Information Theory for Ranking and Selection

Saeid Delshad[a] and Amin Khademi[a]

[a]Department of Industrial Engineering, Clemson University, Clemson, SC, USA.

ARTICLE HISTORY

Compiled March 29, 2020

ABSTRACT

We study the classical ranking and selection problem, where the ultimate goal is to find the unknown best alternative in terms of the probability of correct selection or expected opportunity cost. However, this paper adopts an alternative sampling approach to achieve this goal, where sampling decisions are made with the objective of maximizing information about the unknown best alternative, or equivalently, minimizing its Shannon entropy. This adaptive learning is formulated via a Bayesian stochastic dynamic programming problem, by which several properties of the learning problem are presented, including the monotonicity of the optimal value function in an information-seeking setting. Since the state space of the stochastic dynamic program is unbounded in the Gaussian setting, a one-step look-ahead approach is used to develop a policy. The proposed policy seeks to maximize the one-step information gain about the unknown best alternative, and therefore, it is called Information Gradient (IG). It is also proved that the IG policy is consistent, i.e., as the sampling budget grows to infinity, the IG policy finds the true best alternative almost surely. Later, a computationally efficient estimate of the proposed policy, called Approximated Information Gradient (AIG), is introduced and in the numerical experiments its performance is tested against recent benchmarks alongside several sensitivity analyses. Results show that AIG performs competitively against other algorithms from the literature.

KEYWORDS

Ranking and Selection; Information Gradient; Entropy; Consistency.

Saeid Delshad. Email: sdelsha@clemson.edu
Amin Khademi. Email: khademi@clemson.edu

1. Introduction

1.1. Ranking and Selection

Ranking and selection (R&S) is a class of optimal learning problems in which a decision maker (DM) seeks to efficiently find the unknown best alternative among a finite set of alternatives subject to budgetary constraints. This class of learning problems received significant attention because of its application in many contexts, e.g., optimization via simulation (Chen et al. 2000, Chick et al. 2010, Xu et al. 2013), portfolio selection (Mehrez and Sethi 1989), clinical trials (Berry and Pearson 1985), etc. See Kim and Nelson (2006) for a more comprehensive review.

Throughout the years, different classes of problems have been studied in R&S, each considering a different setting in terms of the objective function, time window, type of decision, frequentist vs. Bayesian view, and the distributions assumed for the prior and measurement. Interested readers are referred to Powell and Ryzhov (2012) for a comprehensive review into different classes of R&S.

R&S is also studied in the multi-armed bandit literature as the best arm identification problem (Bubeck et al. 2011b), as well as sequential adaptive hypothesis testing (Chernoff 1959). In R&S, unlike the standard bandit problem, the DM only seeks to find the best alternative and cumulative rewards are irrelevant. Therefore, R&S is basically different from the standard multi-armed bandit problem. In fact, Bubeck et al. (2011a) showed that policies derived for cumulative regret-minimizing bandits may perform poorly for best arm identification ones.

Many algorithms have been developed for R&S, including procedural approaches (Rinott 1978), algorithms meant for solving large-scale R&S problems using fully sequential procedures (Luo et al. 2015), optimal sampling algorithm (Hunter and Pasupathy 2013), optimal budget allocation methods (Chen et al. 2003), etc. In this study, we use six different algorithms as benchmarks to compare against our proposed policy; namely, knowledge gradient (KG) by Frazier et al. (2008), top-two Thompson sampling (TTTS) by Russo (2016) which promotes more exploration compared to Thompson (posterior) sampling (TS), the most starving sequential Optimal Computing Budget Allocation (OCBA) by Chen and Lee (2011) which is modified for the Bayesian set-

ting, General Bayesian Budget Allocation (GBBA) by Peng et al. (2016) which uses a randomized allocation function for OCBA, and Approximately Optimal Allocation Policy (AOAP) by Peng et al. (2018b) which allocates samples in a myopic sense with the goal of minimizing the largest coefficient of variation of the difference between the best alternative (with the largest posterior mean) and the rest.

1.2. Motivation behind Information Maximization for R&S

The ultimate goal in a R&S problem is to identify the unknown best alternative (with the highest mean) among a given number of alternatives. The DM has a limited budget to explore various alternatives and after exploration the DM recommends a candidate for the best alternative. Traditionally, the DM is evaluated based on the quality of the recommendation in terms of probability of correct selection (PCS) or expected opportunity cost (EOC). Therefore, the DM designs sampling policies that achieve optimum PCS or EOC. Finding optimal sampling procedures, however, is difficult and several approximations are developed.

We propose an alternative sampling approach to minimize the uncertainty regarding the unknown best alternative, measured by the Shannon entropy. The essence of this approach is based on a premise that reducing uncertainty about the unknown best alternative will result in high quality recommendation. In particular, if the uncertainty about the unknown best alternative diminishes to zero, the DM will have identified the best one and the recommendation will be perfect.

Because in our setting the unknown best alternative is a random variable, a measure of uncertainty can be its variance. Therefore, the DM may design a sampling framework to minimize the variance of the unknown best alternative at the end of the sampling period. Such an approach is popular in the literature of experimental design, called D-optimal design, where a statistician seeks to allocate experimental efforts to minimize the variance of a quantity of interest such as an estimator at the end of the experiment (Bhat et al. 2019). However, measuring the uncertainty of a random variable via variance has its own issues because it depends on the parameterization of the parameter space on which the random variable is defined (Kontsevich and Tyler 1999).

Therefore, we use the Shannon entropy to measure uncertainty and the motivation

is that it is more general in theory, especially it is invariant to *all* reparameterization of the parameter space. In fact, this is a remarkable property of entropy that it is parameterization invariant. Although we use entropy minimization in the sampling phase, because the entropy is a number defined over the beliefs about the unknown mean of alternatives, it cannot be used for recommendation purposes. Therefore, we use this approach for sampling purposes and recommend the alternative with the highest posterior mean in the recommendation phase.

In addition, considering an adaptive Bayesian approach with the entropy objective leads to optimization of fundamental concepts in probability and statistics. In particular, our proposed algorithm maximizes the expected information improvement about the best alternative where information is defined as the negative of entropy, which is equivalent to the information-theoretic mutual information given the history up to the current decision epoch. In fact, Kolmogorov (1956) wrote: "I insist that the fundamental concept [is] ... the concept of mutual information." To that end, our proposed policy seeks to *maximize the mutual information between the best alternative and the next sample*, which is fundamentally different from PCS or EOC.

We settle on Bayesian framework where we use the Bayes' theorem to update the probability for a hypothesis as more samples or information is obtained. We formulate the problem of optimal adaptive sampling of alternatives to maximize information (or minimize entropy) about the unknown best alternative at the end of the experiment as a stochastic dynamic program (SDP). Using this formulation, we show several structural properties of the learning problem, e.g., monotonicity of the optimal value function and the value of extra measurement. However, due to the curse of dimensionality, standard techniques fail to produce optimal solutions; therefore, we employ a one-step look-ahead approach in which the DM chooses to observe the alternative whose sampling reduces the expected entropy of the best alternative the most for the next decision epoch. That is, we seek to maximize the one-step mutual information between the best alternative and the next sample. Because this policy maximizes the information about the best alternative in a one-step time frame, we call it information gradient (IG).

We propose approximate information gradient (AIG) as a surrogate for IG because

IG may be computationally demanding. The approximation replaces calculation (by Monte Carlo) of the probability that each alternative is the best at the next period with averaging lower and upper bounds of that probability at the next period. We compare our suggested AIG in terms of PCS and EOC to other policies and observe that AIG is computationally affordable and its performance is competitive compared to other algorithms. In particular, both AIG and IG perform competitively without any initial replications or parameter tuning for a variety of problem settings. Also, the fact that AIG, without any initialization, looks promising at the beginning of the experimentation compared to many other algorithms, makes it a good candidate for low-budget R&S problems. This is significant, especially in early stage clinical trials where the potential number of patients is limited and uncertainty is high. The proposed AIG policy, given the problem, has its own advantages while for small-scale problems is computationally efficient. These advantages include robustness with regards to unknown sampling variances (elaborated in Section 4.3), and also, the prior belief about the alternatives (elaborated in Section 4.4).

In summary, this paper makes the following contributions: (1) We formulate a SDP formulation for sampling purposes in R&S with the objective of maximizing information about the unknown best alternative at the end of the experiment. (2) We show some structural properties about the value function in this class of learning problems. (3) We employ a one-step look-ahead framework to the SDP formulation to design an approximation solution, which results in the IG policy. We show some properties for the IG including its consistency. (4) We propose AIG as an approximation to IG in Gaussian settings and show its competitiveness in an extensive numerical study.

1.3. Background on Entropy-based Objective Functions

Entropy-based objective functions have been used in a variety of problems and for different purposes. Lindley (1956) first used information-theoretic entropy to design and compare statistical experiments. The concept of information gradient has been used in decision making. For instance, Machens (2002) applied an information-maximizing framework in special neurophysiological experiments to identify the best input for the next experiment to achieve a favorable outcome. Pallone et al. (2017) used an

information-maximizing framework for adaptive learning of user preference and with the help of queries, showed several properties of this policy and compared it to KG. Our paper though addresses the properties of the IG policy in the Gaussian R&S.

Maximizing the information gain in a greedy fashion is an active topic in Bayesian optimization. For example, Villemonteix et al. (2009) considered optimization of an unknown function with a limited number of function evaluations and used entropy reduction search. In particular, their setup assumed a Gaussian process (GP) for the function. The approach is to minimize the entropy of the optimal solution for the next function evaluation. Then, the setup is extended to consider noise in sampling. However, computing the conditional entropy is extremely time-consuming. Hennig and Schuler (2012) considered a similar setup to Villemonteix et al. (2009) in which they considered an unknown function and minimized a loss function. Similar to Villemonteix et al. (2009), they assumed a GP prior for the function and considered a relative entropy from a uniform distribution as the loss function. They used a first order approximation and used a greedy approach for sampling. The problem setup in Hernández-Lobato et al. (2014) is similar to Hennig and Schuler (2012) but the setup is completely Bayesian. The setup we study in this work is R&S where the domain of alternatives is discrete and the techniques used in the continuous domain do not apply. In particular, when the domain is continuous, one can use Taylor's expansion and then apply first and second order optimality conditions to find approximate optimal solutions for entropy search. However, in a discrete domain case, this approach is not possible. We use AIG which is based on approximating the probability that an alternative is the best condition on availability of one more sample, which is fundamentally different from above-mentioned entropy search methods.

Russo and Van Roy (2016) introduced a novel framework to learning problems for online learning problems with the objective of minimizing cumulative or simple regret. They proposed a greedy policy that minimizes the ratio of square of expected immediate regret per bit of information gain, and provided approximation methods for calculating this ratio. In particular, Russo and Van Roy (2016, Proposition 9) showed that an upper bound on expected simple regret of information-directed sampling goes to zero, thus the consistency. In our setup, the DM seeks to minimize

the entropy of the unknown best arm as an objective via a stochastic dynamic formulation, which is different from regret objective. In fact, our IG policy is an application of one-step look-ahead framework to our SDP formulation while the information-directed sampling is introduced based on the concept of information ratio. We also provide an approximate version for IG which is computationally more efficient and is competitive in a variety of settings. Finally, the setup in Russo and Van Roy (2016) assumed a known sampling variance but we relax this assumption and show the consistency of IG in a setup where the sampling variance is unknown.

Paper organization. In Section 2, we formulate the decision making problem as a stochastic dynamic program. Section 3 analyzes the formulation and provides several structural properties of the learning problem, including consistency of the IG policy in an almost sure sense. Section 4 provides numerical results and sensitivity analyses for the proposed policy. Section 5 concludes the paper. Also, the proofs of all propositions, corollaries, and theorems are available in the Supplementary Material.

2. Problem Formulation

Suppose we have a finite set of alternatives $\mathcal{A} = \{1, 2, \ldots, J\}$ where the true reward distribution of alternative j is normally distributed with unknown mean θ_j and known variance σ_j^2 for each $j \in \mathcal{A}$. Normality assumption for rewards is usually justified by the central limit theorem and batch sampling. Defining y_j to be the reward value of sampling alternative j, we have $y_j|\theta_j \sim \mathcal{N}(\theta_j, \sigma_j^2)$. Also, y_js are assumed to be independent across the alternatives, i.e., the sample reward of allocating to alternative j is independent of our belief about other alternatives. In problem formulation, we assume that σ_j^2s are known in order to derive structural properties for the learning problem. However, we relax this assumption for showing the consistency of our proposed policy by considering a normal-gamma prior for the mean and variance pair. In Section 4.3, we also numerically analyze the performance of the proposed policy in this more general setting. We define $J^* := \arg\max_{j \in \mathcal{A}} \theta_j$, i.e., the alternatives with the largest mean reward. Consider a DM who has a finite budget of N samples to identify the best

alternative, i.e., at the end of the trial, the DM recommends the alternative that has the largest posterior mean. While this is the most common selection policy, it is not the only selection strategy considered in the literature. See Peng et al. (2016) for more details regarding other selection policies.

Let $n \in \{0, 1, \ldots, N-1\}$ denote decision epochs when the DM decides to sample from an alternative. Assume a Bayesian sequential decision making process in which the DM starts with a joint prior distribution over the means of all alternatives, and at each decision epoch, decides which alternative to sample, then observes the outcome of sampling decision and updates her belief about θ before the next decision epoch. In particular, we assume that the initial prior belief about $\theta = (\theta_1, \theta_2, \ldots, \theta_J)$ is a multivariate normal prior distribution with mean μ^0 and covariance matrix Σ^0. Let a^n be the index of the alternative that will be sampled at time n and y^{n+1} be the sample reward achieved by sampling a^n, which is obtained independently across alternatives. Define filtration $\mathcal{F}^n$ as the sigma algebra generated by $\{\mu^0, \Sigma^0, a^0, y^1, a^1, y^2, \ldots, a^{n-1}, y^n\}$. The DM has a limited budget N to sample alternatives and seeks to gain maximum information about the unknown best alternative.

Let $\mathbb{P}^n(\cdot) = \mathbb{P}(\cdot|\mathcal{F}^n)$ and $\mathbb{E}^n(\cdot) = \mathbb{E}(\cdot|\mathcal{F}^n)$ respectively be the posterior probability mass function and expectation with respect to filtration $\mathcal{F}^n$, i.e., all data available at decision epoch n. Denote $p_j^n = \mathbb{P}^n(J^* = j)$ as the probability of alternative j being the best at decision epoch n. Note that the Shannon entropy of the unknown best alternative is defined as $H_S = -\sum_j p_j^n \log p_j^n = -\mathbb{E}^n\{\log p^n(J^*)\}$ which has a negative sign in front unlike information. That is, the amount of information about J^* is quantified by the negative of the Shannon entropy. The reason is that from a statistical viewpoint, information is maximized when a distribution concentrates on a single value and is minimized when each alternative has an equal probability of being the best, which is in contrast with the situation faced by a communication engineer, where concentration on a single value leaves no choice in the message (Lindley 1956).

The question is what should be the sampling strategy to achieve this goal. Because the decision process is sequential and involves randomness, we formulate it using stochastic dynamic programming. Note that no allocation decision is made at time N when the DM identifies the best alternative. Our selection policy at this time is to

recommend the alternative with the largest posterior mean as the best.

Recall that the DM has a multivariate normal distribution about the true mean of each alternative, θ, and the sample rewards from each alternative are also normally distributed. Therefore, the state of the system at decision epoch n is completely characterized by $s^n = (\mu^n, \Sigma^n)$, where $\mu^n = \mathbb{E}^n(\theta)$ and $\Sigma^n = [\Sigma^n_{ij}]$ with $\Sigma^n_{ij} = \mathbb{E}^n(\theta_i\theta_j) - \mathbb{E}^n(\theta_i)\mathbb{E}^n(\theta_j)$. The action space at each decision epoch is assumed to be $\mathcal{A}$, i.e., $a^n \in \mathcal{A}$ is the action taken for $n = 0, \ldots, N-1$, which refers to the index of the alternative that is sampled. Since at time n we have $\theta \sim \mathcal{N}(\mu^n, \Sigma^n)$ and the sample reward $y^{n+1}\big|\{\theta, a^n = j\} \sim \mathcal{N}(\theta_j, \sigma_j^2)$ the posterior distribution for the true mean vector is given by $\theta \sim \mathcal{N}(\mu^{n+1}, \Sigma^{n+1})$, with

$$\begin{aligned} \mu^{n+1} &= \Sigma^{n+1}\big((\Sigma^n)^{-1}\mu^n + (\sigma_j^2)^{-1}y^{n+1}e_j\big), \\ \Sigma^{n+1} &= \big((\Sigma^n)^{-1} + (\sigma_j^2)^{-1}e_je_j^T\big)^{-1}, \end{aligned} \tag{1}$$

where $(\cdot)^T$ denotes matrix transposition, and e_j is a J-vector of zeros and a single 1 at j^{th} index assuming that Σ^n is invertible. Alternatively, the following updating equations can be used which do not require invertibility of Σ^n

$$\begin{aligned} \mu^{n+1} &= \mu^n + \frac{y^{n+1} - \mu_j^n}{\sigma_j^2 + \Sigma^n_{jj}}\Sigma^n e_j^T, \\ \Sigma^{n+1} &= \Sigma^n - \frac{\Sigma^n e_j e_j^T \Sigma^n}{\sigma_j^2 + \Sigma^n_{jj}}. \end{aligned} \tag{2}$$

In order to ease notation for the dynamic program, define $\tilde{\sigma}$ as a vector-valued function $\tilde{\sigma}\big(\Sigma, j\big) := \dfrac{\Sigma e_j}{\sqrt{\sigma_j^2 + \Sigma_{jj}}}$, and note that $\mathrm{Var}[y^{n+1} - \mu_j^n|\mathcal{F}^n] = \mathrm{Var}[\theta_j + \epsilon_j|\mathcal{F}^n] = \sigma_j^2 + \Sigma^n_{jj}$, where y^{n+1} is the sample reward achieved by sampling a^n and $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ is the measurement error as in $y = \theta + \epsilon$. Define the random variable $X^{n+1} := \dfrac{(y^{n+1} - \mu_j^n)}{\sqrt{\mathrm{Var}[y^{n+1} - \mu_j^n|\mathcal{F}^n]}}$ by which formulation (2) is equivalent to

$$\begin{aligned} \mu^{n+1} &= \mu^n + \tilde{\sigma}(\Sigma^n, j)X^{n+1}, \\ \Sigma^{n+1} &= \Sigma^n - \tilde{\sigma}(\Sigma^n, j)\big(\tilde{\sigma}(\Sigma^n, j)\big)^T, \end{aligned} \tag{3}$$

where random variable X^{n+1} is standard normal when conditioned on $\mathcal{F}^n$.

We consider a DM who seeks to gain maximum information about the best alternative at the end of the decision process. Therefore, the reward at the end of the process is $\mathcal{I}_{J^*}(s^N) = \mathbb{E}^N\{\log p^N(J^*)\}$, where $s^N = (\mu^N, \Sigma^N)$ is the final state of the system at the end of the sampling process. Define $\Pi := \{(a^0, \ldots, a^{N-1}); a^n \in \mathcal{A}, \forall n\}$ to be the set of measurable policies where a^n is $\mathcal{F}^n$-measurable for $n = 0, \ldots, N-1$ and $\pi = (a^0, \ldots, a^{N-1})$ is an element of Π. Let $s^0 = (\mu^0, \Sigma^0)$ be the state of the system before the sampling process begins and $\mathbb{E}^\pi\{\cdot\}$ be the expectation taken with respect to a fixed measurement policy π. The DM solves for

$$V^0(s^0) = \sup_{\pi \in \Pi} \mathbb{E}^\pi \big\{ \mathcal{I}_{J^*}(s^N) \big| s^0 = (\mu^0, \Sigma^0) \big\}, \tag{4}$$

where $V^n(s^n)$ denotes the value function at time n, which is a unique solution to the following Bellman equations

$$\begin{aligned} &V^n(s^n) = \max_{a^n \in \mathcal{A}} \Big\{ \mathbb{E}\big\{ V^{n+1}(s^{n+1}) \big| s^n, a^n \big\} \Big\}, \quad n = 0, 1, \ldots, N-1, \\ &V^N(s^N) = \mathcal{I}_{J^*}(s^N), \end{aligned} \tag{5}$$

and the optimal action is myopic with respect to the optimal value function.

Next, we show some basic properties of the learning process. Let $\eta : (\mathcal{S}, \mathcal{A}, \mathbb{R}) \mapsto \mathcal{S}$ be the transition function determining the next state via equation (3), i.e., $s^{n+1} := \eta(s^n, a^n, X^{n+1})$. Define the Q-factor as

$$Q^n(s, j) := \mathbb{E}\Big\{ V^{n+1}\Big(\eta(s^n, a^n, X^{n+1}) \Big) \Big| s^n = s, a^n = j \Big\},$$

and denote j_*^n as the optimal action at time n, i.e., $j_*^n \in \arg\max_{j \in \mathcal{A}} Q^n(s, j)$, $\forall s \in S$. The following proposition shows that the optimal policy prefers to measure an alternative at each decision epoch rather than not measuring at all. $V^{n+1}(s^n)$ can be interpreted as the value of no measurement while in state s^n.

Proposition 2.1. *For every $s \in S$, $n = 0, 1, \ldots, N-1$, and $j \in \mathcal{A}$, we have $Q^n(s, j) \geq V^{n+1}(s)$.*

As a result of this proposition, we have the following corollaries. The first one implies that if we know the true mean of an alternative, sampling from that alternative will not change our information about the best alternative. The second one implies that not taking a measurement would not improve the value function.

Corollary 2.2. *Let $j, j' \in \mathcal{A}$, $j \neq j'$, $n = 0, 1, \ldots, N-1$, and $s = (\mu, \Sigma)$. If $\Sigma_{jj} = 0$, then $Q^n(s, j) \leq Q^n(s, j')$.*

Corollary 2.3. *For all $s \in S$, $V^n(s) \geq V^{n+1}(s)$.*

3. Solution Proposal and IG Analysis

3.1. Information Gradient

Solving equation (5) to optimality is impractical because the state space is multi-dimensional and continuous. We propose a one-step look-ahead policy where the DM assumes that the next decision epoch is the last sampling opportunity. Since the objective of the DM is to gain maximum information about the unknown best alternative, one may think of a one-step look-ahead policy as a strategy that measures information gain by a sample. Thus, we call this strategy an information gradient (IG) policy.

Specifically, the IG policy at decision epoch n maximizes $\mathbb{E}^n\{V^N(\eta(s^n, a^n, X^{n+1})) - V^N(s^n)\}$ over all possible sampling alternatives, i.e., the IG policy at decision epoch n maximizes $\mathbb{E}^n\{\mathcal{I}_{J^*}(\eta(s^n, a^n, X^{n+1})) - \mathcal{I}_{J^*}(s^n)\}$. The IG policy is stationary, as in any stage, the decision only depends on the current state. Therefore, the action taken by the IG policy denoted as $\mathcal{J}^{IG} : S \to \mathcal{A}$ is given by

$$\mathcal{J}^{IG}(s) \in \arg\max_{j \in \mathcal{A}} \mathbb{E}\Big\{\mathcal{I}_{J^*}\big(\eta(s, j, X)\big) - \mathcal{I}_{J^*}(s)\Big\} = \arg\max_{j \in \mathcal{A}} \mathbb{E}\Big\{\mathcal{I}_{J^*}\big(\eta(s, j, X)\big)\Big\}, \quad (6)$$

where ties are broken randomly, i.e., if there are two or more alternatives with the same expectation in equation (6), then among these tied alternatives, IG picks one by randomizing among them giving each an equal chance.

Note that for $N = 1$, the IG policy is optimal by definition, and IG always chooses an alternative to measure in each epoch. Also, there are two types of decision policies:

1) Allocating samples to alternatives, 2) Selecting an alternative at the end of the sampling process. While IG's allocation policy is to sample an alternative to maximize information (or minimize the Shannon entropy) about the unknown best alternative, for selecting an alternative at the end of the sampling process, we recommend the alternative with the largest posterior mean. Next, we show the monotonicity of the value function on n under IG. Let $V^{IG,n}(s) = \mathbb{E}^{IG}\{V^N(s^N)|s^n = s\}$ denote the value function under the IG policy at n.

Proposition 3.1. *For every $s \in S$, $V^{IG,n}(s) \geq V^{IG,n+1}(s)$.*

Note that Propositions 2.1 and 3.1 closely follow Proposition 3.1 and Theorem 3.1 in Frazier et al. (2008) with one important difference that the objective considered by Frazier et al. (2008) is $\sup_{\pi\in\Pi} \mathbb{E}^{\pi}(\max_x \mu_x^N)$ while our objective is $\sup_{\pi\in\Pi} \mathbb{E}^{\pi}[\mathcal{I}_{J^*}(s^N)]$ which deems its own analyses.

Next, we show a limiting behavior of the information gain under IG. Defining

$$\nu^{IG,n}(s) := \mathbb{E}^n\Big[\mathcal{I}_{J^*}\big(\eta(s, \mathcal{J}^{IG}(s), X^{n+1})\big) - \mathcal{I}_{J^*}(s)\Big], \tag{7}$$

as the information gain at time n following IG, we have the following result.

Proposition 3.2. *Information gain $\nu^{IG,N}(s) \to 0$ as $N \to \infty$ for all s.*

Proposition 3.2 ensures that in long-run the expected information gain vanishes under the IG policy. However, this does not necessarily guarantee that the IG policy finds the best alternative. In fact, the result of Proposition 3.2 holds true when we replace $\mathcal{J}^{IG}(s)$ by any measurement choice $a^n = j$ and it is clear that a policy that samples from the same alternative will not find the best alternative almost surely. Thus, we still need to establish the consistency of the IG policy.

3.2. Consistency

Now, we show that the IG policy is consistent, i.e., it samples each alternative infinitely often, and therefore, it identifies the best alternative when the number of measurements tends to infinity. Consistency of sequential Bayesian sampling policies

is an important feature as the posterior distribution induced by a policy may fail to converge to the true value. In fact, infinite exploration of all alternatives guarantees that when the budget becomes large, the policy finds the true optimal solution in our setting. The lack of consistency may raise concerns about applicability of a proposed policy. The consistency of information-maximizing sampling policies was first addressed by Paninski (2005), where he showed that under some conditions on the parameter space, likelihood function, and the prior, the information-maximizing policy is consistent and its asymptotic performance is better (or not worse) than any non-adaptive strategy. Kujala (2016) relaxed some of the assumptions made in Paninski (2005) and showed its consistency in a more general setting. Since we consider Gaussian rewards, the conditions considered in Kujala (2016) do not apply, because it requires the parameter space to be compact, ruling out the Gaussian R&S. Frazier and Powell (2011) provided a set of sufficient conditions to ensure the consistency of sequential Bayesian sampling policies, but those conditions do not directly apply to our case, as the objective function in this study is maximizing the information about the best alternative, which is different from standard R&S problems. However, we use concepts developed by Frazier and Powell (2011) to show consistency in our setting. In particular, the proof of consistency presented for IG follows the proof of consistency for sequential sampling procedures by Frazier and Powell (2011) and not the proof of consistency for KG in Frazier et al. (2008). This is because the consistency proof for KG in the latter partly relies on the closed form formula found for the Expected Value of Information (EVI) in that paper.

To that end, in this section, we assume independent beliefs about the alternatives, i.e., the prior covariance matrix Σ^0 is diagonal, which given independent sampling, leads to the diagonal posterior covariance matrix Σ^n for all sampling epochs n. Because the posterior distribution about the true means is multivariate normal, the best alternative is known almost surely if and only if the covariance matrix is zero. Because our belief about the true mean of alternative j is normally distributed with $\mathcal{N}(\mu_j, \Sigma_{jj})$ and each time we sample from alternative j, Σ_{jj} will decrease, knowing the true mean of an alternative almost surely is equivalent to measuring it infinitely often. Therefore, in order to show the consistency of the IG policy, one needs to show that the IG policy

samples from each alternative infinitely often if $N \to \infty$. The main idea is to maintain an open neighborhood around knowledge states (introduced later) for which the true means of some alternatives are perfectly known. In this case, the IG policy does not stick to measuring the alternatives that are perfectly known while there are others that are not perfectly known.

Theorem 3.3. *The IG policy is consistent for the R&S with normal prior and rewards given independent beliefs about the alternatives and known sampling variances.*

Next, we relax the assumption that the sampling variances are known and show that the IG policy is consistent in this setting as well. To that end, we assume that the initial prior for each alternative $j \in \mathcal{A}$ has a normal-gamma distribution with parameters μ_j^0, λ_j^0, α_j^0, and κ_j^0 where $\phi_j \sim \Gamma\big(\alpha_j^0, \alpha_j^0/\lambda_j^0\big)$ is the prior distribution for the precision of sampling alternative j and $\theta_j|\phi_j \sim \mathcal{N}\big(\mu_j^0, 1/(\kappa_j^0\phi_j)\big)$. Then, given independent sampling over all alternatives, the conjugate posterior also has a normal-gamma distribution, i.e., $\phi_j|\mathcal{F}^n \sim \Gamma\big(\alpha_j^n, \alpha_j^n/\lambda_j^n\big)$ is the posterior distribution for the precision of sampling alternative j and $\theta_j|(\phi_j, \mathcal{F}^n) \sim \mathcal{N}\big(\mu_j^n, 1/(\kappa_j^n\phi_j)\big)$. We use the following updating equations to obtain the posterior belief about each alternative $j \in \mathcal{A}$,

$$\begin{aligned}
\mu_j^n &= \frac{\kappa_j^0\mu_j^0 + n_j\hat{m}_j^n}{\kappa_j^0 + n_j}, \\
\lambda_j^n &= \big(2\alpha_j^0 + n_j\big)/\Big(\frac{2\alpha_j^0}{\lambda_j^0} + n_j\hat{s}_j^n + \kappa_j^0(\mu_j^0)^2 - \frac{(\kappa_j^0\mu_j^0 + n_j\hat{m}_j^n)^2}{\kappa_j^0 + n_j}\Big), \\
\alpha_j^n &= \alpha_j^0 + \frac{n_j}{2}, \\
\kappa_j^n &= \kappa_j^0 + n_j,
\end{aligned} \tag{8}$$

where n_j, $\hat{m}_j^n$, and $\hat{s}_j^n$ are respectively the number of samples, the sample mean, and the sample variance of reward values observed for alternative j until time n. Then, by redefining η to be the transition function to determine the next state via equation (8), and also, assuming that $\alpha_j^0 > 0.5$ for all $j \in \mathcal{A}$, we have the following result.

Theorem 3.4. *The IG policy is consistent for the R&S with normal prior and rewards given independent beliefs about the alternatives and unknown sampling variances.*

3.3. IG Implementation and Approximate Information Gradient

The challenge of implementing IG is that computing $\mathcal{J}^{IG}$ is not amenable to closed form solutions because even calculating the probability of each alternative being the best at any time given the corresponding multivariate normal distribution, which is the building block for $\mathcal{J}^{IG}$ calculation, does not have a closed form solution. Therefore, we estimate the expectations and probabilities by Monte Carlo. Specifically, noting that IG is stationary, at any time with a given state $s = (\mu, \Sigma)$, we sample M many random vectors from the posterior $\mathcal{N}(\mu, \Sigma)$ and take the j^{th} element each time to create a pool of samples $\theta_j^{(1)}, \theta_j^{(2)}, \ldots, \theta_j^{(M)}$. In the case of independent beliefs across the alternatives, where Σ is diagonal, one can directly sample $\theta_j^{(m)} \sim \mathcal{N}(\mu_j, \Sigma_{jj}), \forall m$. Next, for each $m = 1, 2, \ldots, M$, sample $y_j^{(m)} \big| \{\theta_j^{(m)}, j\} \sim \mathcal{N}(\theta_j^{(m)}, \sigma_j^2)$, construct the new state using equation (2), and then, using Monte Carlo, estimate the posterior probability of each alternative i being the best given sampling alternative j for the m^{th} micro-replication when we are at state s, denoted as $p_i^{(m)} = \mathbb{P}(J^* = i | s, j, y_j^{(m)})$ for $i = 1, 2, \ldots, J$. Finally, find the expected information about J^* by averaging over m, and allocate sampling to the alternative that provides the largest information gain. Note that we exclude $p_i^{(m)}$s that are obtained as zero from the summation. Algorithm 1 describes the sampling procedure, for which $|\cdot|$ denotes cardinality of the set, M is the number of micro-replications simulated to estimate the expected information and K is the number of micro-replications simulated to estimate the expectation for $p_i^{(m)}$.

As discussed, estimating the posterior probability of an alternative being the best is computationally demanding. Therefore, we propose Approximate Information Gradient (AIG) in which we use its lower bound and upper bound to construct an estimation. Recall that $\mathbb{P}\{J^* = i\} = \mathbb{P}\big\{\theta_i > \theta_1, \ldots, \theta_i > \theta_{i-1}, \theta_i > \theta_{i+1}, \ldots, \theta_i > \theta_J\big\}$, and also, when the sampling variances are known, from the Slepians inequality (Branke et al. 2007), we have the following lower bound and upper bound

$$\prod_{i' \neq i} \mathbb{P}(\theta_i > \theta_{i'}) \leq \mathbb{P}\{J^* = i\} \leq \min_{i' \neq i} \mathbb{P}(\theta_i > \theta_{i'}). \tag{9}$$

In order to find $\mathbb{P}(\theta_i > \theta_{i'})$ in general for the case that we have alternatives with

correlated beliefs, we may use the approximation presented in Branke et al. (2007, equation (5)) which follows a general t distribution. For the case of independent beliefs about the alternatives, we use

$$\mathbb{P}(\theta_i > \theta_{i'}) = \Phi\left(\frac{\mu_i - \mu_{i'}}{\sqrt{\Sigma_{ii} + \Sigma_{i'i'}}}\right), \tag{10}$$

where $\theta_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$ and $\theta_{i'} \sim \mathcal{N}(\mu_{i'}, \Sigma_{i'i'})$, and $\Phi(\cdot)$ denotes the cumulative distribution function of standard normal.

Algorithm 1 Computation of the IG and AIG

for each alternative $j \in \mathcal{A}$ **do**

for each $m = 1 : M$ **do**

Sample $\theta^{(m)} \sim \mathcal{N}(\mu, \Sigma)$ and take the j^{th} element $\theta_j^{(m)}$.

Sample $y_j^{(m)} \sim \mathcal{N}(\theta_j^{(m)}, \sigma_j^2)$.

Construct the new state $s^{(m)} = (\mu^{(m)}, \Sigma^{(m)})$ using $y_j^{(m)}$.

if Implementing the IG policy **then**

for $k = 1 : K$ **do**

Sample the vector $\theta^{(k,m)} \sim \mathcal{N}(\mu^{(m)}, \Sigma^{(m)})$.

end for

for $i = 1 : J$ **do**

Let $\mathcal{S}_i^{(m)} := \{k : \theta_i^{(k,m)} > \max_{i' \neq i} \theta_{i'}^{(k,m)}\}$.

Let $p_i^{(m)} \leftarrow \dfrac{|\mathcal{S}_i^{(m)}|}{K}$.

end for

else if Implementing the AIG policy **then**

Let $p_i^{(m)} \leftarrow \beta \prod_{i' \neq i} \Phi\left(\dfrac{\mu_i - \mu_{i'}}{\sqrt{\Sigma_{ii} + \Sigma_{i'i'}}}\right) + (1-\beta) \min_{i' \neq i} \Phi\left(\dfrac{\mu_i - \mu_{i'}}{\sqrt{\Sigma_{ii} + \Sigma_{i'i'}}}\right)$ where $0 \leq \beta \leq 1$.

end if

Let $\mathcal{I}_{J^*}^{(m)} \leftarrow \sum_{i=1}^{J} p_i^{(m)} \log p_i^{(m)}$ excluding $p_i^{(m)}$s that are zero from the summation.

end for

Let $\mathbb{E}\Big\{\mathcal{I}_{J^*}\big(\eta(s,j,X)\big)\Big\} \leftarrow \dfrac{\sum_{m=1}^{M} \mathcal{I}_{J^*}^{(m)}}{M}$.

end for

$\mathcal{J}^{(A)IG}(s) \in \arg\max_{j \in \mathcal{A}} \mathbb{E}\Big\{\mathcal{I}_{J^*}\big(\eta(s,j,X)\big)\Big\}$.

Update the state $s = (\mu, \Sigma)$ by sampling from $\mathcal{J}^{(A)IG}(s)$.

In order to estimate the probability of each alternative being the best, we propose a convex combination of its lower bound and upper bound. In particular, in our numerical experiments, we use the average of lower bound and upper bound to avoid any tuning, and then, we normalize the average values to construct a proper probability measure. Normalization is necessary as we need to use these probabilities in the entropy function, which requires a proper probability measure to calculate expectation. Note that AIG uses this approximation for calculating $p_i^{(m)}$. The rest of the procedure in Algorithm 1 remains unchanged for AIG. Section 4.1 numerically shows the quality of such approximation in a variety of settings.

4. Numerical Experiments

In this section, we first investigate the quality of approximation by AIG compared to the actual IG policy. Next, we analyze AIG's performance and computation time in comparison to available benchmark policies and implement several sensitivity analyses to better understand AIG's performance and check its robustness. Throughout this section, we present the results of simulating AIG against five adaptive policies, alongside the non-adaptive Equal Allocation (EA), which allocates samples to $mod(n+1, J)$ where mod is the modulo operation, n is the sampling epoch, and J is the number of alternatives.

4.1. Quality of Approximation

In order to estimate $p_i^{(m)}$ for AIG one may use any convex combinations of lower and upper bounds, which requires tuning their relative weights. However, our experiments show that the normalized mean of lower and upper bounds is statistically justifiable in estimating these probabilities. Therefore, it is used for all the numerical experiments with AIG. The experiments show that the estimated probabilities with normalization are closer to Monte Carlo probabilities compared to the mean of lower bounds and upper bounds without normalization.

One question is how these approximated probabilities compare against those calculated precisely via Monte Carlo simulations. One way to check whether or not the

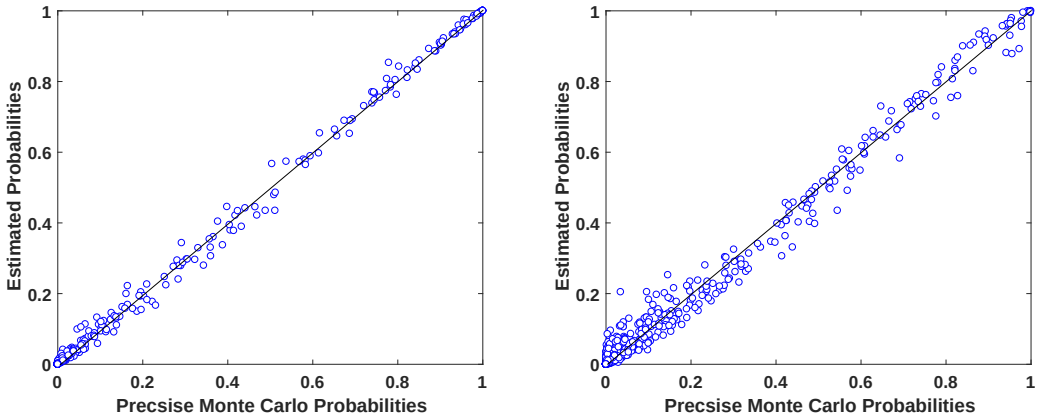
precise Monte Carlo estimations and the normalized mean of lower bounds and upper bounds are almost equal, is to show that for a number of random posterior beliefs they have a linear relationship with each other with the coefficient of slope being 1 and the coefficient of interception being 0.

To that end, we analyze two settings: (i) a R&S where we have ten alternatives with independent beliefs; (ii) a R&S where we have ten alternatives with correlated beliefs. We first conduct a Pearson's correlation test to justify the linearity of this relationship. We then use simple linear regression coefficient tests on both the slope and the interception to argue that there is not a meaningful statistical difference between Monte Carlo and estimated probabilities.

To implement the above-mentioned statistical tests on independent beliefs, a pool of 1000 randomly generated posteriors μ_j and Σ_{jj} is created for all alternatives $j = 1, 2, \ldots, 10$ considering independent beliefs. Specifically, the posterior mean values are generated randomly based on a normal distribution $\mathcal{N}(10, 20)$, and the posterior variances are randomly generated based on gamma distribution $\Gamma(2, 0.1)$. This setup enables us to see all different probabilities ranging from zero to one for each of the alternatives. Then, we carry out the statistical tests on Pearson's correlation and also the regression coefficients for one arbitrary alternative. The results are consistent for all the alternatives.

As we compare the relation between these two, we can see with over 95% confidence that we cannot reject the linearity hypothesis of this relationship as the Pearson's correlation is 0.9993 and p-value of correlation test ($H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$) is less than 10^{-16}. We also implement simple linear regression t-test for the interception coefficient as $H_0 : b_0 = 0$ vs. $H_1 : b_0 \neq 0$, and for the slope coefficient as $H_0 : b_1 = 1$ vs. $H_1 : b_1 \neq 1$, which resulted in p-values of 0.97 and 0.94 for interception and slope respectively. We observe that neither we can reject the linearity hypothesis, nor we can reject the null hypotheses on the regression coefficients.

Therefore, these tests show no meaningful difference between highly precise Monte Carlo probabilities and the normalized mean of upper bounds and lower bounds. The left plot in Figure 1 presents the estimated values against precise Monte Carlo probabilities for an arbitrary alternative, where we have independent beliefs.

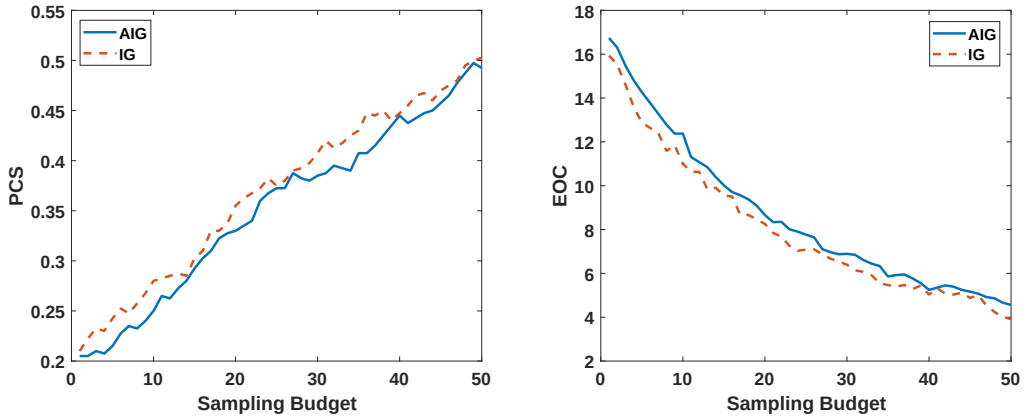


(a) Independent beliefs (b) Correlated beliefs

Figure 1. Precise Monte Carlo probabilities against estimated probabilities via the normalized mean of lower bound and upper bound. The black line resembles 45 degree line.

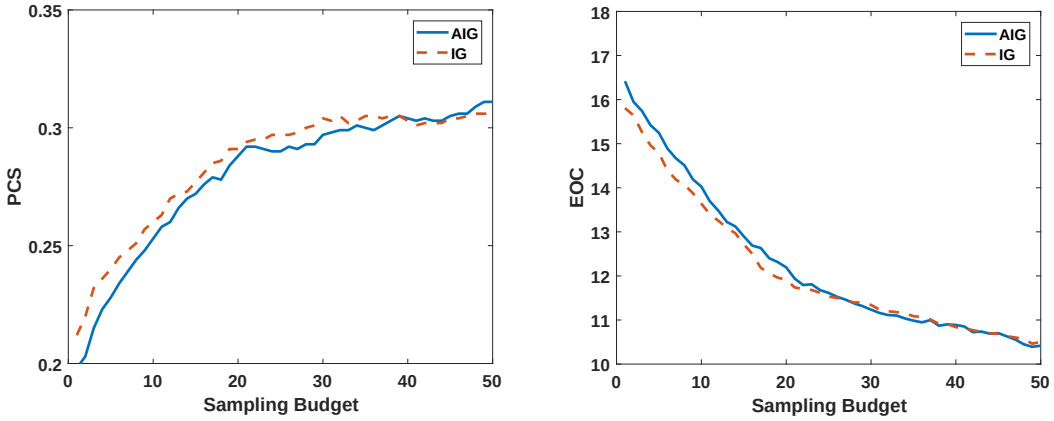
To test the quality of this approximation for correlated beliefs, we carry out a similar test with only one difference that instead of independently generating random posterior variances for each alternative, now we randomly generate 1000 symmetric positive definite covariance matrices Σ. To that end, we first generate a random $J \times J$ matrix A, for which each element is generated according to the uniform distribution $U[0, 5]$. We then create the Gramian matrix $A^T A$ and check its minimum eigenvalue. If positive, we use it in our experiment as a Σ instance, otherwise, we repeat the process. Similar to the previous setting, the random posteriors are generated such that for an arbitrary alternative we see all different probabilities ranging from zero to one. Then, using these correlated posteriors and their multivariate normal distribution, one can find the actual probability of each alternative being the best by Monte Carlo simulations and then compare these precise probabilities against the suggested normalized mean of upper and lower bounds (for which we need the mean vector μ and the diagonal elements of Σ).

The Pearson's correlation in this case turned out to be 0.9947 and also, the p-value of the same correlation test is less than 10^{-16}. Plus, simple t-tests on the linear regression coefficients resulted in p-values of 0.81 and 0.89 for interception and slope respectively. Therefore, similar to the case with independent beliefs, we cannot reject the linearity hypothesis. Plus, neither the hypotheses of regression slope being 1 nor the hypothesis of regression interception being 0 can be rejected. The right plot in Figure 1 presents



(a) Probability of correct selection (b) Expected opportunity cost

Figure 2. IG vs. AIG in a R&S problem with 5 alternatives and known sampling variance of $\sigma_j^2 = 10^3, \forall j$. The policies are implemented without initialization for a total of 10^4 macro-replications each time starting from randomly generated priors and problem instances (precision of ± 0.01 for the PCS plot).



(a) Probability of correct selection (b) Expected opportunity cost

Figure 3. IG vs. AIG in a R&S problem with 5 alternatives and unknown sampling variances over all alternatives. The policies are implemented without initialization for a total of 10^4 macro-replications each time starting from randomly generated priors and problem instances (precision of ± 0.01 for the PCS plot).

the estimated values against precise Monte Carlo probabilities in the correlated case for an arbitrary alternative. Additionally, experiments also show that the quality of this estimation increases as the posterior variances decrease to zero. Therefore, one expects that as the number of samples taken from each alternative increases, AIG behaves more closely to IG.

Finally, we test the performance of AIG and IG in a Bayesian setting, where 100 random initial priors are generated with means coming from a uniform distribution

Table 1. Computation times of one macro-replication of the simulation to allocate 400 samples by each policy.

Allocation Policy	EA	KG	TTTS	AOAP	OCBA	GBBA	AIG	IG
Replication Time (seconds)	0.008	0.074	2.417	0.086	9.826	10.672	62.763	4480.361

$U[0, 50]$ and initial prior variances as $\Sigma^0_{jj} = 100$ for all alternatives $j = 1, 2, \ldots, 5$. Then, from each prior 100 independent problem instances are generated following their normal distribution. Figure 2 presents PCS and EOC plots for the sampling variance of $\sigma^2_j = 10^3$ over all alternatives j, while Figure 3 repeats a similar experiment with random priors and problem instances for unknown sampling variances. For this experiment, the other initial prior hyper-parameters are set to $\kappa^0_j = 1$, $\alpha^0_j = 2$, and $\lambda^0_j = 2 \times 10^5$ for all alternatives $j \in \{1, \ldots, 5\}$. As can be seen from these figures, the performance of AIG in terms of PCS and EOC is similar to that of IG.

4.2. AIG's Performance and Computation Effort

This section designs a variety of settings to evaluate the performance and computation time of AIG compared to benchmark policies. To that end, we start from randomly generated initial prior means over all alternatives, i.e., $\mu^0_j, \forall j \in \mathcal{A}$ is generated from a uniform distribution $U[0, 1]$. The initial prior variance is set to $\Sigma^0_{jj} = 1, \forall j \in \mathcal{A}$. The beliefs about the alternatives are assumed to be independent. Problem instances are randomly generated following the initial prior normal distribution, i.e., $\theta_j \sim \mathcal{N}(\mu^0_j, \Sigma^0_{jj}), \forall j \in \mathcal{A}$. The entire sampling budget of $N = 400$ is allocated by each policy, and the resulting PCS and EOC is presented in Figure 4. We are conducting a simulation with 10^4 macro-replications for each of the policies. This guarantees a 95% confidence interval with the precision length of ± 0.01. As can be seen from Figure 4, given enough allocations, AIG along with other policies produces a PCS close to one.

With regards to computation effort, the parameters that affect the running time of these policies include the number of alternatives and the constant scalars in Algorithm 1 which are set to $J = 10$, $K = 100$, and $M = 15$. The computation times of each policy in one macro-replication (excluding the time for calculating PCS and EOC) of the simulation with a budget of 400 samples are reported in Table 1.

Since the running time of TTTS, OCBA, and GBBA depend on the random sequence of samples and the posterior distributions, Table 1 reports the average running

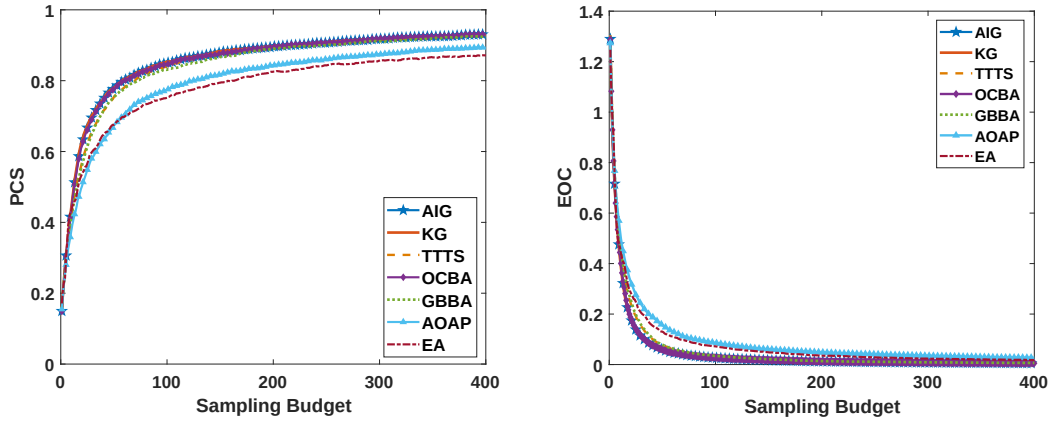
time of a single macro-replication of the simulation. As shown Table 1, AIG is computationally more efficient than IG, yet it falls behind benchmark policies, as AIG's allocation function lacks a closed form. However, AIG is still affordable for small to medium scale problems. We are using a single desktop computer with a Core-i7 processor and 6 Megabytes of cache for all experiments. Using the same computer, due to quadratic time complexity, if we double the number of alternatives, the running time of both IG and AIG will be roughly four times of those reported times in Table 1. For large scale R&S, parallelization and cloud computing methods could mitigate the growth in time as the number of alternatives increases, but we leave the analysis of information-seeking objectives in cloud and parallel computing to future study, because in a parallel computing setting, i.i.d. assumptions fail, and new statistical framework should be developed (Luo et al. 2015, Ni et al. 2017).

Next, we present the results of testing AIG against the benchmark policies in two different settings to better see the performance of AIG in R&S with small budgets and noisy measurements. To that end, here the results of two experiments with ten and twenty alternatives with independent beliefs are reported, respectively.

Following a Bayesian setting, 100 random initial prior means are generated from a uniform distribution $U[0, 50]$ alongside initial prior variances of $\Sigma^0_{jj} = 100$ for all alternatives j. Then, 100 independent problem instances are generated from each prior following their normal distribution to calculate PCS and EOC for times $n = 1, 2, \ldots, 100$. Here again, we are conducting 10^4 simulation macro-replications for each of the policies to guarantee a precision of ± 0.01. Figure 5 presents the PCS and EOC plots for a problem with ten alternatives when the sampling variance is $\sigma^2_j = 10^4$ for all j. Figure 6 presents the same plots for a problem with twenty alternatives when the sampling variance is $\sigma^2_j = 10^3$ for all j. One may conclude that given these plots, AIG is quite competitive against all other algorithms in terms of both PCS and EOC in these specific settings.

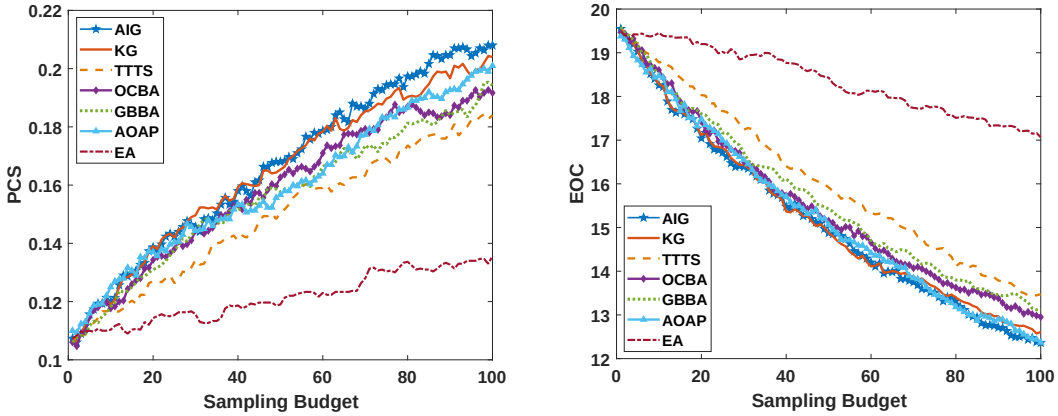
4.3. Sensitivity to Unknown Sampling Variances

To realize the effect of unknown sampling variances, we implement an experiment similar to the one in Section 4.2 for ten alternatives following a Bayesian setting.



(a) Probability of correct selection (b) Expected opportunity cost

Figure 4. Convergence of AIG to the truth vs. the recent benchmarks for a problem with 10 alternatives and a known sampling variance of $\sigma_j^2 = 1, \forall j$. The policies are implemented without initialization for a total of 10^4 macro-replications each starting from randomly generated priors (precision of ± 0.01 for the PCS plot).

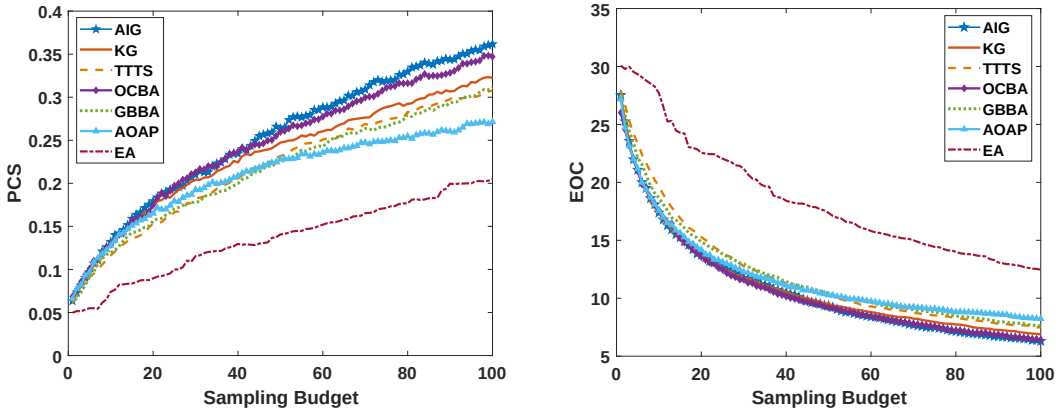


(a) Probability of correct selection (b) Expected opportunity cost

Figure 5. Performance of AIG vs. the recent benchmarks for a problem with 10 alternatives and a known sampling variance of $\sigma_j^2 = 10^4, \forall j$. The policies are implemented without initialization for a total of 10^4 macro-replications each starting from randomly generated priors (precision of ± 0.01 for the PCS plot).

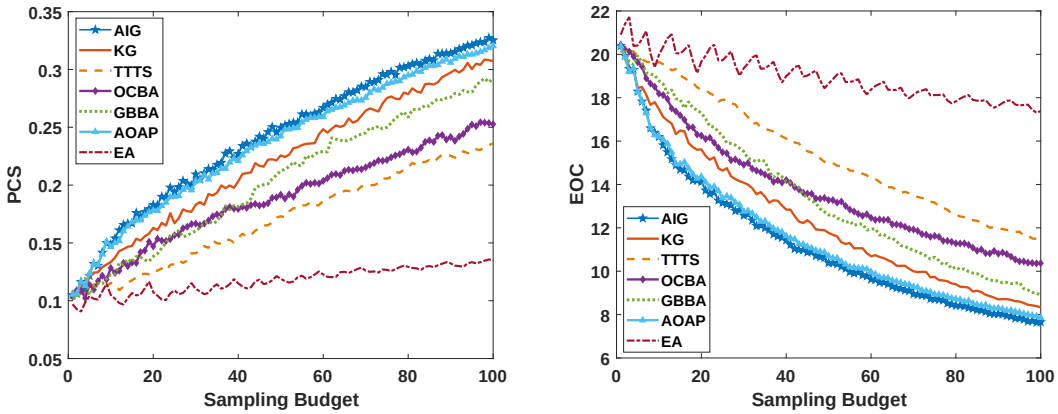
The only difference here is that we consider a case in which all sampling variances are unknown and we use the modified algorithms, to compare AIG with other benchmarks to see the effect of considering unknown variances.

For this experiment in particular, we use a normal-gamma prior with independent beliefs about the alternatives (described in Section 3.2), where at each time, we first sample for the variance and then for the mean value. Note that in this setup where we have independent beliefs about the alternatives but the sampling variances are unknown, equation (10) does not necessarily hold true. Nonetheless, we use it to com-



(a) Probability of correct selection (b) Expected opportunity cost

Figure 6. Performance of AIG vs. the recent benchmarks for a problem with 20 alternatives and a known sampling variance of $\sigma_j^2 = 10^3, \forall j$. The policies are implemented without initialization for a total of 10^4 macro-replications each starting from randomly generated priors (precision of ± 0.01 for the PCS plot).



(a) Probability of correct selection (b) Expected opportunity cost

Figure 7. Performance of AIG vs. the recent benchmarks for a problem with 10 alternatives and unknown sampling variances over all alternatives. The policies are implemented without initialization for a total of 10^4 macro-replications each starting from randomly generated priors (precision of ± 0.01 for the PCS plot).

pute AIG for the case with unknown sampling variances. Although in this case, the approximation for $p_i^{(m)}$ will not necessarily be equal to the average of the actual upper and lower bounds, but instead, it will be the average of their approximation. After all, this approximation becomes more precise as we take more samples and better learn the value of the hyper-parameters.

For this experiment, while μ_j^0 and problem instances are randomly generated according to a Bayesian setting similar to Section 4.2, the other initial prior hyper-parameters are set to $\kappa_j^0 = 1$, $\alpha_j^0 = 2$, and $\lambda_j^0 = 2 \times 10^5$ for all alternatives $j \in \{1, \ldots, 10\}$. Figure

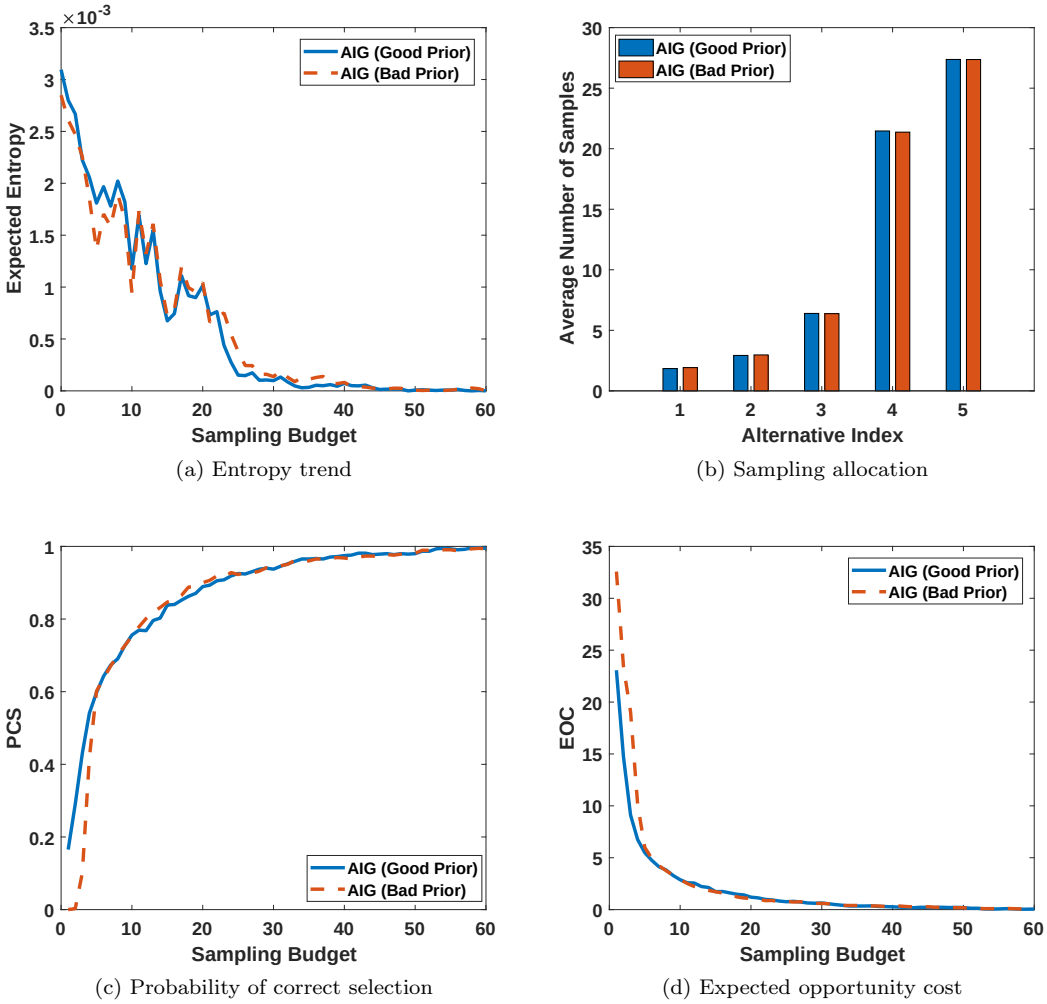
7 presents the PCS and EOC plots for the case that sampling variances are unknown. As can be seen in these plots, AIG still performs competitively compared to other policies, which means it is robust to unknown sampling variances.

4.4. Sensitivity to a Bad Prior

In this section, we analyze the sensitivity of AIG to a bad initial prior belief. In this regard, we test two problem settings in two different cases. Both problems are created in a frequentist setting with a fixed truth $\theta = \{10, 20, 30, 40, 50\}$ for five alternatives with all the parameters the same except for the initial prior mean of the alternatives. In one problem, the initial prior is fixed to a flat and good belief (which is close in values to the truth) $\mu^0 = \{30, 30, 30, 30, 30\}$, and in the other one, we choose to start with a bad (far from truth) initial prior $\mu^0 = \{100, 75, 50, 25, 0\}$ in which the worst alternative is believed to be the best at the beginning of the trial and vice versa. Here, using a simulation with 1000 macro-replications, we analyze PCS, EOC, expected entropy, and the average number of samples taken from each alternative by AIG throughout the trial in two different cases for the initial prior variances.

The first case is where we are dealing with a regular highly uncertain prior with large initial variances ($\Sigma^0_{jj} = 10000$ for all j) at the beginning of the trial for both problem settings with good and bad initial prior means. In this case, when the sampling variance is set to $\sigma^2_j = 200$ over all alternatives j, as presented in Figure 8, the main difference is made at the very beginning in which a bad prior makes us take a few more samples to converge to a good belief, and therefore, in terms of EOC, we observe slightly higher costs for a problem with a bad initial prior. However, in terms of PCS after the few first samples, we can see that the PCS for these two settings become very close and almost the same. It is also safe to say that there is no meaningful difference between the number of samples taken from each alternative when comparing the settings with good and bad prior means, as AIG takes samples from all alternatives to explore enough at the beginning of the trial, and then, it exploits more often from better alternatives. In terms of expected entropy, they both start from a same point, mainly because although the initial prior means are different, the initial prior variances are so large that the probability of each alternative being the best is almost equal for all alternatives in

both settings. From that point on, we see a decrease in the expected entropy plots following a similar pattern for good and bad prior means. Note that the entropy has a decreasing trend but not monotonically, which is similar to non-monotonicity of PCS (Peng et al. 2015, 2018a) and EOC (Qu et al. 2015).



(a) Entropy trend

(b) Sampling allocation

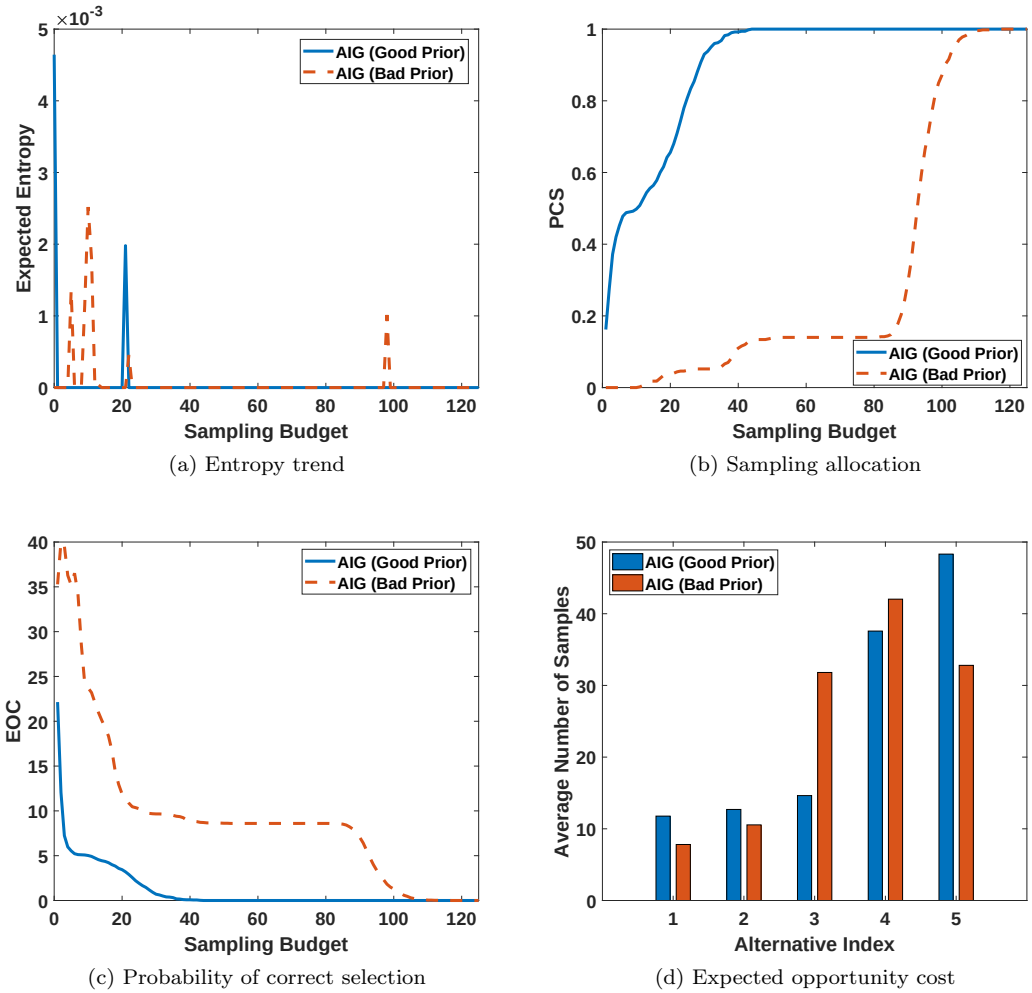
(c) Probability of correct selection

(d) Expected opportunity cost

Figure 8. Sensitivity of AIG to a bad uncertain prior (big initial variance).

The second case is where we choose the initial prior variances to be very small ($\Sigma^0_{jj} = 1$ for all j). So compared to the first case, we are assuming the same good and bad prior means, but the initial variances are assumed to be small. This relates to a DM who has rather a certain initial belief about each of the alternatives. Note that in the second case, starting from bad initial prior mean values and very small initial prior variances, without implementing any initialization, some of the algorithms like

KG and TS (TTTS) will not numerically converge to the true means.



(a) Entropy trend

(b) Sampling allocation

(c) Probability of correct selection

(d) Expected opportunity cost

Figure 9. Sensitivity of AIG to a bad certain prior (small initial variance).

The limit with implementing KG in the setting with bad prior means and small prior variances is the calculation of normal c.d.f. which at the beginning of the trial, given allocating to any alternative, produces extremely small numbers that is detected zero by MATLAB. The calculation of c.d.f. in Frazier et al. (2008, equation (19)) can be avoided if we attempt to find the EVI by Monte Carlo instead; however, we have to drastically increase the precision of this Monte Carlo simulation to the point that KG will be computationally demanding. The issue with TS and TTTS is the calculation of normal p.d.f. given the prior distribution of each alternative, which results in TS and TTTS getting stuck at sampling the alternative which has the greatest posterior mean. This is the issue with TS and TTTS that when the posterior variances are

too small over all alternatives, the exploration becomes highly unlikely. In the setting with bad prior means and small variances, this happens after one or two allocations, and therefore, they both numerically diverge in these settings, even after allocating millions of samples.

However, under such conditions, AIG still converges to the truth with a reasonable budget which is highly dependent on the sampling variance. The reason behind this behavior is in fact the exploration nature of entropy-based algorithms which is quite unique. Other algorithms, e.g., AOAP, OCBA, and GBBA are close in performance to AIG in such a setting. In this case, it can be seen in Figure 9 that with a sampling variance of $\sigma_j^2 = 1$ over all alternatives, starting from a bad prior, it takes more samples for PCS to start improving and therefore, we observe higher EOC levels. Also, AIG starts with taking more samples from alternatives that are believed to be better. In fact, in this case, starting from a bad initial prior, a lot of samples need to be taken from each alternative to bring their posterior mean values closer to reality. Plus, in terms of the expected entropy, under a good prior, we have high entropy levels at time zero, but with a couple of samples it suddenly drops to very small values, while under a bad prior, the expected entropy is small at time zero, and stays small for the first few samples. As we take several samples though, in both settings, we observe increasing shocks in entropy for both settings, but a higher number of shocks under a bad prior, and after a few shocks, they both decrease again and converge to zero.

5. Conclusion

We introduced an entropy-based objective function for ranking and selection and formulated it by a stochastic dynamic program, by which we showed some properties of the underlying learning problem. Because the state space of the stochastic dynamic program is unbounded, standard techniques fail to produce solutions. Therefore, we employed a one-step look-ahead policy, which maximizes the information gain about the best alternative by the next observation, and hence, we called it information gradient. We showed several properties of the IG policy including its consistency, i.e., it identifies the best alternative almost surely when the sampling budget tends to

infinity. Since the IG policy is computationally demanding, we proposed an approximation for the probability of each alternative being the best using the normalized convex combination of its upper bound and lower bound. The resulting algorithm, called AIG which is statistically not different from IG itself, is computationally affordable for small to medium scale problems, and performs competitively compared to non-adaptive EA, as well as five adaptive policies, e.g., KG, TTTS, OCBA, GBBA, and AOAP. Results show that not only AIG performs competitively in terms of both PCS and EOC against other policies, but also, it is quite robust to unknown sampling variances and initial bad priors. Plus, it does not require any parameter tuning or initial replications to perform well. These all make both IG and AIG good candidates in ranking and selection for identifying the best alternative, especially in noisy environments for early stage sampling when the sampling budget is small. As an example, in early stage clinical trials with low budget and high noise, a small (but statistically significant) improvement in PCS may have a significant impact on future costs and patient outcomes.

Acknowledgments: This research is sponsored by the National Science Foundation award number 1651912.

References

Berry, D. A. and Pearson, L. M. (1985). Optimal designs for clinical trials with dichotomous responses. *Statistics in Medicine*, 4(4):497–508.

Bhat, N., Farias, V. F., Moallemi, C. C., and Sinha, D. (2019). Near optimal A-B testing. *Management Science*.

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

Bubeck, S., Munos, R., and Stoltz, G. (2011a). Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. (2011b). X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695.

Chen, C. H., Donohue, K., Yücesan, E., and Lin, J. (2003). Optimal computing budget allocation for Monte Carlo simulation with application to product design. *Simulation Modelling Practice and Theory*, 11(1):57–74.

Chen, C. H. and Lee, L. H. (2011). *Stochastic simulation optimization: an optimal computing budget allocation*, volume 1. World scientific.

Chen, C. H., Lin, J., Yücesan, E., and Chick, S. E. (2000). Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770.

Chick, S. E., Branke, J., and Schmidt, C. (2010). Sequential sampling to myopically maximize the expected value of information. *INFORMS Journal on Computing*, 22(1):71–80.

Frazier, P. I. and Powell, W. B. (2011). Consistency of sequential Bayesian sampling policies. *SIAM Journal on Control and Optimization*, 49(2):712–731.

Frazier, P. I., Powell, W. B., and Dayanik, S. (2008). A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439.

Hennig, P. and Schuler, C. J. (2012). Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(Jun):1809–1837.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, pages 918–926.

Hunter, S. R. and Pasupathy, R. (2013). Optimal sampling laws for stochastically constrained simulation optimization on finite sets. *INFORMS Journal on Computing*, 25(3):527–542.

Kim, S.-H. and Nelson, B. L. (2006). Selecting the best system. *Handbooks in Operations Research and Management Science*, 13:501–534.

Kolmogorov, A. (1956). On the Shannon theory of information transmission in the case of continuous signals. *IEEE Transactions on Information Theory*, 2(4):102–108.

Kontsevich, L. L. and Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, 39(16):2729–2737.

Kujala, J. V. (2016). Asymptotic optimality of myopic information-based strategies for Bayesian adaptive estimation. *Bernoulli*, 22(1):615–651.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005.

Luo, J., Hong, L. J., Nelson, B. L., and Wu, Y. (2015). Fully sequential procedures for large-scale ranking-and-selection problems in parallel computing environments. *Operations Research*, 63(5):1177–1194.

Machens, C. K. (2002). Adaptive sampling by information maximization. *Physical Review Letters*, 88(22):228104.

Mehrez, A. and Sethi, A. P. (1989). Hierarchical planning of project-selection problems with information purchasing. *Journal of the Operational Research Society*, 40(3):267–279.

Ni, E. C., Ciocan, D. F., Henderson, S. G., and Hunter, S. R. (2017). Efficient ranking and selection in parallel computing environments. *Operations Research*, 65(3):821–836.

Pallone, S. N., Frazier, P. I., and Henderson, S. G. (2017). Bayes-optimal entropy pursuit for active choice-based preference learning. *arXiv preprint arXiv:1702.07694*.

Paninski, L. (2005). Asymptotic theory of information-theoretic experimental design. *Neural Computation*, 17(7):1480–1507.

Peng, Y., Chen, C. H., Fu, M. C., and Hu, J.-Q. (2015). Non-monotonicity of probability of correct selection. In *2015 WSC*, pages 3678–3689. IEEE.

Peng, Y., Chen, C. H., Fu, M. C., and Hu, J.-Q. (2016). Dynamic sampling allocation and design selection. *INFORMS Journal on Computing*, 28(2):195–208.

Peng, Y., Chen, C. H., Fu, M. C., and Hu, J.-Q. (2018a). Gradient-based myopic allocation policy: An efficient sampling procedure in a low-confidence scenario. *IEEE Transactions on Automatic Control*, 63(9):3091–3097.

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

Powell, W. B. and Ryzhov, I. O. (2012). *Optimal Learning*. John Wiley & Sons.

Qu, H., Ryzhov, I. O., Fu, M. C., and Ding, Z. (2015). Sequential selection with unknown correlation structures. *Operations Research*, 63(4):931–948.

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

Russo, D. (2016). Simple Bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418.

Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471.

Villemonteix, J., Vazquez, E., and Walter, E. (2009). An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509.

Xu, J., Nelson, B. L., and Hong, L. J. (2013). An adaptive hyperbox algorithm for high-dimensional discrete optimization via simulation problems. *INFORMS Journal on Computing*, 25(1):133–146.