Finite-time analysis for the knowledge-gradient policy and a new testing environment for optimal learning

Yingfei Wang yingfei@cs.princeton.edu
Department of Computer Science
Princeton University
Princeton, NJ 08540, USA

Warren Powell powell@princeton.edu
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08540, USA

Robert Schapire schapire@cs.princeton.edu
Department of Computer Science
Princeton University
Princeton, NJ 08540, USA

Editor:

Abstract

We consider two learning scenarios, the offline Bayesian ranking and selection problem with independent normal rewards and the online multi-armed bandit problem. By reducing the Bayesian ranking and selection problems to the adaptive stochastic multi-set maximization problems, we derive the first finite-time bound of the knowledge-gradient policy for adaptive submodular objective functions. In addition, we introduce the concept of prior-optimality and provide another insight into the performance of the KG policy based on the submodular assumption on the value of information. We demonstrate submodularity for the two-alternative case and provide other conditions for more general problems, filling in a gap in the analysis of the knowledge gradient policy. We then address the relative paucity of empirical testing of learning algorithms (of any type) by introducing a new public-domain, modular optimal learning testing environment (MOLTE) that allows users to draw on a library of algorithms and test problems which makes it easy to add new algorithms and new test problems. We demonstrate the capabilities of MOLTE through a series of comparisons of policies on a starter library of test problems.

1. Introduction

We consider sequential decision problems in which at each time step, we choose one of finitely many alternatives and observe a random reward. The rewards are independent of each other and follow some unknown probability distribution. One goal can be to identify the alternative with the best expected performance within a limited measurement budget, which is the objective of offline ranking and selection. Another goal can be to maximize the expected cumulative sum of rewards obtained in a sequence of allocations, a problem class often addressed under the umbrella of multi-armed bandit problems. Both ranking and selection problems and bandit problems are examples of sequential decision making problems with partial information that address the exploration-exploitation trade-off. Since the learner does not know the true distribution of each alternative, it needs to explore the choices that might give good rewards in the future as well as exploit the alternatives that appear to be better based on previous observations.

Ranking and selection problems arise in many settings. We may have to choose a type of material that has the best performance, the features in a laptop or car that produce the highest sales, or the molecular combination that produces the most effective drug. Often, the cost of a measurement may be substantial. Laboratory or field experiments may take a day or several weeks. For this reason, we assume we have a limited budget for making measurements.

Raiffa and Schlaifer (1961) established the Bayesian framework for R&S problems. Several two-stage and sequential procedures exist for selecting the best alternative. Branke et al. (2007) made a thorough comparison of several fully sequential sampling procedures. They indicate that the optimal computing budget allocation (OCBA) (Chen et al., 1996, 2000; He et al., 2007) and value of information procedures (VIP) (Chick, 2001) perform quite well and better than a deterministic or two-stage policy (Chen et al., 2006). Another single-step Bayesian look-ahead policy first introduced by Gupta and Miescke (1996) and then further studied by Frazier et al. (2008) is called the "knowledge-gradient policy" (KG). It chooses to measure the alternative that maximizes the single-period expected value of information. Whereas the above mentioned policies assumed an independent normal or one-dimensional Wiener process prior on the alternatives' true means, Frazier et al. (2009) modified the knowledge-gradient policy to handle correlated multivariate normal belief on the mean values of these rewards.

The bandit problem was originally studied under Bayesian assumptions (Gittins, 1979). A widely used class of policies for multi-armed bandit problems is called *upper confidence bounding* policies (UCB). Different UCB-type variants have been developed for many types of reward distributions and have provable logarithmic regret bounds (Lai and Robbins, 1985; Agrawal, 1995; Auer et al., 2002; Kleinberg et al., 2010; Bubeck et al., 2012). By contrast, knowledge gradient policies, which enjoy some nice theoretical properties, have never been characterized by the type of regret bounds for which UCB policies are famous.

This paper makes the following contributions: (1) We first reduce the Bayesian ranking and selection problem to adaptive stochastic multi-set function maximization problems where each multi-set corresponds to a set of selected alternatives. The multi-set representation captures our ability to evaluate the same alternative more than once. We derive the first finite-time bound for the knowledge gradient policy for R&S problems under the assumption that the utility function is adaptive submodular. However, pathwise adaptive submodularity can fail in offline learning settings when the utility function itself involves a maximum. To this end, instead of the pathwise behavior analyses of the utility function, we further study its average behavior by taking expectations over the observations given any fixed sample allocation, resulting in a well-know quantity: the value of information. As a result, we introduce the concept of the prior-value of a policy and analyze the prior-optimality of the KG policy to provide another insight into its performance based on the submodular assumption of the value of information that is weaker than adaptive submodularity. To accomplish this, we build on the general structure of the analysis of greedy algorithms given in Nemhauser et al. (1978) and Golovin and Krause (2010). We demonstrate submodularity for the two-alternative case and provide other conditions for more general problems, filling in a gap in the analysis of the knowledge gradient policy. (2) We introduce a new Modular Optimal Learning Testing Environment (MOLTE) for comparing a number of policies on a wide range of learning problems, providing the most comprehensive testbed that has yet appeared in the literature. We draw the conclusion that there is no universal best policy for all problem classes, which means that theoretical guarantees are not by themselves reliable indicators of which policy is best for a particular problem class. We offer MOLTE as an easy-to-use tool for the research community that will make it possible to perform much more comprehensive testing, spanning a broader selection of algorithms and test problems. We also address the problem of tuning and constructing priors that have been largely overlooked in optimal learning literature.

This paper is organized as follows. In section 2, we lay out the mathematical models for R&S problems and multi-armed bandit problems. In section 3, we describe the knowledge gradient policies for offline and online learning. In section 4, we provide theoretical analyses for the KG for offline learning. In section 5, we point out that submodularity does not hold in general. We analyze submodularity for a problem with two alternatives, and present insights for more general problems. Finally, in section 6 we introduce a new Modular Optimal Learning Testing Environment (MOLTE) which gives researchers access to a wide range of test problems and competing policies, within an architecture that makes adding new problems and policies quite easy. In sections 7 and 8, we present performance results and analyses of various policies for both R&S problems and multi-armed bandit problems. These experiments illustrate the features of MOLTE, and help to demonstrate that it is hard to predict how well a particular policy will work on a particular problem class. Our hope is that MOLTE can be used by the research community to simplify the process of doing more comprehensive experimental testing. We close in section 9 with a discussion of the problem of tunable parameters and constructing priors.

2. Model

In this section, we provide formal definitions of the offline ranking and selection problem, and the online multi-armed bandit problem.

2.1 The Offline Ranking and Selection Problem

Suppose we have a collection $\mathcal{X}$ of M alternatives (where M might be quite large), each of which can be measured sequentially to estimate its unknown mean μ_x. We assume normally distributed measurement noise with known variance σ_W^2. We first introduce the model for independent normal beliefs. We begin with a normally distributed Bayesian prior belief on the sampling means that is independent across alternatives, $\mu_x \sim \mathcal{N}(\theta_x^0, \sigma_x^0)$. At the nth iteration, we use some measurement policy π to choose one alternative x^n and observe $W_{x^n}^{n+1} \sim \mathcal{N}(\mu_{x^n}, \sigma_W)$.

For convenience, we introduce the σ-algebras $\mathcal{F}^n$ for any $n = 0, 1, ..., N-1$ which is formed by the previous n measurement choices and outcomes, $x^0, W^1, ..., x^{n-1}, W^n$. We define $\theta_x^n = \mathbb{E}[\mu_x|\mathcal{F}^n]$ and $(\sigma_x^n)^2 = \mathrm{Var}[\mu_x|\mathcal{F}^n]$. Then conditionally on $\mathcal{F}^n$, $\mu_x \sim \mathcal{N}(\theta_x^n, \sigma_x^n)$. Let $\beta_x^n = \frac{1}{(\sigma_x^n)^2}$ be the conditional precision of μ_x and our state of knowledge be $S^n =$

$(\theta_x^n, \beta_x^n)_{x\in\mathcal{X}}$. We will use $\mathcal{F}^n$ and S^n interchangeably. After the nth measurement we update our beliefs using Bayes' rule:

$$\theta_x^{n+1} = \begin{cases} \frac{\beta_x^n \theta_x^n + \beta^W W^{n+1}}{\beta^n + \beta^W} & \text{if } x^n = x \\ \theta_x^n & \text{otherwise,} \end{cases} \qquad \beta_x^{n+1} = \begin{cases} \beta^n + \beta^W & \text{if } x^n = x \\ \beta_x^n & \text{otherwise,} \end{cases}$$

where $\beta^W = 1/\sigma_W^2$.

We may impose correlated beliefs between alternatives in order to strengthen the effect of each measurement. Starting from a prior distribution $\mathcal{N}(\theta^0, \Sigma^0)$ and after measurement W^{n+1} of alternative x, a posterior distribution on the beliefs are calculated by:

$$\theta^{n+1} = \Sigma^{n+1}\left((\Sigma^n)^{-1}\theta^n + \beta^W W^{n+1} e_x\right), \tag{1}$$

$$\Sigma^{n+1} = \left((\Sigma^n)^{-1} + \beta^W e_x e_x^T\right)^{-1}, \tag{2}$$

where e_x is the vector with 1 in the entry corresponding to alternative x and 0 elsewhere. $S^n = (\theta^n, \Sigma^n)$ is then our state of knowledge in this case.

A decision function $X^\pi(S^n)$ is defined as a mapping from the knowledge state to $\mathcal{X}$. We refer to the decision function X^π and the policy π interchangeably.

If we are limited to N measurements, the objective is to maximize the expected reward of the final recommended alternative:

$$\max_{\pi\in\Pi} \mathbb{E}\left[\mu_{x^\pi}\right], \tag{3}$$

where $x^\pi = \arg\max_{x\in\mathcal{X}} \theta_x^N$ and $x^n = X^\pi(S^n)$ for $0 \le n < N$.

2.2 The Multi-armed Bandit Problem

In the multi-armed bandit problem, every arm $x \in \mathcal{X}$ corresponds to an unknown probability distribution with mean μ_x. At each step, we use some policy to choose one arm $x^n = X^\pi(S^n)$ and receive a reward W^{n+1} drawn from that arm's distribution. The goal is to maximize the total expected reward collected over time:

$$\max_{\pi\in\Pi} \mathbb{E}\left[\sum_{n=0}^{N-1} \mu_{X^\pi(S^n)}\right]. \tag{4}$$

3. Knowledge Gradient

For R&S problems, the knowledge gradient is a policy that at the nth iteration chooses its $(n+1)$st measurement from $\mathcal{X}$ to maximize the single-period expected increase in value (Frazier et al., 2008, 2009). To be more specific, the value of being in state S^n is $\max_{x\in\mathcal{X}} \theta_x^n$. If we choose to measure $x^n = x$ right now, allowing us to observe W_x^{n+1}, then we transition to a new state of knowledge $S^{n+1} = (\theta^{n+1}, \Sigma^{n+1})$. At iteration n, θ_x^{n+1} is a random variable since we do not yet know what W^{n+1} is going to be. We would like to choose x at iteration n which maximizes the expected value of $\max_{x\in\mathcal{X}} \theta_x^{n+1}$. We can think of this as choosing an alternative to maximize the incremental value, given by

$$\nu_x^{\mathrm{KG},n} = \mathbb{E}[\max_{x'} \theta_{x'}^{n+1} - \max_{x'} \theta_{x'}^n | x^n = x, S^n]. \tag{5}$$

The knowledge gradient policy $X^{\mathrm{KG}}(S^n)$ is defined by

$$X^{\mathrm{KG}}(S^n) = \arg\max_{x \in \mathcal{X}} \nu_x^{\mathrm{KG},n}. \tag{6}$$

The knowledge gradient policy can handle the presence of a variety of belief models such linear (Negoescu et al., 2011) or nonparametric (Mes et al., 2011; Barut and Powell, 2013).

Next, we present a recent result of the knowledge gradient for undiscounted multi-armed bandit problems. If $\nu_x^{KG,n}$ is the offline knowledge gradient, then the online knowledge gradient $\nu_x^{OLKG,n}$ is given as

$$\nu_x^{OLKG,n} = \theta_x^n + (N-n)\nu_x^{KG,n}. \tag{7}$$

As before, the KG policy chooses the alternative with the largest value, which is to say $X^{\mathrm{OLKG}}(h^n) = \arg\max_{x \in \mathcal{X}} \nu_x^{OLKG,n}$ (Ryzhov et al., 2012). This relationship allows the online knowledge gradient to inherit our ability to handle correlated beliefs (Ryzhov et al., 2012) with no additional computational effort.

The knowledge gradient policy has some nice properties. For offline learning settings, the knowledge gradient policy is optimal (by definition) if the measurement budget $N = 1$. The knowledge gradient is guaranteed to find the best alternative as the measurement budget N tends to infinity. If there are only two choices, the knowledge gradient policy is optimal for any measurement budget. The knowledge gradient policy is the only stationary policy that is both myopically and asymptotically optimal. For online learning problems, the knowledge gradient policy is asymptotically optimal as the discount factor tends to one. However, the KG has not enjoyed the finite-time bounds that have been popular in the UCB policies.

4. Finite-time Bound for the Knowledge Gradient Policy

We follow the general structure of the analysis of greedy approximation (Nemhauser et al., 1978) to develop the first finite-time bound for the knowledge gradient policy for offline learning as follows. In Section 4.1, by reducing the Bayesian R&S problems to the adaptive stochastic multi-set maximization problems, we show that the KG policy inherits precisely the performance guarantees of the greedy algorithm for classic submodular maximization problems if the utility function is adaptive submodular. We show that the adaptive submodular assumption easily fails in the ranking and selection problems. Hence, instead of the pathwise behavior analyses of the utility function, we study its average behavior by taking expectation over the observations in Section 4.2. In Section 4.3, we analyze the *prior-optimality* which provides another insight into the performance of the KG policy based on the submodularity of a well-understood quantity–value of information.

It is important to note that both the submodular maximization reduction and the theoretical analyses on the prior-optimality are not only applied to the specific setup of Gaussian noise in observations and Gaussian prior structure. The theoretical guarantees are more generally applicable to any prior and measurement noise model as long as the adaptive submodular assumption or the submodular value of information assumption holds.

4.1 The Reduction of R&S to Adaptive Stochastic Multi-set Maximization

We first introduce the adaptive stochastic maximization problem. Let E be a finite set of items. Each item $e \in E$ maps to a random outcome of a measurement $\Phi(e)$ in a set O of possible values. We define a realization as a function $\phi : E \mapsto O$ representing the observation of each item in the ground set. Under Bayesian interpretation, we assume that there is a known prior probability distribution $p(\phi) := \mathbb{P}(\Phi = \phi)$ over all possible realizations. The adaptive stochastic optimization problem consists of sequentially picking an item $e \in E$, revealing its outcome $\Phi(e)$ and picking the next item. After each pick, the observations so far can be represented as a partial realization ψ. A partial realization ψ is consistent with realization ϕ, denoted as $\phi \sim \psi$, if all the items selected in ψ have the same outcomes as in ϕ. We use $\mathrm{dom}(\psi)$ to refer to the items observed in ψ. We use the notation $Z^\pi(\phi)$ to denote the set of items chosen by policy π under realization ϕ.

We wish to maximize some utility function $f : 2^E \times O^E \mapsto \mathbb{R}$ that depends on which items we pick and which states they are in. The expected utility of a policy π is $f_{\mathrm{avg}}(\pi) := \mathbb{E}\big[f\big(Z^\pi(\Phi), \Phi\big)\big]$ where the expectation is taken over the prior distribution $p(\phi)$. The goal of adaptive stochastic set maximization problem is to find an optimal policy π^* that maximizes its expected utility under a cardinality constraint,

$$\pi^* \in \arg\max_\pi f_{\mathrm{avg}}(\pi), \text{ subject to } |Z^\pi(\phi)| \leq N,$$

where N is the measurement budget.

It is not obvious to treat the ranking and selection problem in an adaptive stochastic multi-set maximization way of thinking. To see this, define the ground set $E = \mathcal{X}$. The outcomes are real numbers with $O = \mathbb{R}$. Each alternative $e = x$ can be selected multiple times. After each selection, its random outcome $\Phi(e) = W_x \in O$ is revealed.

Since the true values μ_x are random variables, we can let φ be a sample realization of the truth with a (correlated) prior distribution $p(\varphi) = \mathcal{N}(\theta^0, \Sigma^0)$. We use the notation $\phi \in \Phi$ to denote an realization of the random observations in our problem. The prior probability distribution over the realizations is determined by $p(\varphi)$ and the noise distribution $\mathcal{N}(0, \sigma_W)$. For example, if in the ranking and selection problems each alternative can only be selected once, $\phi : E \mapsto O$. For multi-selections, one way of defining the realization is by first making replicas of each item to construct E' and then selecting each $e' \in E'$ at most once.

Consider any sampling allocation $z = (z_x)_{x \in \mathcal{X}}$, by which we measure alternative x for $z_x \in \mathbb{N}$ times. We use Z to represent its corresponding multi-set. We use $Z^\pi(\phi) : \Phi \mapsto (\mathcal{X} \times \mathbb{N})$ to refer to the alternatives selected by π under realization ϕ. Let θ^n be our vector of estimates of the means after n measurements according to allocation Z under realization ϕ, where $|Z| = n$. θ^n can be obtained according to the updating equation (1) and (2), and does not depend on the order of the allocations. It can thus be denoted as $\theta^n(Z, \phi) : (\mathcal{X} \times \mathbb{N}) \times \Phi \mapsto \mathbb{R}^M$. The next lemma states the equivalence of $\mathbb{E}[\mu_{x^\pi}]$ and $\mathbb{E}[\max_x \theta^N_x]$. Hence, the utility function $f : (\mathcal{X} \times \mathbb{N}) \times \Phi \mapsto \mathbb{R}$ can be defined as $\max_x \theta^n_x(Z, \phi)$ and $f_{\mathrm{avg}}(\pi) := \mathbb{E}\Big[\max_x \theta^N_x\big(Z^\pi(\Phi), \Phi\big)\Big]$. The R&S objective (3) can then be re-written as

$$\pi^* \in \arg\max_\pi f_{\mathrm{avg}}(\pi), \text{ subject to } |Z^\pi(\phi)| \leq N.$$

Lemma 1 (Powell and Ryzhov (2012)) *Let π be a policy, and let $x^\pi = \arg\max_x \theta^N_x$ be the alternative selected by the policy. Then*

$$\mathbb{E}[\mu_{x^\pi}] = \mathbb{E}[\max_x \theta^N_x].$$

The definition of the knowledge gradient $\nu_x^{KG,n}$ coincides with the *Conditional Expected Marginal Benefit* $\Delta(e|\psi)$ defined by Golovin and Krause (2010):

$$\Delta(e|\psi) := \mathbb{E}\Big[f\big(\mathrm{dom}(\psi)\cup\{e\},\Phi\big) - f\big(\mathrm{dom}(\psi),\Phi\big)|\Phi\sim\psi\Big].$$

The knowledge gradient policy is thus in fact the adaptive greedy policy with uniform item costs, with a slight difference in the ability of selecting each item more than once. We generalize the definition of adaptive monotonicity and adaptive submodularity for set functions given by Golovin and Krause (2010) to multi-set functions as follows.

Definition 2 (Adaptive Monotonicity) *A function $f : (\mathcal{X}\times\mathbb{N})\times\Phi\mapsto\mathbb{R}$ is adaptive monotone with respect to distribution $p(\phi)$ if the conditional expected marginal benefit of any item is nonnegative: for all ψ and all $x\in\mathcal{X}$.*

$$\Delta(x|\psi)\geq 0.$$

Definition 3 (Adaptive Submodularity) *A function $f : (\mathcal{X}\times\mathbb{N})\times\Phi\mapsto\mathbb{R}$ is adaptive submodular with respect to distribution $p(\phi)$ if for all ψ and ψ' such that dom$(\psi)\subseteq$ dom(ψ') and both ψ,ψ' are consistent with some realization ϕ (i.e. $\psi\subseteq\psi'$), we have the conditional expected marginal benefit of any fixed item $x\in\mathcal{X}$ does not increase as more items are selected and observed,*

$$\Delta(x|\psi)\geq\Delta(x|\psi').$$

Let π^* be the optimal policy to R&S problems. If $f := \max_x\theta_x^n(Z,\phi)$ is adaptive monotone and adaptive submodular with respect to the prior distribution $p(\phi)$, then

$$f_{\mathrm{avg}}(\mathrm{KG}) > (1-e^{-1})f_{\mathrm{avg}}(\pi^*).$$

We next show that the instances generated by ranking and selection problems are adaptive monotone and yet not necessarily submodular.

Lemma 4 *In ranking and selection problems, the utility function $\max_x\theta_x$ is adaptive monotone with any Gaussian prior.*

Proof For any ψ, let $n = |\psi|$. Then for any item $x\in\mathcal{X}$, $\Delta(x|\psi)$ can be rewritten as $\mathbb{E}[\max_{x'}\theta_{x'}^{n+1} - \max_{x'}\theta_{x'}^n|x^n = x,\mathcal{F}^n] = \nu_x^{\mathrm{KG},n}$. Since for any x, $\theta_x^{n+1} = \theta_x^n + \tilde\sigma(\Sigma^n,x^n)Z^{n+1}$, where $\tilde\sigma(\Sigma,x) = \frac{\Sigma e_x}{\sqrt{1/\beta^W+\Sigma_{xx}}}$ and the random variable Z^{n+1} is standard normal when conditioned on $\mathcal{F}^n$ Frazier et al. (2009). Hence we have $\mathbb{E}[\theta_{x'}^{n+1}|x^n = x,\mathcal{F}^n] = \theta_{x'}^n$ for any x'. By Jensen's inequality, we have $\Delta(x|\psi) = \nu_x^{\mathrm{KG},n}\geq 0$. ∎

Even though intuition suggests that the utility function should be adaptive submodular in the amount of information collected, as we collect more information it is natural to expect that the marginal value of this information should decrease, yet it is not always the case. The proof of the next lemma can be found in Appendix A.

Lemma 5 *For any independent normal prior distribution $p(\varphi)$ and non-degenerated noise distribution (i.e. $\sigma^W\neq 0$), there exists ψ, ψ' and $x\in\mathcal{X}$ such that $\psi\subseteq\psi'$ and $\Delta(x|\psi) < \Delta(x|\psi')$.*

It can be seen that the adaptive submodular assumption easily fails in the ranking and selection problems with the special utility function $f = \max_x \theta_x^n(Z,\phi)$ that involves maximization itself. To this end, instead of the above pathwise behavior analyses of the utility function, we would like to study its average behavior by taking the expectation over the observations given any fixed sample allocation Z in the next section.

4.2 The Value of Information

We define the pathwise value of information $\hat{v}(Z,\phi)$ as the incremental improvement over the best expected value that can be obtained without measurement, which is $\max_{x\in\mathcal{X}} \theta_x^0$,

$$\hat{v}(Z,\phi) := \max_{x\in\mathcal{X}} \theta_x^n(Z,\phi) - \max_{x\in\mathcal{X}} \theta_x^0.$$

The value of information $v(Z)$ is then defined to be

$$v(Z) := \mathbb{E}_\Phi[\hat{v}(Z,\Phi)],$$

where the expectation is taken over the prior distribution $p(\phi)$.

The value of information has a long history spanning the literatures of several disciplines. Stigler (1961) considers the value of information in economics when buyers search for the best price. Howard et al. (1966) laid the groundwork for the value of information in a decision-theoretic context and spawned a great deal of work in this area. Yokota and Thompson (2004) give a first comprehensive review of value of information analyses related to health risk management. Raiffa and Schlaifer (1961) poses the Bayesian R&S problem and defines the associated value of information, which marked the beginning of a number of literature on the value of information within Bayesian R&S and the budgeted learning problem (Guttman et al., 1964; Kapoor and Greiner, 2005; Chen et al., 1996; Chick, 2001; Frazier et al., 2008).

Since the value of information is a multi-set function, we first generalize the definitions and properties of submodular set functions described by Nemhauser et al. (1978) to submodular multi-set functions.

Definition 6 *Given a finite set E, a real-valued function g on the set of multi-sets over E is called submodular if for all multi-sets S and T whose elements belong to E,*

$$\rho_x(S) \geq \rho_x(T), \forall S \subseteq T, \forall x \in E,$$

where $\rho_x(S) \triangleq g(S\cup\{x\}) - g(S)$ is the incremental value of adding element x to the multi-set S.

Proposition 1 *Each of the following statements is equivalent and defines a submodular multi-set function (S pathwiseand T are multi-sets on E, x, $y \in E$):*

1. $\rho_x(S) \geq \rho_x(T), \forall S \subseteq T$ *and* $\forall x$.
2. $\rho_x(S) \geq \rho_x(S \cup \{y\}), \forall S, x, y$.
3. $g(T) \leq g(S) + \sum_{x\in T-S} \rho_x(S) - \sum_{x\in S-T} \rho_x(S\cup T - \{x\}), \forall S, T$.

4. $g(T) \le g(S) + \sum_{x \in T-S} \rho_x(S), \forall S \subseteq T$.

This proposition follows from a similar proof of Proposition 2.1 in Nemhauser et al. (1978). For completeness we provide the proof in Appendix B.

It is obvious that if $\theta_x^n(Z, \phi)$ is adaptive monotone or adaptive submodular with respect to $p(\phi)$, then so does $\hat{v}(Z, \phi)$. It is also easy to show that if $\theta_x^n(Z, \phi)$ is adaptive monotone or adaptive submodular with respect to $p(\phi)$, then by the law of total expectation, i.e. $\mathbb{E}[\mathbb{E}[U|V]] = \mathbb{E}[U]$ for any random variables U and V, the value of information $v(Z)$ is monotone or submodular. We close this section by showing the monotonicity of the multi-set function v and leave the analyses of submodularity in Section 5.

Lemma 7 (Monotonicity of the value of information)
For any sampling allocation Z_1 and Z_2, if $Z_1 \subseteq Z_2$, then $v(Z_1) \le v(Z_2)$.

Proof We prove the monotonicity of v by showing $v(Z) \le v(Z \cup \{x^{n+1}\})$ for any allocation Z (with $\sum_{x \in \mathcal{X}} z_x = n$) and any additional measurement x^{n+1}. By the tower property,

$$\begin{aligned}
v(Z \cup \{x^{n+1}\}) - v(Z) &= \mathbb{E}_\Phi[\hat{v}(Z \cup \{x^{n+1}\})] - \mathbb{E}[\hat{v}(Z)] \\
&= \mathbb{E}_\Phi[\max_{x \in \mathcal{X}} \theta_x^{n+1}(Z \cup \{x^{n+1}\})] - \mathbb{E}[\max_{x \in \mathcal{X}} \theta_x^n(Z)] \\
&= \mathbb{E}_\Phi[\mathbb{E}[\max_x \theta_x^{n+1}(Z \cup \{x^{n+1}\}) - \max_x \theta_x^n(Z) | \Phi \sim \psi_Z]] \\
&= \mathbb{E}_\Phi[\nu_x^{\mathrm{KG},n}],
\end{aligned}$$

where ψ_Z is the partial realization with $\mathrm{dom}(\psi_Z) = Z$. The lemma follows from the adaptive monotonicity, $\nu_x^{\mathrm{KG},n} \ge 0$. ■

4.3 Guarantees on the Prior-optimality of the Knowledge Gradient Policy

There are two ways to evaluate the value of a policy. The first, which we call the *posterior view*, conditions on the allocation $Z = Z^\pi(\Phi)$ that would have occurred under policy π for each sample path $\phi \in \Phi$. This is the more conventional approach for evaluating policies. The second, which we call the *prior view*, starts by characterizing the value of an arbitrary allocation Z (before we have seen any sample realizations).

More formally, the classical way to estimate the value of a policy is to calculate the incremental improvement over what we could do before we collect any information, is given by

$$f'_{\mathrm{avg}}(\pi) = \mathbb{E}[f(Z^\pi(\Phi), \Phi)] - \max_x \theta_x^0.$$

We let $\mathbb{P}(\pi \rightsquigarrow Z)$ be the probability that policy π produces allocation Z. Since with a fixed budget of N measurements, there are only finite choices of possible allocations, using the tower property, we can condition on the allocation $Z^\pi = Z$ which gives us

$$f'_{\mathrm{avg}}(\pi) = \sum_{Z \in \mathcal{Z}^N} \mathbb{P}(\pi \rightsquigarrow Z) \Big(\mathbb{E}[\max_x \theta_x^n(Z^\pi(\Phi), \Phi) | Z^\pi = Z] - \max_x \theta_x^0 \Big).$$

We note that in this method for evaluating a policy (which is the standard method), we only consider allocations Z that are actually produced by policy π for the outcomes in ϕ. This approach makes it much more difficult to understand the relationship between the allocation Z and the value of a policy.

For this reason, we adopt a different method of evaluating a policy which we term the *prior view.* Since this idea is new, we define it formally as follows

Definition 8 (The prior-value of a policy) *Let $\mathcal{Z}^n$ be the set of all possible allocations with a limited budget n. The value of a policy π with N measurements is defined as*

$$\begin{aligned} F^\pi &= \sum_{Z \in \mathcal{Z}^N} \mathbb{P}(\pi \rightsquigarrow Z)\left(\mathbb{E}_\Phi[\max_x \theta_x^n(Z,\Phi)] - \max_x \theta_x^0\right) \\ &= \sum_{Z \in \mathcal{Z}^N} \mathbb{P}(\pi \rightsquigarrow Z) v(Z). \end{aligned}$$

In this view, we use the prior probability of an outcome $p(\phi)$ instead of the posterior $p(\phi|Z^\pi(\phi) = Z)$ which is conditioned on an allocation Z. The value of this approach is that it writes the value of a policy directly as a function of $v(Z)$, making it easier to study the effect of the properties of $v(Z)$ on the value of a policy. Intuitively, since a policy could generate different allocations Z for different sample realizations, it is natural to define the value of a policy π as the weighted sum of the expected value of information based on all possible allocations Z and the weight should be the probability of occurrence of Z based on policy π.

We make the following assumption which is weaker than the adaptive submodularity assumption and will analyze it further in Section 5.

Assumption 1 *The value of information v is a submodular multi-set function on the set of alternatives $\mathcal{X}$ with respect to the prior distribution $p(\phi)$.*

Let π^* be the optimal sequential policy under a budget of N measurements in the sense that the prior-value of π^* is the largest. We call it *prior-optimality.* In what follows, we first bound KG's sub-*prior-optimality* in Proposition 12:

$$F^{\pi^*} \leq F^{\mathrm{KG}^{[n]}@\pi^*} \leq F^{\mathrm{KG}^{[n-1]}} + N(F^{\mathrm{KG}^{[n]}} - F^{\mathrm{KG}^{[n-1]}}),\ n = 1, 2, ..., N.$$

Then we derive the worst-case bound for the knowledge gradient policy in Theorem 14:

$$\frac{F^{KG}}{F^{\pi^*}} \geq 1 - (\frac{N-1}{N})^N \geq \frac{e-1}{e} \approx 0.632.$$

Besides the *posterior optimality* bound obtained from adaptive stochastic multi-set maximization, the *prior-optimality* provides another insight into the performance of the KG policy based on a well-understood quantity: value of information.

Definition 9 (Policy concatenation) *(Golovin and Krause, 2010) A concatenated policy $\pi = \pi_1 @ \pi_2$ is constructed by running π_1 to completion, and then running policy π_2 from a fresh start ignoring all the information collected while running π_1.*

To be more specific, suppose π_i has a budget of n_i, $i = 1, 2$, the first phase is to run π_1 for n_1 iterations starting from S^0 and we get a sample realization including decisions and their corresponding measurements. The second phase is to run π_2 for n_2 measurements starting from S^0 and we get another sample realization. Thus the sample realization of the concatenated process is all the decisions and their corresponding measurements collected in two phases. Note here, when running the second policy, we ignore all the information collected during running the first one, but when calculating the value of $\pi_1@\pi_2$, $F^{\pi_1@\pi_2}$, we use all the information collected in two phases.

Definition 10 (Policy truncation) *(Golovin and Krause, 2010) For a policy π, define the j-truncation $\pi^{[j]}$ of π as the policy that runs exactly $(j+1)$ steps under π's decision rule and $\pi^{\{j\}}$ as the single step policy that randomly chooses an alternative according to the probability distribution of policy π's decision for the $(j+1)$-th step.*

We now show that the value of π_1 is no larger than the value of $\pi_1@\pi_2$.

Lemma 11 *$F^{\pi_1} \le F^{\pi_2@\pi_1}$ for all policies π_1 and π_2 under any prior and probability distribution that describes a measurement.*

Proof We first show that $F^{\pi_1@\pi_2} = F^{\pi_2@\pi_1}$. In a concatenated policy, the two phases are independent since no information is shared among the two phases. Hence for a given allocation pair (Z_1, Z_2) where $Z_1 \in \mathcal{Z}^{n_1}$, $Z_2 \in \mathcal{Z}^{n_2}$, we have

$$\begin{aligned}
\mathbb{P}(\pi_1@\pi_2 \rightsquigarrow (Z_1, Z_2)) &= \mathbb{P}(\pi_1 \rightsquigarrow Z_1)\mathbb{P}(\pi_2 \rightsquigarrow Z_2) \\
&= \mathbb{P}(\pi_2 \rightsquigarrow Z_2)\mathbb{P}(\pi_1 \rightsquigarrow Z_1) \\
&= \mathbb{P}(\pi_2@\pi_1 \rightsquigarrow (Z_2, Z_1)).
\end{aligned}$$

$F^{\pi_1@\pi_2} = F^{\pi_2@\pi_1}$ follows immediately from taking the sum over all possible pairs of $(Z_1, Z_2))$ such that $Z_2 \cup Z_1 = Z$ for any fixed allocation Z.

Therefore $F^{\pi_1} \le F^{\pi_1@\pi_2}$ holds if and only if $F^{\pi_1} \le F^{\pi_2@\pi_1}$. We then finish this proof by showing $F^{\pi_1} \le F^{\pi_1@\pi_2}$. We write $F^{\pi_1@\pi_2} - F^{\pi_1}$ as a telescoping sequence

$$\begin{aligned}
& F^{\pi_1@\pi_2} - F^{\pi_1} \\
= & \sum_{Z \in \mathcal{Z}^{n_1+n_2}} v(Z)\mathbb{P}(\pi_1@\pi_2 \rightsquigarrow Z) - \sum_{Z_1 \in \mathcal{Z}^{n_1}} v(Z_1)\mathbb{P}(\pi_1 \rightsquigarrow Z_1) \\
= & \sum_{Z \in \mathcal{Z}^{n_1+n_2}} \sum_{Z_1 \cup Z_2 = Z} v(Z)\mathbb{P}(\pi_1 \rightsquigarrow Z_1)\mathbb{P}(\pi_2 \rightsquigarrow Z_2) \\
- & \sum_{Z_1 \in \mathcal{Z}^{n_1}} \sum_{Z_2 \in \mathcal{Z}^{n_2}} v(Z_1)\mathbb{P}(\pi_1 \rightsquigarrow Z_1)\mathbb{P}(\pi_2 \rightsquigarrow Z_2) \\
= & \sum_{Z_1 \in \mathcal{Z}^{n_1}} \sum_{Z_2 \in \mathcal{Z}^{n_2}} \Big[v(Z_1 \cup Z_2) - v(Z_1)\Big]\mathbb{P}(\pi_1 \rightsquigarrow Z_1)\mathbb{P}(\pi_2 \rightsquigarrow Z_2) \\
\ge & \ 0,
\end{aligned}$$

where the second equality holds due to the same reason as in the proof above for $F^{\pi_1@\pi_2} = F^{\pi_2@\pi_1}$ and the third equality is just the same summation in different orders. The last inequality holds because of the monotonicity of multi-set function v. ∎

Based on the monotonicity of v and a similar argument as in Proposition 11, F is non-decreasing with respect to the number of measurements. Thus the more measurements, the better the policy. Hence π^* has exactly N measurements. We have the following sub-optimality bound on KG's prior-value. For a proof see Appendix C.

Proposition 12 *Let $\rho^{KG,n} = F^{KG^{[n]}} - F^{KG^{[n-1]}}$, then*

$$\begin{aligned} F^{\pi^*} \leq F^{\ KG^{[n-1]}@\pi^*} &\leq F^{\ KG^{[n-1]}} + N\rho^{KG,n} \\ &= \sum_{i=0}^{n-1} \rho^{KG,i} + N\rho^{KG,n},\ n = 0, 1, ..., N-1. \end{aligned} \tag{8}$$

We now derive a bound for the adaptive greedy policy by applying linear programming to the problem of minimizing $\frac{F^{\mathrm{KG}}}{F^{\pi^*}}$ subject to the inequalities (8), which is a worst-case analysis. The following lemma states the linear program and its solution. We use it afterwards to establish the bounds.

Lemma 13 *Given $N \in \mathbb{Z}_+$, consider the following linear program*

$$\min \sum_{i=0}^{N-1} a_i,$$

$$\sum_{i=0}^{t-1} a_i + N a_t \geq 1,\ t = 0, 1, ..., N-1.$$

Then under these N constraints, min $\sum_{i=0}^{N-1} a_i = 1 - \alpha^N$, where $\alpha = \frac{N-1}{N}$.

The proof of this lemma can be found in Nemhauser et al. (1978).

We have the following results, which generalizes the classic result of the greedy algorithm that achieves $(1-1/e)$-approximation to prior-optimality for ranking and selection problems.

Theorem 14 *Assume we have a budget of N measurements. Let π^* denote the optimal sequential policy for the ranking and selection problem, then we have*

$$\frac{F^{KG}}{F^{\pi^*}} \geq 1 - (\frac{N-1}{N})^N.$$

Proof By Proposition 12, we have $F^{\pi^*} \leq \sum_{i=0}^{n-1} \rho^{\mathrm{KG},i} + N\rho^{\mathrm{KG},n},\ n = 0, 1, ..., N-1$. Divide by F^{π^*} on both sides of this inequality, we have

$$1 \leq \sum_{i=0}^{n-1} \frac{\rho^{\mathrm{KG},i}}{F^{\pi^*}} + N\frac{\rho^{\mathrm{KG},n}}{F^{\pi^*}}, n = 0, 1, ..., N-1.$$

Let $a_i = \frac{\rho^{\mathrm{KG},i}}{F^{\pi^*}}$, and then these inequalities are identical to the constraints in Lemma 13. We notice that

$$\min \sum_{i=0}^{N-1} a_i = \min \sum_{i=0}^{N-1} \frac{\rho^{\mathrm{KG},i}}{F^{\pi^*}} \leq \sum_{i=0}^{N-1} \frac{\rho^{\mathrm{KG},i}}{F^{\pi^*}} = \frac{F^{\mathrm{KG}}}{F^{\pi^*}}.$$

By Lemma 13, we have $\min \sum_{i=0}^{N-1} a_i = 1 - \alpha^N$, so $\frac{F^{\mathrm{KG}}}{F^{\pi^*}} \geq 1 - \alpha^N = 1 - (\frac{N-1}{N})^N$. ■

5. Analysis of Submodularity of the Value of Information

The finite-time bounds obtained in the previous sections assume that the value of information is submodular. In general, submodularity does not hold for arbitrary value functions. In this section, we analyze the submodularity of the two-alternative case for independent beliefs.

While submodularity is a property for multi-set functions, we can extend it to any continuous function by making it possible for the increment to take any positive value. It could be easily extended to any continuous function. This allows us to use results from real analysis to study submodularity.

Definition 15 *A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is submodular if for all $x, y \in \mathbb{R}^n$, $x_i \leq y_i$ and $\delta \in \mathbb{R}^n_+$,*

$$f(x+\delta) - f(x) \geq f(y+\delta) - f(y).$$

We show that submodularity of $\mathcal{C}^2$ functions is directly related to its second derivatives and cross-derivatives (the proof is given in Appendix D):

Theorem 16 *$\mathcal{C}^2$ function f: $\mathbb{R}^n \to \mathbb{R}$ is submodular if and only if every element of its Hessian is non-positive.*

The concavity of the value of information has been studied extensively by Frazier and Powell (2010). In this section, we only study the cross-derivatives of the value of information.

Let $M = 2$ and the measurement allocation $z = (z_1, z_2)$. The value of information $v(z) = s(z) f(-\frac{|\theta_1^0 - \theta_2^0|}{s(z)})$, where $s(z) = \sqrt{\tilde{\sigma}_1^2(z_1) + \tilde{\sigma}_2^2(z_2)}$, $\tilde{\sigma}_i^2(z_i) = \frac{\sigma_i^{2,0} z_i}{\sigma_W^2/\sigma_i^{2,0} + z_i}$, $f(a) = a\Phi(a) + \phi(a)$, Φ and ϕ are the standard normal cumulative distribution and density respectively (Frazier and Powell, 2010).

Although the value of information is not concave in general in the two-alternative case, v is concave on the region where all z_i's are large enough (see Frazier and Powell, 2010, Theorem 2).

We directly calculate the first derivative and cross-derivative of v as

$$\begin{aligned}
\frac{\partial v}{\partial z_1} &= \frac{\tilde{\sigma}_1(z_1)\tilde{\sigma}_1'(z_1)}{s(z)}\left[f(-\frac{|\theta_1^0 - \theta_2^0|}{s(z)}) + |\theta_1^0 - \theta_2^0| \frac{\Phi(-\frac{|\theta_1^0 - \theta_2^0|}{s(z)})}{s(z)}\right], \\
\frac{\partial^2 v}{\partial z_1 \partial z_2} &= \frac{\tilde{\sigma}_1(z_1)\tilde{\sigma}_1'(z_1)\tilde{\sigma}_2(z_2)\tilde{\sigma}_2'(z_2)}{s^3(z)} \phi(-\frac{|\theta_1^0 - \theta_2^0|}{s(z)}) \left(\frac{|\theta_1^0 - \theta_2^0|^2}{\tilde{\sigma}_1^2(z_1) + \tilde{\sigma}_2^2(z_2)} - 1 \right).
\end{aligned}$$

Theorem 17 *The value of information is submodular when $M = 2$ and $\theta_1^0 = \theta_2^0$.*

Proof Concavity of $v(z)$ is proven in Remark 2 by Frazier and Powell (2010). Since $\theta_1^0 = \theta_2^0$, $|\theta_1^0 - \theta_2^0| = 0$ and thus $\frac{\partial^2 v}{\partial z_1 \partial z_2} \leq 0$. Therefore, v is submodular in this case. ■

$\frac{\partial^2 v}{\partial z_1 \partial z_2} \leq 0$ is equivalent to $|\theta_1^0 - \theta_2^0|^2 \leq \tilde{\sigma}_1^2(z_1) + \tilde{\sigma}_2^2(z_2)$. Rewriting this inequality, we get

$$\frac{1}{\frac{1}{\sigma_1^{2,0}} + \frac{z_1}{\sigma_W^2}} + \frac{1}{\frac{1}{\sigma_2^{2,0}} + \frac{z_2}{\sigma_W^2}} \leq \sigma_1^{2,0} + \sigma_2^{2,0} - |\theta_1^0 - \theta_2^0|^2. \tag{9}$$

We need $\sigma_1^{2,0} + \sigma_2^{2,0} - |\theta_1^0 - \theta_2^0|^2 \geq 0$, which can be achieved by setting our prior variance large enough or using a uniform prior over all alternatives. This is very reasonable when we have very little information about our problem domain.

Inequality equation (9) defines a region in the $z_1 - z_2$ plane. Specifically, this region has the hyperbolic line $\frac{1}{\frac{1}{\sigma_1^{2,0}} + \frac{z_1}{\sigma_W^2}} + \frac{1}{\frac{1}{\sigma_2^{2,0}} + \frac{z_2}{\sigma_W^2}} = \sigma_1^{2,0} + \sigma_2^{2,0} - |\theta_1^0 - \theta_2^0|^2$ as its boundary and contains infinity. In particular, when z_1 and z_2 are large enough (or equivalently when our measurement is accurate enough), the value of information is submodular.

Since there is no closed-form expression for the value of information under arbitrary allocations, we cannot verify submodularity in a simple way for problems with more than two alternatives and for correlated beliefs. Instead, it can be checked using numerical approximation and is easy to guarantee by running repeated experiments and averaging to reduce measurement noise. A necessary condition is the concavity of the value of information for measuring a fixed alternative x for n times, which can be checked exactly.

Intuitively, we may expect that the marginal value of information should decline as we make more observations. But it is not always the case. It is shown that the value of information for measuring a single alternative may form an S-curve which is concave when there are many measurements, but may be convex at the beginning (Frazier and Powell, 2010). The S-curve behavior arises when the measurement noise is large and thus a single measurement simply contains too little information, leading to algorithmic difficulties and apparent paradoxes. This issue is not related to any specific policy, but rather is an inherent property of learning problems. Although the value of information is not necessarily concave, it can be made concave by measuring each alternative enough times or (equivalently) using sufficiently precise measurements.

6. Modular Optimal Learning Testing Environment (MOLTE)

Since the seminal paper by Lai and Robbins (1985), there has been a long history in the optimal learning literature of proving some sort of bound, supported at times by relatively thin empirical work by comparing a few policies on a small number of randomly generated problems (Audibert et al., 2010; Cappé et al., 2013; Srinivas et al., 2009; Auer et al., 2002; Garivier and Moulines; Audibert et al., 2009). The problem, of course, is that compiling a library of test problems, and then running an extensive set of comparisons, is difficult. The problem is this means that we are analyzing the finite time performance of algorithms using bounds that only apply asymptotically by limited empirical experiments to support the claim of finite time performance.

In this section, we describe a modular optimal learning testing environment (MOLTE)[1] that will make it much easier for researchers to test new policies against a library of test problems, and a library of previously coded policies. The Matlab-based modular architecture, where policies and problems are captured in a set of .m files, makes it easy for researchers to add new policies and new problems.

1. The software is available at `http://www.castlelab.princeton.edu/software.htm`.

6.1 Structural overview

We pre-coded many standard truth functions (problem classes), including standard optimization test functions with additive Gaussian noise (for example, Branin's function in Dixon and Szegö (1978), Goldstein Price function, Rosenbrock function, Griewank function in Hu et al. (2008), Six-hump camel back function in Molga and Smutnicki (2005), etc.), synthetic bandit experiments (Audibert et al., 2010), Gaussian process regression and real-world applications like newsvendor problems and payload delivery. We also pre-coded a number of competing policies, including UCB variants, successive rejects, sequential Kriging (SKO, as a representative of Bayesian global optimization (Jones et al., 1998; Huang et al., 2006; Gutmann, 2001; Jones, 2001)), Thompson sampling, KG variants, etc. Each of the problem classes and policies is organized in its own Matlab file, so that it is easy for a user to add in a new problem or a policy. In order to make a fair comparison, all the observations are pre-generated and shared between competing policies. There may be problems where a domain expert can provide prior knowledge, such as kinetic models in materials science, but we may need to estimate them from data in some cases (optimization test functions). In MOLTE we provide various ways to construct a prior, including user-provided prior distributions, hard-coded default prior distributions, an uninformative prior and MLE estimation (see Section 6.1.3).

The input to the simulator is a spreadsheet which allows users to specify the problem classes and competing policies, as well as the belief models, the objectives, the prior construction and the measurement budgets. We provide a sample input spreadsheet in Table 1. For policies that have tunable parameters, a star included in the parentheses after the policy will initiate an automatic brute force tuning procedure with the optimal value reported. Whereas the user can also specify the value to be used for the policy in the parentheses. All the numerical results presented below are obtained using this environment.

Table 1: Sample input spreadsheet.

Problem class	**Prior**	**Measurement Budget**	**Belief Model**	**Offline/ Online**	**Number of policies**				
Bubeck1	Uninform	10	independent	Online	3	OLKG	IE(*)	UCB	
Branin	MLE	5	independent	Offline	4	UCBE(*)	IE(1.7)	KG	SR
GPR	Default	0.3	correlated	Online	4	KLUCB	EXPL	UCB	TS
NanoDesign	MLE	0.5	correlated	Offline	3	Kriging	EXPT	KG	

While a wide range of problem classes and policies are precoded in MOLTE, in the next two subsections we only briefly summarize the problem classes and policies mentioned in the following numerical experiments of this paper.

Similar libraries have been proposed for Bayesian optimization in different programming languages with different metrics and visualizations, for example, BayesOpt (Martinez-Cantin, 2014) and Spearmint (Snoek et al., 2012). Thanks to its unique modular design, MOLTE allows users to easily specify their own problems or their own algorithms without limitation as long as they follow the general function interface. With various (graphical) metrics, our hope is that MOLTE can be used to facilitate the process of more comprehensive comparisons, on a broader set of test problems and a broader set of policies, so that researchers can more easily draw insights into the behavior of different policies in the context of different problem classes.

6.1.1 Problem Classes

Bubeck's Experiments: (Audibert et al., 2010) We consider Bernoulli distributions with the mean of the best arm always $\mu_1 = 0.5$. M is the number of arms.
Bubeck1: $M = 20$, $\mu_{2:20} = 0.4$.
Bubeck2: $M = 20$, $\mu_{2:6} = 0.42$, $\mu_{7:20} = 0.38$.
Bubeck3: $M = 4$, $\mu_i = 0.5 - (0.37)^i$, $i \in \{2, 3, 4\}$.
Bubeck4: $M = 6$, $\mu_2 = 0.42$, $\mu_{3:4} = 0.4$, $\mu_{5:6} = 0.35$.
Bubeck5: $M = 15$, $\mu_i = 0.5 - 0.025i, i \in \{2, \cdots, 15\}$.
Bubeck6: $M = 20$, $\mu_2 = 0.48$, $\mu_{3:20} = 0.37$.
Bubeck7: $M = 30$, $\mu_{2:6} = 0.45$, $\mu_{7:20} = 0.43$, $\mu_{21:30} = 0.38$.

Asymmetric unimodular function (AUF): x is a controllable parameter ranging from 21 to 120. The objective function is $F(x, \xi) = \theta_1 \min(x, \xi) - \theta_2 x$, where θ_1, θ_2 and the distribution of the random variable ξ are all unknown. ξ is taken as a normal distribution with mean 60. Three noise levels are considered by setting different noise ratios between the standard deviation and the mean of ξ: HNoise–0.5, MNoise–0.4, LNoise–0.3. Unless explicitly pointed out, experiments are taken under LNoise.

Equal-prior: $M = 100$. The true values μ_x are uniformly distributed over $[0, 60]$ and measurement noise $\sigma_W = 100$. $\theta_x^0 = 30$ and $\sigma_x^0 = 10$ for every x.

All the standard optimization test functions are flipped in MOLTE to generate maximization problems instead of minimization in line with R&S and bandit problems. The standard deviation of the additive Guassion noise is set to 20 percent of the range of the function values.

Rosenbrock functions with additive noise:

$$f(x, y, \phi) = 100(y - x^2)^2 + (1 - x)^2 + \phi,$$

where $-3 \le x \le 3$, $-3 \le y \le 3$. x and y are uniformly discretized into 13×13 alternatives.

Pinter's function with additive noise:

$$\begin{aligned} f(x, y, \phi) = {} & \log_{10}\left(1 + (y^2 - 2x + 3y - \cos x + 1)^2\right) + \log_{10}\left(1 + 2(x^2 - 2y + 3x - \cos y + 1)^2\right) \\ & + x^2 + 2y^2 + 20\sin^2(y \sin x - x + \sin y) + 40 \sin^2(x \sin y - y + \sin x) + 1 + \phi, \end{aligned}$$

where $-3 \le x \le 3$, $-3 \le y \le 3$. x and y are uniformly discretized into 13×13 alternatives.

Goldstein-Price's function with additive noise:

$$\begin{aligned} f(x, y, \phi) &= [1 + (x + y + 1)^2(19 - 14x + 3x^2 - 14y + 6xy + 3y^2)] \cdot \\ & \quad [30 + (2x - 3y)^2(18 - 32x + 12x^2 + 48y - 36xy + 27y^2)] + \phi, \end{aligned}$$

where $-3 \le x \le 3$, $-3 \le y \le 3$. x and y are uniformly discretized into 13×13 alternatives.

Branins's function with additive noise:

$$f(x, y, \phi) = (y - \frac{5.1}{4\pi^2}x^2 + \frac{5}{\pi}x - 6)^2 + 10(1 - \frac{1}{8\pi})\cos(x) + 10 + \phi,$$

where $-5 \leq x \leq 10$, $0 \leq y \leq 15$. x and y are uniformly discretized into 15×15 alternatives.

Ackley's function with additive noise:

$$f(x, y, \phi) = -20 \exp\Big(-0.2 \cdot \sqrt{\frac{1}{2}(x^2 + y^2)}\Big) - \exp\Big(\frac{1}{2}(\cos(2\pi x) + \cos(2\pi y))\Big) + 20 + \exp(1) + \phi,$$

where $-3 \leq x \leq 3$, $-3 \leq y \leq 3$. x and y are uniformly discretized into 13×13 alternatives.

Hyper Ellipsoid function with additive noise:

$$f(x, y, \phi) = x^2 + 2y^2 + \phi.$$

where $-3 \leq x \leq 3$, $-3 \leq y \leq 3$. x and y are uniformly discretized into 13×13 alternatives.

Rastrigin function with additive noise:

$$f(x, y, \phi) = 20 + \big[x^2 - 10\cos(2\pi x)\big] + \big[x^2 - 10\cos(2\pi y)\big] + \phi,$$

where $-3 \leq x \leq 3$, $-3 \leq y \leq 3$. x and y are uniformly discretized into 11×11 alternatives.

Six-hump camel back function with additive noise:

$$f(x, y, \phi) = (4 - 2.1x^2 + \frac{x^4}{3})x^2 + xy + (-4 + 4y^2)y^2 + \phi,$$

where $-2 \leq x \leq 2$, $-1 \leq y \leq 1$. x and y are uniformly discretized into 13×13 alternatives.

6.1.2 Policies Considered

In addition to the KG policies defined in (6) and (7), we shall consider the following policies π, which differ according to their decision $X^{\pi,n}(S^n)$ of the alternative to measure at time n given state S^n.

Interval Estimation (IE): (Kaelbling, 1993)

$$X^{\mathrm{IE},n}(S^n) = \arg\max_x \theta_x^n + z_{\alpha/2}\sigma_x^n,$$

where $z_{\alpha/2}$ is a tunable parameter.

Kriging: Huang et al. (2006)

Let $x^* = \arg\max_x(\theta_x^n + \sigma_x^n)$, then

$$X^{\mathrm{Kriging},n}(S^n) = \arg\max_x(\theta_x^n - \theta_{x^*}^n)\Phi(\frac{\theta_x^n - \theta_{x^*}^n}{\sigma_x^n}) + \sigma_x^n\phi(\frac{\theta_x^n - \theta_{x^*}^n}{\sigma_x^n}),$$

where ϕ and Φ are the standard normal density and cumulative distribution functions.

Thompson sampling (TS): (Thompson, 1933)

$$X^{\mathrm{TS},n}(S^n) = \arg\max_x \hat{\theta}_x^n,$$

where $\hat{\theta}_x^n \sim \mathcal{N}(\theta_x^n, \sigma_x^n)$ for independent beliefs or $\hat{\theta}_x^n \sim \mathcal{N}(\theta^n, \Sigma^n)$ for correlated beliefs.

UCB: (Auer et al., 2002)

$$X^{\mathrm{UCB},n}(S^n) = \arg\max_x \hat{\mu}_x^n + \sqrt{\frac{2V_x^n \log n}{N_x^n}},$$

where $\hat{\mu}_x^n$, V_x^n, N_x^n are the sample mean of μ_x, sample variance of μ_x, and number of times x has been sampled up to time n, respectively. The quantity $\hat{\mu}_x^0$ is initialized by measuring each alternative once. These are similarly defined in the following variants of UCB.

UCB-E: (Audibert et al., 2010)

$$X^{\text{UCB-E},n}(S^n) = \arg\max_x \hat{\mu}_x^n + \sqrt{\frac{\alpha}{N_x^n}},$$

where α is a tunable parameter.

UCB-V: (Audibert et al., 2009)

$$X^{\text{UCB-V},n}(S^n) = \arg\max_x \hat{\mu}_x^n + \sqrt{\frac{V_x^n \log n}{N_x^n}} + 1.5\frac{\log n}{N_x^n}.$$

SR: (Audibert et al., 2010) Let $A_1 = \mathcal{X}$, $\overline{\log}(M) = \frac{1}{2} + \sum_{i=2}^{M} \frac{1}{i}$,

$$n_m = \left\lceil \frac{1}{\overline{\log}(M)} \frac{n-M}{M+1-m} \right\rceil.$$

For each phase $m = 1, ..., M-1$:

1. For each $x \in A_m$, select alternative x for $n_m - n_{m-1}$ rounds.
2. Let $A_{m+1} = A_m \setminus \arg\min_{x \in A_m} \hat{\mu}_x$.

KLUCB: (Cappé et al., 2013)

$$X^{\text{KLUCB},n}(S^n) = \arg\max_x \hat{\mu}_x^n + \sqrt{\frac{2V_x^n(\log n + 3\log\log(n))}{N_x^n}}.$$

EXPL: A pure exploration strategy that tests each alternative equally often.

EXPT: A pure exploitation strategy.

$$X^{\text{EXPT},n}(S^n) = \arg\max_x \hat{\mu}_x^n.$$

6.1.3 Prior Generation

If an uninformative prior is specified by the user for independent beliefs, a uniform prior will be used with $\theta_x^0 = 0$ and $\sigma_x^0 = inf$ for every x. In such case, same as with frequentist approaches (for example, UCBs), Bayesian approaches will measure each alternative once at the very beginning.

If maximum likelihood estimation (MLE) is chosen to obtain the prior distribution for either independent beliefs or correlated beliefs, we follow Jones et al. (1998) and Huang et al. (2006) to use Latin hypercube designs for initial fit. For independent beliefs, we adopt a uniform prior with the same mean value θ_x^0 and standard deviation σ_x^0 for all alternatives. For correlated beliefs, we use a constant mean value θ_x^0 for all alternatives and a prior covariance matrix of the form

$$\Sigma_{xx'}^0 = \sigma e^{-\sum_{i=1}^{d} \lambda_i (x_i - x_i')^2},$$

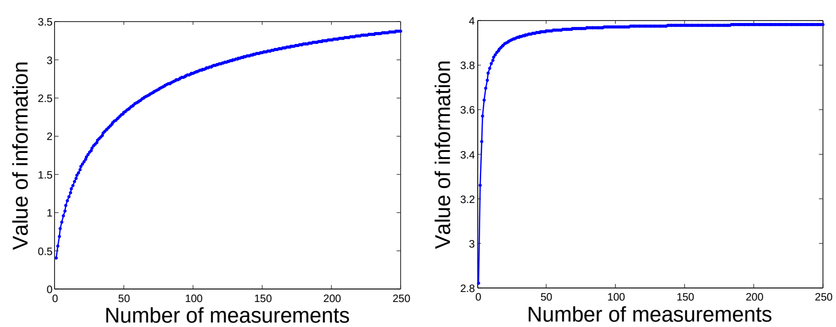
where each arm x is a d-dimensional vector and σ, λ_i are constant. We adopt the rule of thumb by Jones et al. (1998) for the default number $(10 \times p)$ of points, where p is the number of parameters to be estimated. In addition, as suggested by Huang et al. (2006), to estimate the random errors, after the first $10 \times p$ points are evaluated, we add one replicate at each of the locations where the best p responses are found. Maximum likelihood estimation is then used to estimate the parameters based on the points in the initial design.

7. Numerical Experiments for Offline Ranking and Selection Problems

In this section we report on a series of experiments with the goal of illustrating the use of MOLTE and the types of reports that it produces. We do not attempt to demonstrate that any policy is better than another, but our experiments support the hypothesis that different policies work well on different problem classes. This observation supports the claim more careful empirical work is needed to develop a better understanding of which policies work best, and under what conditions.

7.1 Illustration of the submodularity assumption

We consider two offline learning settings, Equal-prior and AUF, to illustrate the necessary condition of our submodularity assumption: the concavity of the value of information for measuring a fixed alternative x for n times. Assuming we are measuring a single alternative x for n_x times and $n_{x'} = 0$ for $x' \neq x$. Figure 1 shows the value of n_x measurements as n_x ranges from 1 to 250 for equal-prior problems and the AUF problem with $\theta_2 = 0.2\theta_1$. The plots for AUF problem with $\theta_2 = 0.5\theta_1$ and $\theta_2 = 0.8\theta_1$ are similar to the one with $\theta_2 = 0.2\theta_1$. These plots show that the value of information is concave.



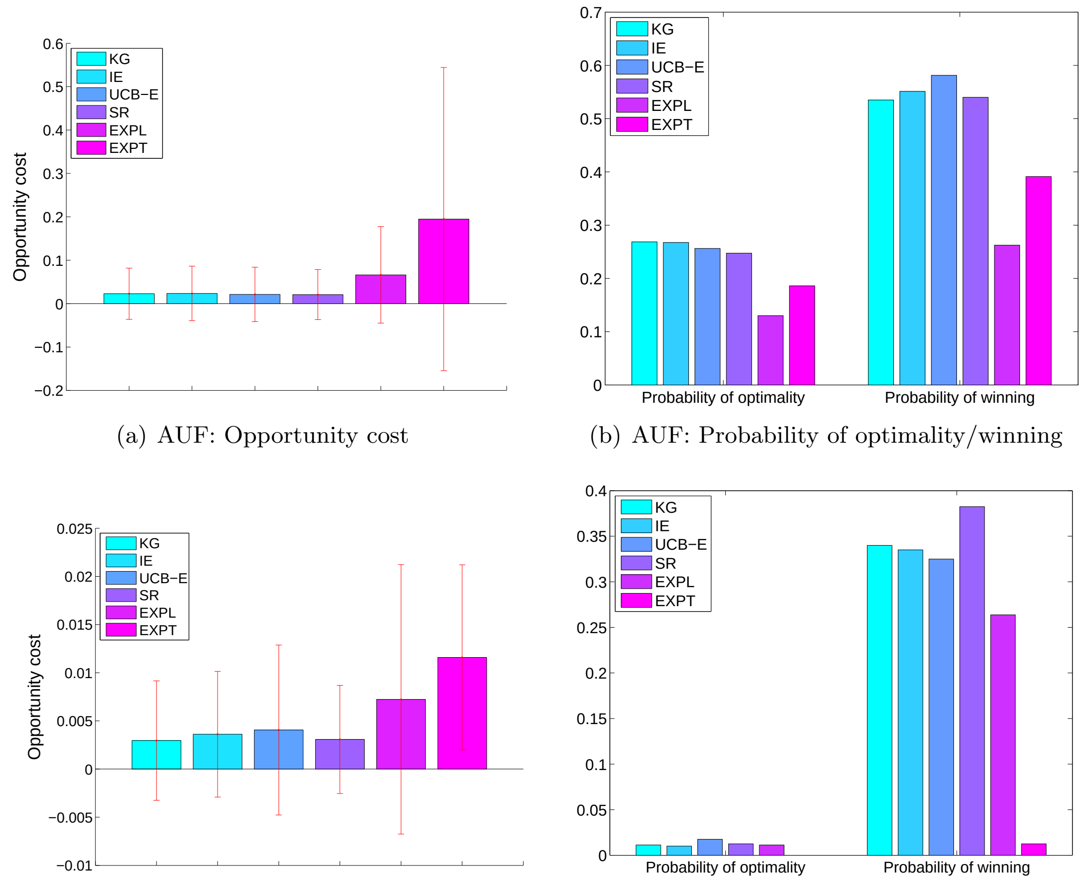
(a) Equal-prior

(b) AUF with $\theta_2 = 0.2\theta_1$

Figure 1: Value of measuring a single alternative.

7.2 Experiments with independent beliefs

We first compare the performance of KG, IE with tuning, UCB-E with tuning, SR, EXPL and EXPT for offline ranking and selection problems. MLE is used to construct the prior distribution for KG and IE. Figure 2 shows the performance in problem classes AUF and



(a) AUF: Opportunity cost

(b) AUF: Probability of optimality/winning

(c) Goldstein: Opportunity cost

(d) Goldstein:Probability of optimality/winning

Figure 2: Comparisons for AUF and Goldstein. (a) and (c) depict the mean opportunity cost with error bars indicating the standard deviation of each policy. The first bar group in (b) and (d) demonstrates the probability that the final recommendation of each policy is the optimal one. The second bar group in (b) and (d) illustrates the probability that the opportunity cost of each policy is the lowest.

Goldstein with independent beliefs under a measurement budget five times the number of alternatives.

We run each policy for 1000 times. In each run, we pre-generate all the observations and share across different policies in order to make a fair comparison. We illustrate in the first column of Figure 2 the mean opportunity cost and the standard deviation of each policy over 1000 runs, with the opportunity cost (OC^π) defined as:

$$\mathrm{OC}^\pi = \max_x \mu_x - \mu_{x^N},$$

where $x^N = \arg\max_x \theta_x^N$.

In order to give a more comprehensive comparison of different policies, we also calculate the probability that the final recommendation of each policy is the optimal one and the probability that the opportunity cost of each policy is the lowest, as illustrated in the figures on the right hand side of Figure 2.

The three criteria characterize the behavior of policies in different aspects. For example, under AUF, if one cares about the average performance of the policy and its stability, SR is the best choice concluding from Figure 2 (a). Yet, if one can only run one trial (as in most cases of experimental science) and want to identify the best alternative, KG might be a better choice since it has the highest probability of finding the optimal alternative. Or if one can live with fairly good alternatives other than the optimal one, UCB-E could be the choice (although it has to be carefully tuned).

One observation is that there is no universal best policy for all problem classes or under all criteria. In practice, a useful guidance could be abstracting the real world problem and running synthetic simulations to find the best simulated policy under some desired criterion before conducting the real experiments.

7.3 Experiments with correlated beliefs

In this section, we exploit correlated beliefs between alternatives in order to strengthen the effect of each measurement so that one measurement of some alternative can provide information for other alternatives.

First, we present the performance of different policies as time goes by under AUF ($\theta_2 = 0.5\theta_1$) in Figure 3. We run each policy on 1000 different sample paths and compute the mean OC obtained after each measurement. We tune z_α for IE and α for UCB for $N = 400$ measurements and the optimal values are $z_\alpha = 0.969$ and $\alpha = 6.657$. Since UCB-E needs to measure each alternative once, we omit the OC for its first 100 (which is the number of alternatives) steps. KG uses independent beliefs while KGCB, IE and Kriging start from MLE fitted correlated beliefs. When incorporating correlated beliefs, a measurement of one alternative tells us something about other alternatives. As a result, KGCB learns faster than KG. We draw the conclusion that correlated beliefs make learning faster and make learning possible for the case where the measurement budget is smaller (and potentially much smaller) than the number of alternatives.

In order to better understand the behavior of each policy, a useful way is to examine the sampling pattern of each policy. We present an example of the frequency of measuring each alternative for each competing policy for Branin functions with a measurement budget of 100. To take advantage of correlated beliefs, rather than measuring each alternative

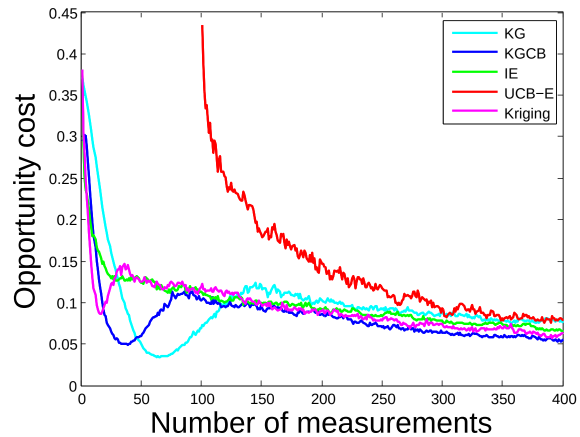


Figure 3: OC obtained by each policy after each measurement under AUF ($\theta_2 = 0.5\theta_1$).

once to initialize the empirical mean, we use the prior mean as the starting point and use the posterior mean θ^n in place of the empirical mean $\hat{\mu}^n$ for UCB-E. In the left column of Figure 4, the sampling pattern of each policy is displayed together with the contour of the Branin objective function which exhibits one global maximum at $(-3, 12)$ and other two local maxima at $(9, 3)$ and $(16, 4)$. The frequency that each alternative is measured is marked in numbers. The right column depicts the final prediction under each policy. All the observations are pre-generated and shared for all policies. We see from the figures that since KGCB and Kriging take correlation into consideration in the decision functions, they need less exploration and rely on the correlation to provide information for less explored alternatives. They quickly begin to focus on the alternatives that have the best values. Yet Kriging wanders around local minima for a while before it heads toward the global maximum. Note that the prediction of KGCB gives a good match in general. The function value at the true maximum alternative is well approximated, while moderate error in the estimate is located away from this region of interest. UCB-E is exploring more than necessary and wasting time on less promising regions. But when the budget is big enough, the exploration will contribute to better prediction of the surface, leading to a potentially larger final outcome in the long run. Pure exploitation gets stuck in a seemingly good alternative and the sampling pattern is not reasonable nor meaningful.

8. Numerical Experiments for Online Multi-armed Bandit Problems

In this section, we provide sample comparisons of different policies using the online objective function. The performance measure that we use to evaluate a policy π in online setting is $\frac{\bar{R}_N^\pi}{N}$, where the pseudo-regret $\bar{R}_N^\pi$ is defined as

$$\bar{R}_N^\pi = N \max_{x \in \mathcal{X}} \mu_x - \sum_{n=0}^{N} \mathbb{E}[\mu_{X^{\pi,n}(s^n)}].$$

The opportunity cost (OC) between two policies in online setting is defined as the difference of their pseudo-regrets.

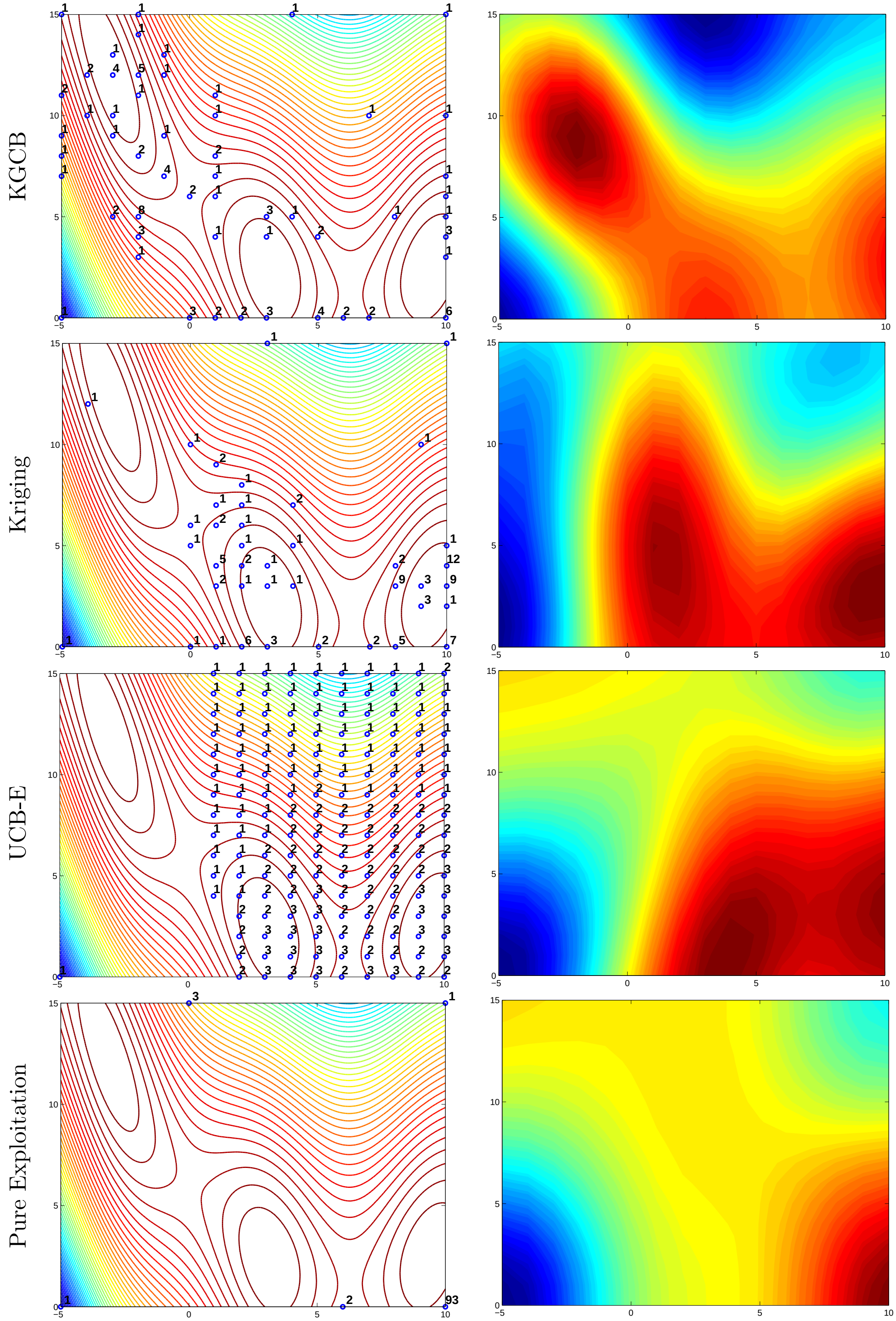


Figure 4: Left column: sampling distribution. Right column: posterior distribution.

8.1 Experiments with independent beliefs

In real world problems, especially in experimental science, frequentist techniques cannot incorporate prior knowledge from domain experts, relying instead on the training from vast pools of data. This may be infeasible to perform in reality since running one experiment might be very expensive. The advantage of a Bayesian approach is unarguable in such cases. However, if we use MLE to fit the prior instead of using domain knowledge, it seems that the comparisons are in favor of Bayesian approaches by using an extra $11 \times p$ measurements. In order to make a seemingly more fair comparison in our synthetic experimental setting, we also experiment with uninformative priors with no additional information provided for Bayesian approach.

Tables 2, 3 and 4 provide comparisons of OLKG, IE with tuning, UCB-E with tuning, UCB, KLUCB, pure exploration (EXPL) using the Bubeck problems with uninformative prior. The measurement budgets are set to 10, 100 and 500 times the number of alternatives of each problem class in Tables 2, 3 and 4, respectively. IE and UCB-E are carefully tuned for each problem class. Under each problem class, we ran each policy for 1000 times. In each run, all the measurements are pre-generated and shared across all the policies. For each policy we record the normalized opportunity cost between OLKG and other competing policies, where the normalized opportunity cost is defined as the ratio between the opportunity cost $\frac{\bar{R}_N^{\pi}}{N} - \frac{\bar{R}_N^{\mathrm{OLKG}}}{N}$ and the range of the truth μ. Positive values of OC indicate that the corresponding policy underperforms OLKG on average. Other than the interest of average performance measured by pseudo-regret, only one sample path will be realized in real world experiments and it is meaningful to find out which policy is most likely to perform the best in one sample run. Thus we also report the probability that each of the other policy outperforms (obtains a lower regret than) OLKG within 1000 realizations. Any policy can be set as a benchmark by placing it as the first policy in the input spreadsheet.

We see from the three tables that the probability of any other policy that outperforms OLKG is in general much less than 0.5. If this criterion is what an experimenter anticipates, then OLKG is a safe choice in most situations. We then discuss the performance of each policy in terms of OC. At the beginning of each trial, IE and UCB-E are more exploiting than exploring while OLKG tends to explore before it moves toward the best estimates. This contributes to good performance (measured by OC) of IE and UCB-E in Table 2 with a small measurement budget. The tuned values of parameters further sharpen this effect by utilizing smaller values compared to those under larger measurement budgets as reported in Table 5 which summarizes the optimally tuned values for each parameter. Since UCB policies tend to explore more than necessary (which can be seen from the sampling pattern, for example, Figure 4), the performance degenerates with a moderate measurement budget as shown in Table 3. In this case, OLKG yields the best performance since after an exploration period, it begins to focus on the alternatives that have the best estimates while looking for alternatives whose estimates are less certain. Yet exploration benefits in the long run. Thus the performance of UCB policies and IE improves if allowed to explore for a sufficiently long time as reported in Table 4.

Table 2: The difference between each policy and OLKG (OC), and the probability that each policy outperforms OLKG, using uninformative priors with a measurement budget 10 times the number of alternatives.

Problem Class	IE		UCBE		UCBV		UCB		KLUCB		EXPL	
	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.
Bubeck1	-0.031	0.43	-0.032	0.43	0.073	0.51	0.016	0.35	0.054	0.50	0.078	0.50
Bubeck2	-0.032	0.55	-0.031	0.52	0.097	0.30	0.025	0.43	0.070	0.35	0.105	0.29
Bubeck3	-0.000	0.29	0.006	0.30	0.068	0.26	0.021	0.53	0.020	0.34	0.095	0.23
Bubeck4	-0.004	0.39	-0.003	0.57	0.100	0.36	0.029	0.48	0.040	0.40	0.124	0.33
Bubeck5	-0.019	0.71	-0.020	0.71	0.213	0.01	0.018	0.48	0.087	0.11	0.255	0.00
Bubeck6	-0.034	0.49	-0.035	0.48	0.139	0.34	0.034	0.41	0.098	0.37	0.151	0.33
Bubeck7	-0.036	0.70	-0.036	0.71	0.065	0.17	0.009	0.48	0.043	0.22	0.073	0.15

Table 3: The difference between each policy and OLKG (OC), and the probability that each policy outperforms OLKG, using uninformative priors with a measurement budget 100 times the number of alternatives.

Problem Class	IE		UCBE		UCBV		UCB		KLUCB		EXPL	
	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.
Bubeck1	0.006	0.34	0.015	0.32	0.387	0.36	0.245	0.14	0.311	0.37	0.431	0.36
Bubeck2	0.006	0.31	0.017	0.35	0.399	0.09	0.226	0.17	0.309	0.22	0.458	0.06
Bubeck3	0.002	0.32	0.007	0.31	0.111	0.18	0.077	0.39	0.052	0.25	0.214	0.07
Bubeck4	-0.014	0.31	-0.005	0.30	0.232	0.27	0.156	0.32	0.114	0.30	0.365	0.17
Bubeck5	-0.003	0.39	0.003	0.34	0.228	0.01	0.064	0.26	0.094	0.15	0.425	0.00
Bubeck6	0.014	0.38	0.025	0.38	0.522	0.10	0.274	0.12	0.380	0.10	0.619	0.09
Bubeck7	0.015	0.52	0.016	0.44	0.260	0.00	0.158	0.21	0.215	0.09	0.303	0.00

Table 4: The difference between each policy and OLKG (OC), and the probability that each policy outperforms OLKG, using uninformative priors with a measurement budget 500 times the number of alternatives.

Problem Class	IE		UCBE		UCBV		UCB		KLUCB		EXPL	
	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.
Bubeck1	-0.105	0.30	-0.098	0.30	0.296	0.26	0.288	0.10	0.175	0.27	0.634	0.26
Bubeck2	-0.089	0.28	-0.080	0.26	0.253	0.31	0.226	0.15	0.139	0.32	0.609	0.02
Bubeck3	-0.009	0.34	-0.006	0.31	0.069	0.18	0.077	0.39	0.035	0.29	0.268	0.03
Bubeck4	-0.075	0.28	-0.069	0.27	0.091	0.26	0.174	0.24	0.014	0.26	0.462	0.12
Bubeck5	-0.030	0.33	-0.026	0.31	0.066	0.28	0.050	0.23	0.012	0.34	0.462	0.00
Bubeck6	-0.024	0.26	-0.022	0.24	0.310	0.05	0.227	0.16	0.190	0.06	0.771	0.05
Bubeck7	-0.045	0.33	-0.045	0.34	0.262	0.11	0.152	0.23	0.200	0.27	0.430	0.00

Table 5: Tuned parameters of IE and UCB-E under different problem classes and measurement budgets. The second row indicates the ratio between the measurement budget and the number of alternatives.

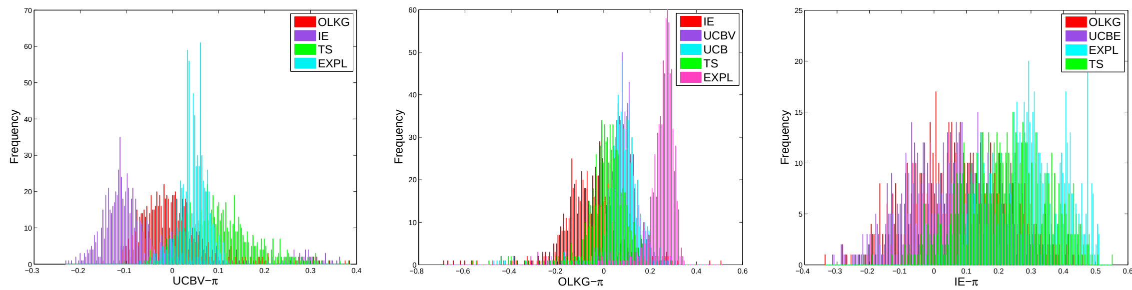
Problem Class	IE			UCBE		
	10	100	500	10	100	500
Bubeck1	0.0007079	1.295	2.036	0.0008991	0.3934	1.103
Bubeck2	0.1675	1.295	2.169	0.002359	0.337	0.9063
Bubeck3	0.8991	1.395	1.878	0.1206	0.4562	0.8635
Bubeck4	0.8991	1.571	2.196	0.004392	0.5332	1.197
Bubeck5	0.004566	1.395	2.169	0.0003102	0.3518	1.002
Bubeck6	0.09063	1.197	1.642	0.000505	0.3201	0.7748
Bubeck7	0.002773	0.8991	1.878	0.0005936	0.2169	0.8007

8.2 Experiments with correlated beliefs

In this section, we summarize numerical experiments on problems with correlated beliefs between different policies, including OLKG, IE with tuning, UCBE, UCBV, Kriging, UCB, Thompson Sampling (TS) and pure exploration (EXPL). To take advantage of correlated beliefs, we use the prior mean as the starting point and use posterior mean θ^n in place of the empirical mean for UCBV and UCB policies.

In order to gain a good understanding of the performance of the policies, MOLTE produces histograms illustrating the distribution of the difference between the normalized OC of a benchmark policy and either of the other policies over 1000 runs. Whichever policy that is listed as the first policy is treated as the benchmark. The measurement budget is set to 0.2 times the number of alternatives of each problem class. Figure 4 compares the performance of several policies under various problem classes with different benchmark policies. A distribution centered around a positive value implies the policy underperforms the benchmark policy, while one centered around a negative number means the policy outperforms the benchmark. For example, Figure 4(a) compares the performance of UCBV, OLKG, IE, TS and EXPL under Goldstein with UCBV as the benchmark policy. We can see that the tuned IE and OLKG are outperforming UCBV and others are underperforming.

We close this section by providing more comparisons between other policies with OLKG under various problem classes. The measurement budget is set to 0.2 times the number of alternatives of each problem class. Table 6 reports the normalized mean OCs and the probability that each of the other policy outperforms OLKG under 1000 runs. IE and UCB-E are carefully tuned for each problem classes with the optimal value shown in Table 7. IE and UCB-E after tuning works generally well. Yet the optimal values of the tuned parameters are quite different for different problems as shown in Table 5 and 7. In addition, the performance of the policies are sensitive to the value of the tunable parameters. In light of this issue, we can conclude that OLKG and Kriging have one attractive advantage over IE and UCB-E: they require no tuning at all, while yielding comparable performance to a



(a) Goldstein (b) HyperEllipsoid (c) Rastrigin

Figure 4: Normalized opportunity cost between different policies.

Table 6: Comparisons with OLKG for correlated beliefs with the measurement 0.2 times the number of alternatives of each problem class.

Problem Class	IE		UCBE		UCBV		Kriging		TS		EXPL	
	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.
Goldstein	-0.061	0.81	-0.097	0.92	-0.003	0.45	-0.031	0.73	0.100	0.09	0.041	0.16
AUF_HNoise	0.058	0.40	0.022	0.43	0.037	0.54	0.031	0.39	0.073	0.22	0.047	0.48
AUF_MNoise	0.043	0.29	0.027	0.42	0.343	0.21	0.023	0.28	0.173	0.21	-0.057	0.52
AUF_LNoise	-0.043	0.73	-0.013	0.64	0.053	0.51	0.005	0.53	0.038	0.20	0.003	0.62
Branin	-0.027	0.76	0.025	0.24	0.026	0.26	0.004	0.54	0.041	0.07	0.123	0.00
Ackley	0.007	0.42	0.04	0.41	0.106	0.20	0.037	0.42	0.100	0.23	0.344	0.00
HyperEllipsoid	-0.059	0.73	0.064	0.12	0.08	0.07	0.146	0.22	0.011	0.38	0.243	0.03
Pinter	-0.028	0.56	-0.003	0.51	0.029	0.42	-0.055	0.65	0.122	0.19	0.177	0.04
Rastrigin	-0.082	0.70	-0.03	0.56	0.162	0.04	-0.026	0.57	0.136	0.08	0.203	0.01

finely tuned IE or UCB-E policy. A detailed study on the issue of tuning is presented in Section 9.1.

Table 6 together with the comparisons shown in previous sections suggests that there is no universal best policy for all problem classes and one could possibly design toy problems for either policy to perform the best. Similar observations have also been reported by Kuleshov and Precup (2000) for different bandit problems on different metrics. Besides, there are theoretical guarantees proved for each of the policy mentioned above, but the existence of these bounds does not appear to provide reliable guidance regarding which policy works best. An asymptotic bound does not provide any assurance that an algorithm will work well on a particular problem in finite time. In practice, we believe that more useful guidance could be obtained by abstracting a real world problem, running simulations and using these to indicate which policy works best.

Table 7: Tuned parameters of IE and UCB-E under different problem classes.

Problem Class	IE	UCBE
Goldstein	0.009939	2571
AUF_HNoise	0.01497	0.319
AUF_MNoise	0.01871	1.591
AUF_LNoise	0.01095	6.835
Branin	0.2694	0.0003664
Ackley	1.197	1.329
HyperEllipsoid	0.8991	21.21
Pinter	0.9989	0.0001636
Rastrigin	0.2086	0.001476

9. Discussion

We close our presentation by discussing two issues that tend to be overlooked in comparisons of learning algorithms: the tuning of heuristic parameters (widely used in frequentist UCB policies) and priors (used in all Bayesian policies such as knowledge gradient).

9.1 The issue of tuning

Previous experimental results show that tuned version of IE and UCB-E yield good performance in general and yet the optimal value for IE and UCB-E may be highly problem dependent. Our experiments also suggest that the performance of a policy is sensitive to the value of the tuned parameter. For example, Figure 8 provides the comparisons between the performances of IE with different parameter values (provided in the parentheses) with the online objective function under various problem classes. The measurement budget is set to five times the number of alternatives for each problem class experimented with independent beliefs and 0.3 times the number of alternatives for each problem class experimented with correlated beliefs. 'OC' is the mean opportunity cost comparing tuned IE with others $OC^{\text{IE}} - OC^{\pi}$, with a positive value indicating a win for tuned IE. 'Prob.' is the probability that other policies outperform the tuned IE. We see from the table that z_α is highly problem dependent and the performance degrades quickly away from the optimal value. For some experimental applications, tuning can require running physical experiments, which may be very expensive or even entirely infeasible.

9.2 The issue of constructing priors

In MOLTE, we use MLE to fit the prior for test functions based on sampling measurements, which seems like a tuning process. Yet designing a Bayesian prior is not necessarily the same as tuning parameters. In real world problems, such as applications in experimental sciences (although there are many other examples from other problem domains), the Bayesian prior may be based on an understanding of the physical system and might be based on the underlying chemistry/physics of the problem, a review of the literature, or past experience. This information might be qualitative in nature and is not easily incorporated by frequentist

Problem Class	B	z_{α}^{*}	IE(1)		IE(2)		IE(3)		IE(4)		IE(5)	
			OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.	OC	Prob.
Bubeck4	I	2.086	0.002	0.40	0.001	0.45	0.002	0.46	0.015	0.47	0.017	0.47
Bubeck6	I	2.01	0.003	0.44	0.001	0.48	0.004	0.43	0.013	0.23	0.028	0.13
AUF_MNoise	I	1.1305	0.004	0.38	0.041	0.04	0.071	0.00	0.095	0.00	0.114	0.09
CamelBack	I	1.295	0.006	0.35	0.069	0.32	0.108	0.03	0.145	0.00	0.172	0.00
AUF_LNoise	C	0.9498	0.043	0.00	0.080	0.00	0.105	0.00	0.123	0.03	0.136	0.00
Branin	C	0.4438	0.001	0.25	0.005	0.32	0.014	0.07	0.023	0.01	0.032	0.01
Goldstein	C	0.079	0.071	0.00	0.090	0.00	0.101	0.00	0.108	0.00	0.113	0.00
Rosenbrock	C	0.9989	0.007	0.18	0.060	0.08	0.093	0.05	0.120	0.04	0.143	0.03

Table 8: Comparisons between tuned IE and IEs with fixed parameter values. The second column indicates the belief model, with I for independent belief and C for correlated belief. z_{α}^{*} is the tuned value for each problem class. The number included in the parenthesis is the parameter value used by each IE policy.

approaches. When this domain knowledgeable is available, and especially when experiments are expensive, Bayesian approaches are strongly preferred.

10. Conclusion

In this paper we present the first finite-time bounds for the knowledge gradient policy applied to offline ranking and selection problems from two directions. First by reducing the Bayesian ranking and selection problems to the adaptive stochastic multi-set maximization problems, we show that the KG policy inherits precisely the performance guarantees of the greedy algorithm for classic submodular maximization problems if the utility function is adaptive submodular. We also analyze the prior-optimality of the KG policy under a weaker assumption on the value of information to provide another insight into the performance of the KG policy. We point out that in general, submodularity does not hold for arbitrary value functions and analyze the submodularity of the two-alternative case. We introduce a new modular optimal learning testing environment (MOLTE) and present its ability to compare different policies under various problem classes. We draw the conclusion that there is no universal best policy for all problem classes, and bounds, by themselves, do not provide reliable guidance to the policy that will work the best. We offer MOLTE as a public-domain test environment to facilitate the process of more comprehensive comparisons, on a broader set of test problems and a broader set of policies, so that researchers can more easily draw insights into the behavior of different policies in the context of different problem classes.

Acknowledgments

This research was supported by grant FA9550-12-1-0200 from the Air Force Office of Scientific Research, Program for Natural Materials, Systems and Extremophiles.

Appendix A: Proof of Lemma 5.

For any ψ with $|\psi| = n$, we consider the resulting knowledge state $S^n = (\theta_x^n, \beta_x^n)_{x\in\mathcal{X}}$. Since $\sigma^W \neq 0$, there exists such ψ that $\max_x \theta_x^n > \max_{x\neq x'} \theta_x^n$ with positive probability. Now consider another realization ψ' with $\mathrm{dom}(\psi') = \mathrm{dom}(\psi) \cup \{x_2\}$, where x_2 is the second largest alternative of θ_x^n. We denote the observation of x_2 in ψ' as W_2 and the resulting S^{n+1} as $(\theta_x^{n+1}, \beta_x^{n+1})_{x\in\mathcal{X}}$ according to Bayes' rule. The knowledge gradient $\Delta(x|\psi) = \nu_x^{\mathrm{KG},n}$ can be analytically expressed by

$$\nu_x^{\mathrm{KG},n} = \tilde{\sigma}_x^n f(\zeta_x^n),$$

where $\tilde{\sigma}_x^n = \sqrt{(\beta_x^n)^{-1} - (\beta_x^n + \beta^W)^{-1}}$, $\zeta_x^n = -\left|\frac{\theta_x^n - \max_{x'\neq x}\theta_{x'}^n}{\tilde{\sigma}_x^n}\right|$ and $f(\zeta) = \zeta\Phi(\zeta) + \phi(\zeta)$. $\Phi(\zeta)$ and $\phi(\zeta)$ are, respectively, the cumulative standard normal distribution the standard normal density (Frazier et al., 2008). We first notice that $f'(\zeta) = \Phi(\zeta) \geq 0$ for any $\zeta \in \mathbb{R}$ so that $f(\zeta)$ is non-decreasing. We next compare $\nu_{x_1}^{\mathrm{KG},n}$ and $\nu_{x_1}^{\mathrm{KG},n+1}$ for $x_1 = \arg\max_x \theta_x^n$. According to Bayes' rule, the precision β of x_2 changes only when x_2 is measured. So we have $\tilde{\sigma}_x^n = \tilde{\sigma}_x^{n+1}$. Similarly we have all the θ_x^{n+1} unchanged except for alternative x_2. By some algebra, it can be shown that for any W_2 such that $\theta_{x_2}^n < W_2 \leq \frac{\beta_{x_2}^n}{\beta^W}(\theta_{x_1}^n - \theta_{x_2}^n) + \theta_{x_1}^n$, we have $\nu_{x_1}^{\mathrm{KG},n} < \nu_{x_1}^{\mathrm{KG},n+1}$. Since $\theta_{x_1}^n > \theta_{x_2}^n$ by construction, such W_2 can be obtained with positive probability.

Appendix B: Proof of Proposition 1.

In this appendix, we prove the properties of submodular multi-set functions. We prove the equivalence by showing 2) $\Rightarrow$ 1) $\Rightarrow$ 3) $\Rightarrow$ 4) $\Rightarrow$ 2).

- 2) $\Rightarrow$ 1). Take $S \subseteq T$ and $T - S = \{x_1, x_2, ..., x_r\}$. Then from 3) we have $\rho_x(S) \geq \rho_x(S \cup \{x_1\})$, $\rho_x(S \cup \{x_1\}) \geq \rho_x(S \cup \{x_1, x_2\})$,..., $\rho_x(S \cup \{x_1, x_2, ..., x_{r-1}\}) \geq \rho_x(T)$. Summing these r inequalities yields 1).

- 1) $\Rightarrow$ 3). For arbitrary S and T with $T - S = \{x_1, x_2, ..., x_r\}$ and $S - T = \{y_1, y_2, ..., y_q\}$, from 1) we have

$$\begin{aligned} g(S \cup T) - g(S) \quad &= \sum_{t=1}^{r} [g(S \cup \{x_1, ..., x_t\}) - g(S \cup \{x_1, ..., x_{t-1}\})] \\ &= \sum_{t=1}^{r} \rho_{x_t}(S \cup \{x_1, ..., x_{t-1}\}) \\ &\leq \sum_{t=1}^{r} \rho_{x_t}(S) = \sum_{x\in T-S} \rho_x(S). \end{aligned} \tag{10}$$

And

$$
\begin{aligned}
g(S \cup T) - g(T) \quad &= \sum_{t=1}^{q}[g(T \cup \{y_1, ..., y_t\}) - g(T \cup \{y_1, ..., y_{t-1}\})] \\
&= \sum_{t=1}^{q} \rho_{y_t}(T \cup \{y_1, ..., y_t\} - \{y_t\}\}) \\
&\geq \sum_{t=1}^{q} \rho_{y_t}(T \cup S - \{y_t\}) = \sum_{x \in S-T} \rho_x(S \cup T - \{x\}). \qquad (11)
\end{aligned}
$$

Subtracting equation (11) from equation (10) we get 3).

- 3) $\Rightarrow$ 4). If $S \subseteq T$, $S - T = \emptyset$, and therefore the last term in 3) vanishes.
- 4) $\Rightarrow$ 2). Substitute $T = S \cup \{x, y\}$ into 4) to obtain

$$g(S \cup \{x, y\}) \leq g(S) + \rho_x(S) + \rho_y(S) = \rho_x(S) + g(S \cup \{y\}).$$

Rearrange this inequality, we get

$$\rho_x(S \cup \{y\}) = g(S \cup \{x, y\}) - g(S \cup \{y\} \leq \rho_x(S).$$

Appendix C: Proof of Proposition 12.

Let $z^*(Z, \pi, \Phi)$ be the next adaptive greedy choice that maximizes the expected marginal increment given that policy π has generated Z. We first show that

$$F^{\pi_2 @ \pi_1} \leq F^{\pi_2} + n_1 \sum_{Z \in \mathcal{Z}^n} \mathbb{P}(\pi_2 \rightsquigarrow Z)\Big(\mathbb{E}\big[\hat{v}(Z \cup \{z^*(Z, \pi_2, \Phi)\}, \Phi)\big] - v(Z)\Big)$$

for all policies π_1 with a measurement budget n_1 and π_2 with a budget n_2 under any prior and probability distribution that describes a measurement.

Proof Let $\pi^{[j]}$ denote the first j measurement decisions under some policy π. First of all we break $F^{\pi_2 @ \pi_1} - F^{\pi_2}$ into n_1 consecutive differences,

$$F^{\pi_2 @ \pi_1} - F^{\pi_2} = \sum_{j=1}^{n_1} \left(F^{\pi_2 @ \pi_1^{[j]}} - F^{\pi_2 @ \pi_1^{[j-1]}}\right).$$

Similar to what we did in the last lemma, for each difference we have

$$
\begin{aligned}
& F^{\pi_2 @ \pi_1^{[j]}} - F^{\pi_2 @ \pi_1^{[j-1]}} \\
= & \sum_{Z_1 \in \mathcal{Z}^{n_2+j}} \mathbb{P}(\pi_2 @ \pi_1^{[j]} \rightsquigarrow Z_1) v(Z_1) - \sum_{Z_2 \in \mathcal{Z}^{n_2+j-1}} \mathbb{P}(\pi_2 @ \pi_1^{[j-1]} \rightsquigarrow Z_2) v(Z_2) \\
= & \sum_{Z_1 \in \mathcal{Z}^{n_2+j}} \sum_{Z_2 \in \mathcal{Z}^{n_2+j-1}, Z_2 \cup Z_3 = Z_1} \mathbb{P}(\pi_2 @ \pi_1^{[j-1]} \rightsquigarrow Z_2) \mathbb{P}(\pi_1^{\{j\}} \rightsquigarrow Z_3 | \pi_2 @ \pi_1^{[j-1]} \rightsquigarrow Z_2) v(Z_1) \\
- & \sum_{Z_2 \in \mathcal{Z}^{n_2+j-1}} \sum_{Z_3 \in \mathcal{Z}^1} \mathbb{P}(\pi_2 @ \pi_1^{[j-1]} \rightsquigarrow Z_2) \mathbb{P}(\pi_1^{\{j\}} \rightsquigarrow Z_3 | \pi_2 @ \pi_1^{[j-1]} \rightsquigarrow Z_2) v(Z_2) \\
= & \sum_{Z_2 \in \mathcal{Z}^{n_2+j-1}} \sum_{Z_3 \in \mathcal{Z}^1} \mathbb{P}(\pi_2 @ \pi_1^{[j-1]} \rightsquigarrow Z_2) \mathbb{P}(\pi_1^{\{j\}} \rightsquigarrow Z_3 | \pi_2 @ \pi_1^{[j-1]} \rightsquigarrow Z_2) \big(v(Z_2 \cup Z_3) - v(Z_2)\big).
\end{aligned}
$$

Now we consider all possible pair (Z_4, Z_5) such that $Z_4 \in \mathcal{Z}^{n_2}$, $Z_5 \in \mathcal{Z}^{j-1}$ and $Z_4 \cup Z_5 = Z_2$. Notice that the policy $\pi_2 @ \pi_1^{[j]}$ employs a fresh start at the time n_2, therefore the events before and after time n_2 are independent. Then we have

$$\begin{aligned}
& \sum_{Z_2 \in \mathcal{Z}^{n_2+j-1}} \sum_{Z_3 \in \mathcal{Z}^1} \mathbb{P}(\pi_2 @ \pi_1^{[j-1]} \rightsquigarrow Z_2) \mathbb{P}(\pi_1^{\{j\}} \rightsquigarrow Z_3 | \pi_2 @ \pi_1^{[j-1]} \rightsquigarrow Z_2) \big(v(Z_2 \cup Z_3) - v(Z_2) \big) \\
= \quad & \sum_{Z_2 \in \mathcal{Z}^{n_2+j-1}} \sum_{Z_4 \cup Z_5 = Z_2} \sum_{Z_3 \in \mathcal{Z}^1} \mathbb{P}(\pi_2 \rightsquigarrow Z_4) \mathbb{P}(\pi_1^{[j-1]} \rightsquigarrow Z_5) \mathbb{P}(\pi_1^{\{j\}} \rightsquigarrow Z_3 | \pi_2 @ \pi_1^{[j-1]} \rightsquigarrow Z_2) \\
& \times \big(v(Z_2 \cup Z_3) - v(Z_2) \big).
\end{aligned}$$

Based on the submodular property of function v, we have

$$v(Z_2 \cup Z_3) - v(Z_2) \leq v(Z_4 \cup Z_3) - v(Z_4).$$

Then from the definition of z^*, we have

$$\begin{aligned}
v(Z_4 \cup Z_3) - v(Z_4) &= \mathbb{E}[\hat{v}(Z_4 \cup Z_3, \Phi) - \hat{v}(Z_4, \Phi)] \\
&= \mathbb{E}_\Phi \big[\mathbb{E}[\hat{v}(Z_4 \cup Z_3, \Phi) - \hat{v}(Z_4, \Phi) | Z^{\pi_2}(\Phi) = Z_4] \big] \\
&\leq \mathbb{E}_\Phi \big[\mathbb{E}[\hat{v}(Z_4 \cup \{z^*(Z_4, \pi_2, \Phi)\}, \Phi) - \hat{v}(Z_4, \Phi) | Z^{\pi_2}(\Phi) = Z_4] \big] \\
&= \mathbb{E}_\Phi [\hat{v}(Z_4 \cup \{z^*(Z_4, \pi_2, \Phi)\}, \Phi)] - v(Z_4).
\end{aligned}$$

Combining the last two inequalities, we have

$$\begin{aligned}
& \sum_{Z_2 \in \mathcal{Z}^{n_2+j-1}} \sum_{Z_4 \cup Z_5 = Z_2} \sum_{Z_3 \in \mathcal{Z}^1} \mathbb{P}(\pi_2 \rightsquigarrow Z_4) \mathbb{P}(\pi_1^{[j-1]} \rightsquigarrow Z_5) \mathbb{P}(\pi_1^{\{j\}} \rightsquigarrow Z_3 | \pi_2 @ \pi_1^{[j-1]} \rightsquigarrow Z_2) \\
& \times \big(v(Z_2 \cup Z_3) - v(Z_2) \big) \\
\leq \quad & \sum_{Z_2 \in \mathcal{Z}^{n_2+j-1}} \sum_{Z_4 \cup Z_5 = Z_2} \sum_{Z_3 \in \mathcal{Z}^1} \mathbb{P}(\pi_2 \rightsquigarrow Z_4) \mathbb{P}(\pi_1^{[j-1]} \rightsquigarrow Z_5) \mathbb{P}(\pi_1^{\{j\}} \rightsquigarrow Z_3 | \pi_2 @ \pi_1^{[j-1]} \rightsquigarrow Z_2) \\
& \times \big(\mathbb{E}\hat{v}(Z_4 \cup \{z^*(Z_4, \pi_2, \Phi)\}, \Phi) - v(Z_4) \big) \\
= \quad & \sum_{Z_2 \in \mathcal{Z}^{n_2+j-1}} \sum_{Z_4 \cup Z_5 = Z_2} \mathbb{P}(\pi_2 \rightsquigarrow Z_4) \mathbb{P}(\pi_1^{[j-1]} \rightsquigarrow Z_5) \big(\mathbb{E}\hat{v}(Z_4 \cup \{z^*(Z_4, \pi_2, \Phi)\}, \Phi) - v(Z_4) \big) \\
= \quad & \sum_{Z_4 \in \mathcal{Z}^{n_2}} \sum_{Z_5 \in \mathcal{Z}^{j-1}} \mathbb{P}(\pi_2 \rightsquigarrow Z_4) \mathbb{P}(\pi_1^{[j-1]} \rightsquigarrow Z_5) \big(\mathbb{E}\hat{v}(Z_4 \cup \{z^*(Z_4, \pi_2, \Phi)\}, \Phi) - v(Z_4) \big) \\
= \quad & \sum_{Z_4 \in \mathcal{Z}^{n_2}} \mathbb{P}(\pi_2 \rightsquigarrow Z_4) \big(\mathbb{E}\hat{v}(Z_4 \cup \{z^*(Z_4, \pi_2, \Phi)\}, \Phi) - v(Z_4) \big),
\end{aligned}$$

and this ends the proof. ■

Set $\pi_1 = \pi^*$ and $\pi_2 =$ KG$^{[n-1]}$ in Lemma 11 and the above proposition then what left to show is that

$$F^{\mathrm{KG}^{[n]}} - F^{\mathrm{KG}^{[n-1]}} \geq \sum_{Z \in \mathcal{Z}^n} \mathbb{P}(\pi_2 \rightsquigarrow Z) \Big(\mathbb{E}\hat{v}(Z \cup \{z^*(Z, \mathrm{KG}^{[n-1]}, \Phi)\}, \Phi) - v(Z) \Big).$$

From the definition, the left hand side of the last equation:

$$\begin{aligned} F^{\mathrm{KG}^{[n]}} - F^{\mathrm{KG}^{[n-1]}} &= \sum_{Z_1 \in \mathcal{Z}^{n+1}} \mathbb{P}(\mathrm{KG} \rightsquigarrow Z_1) v(Z_1) - \sum_{Z_2 \in \mathcal{Z}^n} \mathbb{P}(\mathrm{KG} \rightsquigarrow Z_2) v(Z_2) \\ &= \sum_{Z_2 \in \mathcal{Z}^n} \sum_{Z_3 \in \mathcal{Z}^1} \mathbb{P}(\mathrm{KG} \rightsquigarrow Z_2) \mathbb{P}(\mathrm{KG} \rightsquigarrow Z_3 | \mathrm{KG} \rightsquigarrow Z_2) v(Z_2 \cup Z_3) \\ &\quad - \sum_{Z_2 \in \mathcal{Z}^n} \mathbb{P}(\mathrm{KG} \rightsquigarrow Z_2) v(Z_2). \end{aligned}$$

Now it is enough to show that

$$\sum_{Z_3 \in \mathcal{Z}^1} \mathbb{P}(\mathrm{KG} \rightsquigarrow Z_3 | \mathrm{KG} \rightsquigarrow Z_2) v(Z_2 \cup Z_3) - v(Z_2) \geq \mathbb{E}\hat{v}(Z_2 \cup \{z^*(Z_2, \mathrm{KG}^{[n-1]}, \Phi)\}, \Phi) - v(Z_2).$$

We could group together the partial realizations ψ that lead to the same single step optimal decision $z^*(Z_2, \mathrm{KG}^{[n-1]}, \Phi)$, and then the last inequality follows from the adaptive greedy nature of the KG policy.

Appendix D: Proof of Theorem 16.

First of all, we consider the case when f is a two dimensional function and the four points we pick form a rectangle. Assume $f(x, y)$ is submodular. For any given point (x_0, y_0), we have $f(x_0 + t + s, y_0) - f(x_0 + t, y_0) \leq f(x_0 + s, y_0) - f(x_0, y_0)$ and $f(x_0 + t, y_0) - f(x_0, y_0) \leq f(x_0 + t, y_0 + s) - f(x_0, y_0 + s)$ for any $s, t > 0$. From the first inequality we get $f_{xx}(x_0, y_0) \leq 0$ directly. From the second inequality, we have $f_x(x_0, y_0) \leq f_x(x_0, y_0 + s)$, and finally $f_{x,y}(x_0, y_0) \leq 0$. On the other hand, if we have $f_{xy} \leq 0$, $f_{xx} \leq 0$, for any (x, y), then due to the fact that $f(x_0 + t, y_0 + s) - f(x_0 + t, y_0) - \big(f(x_0, y_0 + s) - f(x_0, y_0)\big) = \int_{x_0}^{x_0+t} \int_{y_0}^{y_0+s} f_{xy}(u, v) \mathrm{d}u \mathrm{d}v \leq 0$, $f(x_0 + t + s, y_0) - f(x_0 + t, y_0) - \big(f(x_0 + s, y_0) - f(x_0, y_0)\big) = st f_{xx}(x_0 + \xi, y_0) \leq 0$, we obtain the submodularity.

We next consider the general case when f is n dimensional and the four points only form a parallelogram. Since the difference between the two marginal values can be decomposed into summation of several marginal value differences whose reference points form rectangles that parallel to coordinate planes, the result for the general case is straightforward from the two dimensional case.

References

Rajeev Agrawal. Sample mean based index policies with $o(logn)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, pages 1054–1078, 1995.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410(19), 2009.

Jean-Yves Audibert, Sébastien Bubeck, et al. Best arm identification in multi-armed bandits. *COLT 2010-Proceedings*, 2010.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Emre Barut and Warren B Powell. Optimal learning for sequential sampling with non-parametric beliefs. *Journal of Global Optimization*, pages 1–27, 2013.

Jürgen Branke, Stephen E Chick, and Christian Schmidt. Selecting a selection procedure. *Management Science*, 53(12):1916–1932, 2007.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *arXiv preprint arXiv:1209.1727*, 2012.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

Chun-Hung Chen, Hsiao-Chang Chen, and Liyi Dai. A gradient approach for smartly allocating computing budget for discrete event simulation. In *Proceedings of the 28th conference on Winter simulation*, pages 398–405. IEEE Computer Society, 1996.

Chun-Hung Chen, Jianwu Lin, Enver Yücesan, and Stephen E Chick. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270, 2000.

Chun-Hung Chen, Donghai He, and Michael Fu. Efficient dynamic simulation allocation in ordinal optimization. *Automatic Control, IEEE Transactions on*, 51(12):2005–2009, 2006.

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

LCW Dixon and GP Szegö. The global optimization problem: an introduction. *Towards global optimization*, 2:1–15, 1978.

Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(4):599–613, 2009.

Peter I. Frazier and Warren B. Powell. Paradoxes in learning and the marginal value of information. *Decision Analysis*, 7(4):378–403, 2010.

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

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems.

John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodoogical)*, 1979.

Daniel Golovin and Andreas Krause. Adaptive submodularity: A new approach to active learning and stochastic optimization. In *COLT*, pages 333–345, 2010.

Shanti S Gupta and Klaus J Miescke. Bayesian look ahead one-stage sampling allocations for selection of the best population. *Journal of statistical planning and inference*, 54(2):229–244, 1996.

H-M Gutmann. A radial basis function method for global optimization. *Journal of Global Optimization*, 19 (3):201–227, 2001.

Irwin Guttman, George C Tiao, et al. A bayesian approach to some best population problems. *The Annals of Mathematical Statistics*, 35(2):825–835, 1964.

Donghai He, Stephen E Chick, and Chun-Hung Chen. Opportunity cost and OCBA selection procedures in ordinal optimization for a fixed number of alternative systems. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 37(5):951–961, 2007.

Ronald Howard et al. Information value theory. *Systems Science and Cybernetics, IEEE Transactions on*, 2(1):22–26, 1966.

Jiaqiao Hu, Michael C Fu, and Steven I Marcus. A model reference adaptive search method for stochastic global optimization. *Communications in Information & Systems*, 8(3):245–276, 2008.

Deng Huang, Theodore T Allen, William I Notz, and N Zeng. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of global optimization*, 34(3):441–466, 2006.

Donald R Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383, 2001.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Leslie Pack Kaelbling. *Learning in embedded systems*. 1993.

Aloak Kapoor and Russell Greiner. *Learning and classifying under hard budgets*. Springer, 2005.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2-3):245–272, 2010.

Volodymyr Kuleshov and Doina Precup. Algorithms for multi-armed bandit problems. *Journal of Machine Learning Research*, 2000.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Ruben Martinez-Cantin. Bayesopt: a bayesian optimization library for nonlinear optimization, experimental design and bandits. *The Journal of Machine Learning Research*, 15(1):3735–3739, 2014.

Martijn RK Mes, Warren B Powell, and Peter I Frazier. Hierarchical knowledge gradient for sequential sampling. *The Journal of Machine Learning Research*, 12:2931–2974, 2011.

Marcin Molga and Czesław Smutnicki. Test functions for optimization needs. *Test functions for optimization needs*, 2005.

Diana M Negoescu, Peter I Frazier, and Warren B Powell. The knowledge-gradient algorithm for sequencing experiments in drug discovery. *INFORMS Journal on Computing*, 23(3):346–363, 2011.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

Warren B Powell and Ilya O Ryzhov. *Optimal learning*, volume 841. 2012.

Howard Raiffa and Robert Schlaifer. Applied statistical decision theory. *Harvard Business School Publications*, 1961.

Ilya O Ryzhov, Warren B Powell, and Peter I Frazier. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195, 2012.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

George J Stigler. The economics of information. *The journal of political economy*, pages 213–225, 1961.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.

Fumie Yokota and Kimberly M Thompson. Value of information literature analysis: a review of applications in health risk management. *Medical Decision Making*, 24(3):287–298, 2004.