\sqrt{10}$	3000
N.2	10	$9-3\sqrt{10-i}$	6	1000	N.6	20	10	0.1	$\sqrt{10}$	5000
N.3	10	$9-\left(\frac{10-i}{3}\right)^2$	6	2000	N.7	100	10	0.1	10	10000
N.4	100	$(i-1)/10$	1	2000	N.8	10	$i/10$	$\sqrt{i}/10$	$5\sqrt{2i}$	5000

The parameter settings in Examples N.1-N.4 are the same as in numerical experiments in Chen et al. (2000). In Examples N.5-N.8, σ_i^2 is chosen to be large relative to $\sigma_i^{(0)}$, so that the differences in μ_i are relatively small and the true variances of alternatives are large, which is referred to as low-confidence scenarios in Peng et al. (2018). Figures 2 and 3 show the performances of different sampling procedures in the eight numerical examples.

In four experiments shown in Figure 2, the performance of IZ is comparable to that of EA, and TTTS 10 and TTTS 100 have comparable performance. As simulation budget grows large, UCB is surpassed by EA in Figure 2 (a) – (c), and EI is surpassed by EA in Figure 2 (a) – (b). OCBA, KG, BOLD, AOAP, and TTTS are competitive in four experiments. In Figure 2 (a), OCBA and AOAP have indistinguishable performance, and they have a slight edge over BOLD. AOAP performs the best among all sampling procedures at the beginning and is surpassed by KG and TTTS as simulation budget increases. In Figure 2 (b), the performance rank of all sampling procedures is similar to that in Figure 2 (a). In Figure 2 (c), AOAP performs the best among all sampling procedures at the beginning and has a performance comparable to OCBA and BOLD as simulation budget grows; KG has a slight edge over TTTS and becomes the best among all sampling procedures as simulation budget grows large. In Figure 2 (d), EI is surpassed by OCBA as simulation budget grows; UCB surpasses BOLD as simulation budget increases, and it is inferior to AOAP; AOAP performs the best among all sampling procedures at the beginning, but it is surpassed by TTTS and KG as simulation budget increases; TTTS has the best performance as simulation budget grows. Although KG has a better performance in the four experiments N.1-N.4 , it is much more time-consuming than other sampling procedures, since it involves matrix inversion and the matrix size increases.

In Figure 3 (a), KG and AOAP have comparable performance and are superior to the other sampling procedures; OCBA, EI, BOLD, and UCB outperform TTTS 100 and TTTS 10 at the beginning but are surpassed by TTTS 100 and TTTS 10 as simulation budget grows; TTTS 100 is slightly better than TTTS 10; OCBA has a performance comparable to EI, which is better than BOLD and UCB; EA surpasses UCB as simulation budget increases. In Figure 3 (b), TTTS 100 and TTTS 10 have comparable performance significantly better than EA; EI is better than UCB at the beginning but is surpassed by the latter as simulation budget grows; AOAP is slightly better than OCBA, KG, and BOLD, whose performance is indistinguishable. In Figure 3 (c), TTTS 10 and TTTS 100 have indistinguishable performance only slightly better than EA; OCBA has a performance comparable to BOLD, which is superior to EI and UCB; EI is better than UCB at the beginning, but it is surpassed by the latter as simulation budget increases; KG performs worse than EI as simulation budget grows; AOAP performs the best among all sampling procedures. In Figure 3 (d),



(a) Example N.1

(b) Example N.2

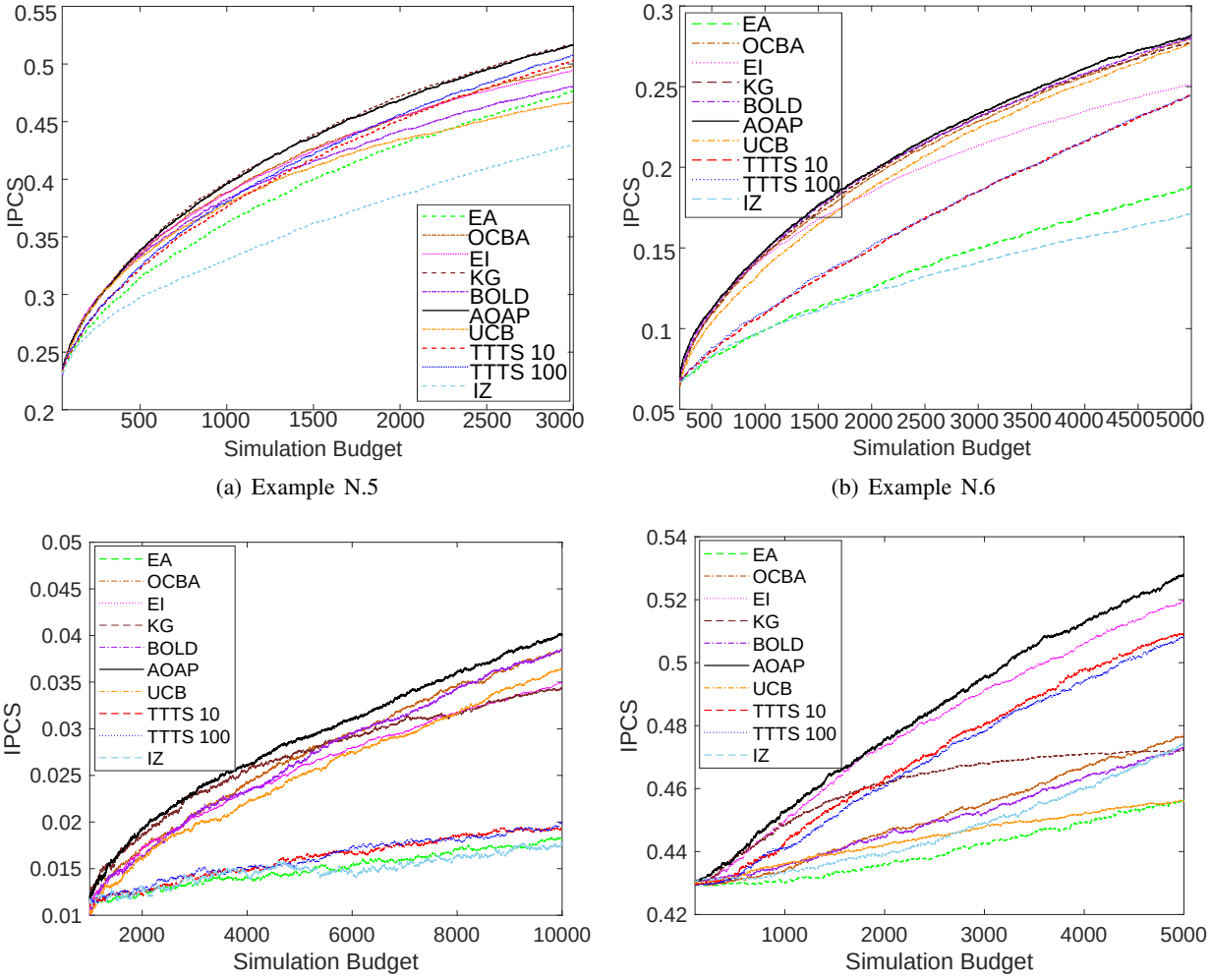
(c) Example N.3

(d) Example N.4

Figure 2: IPCS of the ten sampling procedures with parameter settings (a) Example N.1; (b) Example N.2; (c) Example N.3; (d) Example N.4 on the example of Section 3.1.

EA surpasses UCB in the end; OCBA is better than IZ and has a slight edge over BOLD as simulation budget grows; KG performs better than TTTS 10 and TTTS 100 at the beginning but its IPCS flattens out as simulation budget increases; TTTS 10 has a slight edge over TTTS 100; AOAP and EI perform better than other sampling procedures, and the advantage of AOAP over EI become apparent as simulation budget grows.

From the numerical observations above, we can summarize some rules-of-thumb: AOAP performs the best among all sampling procedures when the number of simulation budget is relatively small, and KG, TTTS 10, and TTTS 100 are recommended when simulation budget is large; AOAP is the best choice in low-confidence scenarios, i.e., differences in the performance of competing alternatives are small, variances are large, the number of alternatives is large, and simulation budget is relatively small. BOLD and TTTS, two other asymptotically optimal sampling procedures besides AOAP, might not perform well in low-confidence scenarios. The choice of cutoff value for TTTS may have an influence on the performance.



(a) Example N.5 (b) Example N.6

(c) Example N.7 (d) Example N.8

Figure 3: IPCS of the ten sampling procedures with parameter settings (a) Example N.5; (b) Example N.6; (c) Example N.7; (d) Example N.8 on the example of Section 3.1.

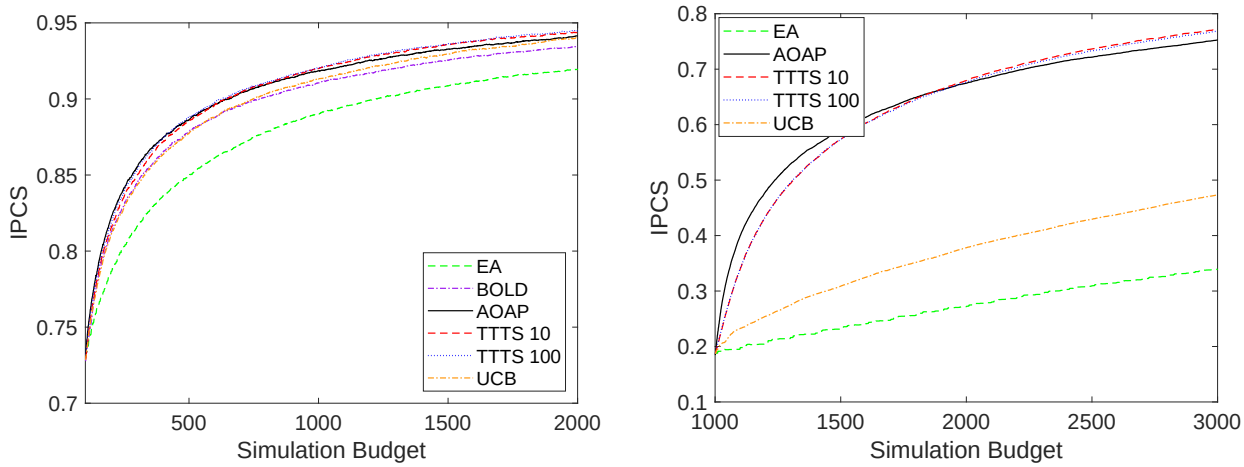
3.2 Bernoulli Sampling Distribution

For Bernoulli sampling distribution, we implement six sampling procedures: EA; BOLD; AOAP in Li et al. (2020), which uses normal distribution to approximate the posterior Beta distribution; TTTS 10 and TTTS 100; UCB. For AOAP, a distance parameter is chosen to be $\varepsilon = 10^{-5}$. We conduct two numerical experiments with Bernoulli sampling distributions. In example B.1, the number of initial simulation replications is chosen to be $n_0 = 20$, whereas in example B.2, the number of initial simulation replications is chosen to be $n_0 = 10$. We do not implement BOLD for comparison in example B.2 because inaccurate initial parameter estimation hinder its implementation. In each macro experiment, μ_i, $i = 1, \cdots, k$ are generated from Beta conjugate priors $Beta(\alpha_i^{(0)}, \beta_i^{(0)})$, where $\alpha_i^{(0)}$ and $\beta_i^{(0)}$ are drawn independently from uniform distribution, and simulation replications are generated independently from Bernoulli distribution $Ber(\mu_i)$. The parameter settings are summarized in Table 2.

Example B.1 is designed to show performances of different sampling procedures when the number of alternatives is small, whereas Example B.2 has a much larger number of alternatives. Figure 4 shows the performances of different sampling procedures in the two numerical examples.

Table 2: Parameter settings of the numerical examples of Section 3.2.

Examples	k	$\alpha_i^{(0)}$	$\beta_i^{(0)}$	T
B.1	5	$U(1,20)$	$U(1,20)$	2000
B.2	100	$U(5,10)$	$U(1,5)$	3000



(a) Example B.1 (b) Example B.2

Figure 4: IPCS of the ten sampling procedures with parameter settings (a) Example B.1; (b) Example B.2 on the example of Section 3.2.

In Figure 4 (a), EA has the worst performance among all sampling procedures; BOLD and UCB have a comparable performance at the beginning, and UCB surpasses the BOLD when simulation budget reaches around 560; UCB catches up with AOAP as simulation budget increases; AOAP has a slight edge over TTTS 10 and TTTS 100 at the beginning, but it is surpassed by TTTS 10 and TTTS 100 when simulation budget reaches around 575 and 420, respectively. In Figure 4 (b), UCB has an edge over EA which performs worst among all sampling procedures. At the beginning, AOAP has an edge over TTTS 10 and TTTS 100, but it is surpassed by both TTTS 10 and TTTS 100 when simulation budget reaches around 1930. Compared with Example B.1, the advantage of AOAP appears more apparent when the number of competing alternatives is large and simulation budget is small.

3.3 Exponential Sampling Distribution

For exponential sampling distribution, we implement seven sampling procedures: EA; OCBA for exponential sampling distribution (OCBA-exp) in Gao and Gao (2016); BOLD; TTTS 10 and TTTS 100; AOAP in Zhang et al. (2020), which uses normal distribution to approximate the posterior gamma distribution; UCB.

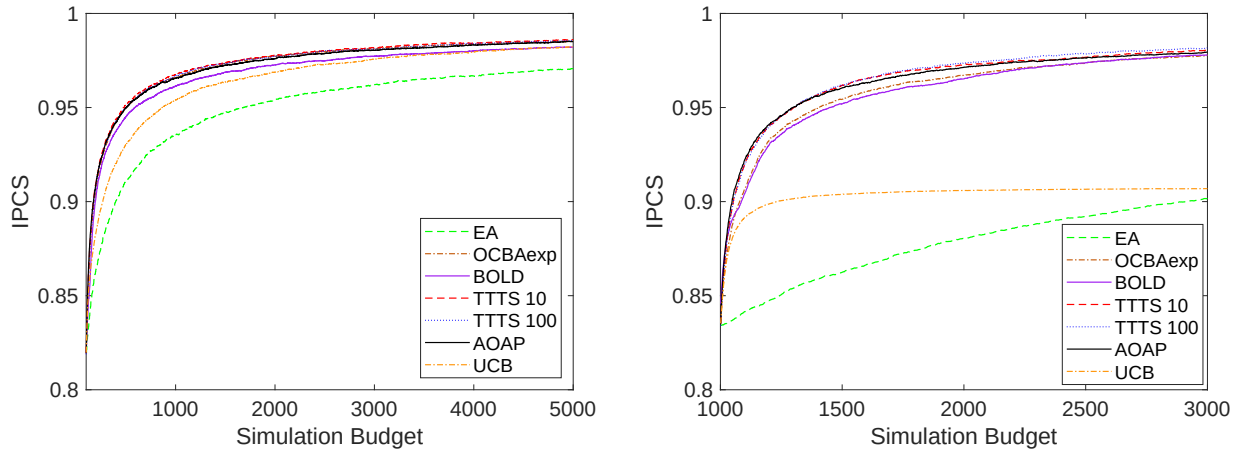
We conduct two numerical experiments with exponential sampling distribution. In both experiments, the number of initial simulation replications for parameter estimation is chosen to be $n_0 = 10$. In each macro experiment, $\lambda_i = 1/\mu_i$, $i = 1, \cdots, k$, are generated from Gamma conjugate priors $Gamma(k_i^{(0)}, \theta_i^{(0)})$, and simulation replications are generated independently from exponential distribution $Exp(\lambda_i)$. The parameter settings are summarized in Table 3.

The number of alternatives in Example E.1 is small, and the number of alternatives in Example E.2 is large. Figure 5 shows the performances of different sampling procedures in the two numerical examples.

In Figure 5 (a), UCB has a significant edge over EA, but it lags behind BOLD at the beginning and catches up as simulation budget grows; AOAP, TTTS 10, TTTS 100 have comparable performance and are

Table 3: Parameter setting of the numerical examples of Section 3.2.

Examples	k	$k_i^{(0)}$	$\theta_i^{(0)}$	T
E.1	10	$U(2,10)$	$U(1,2)$	5000
E.2	100	1	2	3000



(a) Example E.1 (b) Example E.2

Figure 5: IPCS of the ten sampling procedures with parameters settings (a) Example B.1; (b) Example B.2 on the example of Section 3.3.

better than other sampling procedures. In Figure 5 (b), EA is inferior to other sampling procedures at the beginning, but it catches up with UCB whose IPCS flattens out as simulation budget grows; OCBA-exp has a slight edge over BOLD at the beginning, and BOLD catches up with OCBA-exp at the end; AOAP, TTTS 10, TTTS 100 have comparable performance and are better than other sampling procedures.

4 CONCLUSIONS

TS is originated from MAB, and its variant TTTS appears to be an appealing sampling procedure for R&S based on comprehensive comparisons between TTTS and some popular methods in R&S from both theoretical and numerical perspectives in this work. Together with algorithms that can generate samples from a posterior distribution, TTTS applies to general sampling distribution and performs reasonably well. In R&S, the developments of sampling procedures are predominately based on normal sampling distribution, and it usually requires extra work to adapt the sampling procedures to a non-normal sampling distribution. For normal sampling distribution, some existing sampling procedures in R&S are still quite competitive relative to TTTS, and particularly in low-confidence scenarios, AOAP seems to have an advantage over TTTS and others. In the literature of R&S and MAB, there are works on selecting top-m alternatives and context-dependent selection of the best (Chen et al. 2008; Shen et al. 2021; Han et al. 2020), whereas TS-like schemes have not been developed for these extensions. In implementation, the need to choose an appropriate cutoff value that may influence the performance for TTTS is troublesome.

The recent progresses made by different communities such as simulation and machine learning coming together to address same problems like R&S might inspire researchers to borrow ideas and tools from both sides to come up with new or better solutions for big problems. AlphaGo is designed with a MCTS backbone using a UCB-like node selection policy. Recently, Li et al. (2021) and Zhang et al. (2022) demonstrate that OCBA and AOAP could lead to better performance in MCTS. Since the dynamic sampling

decisions can be formulated as MDP, leveraging the state-of-art techniques in reinforcement learning would be a promising future direction to elevate the standard of the research in R&S.

REFERENCES

Auer, P., N. Cesa-Bianchi, and P. Fischer. 2002. "Finite-time analysis of the multiarmed bandit problem". *Machine learning* 47(2):235–256.

Bechhofer, R. E. 1954. "A single-sample multiple decision procedure for ranking means of normal populations with known variances". *The Annals of Mathematical Statistics* 25(1):16–39.

Bechhofer, R. E., T. J. Santner, and D. M. Goldsman. 1995. *Design and Analysis for Statistical Selection, Screening, and Multiple Comparisons*. New York: John Wiley and Sons.

Bertsekas, D. P. 1995. *Dynamic Programming and Optimal Control*, Volume 1. Belmont, MA: Athena Scientific.

Branke, J., S. E. Chick, and C. Schmidt. 2007. "Selecting a selection procedure". *Management Science* 53(12):1916–1932.

Chen, C.-H., D. He, M. Fu, and L. H. Lee. 2008. "Efficient Simulation Budget Allocation for Selecting an Optimal Subset". *INFORMS Journal on Computing* 20(4):579–595.

Chen, C.-H., and L. H. Lee. 2011. *Stochastic Simulation Optimization: An Optimal Computing Budget Allocation*, Volume 1. Toh Tuck Link, Singapore: World Scientific Publishing Company.

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(3):251–270.

Chen, Y., and I. O. Ryzhov. 2019. "Balancing optimal large deviations in ranking and selection". In *2019 Winter Simulation Conference*, edited by N. Mustafee, K.-H. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, 3368–3379. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Chick, S. E., and K. Inoue. 2001a. "New procedures to select the best simulated system using common random numbers". *Managemetn Science* 47(8):1133–1149.

Chick, S. E., and K. Inoue. 2001b. "New two-stage and sequential procedures for selecting the best simulated system". *Operations Research* 49(5):732–743.

Eckman, D. J., and S. G. Henderson. 2022. "Posterior-Based Stopping Rules for Bayesian Ranking-and-Selection Procedures". *INFORMS Journal on Computing* 34(3):1711–1728.

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. 2008. "A knowledge-gradient policy for sequential information collection". *SIAM Journal on Control and Optimization* 47(5):2410–2439.

Gao, F., and S. Gao. 2016. "Optimal computing budget allocation with exponential underlying distribution". In *2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 682–689. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Glynn, P. W., and S. Juneja. 2004. "A Large Deviations Perspective on Ordinal Optimization". In *Proceedings of the 2004 Winter Simulation Conference*, edited by R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 577–585. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Goldsman, D., and B. L. Nelson. 1998. "Comparing systems via simulation". In *Chapter 8 in Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice, ed.* J. Banks. New York: John Wiley and Sons.

Han, Y., Z. Zhou, Z. Zhou, J. Blanchet, P. W. Glynn, and Y. Ye. 2020. "Sequential batch learning in finite-action linear contextual bandits". *arXiv preprint arXiv:2004.06321*.

Hong, L. J., W. Fan, and J. Luo. 2021. "Review on ranking and selection: A new perspective". *Frontiers of Engineering Management* 8(3):321–343.

Kim, S.-H. 2013. "Statistical ranking and selection". In *In Encyclopedia of Operations Research and Management Science, 3rd edition, S.I. Gass and M.C. Fu (ed.),*, Volume 2, 1459–1469. Springer.

Kim, S.-H., and B. L. Nelson. 2006. "Selecting the best system". In *Chapter 17 in Handbooks in Operations Research and Management Science: Simulation*, Volume 13, 501–534. Elsevier.

Lai, T. L. 1987. "Adaptive treatment allocation and the multi-armed bandit problem". *The Annals of Statistics*:1091–1114.

Li, H., X. Xu, Y. Peng, and C.-H. Chen. 2020. "Efficient Learning for Selecting Important Nodes in Random Network". *IEEE Transactions on Automatic Control* 66(3):1321–1328.

Li, Y., M. C. Fu, and J. Xu. 2021. "An optimal computing budget allocation tree policy for Monte Carlo tree search". *IEEE Transactions on Automatic Control* 67(6):2685–2699.

Peng, Y., C.-H. Chen, E. K. 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*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 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., C.-H. Chen, M. C. Fu, and J.-Q. Hu. 2018. "Gradient-Based Myopic Allocation Policy: An Efficient Sampling Procedure in a Low-Confidence Scenario". *IEEE Transaction Automatic Control* 63(9):3091–3097.

Peng, Y., E. K. Chong, C.-H. Chen, and M. C. Fu. 2018a. "Ranking and selection as stochastic control". *IEEE Transactions on Automatic Control* 63(8):2359–2373.

Powell, W. B., and I. O. Ryzhov. 2012. "Ranking and selection". In *Chapter 4 in Optimal Learning*, 71–88: New York: John Wiley and Sons.

Rinott, Y. 1978. "On two-stage selection procedures and related probability-inequalities". *Communications in Statistics-Theory and methods* 7(8):799–811.

Robbins, H. 1952. "Some aspects of the sequential design of experiments". *Bulletin of the American Mathematical Society* 58(5):527–535.

Russo, D. 2020. "Simple Bayesian Algorithms for Best-Arm Identification". *Operations Research* 68(6):1625–1647.

Russo, D. J., B. Van Roy, A. Kazerouni, I. Osband, Z. Wen et al. 2018. "A tutorial on thompson sampling". *Foundations and Trends® in Machine Learning* 11(1):1–96.

Ryzhov, I. O. 2016. "On the convergence rates of expected improvement methods". *Operations Research* 64(6):1515–1528.

Shen, H., L. J. Hong, and X. Zhang. 2021. "Ranking and selection with covariates for personalized decision making". *INFORMS Journal on Computing* 33(4):1500–1519.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. V. D. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, and M. Lanctot. 2016. "Mastering the game of Go with deep neural networks and tree search". *Nature* 529(7587):484–489.

Thompson, W. R. 1933. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". *Biometrika* 25(3-4):285–294.

Zhang, G., H. Li, and Y. Peng. 2020. "Sequential sampling for a ranking and selection problem with exponential sampling distributions". In *2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 2984–2995. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zhang, G., Y. Peng, and Y. Xu. 2022. "An efficient dynamic sampling policy for Monte Carlo tree search". *arXiv preprint arXiv: 2204.12043*.

AUTHOR BIOGRAPHIES

YIJIE PENG is an Associate Professor in Guanghua School of Management at Peking University. His research interests include simulation and AI. 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.

GONGBO ZHANG is a Ph.D. candidate in the Department of Management Science and Information Systems in Guanghua School of Management at Peking University, Beijing, China. He received the B.S. degree in mathematics and applied mathematics from College of Sciences, Northeastern University, Shenyang, China, in 2018. His research interests include simulation and AI. His email address is gongbozhang@pku.edu.cn.