The knowledge gradient algorithm for a general class of online learning problems

Ilya O. Ryzhov Warren B. Powell Peter I. Frazier

April 19, 2011

Abstract

We derive a one-period look-ahead policy for finite- and infinite-horizon online optimal learning problems with Gaussian rewards. Our approach is able to handle the case where our prior beliefs about the rewards are correlated, which is not handled by traditional multi-armed bandit methods. Experiments show that our KG policy performs competitively against the best known approximation to the optimal policy in the classic bandit problem, and outperforms many learning policies in the correlated case.

1 Introduction

We consider a class of optimal learning problems in which sequential measurements are used to gradually improve estimates of unknown quantities. In each time step, we choose one of finitely many alternatives and observe a random reward whose expected value is the unknown quantity corresponding to that alternative. The rewards are independent of each other and follow a Gaussian distribution with known variance. We maximize the total expected reward collected over time, a problem class often addressed under the umbrella of multi-armed bandit problems. We allow several variations of this basic setup: the rewards may be discounted over time, the time horizon may be finite or infinite, and our beliefs about the unknown rewards may be correlated. Correlated beliefs are not handled by the traditional bandit literature, but are significant in practice.

Applications arise in many fields where we need to sequentially allocate measurements to alternatives in order to eliminate less valuable alternatives as we go. We deal with online learning in this paper, so we consider applications in which we are interested not only in finding the best alternative, but in maximizing the *total* expected reward collected over the entire time horizon. Several situations where this distinction is important are:

1. *Clinical trials.* Experimental drug treatments are tested on groups of human patients. Each treatment has a different, unknown expected effectiveness. We are interested in the well-being of the patients as well as in finding the best treatment, so the problem is online. If the treatments consist of overlapping sets of drugs, our beliefs about them will be correlated.

2. *Energy management.* We are applying sets of energy-saving technologies (e.g. insulation, computer-controlled thermostats, tinted windows) to identical industrial buildings. Different technologies interact in an unknown way which can only be measured by actually implementing portfolios of technologies and measuring their combined performance. We maximize total performance over all buildings.

3. *Sensor management.* In this area, a sensor (airport inspector, radiation detector, medical clinic) is used to collect information about the environment. We often have the ability to control the use of a sensor which allows us to not only better learn the state of the system, but also to learn relationships among different variables. See Mahajan & Teneketzis (2008) and Washburn (2008) for more on applications.

The dimension of correlated beliefs generalizes the well-known multi-armed bandit problem, which assumes that our beliefs about the rewards are independent of each other. Much of the literature on multi-armed bandits has focused on the development of index policies. An index policy decomposes the problem by considering each alternative separately from the others, and computing a value for each alternative that depends only on our beliefs about that alternative, and not on our beliefs about other alternatives. The most famous of these is the policy based on Gittins indices (Gittins & Jones (1974)), which is optimal for the classic infinite-horizon bandit problem. Alternatives to Gittins indices include upper confidence bounding (Lai (1987)) and interval estimation (Kaelbling (1993)). These methods construct an interval around our current estimate of the value of an alternative such that the true value is in the interval with high probability, and then measure the alternative whose interval has the highest upper bound.

One problem with Gittins indices is that they are hard to compute exactly when the space of possible beliefs is infinite. The computation of Gittins indices is discussed by Katehakis & Veinott (1987) and Duff (1995). An LP-based computational method was developed by Bertsimas & Nino-Mora (2000), but it is founded on a Markov decision process framework (see also Goel et al. (2009) for more work in this setting), where the prior beliefs are limited to a finite set of values, whereas the Gaussian beliefs in our problem are characterized by continuous parameters. There have been

several studies on approximating Gittins indices for the continuous case (Brezzi & Lai (2002), Yao (2006), Chick & Gans (2009)), but such approximations rely on a continuous-time analogy that is subject to errors in the discrete-time bandit model. In addition to the optimality of the Gittins policy, there is also a body of work on theoretical performance guarantees for certain classes of index policies. General bounds on the performance of upper confidence bound policies are presented by Lai & Robbins (1985), and by Auer et al. (2002) for the case of rewards with bounded support. The upper confidence bound approach has also been extended to more complex optimal learning problems, such as Markov decision processes with unknown transition functions (Tewari & Bartlett (2007)) and response-surface bandits (Ginebra & Clayton (1995)).

There are also many general heuristics (descriptions can be found in Powell (2007) and Sutton & Barto (1998)) that can be applied to broad classes of optimal learning problems, including multi-armed bandits. Examples include Boltzmann exploration, pure exploitation, and the equal-allocation policy. Empirical comparisons of some policies in certain settings are available in Vermorel & Mohri (2005).

Our approach applies to the classic bandit problem, but is also able to handle problems where our prior belief about the reward of one alternative is correlated with our beliefs about other rewards. For example, the first two applications considered above are instances of the subset selection problem: we have to investigate a medical treatment (consisting of one or more drugs) or an energy portfolio (of multiple energy-efficient technologies). Correlated beliefs allow us to learn about many subsets with common elements by measuring only a single one. It is logical to suppose that implementing a particular energy portfolio teaches us about the value of other portfolios containing the same technologies. If our beliefs are highly correlated, we can consider problems where the number of choices is much larger than the measurement budget, because a single measurement can now provide information about many or even all the alternatives.

The classical literature on index policies generally does not handle correlated beliefs. Gittins indices are no longer optimal in this setting. Some studies such as Feldman (1962) and Keener (1985) have considered correlated beliefs in a simple setting with only two possible values for a single unknown parameter. Recent work has considered more complex correlated problems under various structural assumptions. For example, the study by Pandey et al. (2007) considers correlated binomial rewards. An important step forward in the study of correlated bandits was made by Mersereau et al. (2008) and Mersereau et al. (2009). These studies assume a particular structure

in which the rewards are linear functions of random variables, and the correlations come from a single random variable shared by every reward. In this case, a greedy policy that always chooses the alternative that we believe to be the best, with no regard for the uncertainty in this belief, can be shown to perform very well. Our work, however, considers a more general correlation structure in the form of a multivariate Gaussian prior.

Our analysis is motivated by the knowledge gradient (KG) concept, developed by Gupta & Miescke (1994) and further analyzed by Frazier et al. (2008) and Chick et al. (2010) for the ranking and selection problem. This problem is the offline version of the multi-armed bandit problem: we must find the best out of M alternatives with unknown rewards, given N chances to learn about them first. The KG policy for ranking and selection chooses the measurement that yields the greatest expected single-period improvement in the estimate of the best reward, a quantity that can be computed exactly. More recently, the KG concept was extended by Frazier et al. (2009) to the ranking and selection problem with correlated priors, and by Chick et al. (2010) to the case of unknown measurement noise.

The knowledge gradient offers an important practical advantage: it is easily computable, in contrast with the far more difficult calculations required for Gittins indices. We present experimental evidence that our KG policy is competitive against the best available Gittins index approximation, given by Chick & Gans (2009). Furthermore, the knowledge gradient methodology can be applied to other distributions, although these require the development of different computational formulas.

This paper makes the following contributions: 1) We propose a new type of online learning policy, based on the knowledge gradient concept. This policy is not an index policy, but rather a one-step look-ahead that computes the marginal value of a single measurement. This quantity is much easier to compute than Gittins indices, with a natural derivation that is easy to understand. 2) We show how this method can handle important variations, such as both finite and infinite time horizons and discount factors. 3) We show that, as the discount factor becomes large, the infinite-horizon KG policy achieves the best possible estimate of the value of the best alternative. Furthermore, only one alternative can be measured infinitely often by the policy, and the probability that it will be the true best alternative converges to 1 as the discount factor becomes large. 4) We conduct a thorough experimental study of the performance of KG for problems with both independent and correlated beliefs. We find that KG is competitive against the best known Gittins approximation on classic bandit problems, and outperforms other index policies and heuristics on

problems with correlated beliefs, without any tunable parameters.

We proceed as follows. In Section 2, we lay out a dynamic programming-based model of the problem. In Section 3, we derive the KG measurement policy for problems with independent beliefs, with both discounted and undiscounted, finite- and infinite-horizon variations. In Section 4, we derive convergence results for the infinite-horizon discounted KG policy as the discount factor increases to 1. In Section 5, we extend KG to problems with correlated beliefs. Finally, we present numerical results comparing online KG to existing policies. We emphasize KG as a general approach to different kinds of optimal learning problems, with the intent of eventually extending it to more complicated problem classes.

2 Mathematical model for learning

Suppose that there are M objects or alternatives. In every time step, we can choose any alternative to measure. If we measure alternative x, we will observe a random reward $\hat{\mu}_x$ that follows a Gaussian distribution with mean μ_x and variance σ_ε^2. The measurement error σ_ε^2 is known, and we use the notation $\beta_\varepsilon = \sigma_\varepsilon^{-2}$ to refer to the measurement precision. Although μ_x is unknown, we assume that $\mu_x \sim \mathcal{N}\left(\mu_x^0, \left(\sigma_x^0\right)^2\right)$, where μ_x^0 and σ_x^0 represent our prior beliefs about μ_x. We also assume that the rewards of the objects are mutually independent, conditioned on μ_x, $x = 1, ..., M$.

We use the random observations we make while measuring to improve our beliefs about the rewards of the alternatives. Let $\mathcal{F}^n$ be the sigma-algebra generated by our choices of the first n objects to measure, as well as the random observations we made of their rewards. We say that something happens "at time n" if it happens after we have made exactly n observations. Then,

$$\mu_x^n = \mathbb{E}^n\left(\mu_x\right),$$

where $\mathbb{E}^n\left(\cdot\right) = \mathbb{E}\left(\cdot\,|\mathcal{F}^n\right)$, represents our beliefs about μ_x after making n measurements. Then, $\left(\sigma_x^n\right)^2$ represents the conditional variance of μ_x given $\mathcal{F}^n$, which can be viewed as a measure of how confident we are about the accuracy of μ_x^n. We also use the notation $\beta_x^n = \left(\sigma_x^n\right)^{-2}$ to denote the conditional precision of μ_x. Thus, at time n, we believe that $\mu_x \sim \mathcal{N}\left(\mu_x^n, \left(\sigma_x^n\right)^2\right)$, and our beliefs are updated after each measurement using Bayes' rule:

$$\mu_x^{n+1} = \begin{cases} \frac{\beta_x^n \mu_x^n + \beta_\varepsilon \hat{\mu}_x^{n+1}}{\beta_x^n + \beta_\varepsilon} & \text{if } x \text{ is the } (n+1)\text{st object measured} \\ \mu_x^n & \text{otherwise.} \end{cases} \tag{1}$$

The rewards of the objects are independent, so we update only one set of beliefs about the object we have chosen. The precision of our beliefs is updated as follows:

$$\beta_x^{n+1} = \begin{cases} \beta_x^n + \beta_\varepsilon & \text{if } x \text{ is the } (n+1)\text{st object measured} \\ \beta_x^n & \text{otherwise.} \end{cases} \tag{2}$$

We use the notation $\mu^n = (\mu_1^n, ..., \mu_M^n)$ and $\beta^n = (\beta_1^n, ..., \beta_M^n)$. We also let

$$(\tilde{\sigma}_x^n)^2 = Var_x^n\left(\mu_x^{n+1}\right) = Var_x^n\left(\mu_x^{n+1}\right) - Var\left(\mu_x^n | \mathcal{F}^n\right)$$

be the reduction in the variance of our beliefs about x that we achieve by measuring x at time n. The notation Var_x^n denotes the conditional variance given $\mathcal{F}^n$ and given that x is the $(n+1)$st alternative measured. The quantity μ_x^n is $\mathcal{F}^n$-measurable, and hence $Var\left(\mu_x^n | \mathcal{F}^n\right) = 0$. It can be shown that

$$\tilde{\sigma}_x^n = \sqrt{\left(\sigma_x^n\right)^2 - \left(\sigma_x^{n+1}\right)^2} = \sqrt{\frac{1}{\beta_x^n} - \frac{1}{\beta_x^n + \beta_\varepsilon}}.$$

It is known, e.g. from DeGroot (1970), that the conditional distribution of μ_x^{n+1} given $\mathcal{F}^n$ is $\mathcal{N}\left(\mu_x^n, (\tilde{\sigma}_x^n)^2\right)$. In other words, given $\mathcal{F}^n$, we can write

$$\mu_x^{n+1} = \mu_x^n + \tilde{\sigma}_x^n \cdot Z \tag{3}$$

where Z is a standard Gaussian random variable.

We can define a *knowledge state*

$$s^n = (\mu^n, \beta^n)$$

to represent our beliefs about the alternatives after n measurements. If we choose to measure an object x^n at time n, we write

$$s^{n+1} = K^M\left(s^n, x^n, \hat{\mu}_{x^n}^{n+1}\right)$$

where the transition function K^M is described by (1) and (2). For notational convenience, we often suppress the dependence on $\hat{\mu}_{x^n}^{n+1}$ when we write K^M. The term "knowledge state" has numerous analogues in other communities. The stochastic control literature uses the term "information state" to denote the same concept, whereas the reinforcement learning community often uses the term "belief state."

We assume that we collect rewards as we measure them. For the time being, we also assume that the rewards are not discounted over time. Thus, if we have N measurements to make, followed

by one final chance to collect a reward, our objective is to choose a measurement policy π that achieves

$$\sup_{\pi} \mathbb{E}^{\pi} \sum_{n=0}^{N} \mu_{X^{\pi,n}(s^n)}, \tag{4}$$

where $X^{\pi,n}(s^n)$ is the alternative chosen by policy π at time n given a knowledge state s^n. The value of following a measurement policy π, starting at time n in knowledge state s^n, is given by Bellman's equation for dynamic programming (applied to optimal learning by DeGroot (1970)):

$$V^{\pi,n}(s^n) = \mu^n_{X^{\pi,n}(s^n)} + \mathbb{E}^n V^{\pi,n+1}\left(K^M(s^n, X^{\pi,n}(s^n))\right) \tag{5}$$

$$V^{\pi,N}\left(s^N\right) = \max_x \mu^N_x. \tag{6}$$

At time N, we can collect only one more reward. Therefore, we should simply choose the alternative that looks the best given everything we have learned, because there are no longer any future decisions that might benefit from learning. At time $n < N$, we collect an immediate reward for the object we choose to measure, plus an expected downstream reward for future measurements. The optimal policy satisfies a similar equation

$$V^{*,n}(s^n) = \max_x \left[\mu^n_x + \mathbb{E}^n V^{*,n+1}\left(K^M(s^n, x)\right)\right] \tag{7}$$

$$V^{*,N}\left(s^N\right) = \max_x \mu^N_x \tag{8}$$

with the only difference being that the optimal policy always chooses the best possible measurement, the one that maximizes the sum of the immediate and downstream rewards. By the dynamic programming principle, the function $V^{*,n}$ represents the optimal value that can be collected from time n onward, and only depends on the past through the starting state s^n. Thus, the expectation of $V^{*,n+1}$ given $\mathcal{F}^n$ is over the single random transition from s^n to s^{n+1}.

3 The online knowledge gradient policy

We derive an easily computable online decision rule for an undiscounted, finite-horizon online problem using the KG principle. We then show that it is always better to measure under this policy than to not measure at all. Finally, we derive KG decision rules for discounted and infinite-horizon problems.

3.1 Derivation

Suppose that we have made n measurements, reached the knowledge state s^n, and then stopped learning entirely. That is, we would still collect rewards after time n, but we would not be able to use those rewards to update our beliefs. Then, we should follow the "stop-learning" (SL) policy of always choosing the alternative that looks the best based on the most recent information. The expected total reward obtained after time n under these conditions is

$$V^{SL,n}(s^n) = (N-n+1)\max_x \mu_x^n. \tag{9}$$

This quantity is somewhat analogous to the "retirement reward" of Whittle (1980), as it represents a fixed reward that we collect after retiring from the learning problem.

The knowledge gradient concept, first described by Gupta & Miescke (1994) and later developed by Frazier et al. (2008), can be stated as "choosing the measurement that would be optimal if it were the last measurement we were allowed to make." Suppose we are at time n, with $N-n+1$ more rewards to collect, but only the $(n+1)$st reward will be used to update our beliefs. Then, we need to make an optimal decision at time n, under the assumption that we will switch to the SL policy starting at time $n+1$. The KG decision rule that follows from this assumption is

$$X^{KG,n}(s^n) = \arg\max_x \mu_x^n + \mathbb{E}^n V^{SL,n+1}\left(K^M(s^n,x)\right). \tag{10}$$

If ties occur, they can be broken by randomly choosing one of the alternatives that achieve the maximum.

The expectation on the right-hand side of (10) can be written as

$$\begin{aligned}
\mathbb{E}^n V^{SL,n+1}\left(K^M(s^n,x)\right) &= (N-n)\,\mathbb{E}^n \max_{x'} \mu_{x'}^{n+1} \\
&= (N-n)\,\mathbb{E}\max\left\{\max_{x'\neq x}\mu_{x'}^n, \mu_x^n + \tilde{\sigma}_x^n \cdot Z\right\} \\
&= (N-n)\left(\max_{x'}\mu_{x'}^n\right) + (N-n)\,\nu_x^{KG,n}
\end{aligned} \tag{11}$$

where the computation of $\mathbb{E}^n \max_{x'} \mu_{x'}^{n+1}$ comes from Frazier et al. (2008). The quantity $\nu_x^{KG,n}$ is called the *knowledge gradient* of alternative x at time n, and is defined by

$$\nu_x^{KG,n} = \mathbb{E}_x^n\left[\left(\max_{x'}\mu_{x'}^{n+1}\right) - \left(\max_{x'}\mu_{x'}^n\right)\right], \tag{12}$$

where $\mathbb{E}_x^n$ is a conditional expectation given $\mathcal{F}^n$ and given that x is the $(n+1)$st alternative measured. The knowledge gradient can be computed exactly using the formula

$$\nu_x^{KG,n} = \tilde{\sigma}_x^n \cdot f\left(-\left|\frac{\mu_x^n - \max_{x' \neq x} \mu_{x'}^n}{\tilde{\sigma}_x^n}\right|\right) \tag{13}$$

where $f(z) = z\Phi(z) + \phi(z)$ and ϕ, Φ are the pdf and cdf of the standard Gaussian distribution. We know from Gupta & Miescke (1994) and Frazier et al. (2008) that (12) and (13) are equivalent in this problem, and that ν^{KG} is always positive. The term "knowledge gradient" arises from (12), where the quantity $\nu_x^{KG,n}$ is the marginal value of the information gained by measuring x.

It is easy to see that (10) can be rewritten as

$$X^{KG,n}(s^n) = \arg\max_x \mu_x^n + (N-n)\,\nu_x^{KG,n}. \tag{14}$$

The term $(N-n)\max_{x'} \mu_{x'}^n$ in (11) is dropped because it does not depend on the choice of x and thus does not affect which x achieves the maximum in (10). The value of this policy follows from (5) and is given by

$$V^{KG,n}(s^n) = \mu^n_{X^{KG,n}(s^n)} + \mathbb{E}^n V^{KG,n+1}\left(K^M\left(s^n, X^{KG,n}(s^n)\right)\right). \tag{15}$$

Instead of choosing the alternative that looks the best, the KG policy adds an uncertainty bonus of $(N-n)\,\nu_x^{KG,n}$ to the most recent beliefs μ_x^n, and chooses the alternative that maximizes this sum. In this way, the KG policy finds a balance between exploitation (measuring alternatives that are known to be good) and exploration (measuring alternatives that might be good), with the uncertainty bonus representing the value of exploration.

Remark 3.1. *Like the KG policy for ranking and selection, the online KG policy is optimal for $N = 1$. This follows from (7) and (8), because*

$$\begin{aligned}
V^{*,N-1}\left(s^{N-1}\right) &= \max_x \left[\mu_x^{N-1} + \mathbb{E}^{N-1} V^{*,N}\left(K^M\left(s^{N-1}, x\right)\right]\right) \\
&= \max_x \mu_x^{N-1} + \mathbb{E}^{N-1}\left(\max_{x'} \mu_{x'}^N\right) \\
&= \mu^{N-1}_{X^{KG,N-1}(s^{N-1})} + \mathbb{E}^{N-1} V^{SL,N}\left(K^M\left(s^{N-1}, X^{KG,N-1}\left(s^{N-1}\right)\right)\right) \\
&= \mu^{N-1}_{X^{KG,N-1}(s^{N-1})} + \mathbb{E}^{N-1} V^{KG,N}\left(K^M\left(s^{N-1}, X^{KG,N-1}\left(s^{N-1}\right)\right)\right) \\
&= V^{KG,N-1}\left(s^{N-1}\right).
\end{aligned}$$

The last measurement is chosen optimally, so if there is only one measurement in the problem, then the online KG algorithm is optimal.

The KG policy is analogous to a class of algorithms in the stochastic control literature known as roll-out policies. These methods choose an action by approximating the value obtained by following some policy after taking the action. For example, the work by Tesauro & Galperin (1996) uses Monte Carlo simulation to approximate the value of the policy in a discrete-state Markov decision process setting. The KG policy can be viewed as a one-step roll-out algorithm in which we take a single action and then follow the SL policy for the rest of the time horizon. Although the state space (the space of all knowledge states) is multi-dimensional and continuous, a one-step look-ahead can be computed exactly, yielding a closed-form decision rule, with no need for simulation-based approximation. This is a strength of the KG approach, in a setting that would otherwise be difficult to handle (because of the continuous state space) using classical dynamic programming techniques.

Much of the traditional bandit literature (e.g. the work on upper confidence bound policies by Lai & Robbins (1985) and Lai (1987)) has focused on index policies, with decision rules of the form $X^{\pi,n}(s^n) = \arg\max_x I_x^\pi(\mu_x^n, \sigma_x^n)$. In an index policy, the index I_x^π used to determine the value of measuring x can only depend on our beliefs μ_x^n, σ_x^n about x, and not on our beliefs about any other alternatives. The KG policy, however, is not an index policy, because the formula for $\nu_x^{KG,n}$ in (13) depends on $\max_{x' \neq x} \mu_{x'}^n$ as well as on μ_x^n. Thus, the theoretical advantages of index policies do not apply to KG; however, in Section 5 we consider an important problem class where index policies are not well-suited, but the KG reasoning still holds.

An expected structural result is that it is better to measure under the KG policy than to not measure at all. More formally, the value obtained by the KG policy is greater than the SL value of (9). The proof is given in the Appendix.

Proposition 3.1. *For any s and any n,*

$$V^{KG,n}(s) \geq V^{SL,n}(s).$$

3.2 Discounted problems

Let us now replace the objective function in (4) with the discounted objective function

$$\sup_\pi \mathbb{E}^\pi \sum_{n=0}^{N} \gamma^n \mu_{X^{\pi,n}(s^n)}$$

where $\gamma \in (0,1)$ is a given parameter and $\mathbb{E}^\pi$ denotes an expectation over the outcomes of the N measurements, given that the measurement decisions are made according to policy π. In this

section, we show the KG decision rule for the discounted problem, for both finite- and infinite-horizon settings. We show that the infinite-horizon KG policy is guaranteed to eventually find the true best alternative in the limit as $\gamma \nearrow 1$.

The knowledge gradient policy for this problem is derived the same way as in Section 3. First, in the discounted setting,

$$V^{SL,n}(s^n) = \frac{1-\gamma^{N-n+1}}{1-\gamma} \max_x \mu_x^n.$$

Then, (10) is computed as

$$\begin{aligned} X^{KG,n}(s^n) &= \arg\max_x \mu_x^n + \gamma \cdot \mathbb{E}^n V^{SL,n+1}\left(K^M(s^n, x)\right) \\ &= \arg\max_x \mu_x^n + \gamma \frac{1-\gamma^{N-n}}{1-\gamma} \nu_x^{KG,n} \end{aligned} \tag{16}$$

where $\nu_x^{KG,n}$ is as in (13). Taking $N \to \infty$, we obtain the infinite-horizon KG rule

$$X^{KG,n}(s^n) = \arg\max_x \mu_x^n + \frac{\gamma}{1-\gamma} \nu_x^{KG,n}. \tag{17}$$

Both (16) and (17) look similar to (14), with a different multiplier in front of the knowledge gradient.

This discussion illustrates the flexibility of the KG approach. We can derive a KG decision rule for both finite and infinite horizons, in both discounted and undiscounted problems. As the discount factor γ increases to 1, we can obtain certain convergence results for the online KG policy. These results are discussed in the next section.

Our paper focuses on a Gaussian learning model because of its generality. In Section 5, we show how KG can be used in problems with multivariate Gaussian priors, allowing us to learn about multiple alternatives from a single measurement. However, it is important to note that the KG approach is not limited to Gaussian models, and in fact represents a general methodology that is applicable to many broad classes of learning problems. To streamline our presentation, we maintain a focus on Gaussian models in the main body of our paper. However, interested readers can see the Appendix for a discussion of how KG can be used in a non-Gaussian setup.

4 Convergence properties of infinite-horizon KG

Our asymptotic analysis of the KG rule in (17) depends on the concept of convergence. We begin by showing that only one alternative can be measured infinitely often by infinite-horizon KG. Thus,

we can say that KG *converges* to x if it measures x infinitely often. All proofs in this section are given in the Appendix.

Proposition 4.1. *For almost every sample path, only one alternative will be measured infinitely often by the infinite-horizon discounted KG policy.*

The particular alternative to which KG converges depends on the sample path, and is not guaranteed to be the true best alternative $\arg\max_x \mu_x$. However, even the Gittins index policy, which is known to be optimal, is not guaranteed to converge to the best alternative either (Brezzi & Lai (2000)). The Gittins policy is optimal in the sense that it learns efficiently, but it is not certain to find the true best alternative.

However, we can establish theoretical guarantees for the KG policy in the limiting case as $\gamma \nearrow 1$. The remainder of this section presents two key results. First, KG achieves an optimal estimate of the true best reward in the limit as $\gamma \nearrow 1$. Second, the probability that KG converges to the true best alternative $\arg\max_x \mu_x$ converges to 1 as $\gamma \nearrow 1$. That is, the convergence behaviour of KG becomes optimal in the limiting case.

Our argument is based on a connection to the ranking and selection problem and the offline KG policy (Gupta & Miescke (1996)), given by

$$X^{Off,n}(s^n) = \arg\max_x \mathbb{E}_x^n \left[\left(\max_{x'} \mu_{x'}^{n+1} \right) - \left(\max_{x'} \mu_{x'}^{n} \right) \right] = \arg\max_x \nu_x^{KG,n}, \tag{18}$$

where $\nu_x^{KG,n}$ is as in (13). Observe that, for large γ, the infinite-horizon online KG rule in (17) becomes similar to (18). If γ is large enough, the effect of μ_x^n in (17) becomes negligible, and the choice of measurement comes to be determined by the KG factor, just as in the offline KG rule. However, the work by Frazier et al. (2008) shows that offline KG is guaranteed to find the true best alternative in an infinite horizon. It stands to reason that online KG should have the same property if $\gamma \to 1$.

Denote by $KG(\gamma)$ the infinite-horizon online KG policy for a fixed discount factor γ. We define the stopping time

$$N_\gamma = \min \left\{ n \geq 0 \,|\, X^{Off,n}(s^n) \neq X^{KG(\gamma),n}(s^n) \right\}$$

to be the first time when the offline and online KG policies choose different alternatives to measure ("disagree"). This time is allowed to be zero, in the event that they choose different alternatives

in the very first time step. In our analysis, we assume without loss of generality that no two alternatives will ever be tied under either policy. This is because the outcome of each measurement is continuous, so the probability that two KG factors or sets of beliefs will be equal as a result of a measurement is zero. Ties can only occur in the early stages, if those particular alternatives are tied under the prior s^0. However, in that case, the ties will disappear after a finite number of measurements, with no effect on the asymptotic behaviour.

Proposition 4.2. *Under the probability measure induced by the distribution of μ and $\hat{\mu}_x^{n+1}$ for all $n \geq 0$ and all x,*

$$\lim_{\gamma \nearrow 1} N_\gamma = \infty \qquad a.s.$$

We next show that, by measuring infinitely many times under the KG policy, we will obtain a better estimate of the value of the best alternative than the estimate at time N_γ. This is an intuitive idea. We already know that we expect our time-$(n+1)$ estimate to be better than our time-n estimate; the next proposition allows us to replace n with the stopping time N_γ.

Proposition 4.3. $\lim_{n\to\infty} \mathbb{E}^{KG(\gamma)} \left(\max_x \mu_x^n\right) \geq \mathbb{E}^{KG(\gamma)} \left(\max_x \mu_x^{N_\gamma}\right)$.

The next step shows that our estimate of the best value at time N_γ becomes accurate in expectation as $\gamma \nearrow 1$. By definition, the online and offline KG policies agree on all measurements up to time N_γ. The proof uses the connection to the offline KG policy.

Proposition 4.4. $\lim_{\gamma \nearrow 1} \mathbb{E}^{KG(\gamma)} \left(\max_x \mu_x^{N_\gamma}\right) = \mathbb{E} \left(\max_x \mu_x\right)$.

We can now state our first key result. As $\gamma \nearrow 1$, the infinite-horizon limit of our estimate of the best value under the KG policy converges to the true value of the best alternative.

Theorem 4.1. $\lim_{\gamma \nearrow 1} \lim_{n\to\infty} \mathbb{E}^{KG(\gamma)} \left(\max_x \mu_x^n\right) = \mathbb{E} \left(\max_x \mu_x\right)$.

In general, the result of Theorem 4.1 does not require convergence to the true best alternative. However, in the specific case of the online KG policy, it can be shown that the probability of the policy converging to a suboptimal alternative vanishes as $\gamma \nearrow 1$. The remainder of this section is dedicated to showing this result.

Let B be the event that $\arg\max_x \mu_x$ is measured infinitely often and denote $P^{KG(\gamma)}(B) = \mathbb{E}^{KG(\gamma)} 1_B$. For notational convenience, let $x_* = \arg\max_x \mu_x$ and $x_*^\gamma = \arg\max_x \mu_x^{N_\gamma}$. As before,

we assume without loss of generality that no ties will occur. Observe that, for fixed γ,

$$\begin{aligned} P^{KG(\gamma)}(B) &= P^{KG(\gamma)}\left(\sum_{n=0}^{\infty} 1_{\{x^n=x_*\}} = \infty\right) \\ &\geq P^{KG(\gamma)}\left(\sum_{n=0}^{\infty} 1_{\{x^n=x_*\}} = \infty,\ x_* = x_*^\gamma\right). \end{aligned}$$

We will continue to place lower bounds on $P^{KG(\gamma)}(B)$, in order to eventually arrive at a lower bound that converges to 1 as $\gamma \nearrow 1$. The next result is one step in this process.

Proposition 4.5. *For fixed γ,*

$$P^{KG(\gamma)}\left(\sum_{n=0}^{\infty} 1_{\{x^n=x_*\}} = \infty,\ x_* = x_*^\gamma\right) \geq P^{KG(\gamma)}\left(x_* = x_*^\gamma,\ \arg\max_x \mu_x^n = x_*^\gamma\ \forall n \geq N_\gamma\right).$$

Now, for every alternative x, define a process B^x as follows. Given $\mathcal{F}^{N_\gamma}$, B^x is a Brownian motion with volatility $\sigma_x^{N_\gamma}$ and initial value $B_0^x = \mu_x^{N_\gamma}$. Furthermore, for any $x \neq y$, B^x and B^y are conditionally independent of each other given $\mathcal{F}^{N_\gamma}$. We interpret B^x as an interpolation of the values of $\mu_x^{N_\gamma+n}$ that we would observe by making $n = 0, 1, ...$ measurements of x starting from time N_γ. In particular, $\mu_x^{N_\gamma+n}$ has the same distribution as $B_{t^n}^x$ where

$$t^n = \frac{Var\left(\mu_x^{N_\gamma+n} \mid \mathcal{F}^{N_\gamma}\right)}{\left(\sigma_x^{N_\gamma}\right)^2}.$$

Observe that the conditional distribution of $\mu_x^{N_\gamma+(n+1)}$, given $\mathcal{F}^{N_\gamma}$ and $\mu_x^{N_\gamma+n}$, is Gaussian with mean $\mu_x^{N_\gamma+n}$ and variance $\left(\tilde{\sigma}_x^{N_\gamma+n}\right)^2$. This is precisely the distribution of $B_{t^{n+1}}^x$ given $B_{t^n}^x$. Furthermore, $\mu_x^{N_\gamma+(n+1)}$ is conditionally independent of $\mu^{N_\gamma+n'}$ for $n' < n$, given $\mathcal{F}^{N_\gamma}$ and $\mu^{N_\gamma+n}$. Thus, the processes $(B_{t^n}^x)_{n=0}^\infty$ and $\left(\mu_x^{N_\gamma+n}\right)_{n=0}^\infty$ are both Markov processes with the same distribution given $\mathcal{F}^{N_\gamma}$. By the continuity of Brownian motion, $\lim_{n\to\infty} \mu_x^n = \mu_x$ corresponds to B_1^x.

Proposition 4.6. *Let $L_x = \min_{0\leq t\leq 1} B_t^x$ and $U_x = \max_{0\leq t\leq 1} B_t^x$. Then,*

$$P^{KG(\gamma)}\left(x_* = x_*^\gamma,\ \arg\max_x \mu_x^n = x_*^\gamma\ \forall n \geq N_\gamma\right) \geq P^{KG(\gamma)}\left(L_{x_*^\gamma} > \max_{x\neq x_*^\gamma} U_x\right).$$

For each x, we can write $B_t^x = B_0^x + \sigma_x^{N_\gamma} W_t^x$, where W^x is a Wiener process. A standard result from stochastic analysis (see e.g. Steele (2000)) tells us that $\max_{0\leq t\leq 1} W_t^x$ has the same distribution as $|W_1^x|$. Analogously, $\min_{0\leq t\leq 1} W_t^x$ has the same distribution as $-|W_1^x|$. Consequently, L_x has the same distribution as $B_0^x - \sigma_x^{N_\gamma}|W_1^x|$, and U_x has the same distribution as $B_0^x + \sigma_x^{N_\gamma}|W_1^x|$. We use these facts to derive the next lower bound.

Proposition 4.7. *Define*

$$h\left(\mu^{N_\gamma}, \sigma^{N_\gamma}\right) = P^{KG(\gamma)}\left(L_{x_*^\gamma} > \max_{x \neq x_*} U_x \,|\, \mathcal{F}^{N_\gamma}\right),$$

where h depends only on μ^{N_γ} and σ^{N_γ}. Then,

$$h\left(\mu^{N_\gamma}, \sigma^{N_\gamma}\right) \geq g\left(\frac{\mu_{x_*^\gamma}^{N_\gamma} - \max_{x \neq x_*^\gamma} \mu_x^{N_\gamma}}{2}, \sigma^{N_\gamma}\right)$$

where $g : \mathbb{R}_{++} \times \mathbb{R}_+^M$ is defined to be

$$g\left(a, b\right) = \prod_{\{x : b_x > 0\}} \left[2\Phi\left(\frac{a}{b_x}\right) - 1\right]$$

and $g\left(a, 0\right) = 1$.

Recall that B is the event that $\arg\max_x \mu_x$ is measured infinitely often. All the elements are now in place to show our second key result, namely that the probability that B occurs under the $KG\left(\gamma\right)$ policy converges to 1 as $\gamma \nearrow 1$.

Theorem 4.2. $\lim_{\gamma \nearrow 1} P^{KG(\gamma)}\left(B\right) = 1$.

Together, Theorems 4.1 and 4.2 add an important detail to our understanding of the online KG policy. From Brezzi & Lai (2000), we know that even the optimal Gittins index policy has a positive probability of converging to a suboptimal alternative for any $\gamma < 1$. However, under the KG policy, this probability vanishes to zero in the limit as $\gamma \nearrow 1$, and our estimate of the best value under KG converges to the true best value.

5 Problems with correlated normal priors

Let us now return to the undiscounted setting, and the objective function from (4). However, we now assume a covariance structure on our prior beliefs about the different alternatives. We now have a multivariate normal prior distribution on the vector $\mu = (\mu_1, ..., \mu_M)$ of true rewards. Initially, we assume that $\mu \sim \mathcal{N}\left(\mu^0, \Sigma^0\right)$, where $\mu^0 = \left(\mu_1^0, ..., \mu_M^0\right)$ is a vector of our beliefs about the mean rewards, and Σ^0 is an $M \times M$ matrix representing the covariance structure of our beliefs about the true mean rewards. As before, if we choose to measure alternative x at time n, we observe a random reward $\hat{\mu}_x^n \sim \mathcal{N}\left(\mu_x, \sigma_\varepsilon^2\right)$. Conditioned on $\mu_1, ..., \mu_M$, the rewards we collect are independent of each

other. After n measurements, our beliefs about the mean rewards are expressed by a vector μ^n and a matrix Σ^n, representing the conditional expectation and conditional covariance matrix of the true rewards given $\mathcal{F}^n$.

The updating equations, given by (1) and (2) in the uncorrelated case, now become

$$\mu^{n+1} = \mu^n + \frac{\hat{\mu}^{n+1}_{x^n} - \mu^n_{x^n}}{\sigma^2_\varepsilon + \Sigma^n_{x^n x^n}} \Sigma^n e_{x^n} \tag{19}$$

$$\Sigma^{n+1} = \Sigma^n - \frac{\Sigma^n e_{x^n} e^T_{x^n} \Sigma^n}{\sigma^2_\varepsilon + \Sigma^n_{x^n x^n}} \tag{20}$$

where $x^n \in \{1, ..., M\}$ is the alternative chosen at time n, and e_{x^n} is a vector with 1 at index x^n, and zeros everywhere else. Note that a single measurement now leads us to update the entire vector μ^n, not just one component as in the uncorrelated case. Furthermore, (3) now becomes a vector equation

$$\mu^{n+1} = \mu^n + \tilde{\sigma}^{corr,n}(x^n) \cdot Z$$

where Z is standard Gaussian and

$$\tilde{\sigma}^{corr,n}(x^n) = \frac{\Sigma^n e_{x^n}}{\sqrt{\sigma^2_\varepsilon + \Sigma^n_{x^n x^n}}}.$$

The SL policy, which we follow if we are unable to continue learning after time n, is still given by (9). The derivation of the online KG policy remains the same. However, the formula for computing $\nu^{KG,n}$ in (13) no longer applies. In the correlated setting, we have

$$\mathbb{E}^n_x \max_{x'} \mu^{n+1}_{x'} = \mathbb{E}^n \left[\max_{x'} \left(\mu^n_{x'} + \tilde{\sigma}^{corr,n}_{x'}(x) \cdot Z \right) \right].$$

We are computing the expected value of the maximum of a finite number of piecewise linear functions of Z. Let

$$\nu^{KGC,n}_x = \mathbb{E}^n_x \left[\left(\max_{x'} \mu^{n+1}_{x'} \right) - \left(\max_{x'} \mu^n_{x'} \right) \right]$$

be the analog of (12) in the correlated setting. From the work by Frazier et al. (2009), it is known that

$$\nu^{KGC,n}_x = \sum_{y \in A} \left(\tilde{\sigma}^{corr,n}_{y+1}(x) - \tilde{\sigma}^{corr,n}_y(x) \right) f(-|c_y|) \tag{21}$$

where A is the set of all alternatives y for which we can find numbers $c_{y-1} < c_y$ for which $y = \arg\max_{x'} \mu^n_{x'} + \tilde{\sigma}^{corr,n}_{x'}(x) \cdot z$ for $z \in (c_{y-1}, c_y)$, with ties broken by the largest-index rule. These

quantities c_y are also used in (21). We number the alternatives in the set A in order of increasing $\tilde{\sigma}_y^{corr,n}$. The function f is as in (13).

The online KG decision rule for the correlated case is given by

$$X^{KGC,n}(s^n) = \arg\max_x \mu_x^n + (N-n)\,\nu_x^{KGC,n}. \tag{22}$$

If we introduce a discount factor into the problem, the decision rule becomes as in (16) or (17), using ν^{KGC} instead of ν^{KG}. An algorithm for computing ν^{KGC} exactly is presented in Frazier et al. (2009), and can be used to solve this decision problem. The computational complexity of the algorithm is $\mathcal{O}\left(M^2 \log M\right)$, but the following result (see the Appendix for the proof) allows us to reduce the computation time.

Proposition 5.1. *Let s^n be the knowledge state at time n. If alternative x satisfies the inequality*

$$\mu_x^n + (N-n)\,\frac{1}{2\pi}\max_{x'} \tilde{\sigma}_{x'}^{corr,n}(x) < \max_{x'} \mu_{x'}^n, \tag{23}$$

then alternative x will not be chosen by the KG policy at time n.

The significance of Proposition 5.1 is practical. We require $\mathcal{O}\left(M^2\right)$ operations to ascertain whether or not (23) holds for every alternative. Let $x_1, ..., x_K$ represent the alternatives for which (23) does not hold at time n (where K is the total number of such alternatives). Then we can define a matrix A^n of size $M \times K$ by

$$A^n = \left[e_{x_1} \quad ... \quad e_{x_K}\right].$$

The time-n marginal distribution of $(\mu_{x_1}, ..., \mu_{x_K})$ is Gaussian with mean vector $(A^n)^T \mu^n$ and covariance matrix $(A^n)^T \Sigma^n A^n$. As a consequence of Proposition 5.1, the KG decision rule in (22) can be rewritten as

$$X^{KGC,n}(s^n) = \arg\max_k \mu_{x_k}^n + (N-n)\,\nu_{x_k}^{KGC,n}$$

where $\nu^{KGC,n}$ can be computed by running the correlated KG algorithm from Frazier et al. (2009) on the reduced choice set $\{x_1, ..., x_K\}$ with the marginal mean vector and covariance matrix given above. Typically, in practice, K is close to M for small values of n, but becomes dramatically smaller as n increases. Consequently, we only need to compute KG factors for a choice set whose size can be much smaller than M.

6 Computational experiments: independent beliefs

Our experimental study presents evidence that online KG is competitive against the best known approximation to the optimal Gittins policy on classic multi-armed bandit problems (no correlations). At the same time, we single out key parameters that may cause KG to perform less efficiently for certain values. In Section 7, we also consider the correlated case, and show that KG outperforms many well-known index policies in that setting.

The performance measure that we use to evaluate a policy is the opportunity cost. For a learning policy π, the opportunity cost for a discounted problem is given by

$$C^{\pi} = \sum_{n=0}^{N} \gamma^n \left[\left(\max_x \mu_x \right) - \mu_{X^{\pi,n}(s^n)} \right].$$

To obtain an accurate assessment of the quality of a policy, we calculate opportunity costs using the true values μ. However, in order to do this, we must know what the true values are. Thus, we test a policy by first fixing a particular truth μ, then evaluating the ability of the policy to find that truth. For this reason, the starting data for all our experiments were randomly generated.

For two policies π_1 and π_2, the difference

$$C^{\pi_2} - C^{\pi_1} = \sum_{n=0}^{N} \gamma^n \left(\mu_{X^{\pi_1,n}(s^n)} - \mu_{X^{\pi_2,n}(s^n)} \right) \tag{24}$$

gives us the amount by which π_1 outperformed (or was outperformed by) π_2. For a given set of initial data, we run each policy 10^4 times, thus obtaining 10^4 samples of the opportunity cost. We then divide the 10^4 sample paths into groups of 500 in order to obtain approximately normal samples of opportunity cost and the standard errors of those averages. The standard error of the difference in (24) is the square root of the sum of the squared standard errors of C^{π_1}, C^{π_2}.

In the classic multi-armed bandit problem, with $N \to \infty$ and $0 < \gamma < 1$, there is a clear, natural competitor for KG in the form of the optimal Gittins policy (Gittins (1989)). The Gittins decision rule is given by

$$X^{Gitt,n}(s^n) = \arg\max_x \Gamma\left(\mu_x^n, \sigma_x^n, \sigma_\varepsilon, \gamma\right) \tag{25}$$

where $\Gamma\left(\mu_x^n, \sigma_x^n, \sigma_\varepsilon, \gamma\right)$ is the Gittins index based on our current beliefs about an alternative, the measurement error, and the discount factor γ. To simplify the computation of Gittins indices, we

use the identity

$$\Gamma\left(\mu_x^n, \sigma_x^n, \sigma_\varepsilon, \gamma\right) = \mu_x^n + \sigma_\varepsilon \cdot \Gamma\left(0, \frac{\sigma_x^n}{\sigma_\varepsilon}, 1, \gamma\right).$$

From Brezzi & Lai (2002), we know that

$$\Gamma\left(0, s, 1, \gamma\right) = \sqrt{-\log\gamma} \cdot b\left(-\frac{s^2}{\log\gamma}\right)$$

where the function b must be approximated. The current state of the art in Gittins approximation is the work by Chick & Gans (2009), which builds on Brezzi & Lai (2002) and Yao (2006). It is shown that $b \approx \tilde{b}$ where

$$\tilde{b}\left(s\right) = \begin{cases} \frac{s}{\sqrt{2}} & s \leq \frac{1}{7} \\ e^{-0.02645(\log s)^2 + 0.89106 \log s - 0.4873} & \frac{1}{7} < s \leq 100 \\ \sqrt{s}\left(2\log s - \log\log s - \log 16\pi\right)^{\frac{1}{2}} & s > 100. \end{cases}$$

Thus, the approximation to (25) is given by

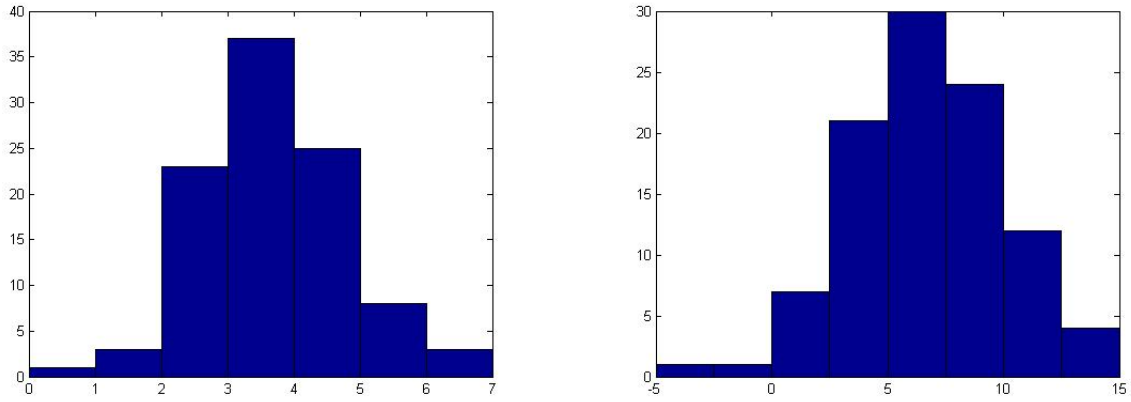
$$X^{Gitt,n}\left(s^n\right) \approx \arg\max_x \mu_x^n + \sigma_\varepsilon \sqrt{-\log\gamma} \cdot \tilde{b}\left(-\frac{\left(\sigma_x^n\right)^2}{\sigma_\varepsilon^2 \log\gamma}\right). \tag{26}$$

For many learning problems, it is more difficult to approximate Gittins indices when γ is close to 1. However, the particular approximation $\tilde{b}$ given above uses a better fit for the range $\frac{1}{7} < s \leq 100$ compared to previous approximations. As the posterior variance decreases, this range will be exercised further on in the time horizon when γ is large.

We compared the infinite-horizon discounted online KG rule from (17) against the Gittins approximation in (26). The remainder of this section describes the methodology and results of this comparison.

6.1 Effect of prior structure on KG performance

We first consider a set of experiments where our modeling assumption $\mu_x \sim \mathcal{N}\left(\mu_x^0, \left(\sigma_x^0\right)^2\right)$ is satisfied. These experiments are referred to as *truth-from-prior* experiments. We generated 100 problems with $M = 100$, where $\sigma_x^0 = 10$ for all x and each μ_x^0 is a random sample from the distribution $\mathcal{N}\left(0, 100\right)$. We followed Vermorel & Mohri (2005) in using centered Gaussian distributions to generate the initial data. The measurement noise was chosen to be $\sigma_\varepsilon = 10$, and the discount factor was chosen to be $\gamma = 0.9$.



(a) (b)

Figure 1: Histograms of the sampled difference in opportunity cost between KG and Gittins across (a) 100 truth-from-prior experiments, and (b) 100 equal-prior experiments.

For every experiment, we ran the KG and Gittins policies on 10^4 different sample paths. In every sample path, the truths μ_x are generated from the prior distribution corresponding to the given experiment. The outcomes of the measurements are then generated from those truths. Thus, for each problem out of our set of 100, we have a sample of the difference $C^{Gitt} - C^{KG}$, averaged over 10^4 different truths. Figure 1(a) shows the distribution of these differences across 100 problems. Bars to the right of zero indicate that KG outperformed the Gittins approximation, and bars to the left of zero indicate the converse. KG outperformed the Gittins approximation on every problem. The average margin of victory was 3.6849, with average standard error 0.7215.

The victory of KG in Figure 1(a) is due to the fact that we are using an approximation of Gittins indices. While the margin of victory is small, it indicates that the approximation is not completely accurate. We can see from these results that KG is a worthwhile alternative to the best known approximation of the optimal policy.

We also consider a situation in which the main modeling assumptions do not hold. In the *equal-prior* experiments, we let $\mu_x^0 = 0$ and $\sigma_x^0 = 10$ for every x. The true values μ_x come from a uniform distribution on the interval $[-30, 30]$. A new set of truths is generated for each experiment, but not for each sample path. Each equal-prior problem has a fixed truth, and we run the KG and Gittins policies on 10^4 outcomes of the measurements. This represents a situation that is common in real-world applications: we do not have much information about the true values, and our prior only gives us a general range of values for the truths, without telling us anything about

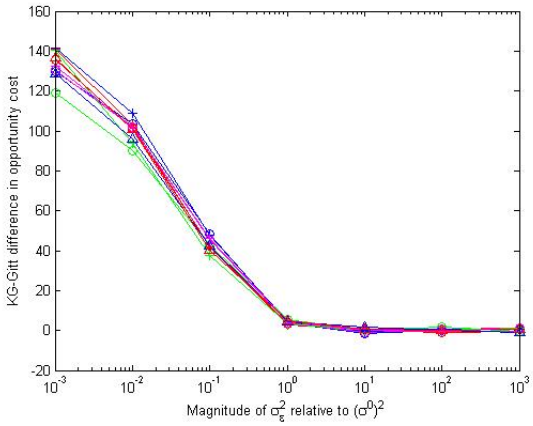


Figure 2: Effect of the measurement noise σ_ε^2 on the performance of KG relative to Gittins.

which alternative is better. Figure 1(b) shows the results of the comparison. The average margin of victory of KG is 6.7885, with average standard error 0.8368.

Since μ does not come from the prior distribution in the equal-prior experiments, the Gittins policy loses its optimality properties. While there are no theoretical guarantees in a situation where the main modeling assumption is violated, Figure 1(b) suggests that the KG heuristic may retain its practical usefulness in situations where we are not certain that the truths really do come from the prior distribution.

6.2 Effect of σ_ε^2 and γ on KG performance

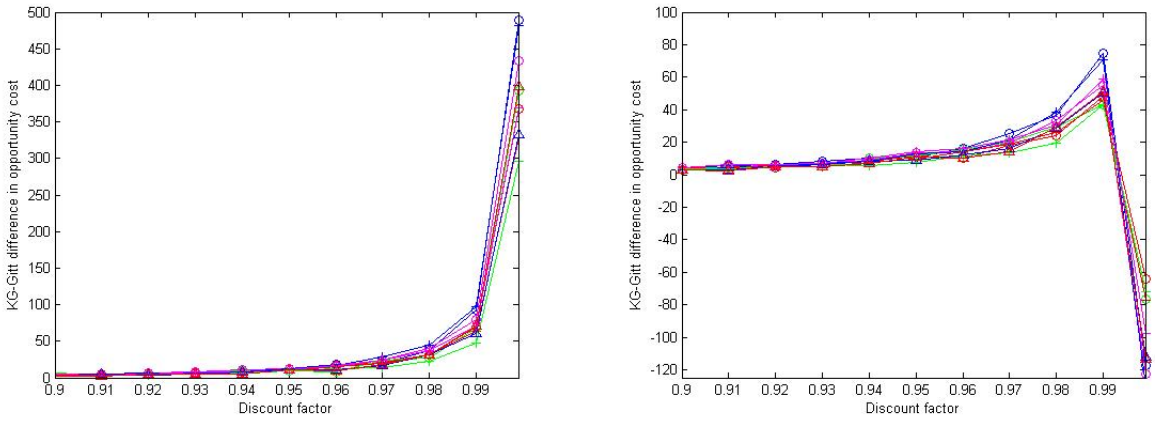
Two key parameters in the online problem are the measurement noise σ_ε^2 and the discount factor γ. We varied these parameters in ten randomly chosen truth-from-prior problems, that is, we considered ten different sets of initial priors. For each parameter value in each problem, we simulated KG and Gittins across 10^4 truths generated from the initial prior (as in Figure 1(a)).

Figure 2 shows the effect of measurement noise on performance. We varied σ_ϵ^2 relative to the fixed prior variance $\left(\sigma^0\right)^2 = 100$. For instance, the point 10^0 on the horizontal axis of Figure 2 indicates that $\sigma_\varepsilon^2 = \left(\sigma^0\right)^2$, the point 10^{-1} indicates that $\sigma_\varepsilon^2 = 0.1 \cdot \left(\sigma^0\right)^2$, and so on. Points to the left of 10^0 represent situations in which the measurement noise is smaller than the prior variance, enabling us to come close to the true value of an alternative in relatively few measurements. Each line in Figure 2 corresponds to one of the ten truth-from-prior problems considered; we do not label the individual lines, since they all exhibit the same behaviour.

We see that the Gittins approximation performs poorly compared to KG for low measurement noise. As σ_ε^2 increases, the KG policy's margin of victory shrinks. However, the Gittins policy also becomes less effective when the measurement noise gets too high. We see that, for very large values of σ_ε^2, the difference $C^{Gitt} - C^{KG}$ goes to zero for all ten problems under consideration.

A different relationship holds for the discount factor. For large values of γ, we see a distinction between the short-term and long-term performance of KG. Figure 3(a) compares KG to Gittins with a time horizon of $N = 150$, and Figure 3(b) shows the results with the time horizon chosen to be large enough for $\gamma^N < 10^{-3}$. We see that, in both cases, KG and Gittins perform comparably well for $\gamma = 0.9$, and KG begins to significantly outperform Gittins for values of γ up to 0.99. This lead of KG over Gittins is preserved in the infinite-horizon case. However, for larger values ($\gamma = 0.999$), we see that Gittins catches up and significantly outperforms KG in the long run, although KG does much better in the first 150 iterations. This result suggests that the KG policy may perform especially well in problems with relatively short time horizons, where the budget is too small for Gittins to overtake KG. Our study of correlated problems in the next section explores this issue further.

Our analysis reveals some of the strengths and weaknesses of the KG policy. First, KG does much better than approximate Gittins if σ_ε^2 is low, and continues to match Gittins as the measurement noise increases. Second, KG is significantly outperformed by Gittins for $\gamma = 0.999$, for the set of parameters we chose. However, for moderately high values of the discount factor such as 0.99,



(a) (b)

Figure 3: Effect of the discount factor γ on the performance of KG relative to Gittins for (a) $N = 150$ and (b) an infinite horizon.

KG achieves a significant lead over approximate Gittins. For many instances of the classic bandit problem, KG is competitive against the best known Gittins approximation.

7 Computational experiments: correlated beliefs

This section describes the experiments we conducted on problems with correlated beliefs. In Section 7.1, we explain the setup of the experiments and present the main results. The remainder of the section studies particular aspects of correlated problems, such as the effect of correlation on performance and the benefits obtained from incorporating correlations into the KG decision rule.

7.1 Setup and main results

The class of correlated online problems is very large. We tested KG on a subset of these problems, in which the prior covariances are given by the power-exponential rule:

$$\Sigma^0_{ij} = 100 \cdot e^{-\lambda(i-j)^2}, \tag{27}$$

where λ is a constant. An example of a problem where this covariance structure can be used is the problem of learning a scalar function, where the covariance between i and j is smaller when i and j are farther apart. We used this covariance structure together with the values of μ^0 generated for the truth-from-prior experiments in Section 6. The true values were taken from the prior distribution $\mathcal{N}\left(\mu^0, \Sigma^0\right)$. In every problem, a new truth was generated in each of 10^4 sample paths, allowing us to compare learning policies in expectation over the entire prior distribution. The parameter λ was set to 0.01.

Because there is no known optimal policy for correlated bandit problems (to our knowledge, KG is the first policy to be proposed for this problem), our experiments for the correlated case assume a finite horizon of $N = 50$ with $\gamma = 1$. In this setting, a convenient measure of performance is the difference in average opportunity cost

$$C^{\pi_2} - C^{\pi_1} = \frac{1}{N+1} \sum_{n=0}^{N} \mu_{X^{\pi_1,n}(s^n)} - \mu_{X^{\pi_2,n}(s^n)}, \tag{28}$$

which represents the amount by which π_1 outperformed or was outperformed by π_2 on average in each time step. We compared the correlated KG rule given in (22) to several representative index policies, as well as an approximate two-step look-ahead. We briefly describe the implementation of the competing policies.

Approximate two-step look-ahead (2Step). A natural choice of competition for the KG policy, which looks ahead one time step into the future, is a multi-step look-ahead. Such a policy, however, is much more difficult to implement and compute than the KG policy. The decision rule for the one-step look-ahead can be computed exactly using (22), whereas there is no known closed-form expression for a multi-step look-ahead rule.

We approximated a two-step look-ahead policy in the following manner. The outcome of the measurement in the first step was discretized into K branches by dividing the conditional distribution of the measurement into $K+1$ intervals of equal probability. In the second step, KG factors were computed based on the new beliefs resulting from the outcome on each branch. Thus, in order to make a single decision at time n, we must compute a total of $K \cdot M$ correlated KG factors. The computational complexity of this procedure is $\mathcal{O}\left(KM^2 \log M\right)$, which is already noticeably costly for $M = 100$ alternatives. More generally, an approximate d-step look-ahead would have complexity $\mathcal{O}\left(K^{d-1}M^2 \log M\right)$ per decision, making it prohibitive to roll out for more than two time steps.

In our experiments, we used $K = 10$. The accuracy of the approximation can be improved by increasing K, however this adds greatly to the computational cost. By contrast, the KG policy can be computed exactly, given extremely precise approximations of the Gaussian cdf.

Approximate Gittins (Gitt). We use the Gittins approximation from (26) with $\sigma_x^n = \sqrt{\Sigma_{xx}^n}$. Gittins indices do not retain their optimality properties in the correlated setting. We use this policy as a heuristic and treat γ as a tunable parameter. In our experiments, $\gamma = 0.9$ yielded the best performance.

Interval estimation (IE). The IE decision rule by Kaelbling (1993) is given by

$$X^{IE,n}(s^n) = \arg\max_x \mu_x^n + z_{\alpha/2} \cdot \sigma_x^n,$$

where $z_{\alpha/2}$ is a tunable parameter. We found that $z_{\alpha/2} = 1.5$ yielded the best performance on average across 100 problems. However, the IE policy is sensitive to the choice of tuning parameter. We discuss this issue in Section 7.3.

Upper confidence bound (UCB). The UCB decision rule by Lai (1987) is given by

$$X^{UCB,n}(s^n) = \mu_x^n + \sqrt{\frac{2}{N_x^n} g\left(\frac{N_x^n}{N}\right)}$$

where N_x^n is the number of times x has been measured up to and including time n, and

$$g(t) = \log\frac{1}{t} - \frac{1}{2}\log\log\frac{1}{t} - \frac{1}{2}\log 16\pi.$$

UCB1-Normal (UCB1). The study by Auer et al. (2002) proposes a different UCB-style policy for problems with Gaussian rewards. The UCB1 decision rule is given by

$$X^{UCB1,n}(s^n) = \mu_x^n + 4\sigma_\varepsilon\sqrt{\frac{\log n}{N_x^n}}. \tag{29}$$

The original presentation of the policy uses a frequentist estimate of the measurement noise σ_ε. Because we assume that this quantity is known, we can simplify the decision rule, resulting in an interesting parallel to (26), where the uncertainty bonus also has σ_ε out in front. We can improve performance by treating the coefficient 4 in the UCB1 decision rule as a tunable parameter; we found that a value of 0.5 produced the best results for the problems we considered.

Note also that the quantity $\frac{\sigma_\varepsilon}{\sqrt{N_x^n}}$ in (29) can be viewed as a frequentist analog of the posterior variance σ_x^n. In fact, if we begin with a non-informative prior on x (with $\sigma_x^0 = \infty$), then $\sigma_x^n = \frac{\sigma_\varepsilon}{\sqrt{N_x^n}}$ exactly. We considered a version of the policy with the decision rule $X^{UCB1,n}(s^n) = \arg\max_x \mu_x^n + 4\sigma_x^n\sqrt{\log n}$, but found that this modification did not substantially change the policy's performance.

Pure exploitation (Exp). This decision rule is given by $X^{Exp,n}(s^n) = \arg\max_x \mu_x^n$. It has no uncertainty bonus and no tunable parameters.

Table 1 gives the means and average standard errors of our estimates of (28) for each relevant comparison. As before, positive numbers indicate that KG outperformed the other policy in the comparison, and negative numbers indicate the converse. We see that, on average, KG outperformed approximate Gittins indices, UCB, UCB1, and pure exploitation by a statistically significant amount. The 2Step policy yielded virtually the same results as KG. Interval estimation slightly outperformed KG on average, but the margin of victory was not statistically significant. In Section 7.3, we address the issue of the sensitivity of IE to its tunable parameter.

Figure 4 shows the distribution of the sampled differences in opportunity cost across 100 truth-from-prior problems. We see that KG outperforms IE 23/100 times, and usually loses by a small

	KG-2Step	KG-Gitt	KG-IE	KG-UCB	KG-UCB1	KGC-Exp
Mean	0.0599	0.7076	-0.0912	44.4305	1.2091	5.5413
Avg. SE	0.0375	0.0997	0.0857	0.6324	0.1020	0.1511

Table 1: Means and standard errors for the correlated experiments.

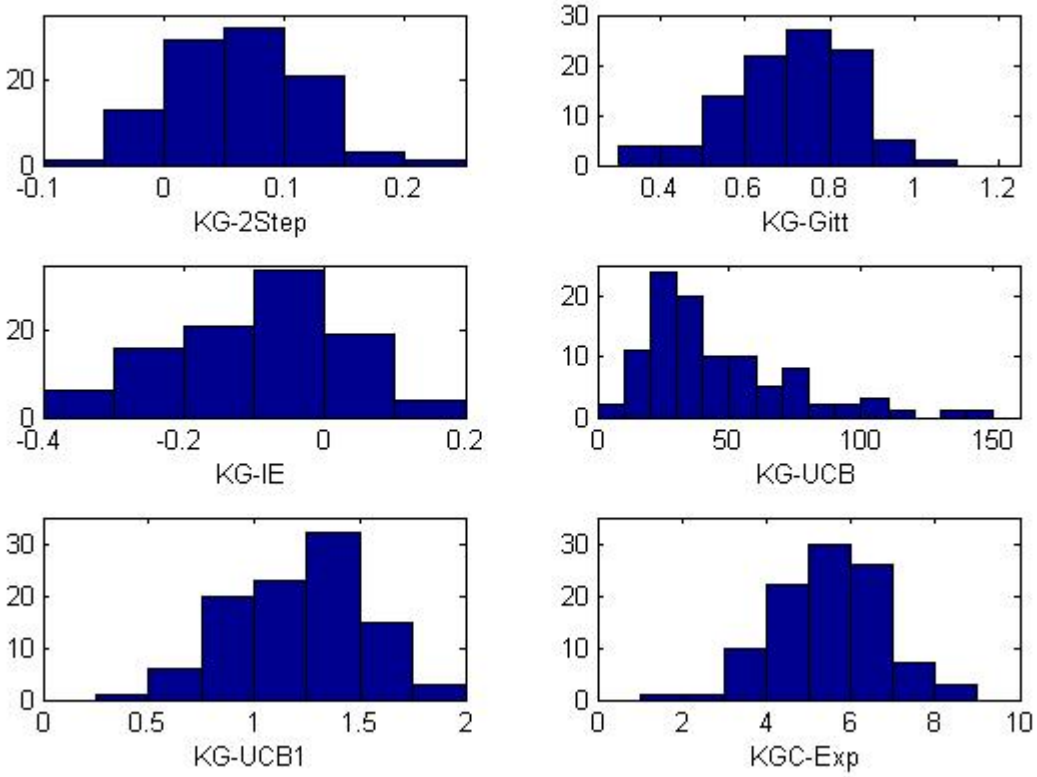


Figure 4: Histograms of the sampled difference in opportunity cost between KG and other policies across 100 correlated truth-from-prior problems.

margin when $z_{\alpha/2}$ is carefully tuned. The 2Step policy outperformed KG 14/100 times, and was outperformed by KG the remaining times. However, the differences are not statistically significant in most cases. The reason why we do not see a significant advantage to using the 2Step policy is because we are required to approximate the two-step look-ahead, whereas the one-step look-ahead used by the KG policy can be computed exactly.

All other policies are outperformed by KG in every experiment. In particular, the UCB policy displays a very large positive tail. This policy suffers especially, compared to the other index policies, because it does not have any tunable parameters, and cannot be tweaked to yield better results in the correlated case. The UCB1 policy yields better performance once tuned, but is nonetheless outperformed by both KG and interval estimation. In the case of IE, the tunable parameter $z_{\alpha/2}$ can be adjusted to make the policy yield good performance. However, we observed that the performance of IE was very sensitive to the choice of tuning parameter (discussed further down in Section 7.3). In a large problem where the distribution of the rewards is not obvious, it

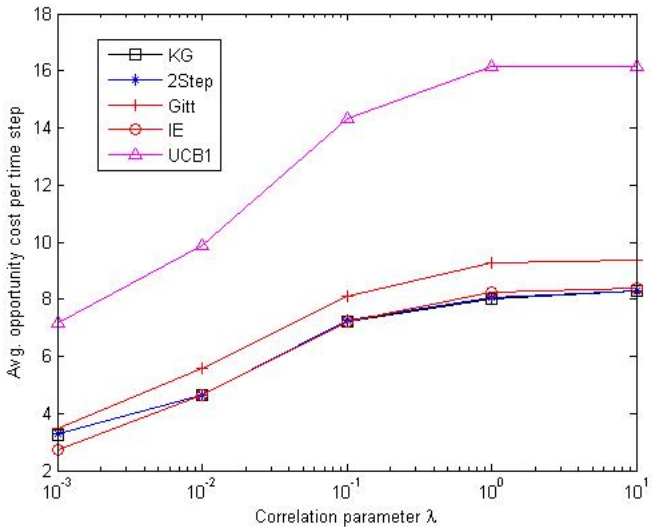


Figure 5: Opportunity cost as a function of the correlation parameter λ.

may be difficult to tune this policy sufficiently.

7.2 Effect of correlation on KG performance

We varied the correlation parameter λ for a single randomly chosen problem out of the truth-from-prior set. Figure 5 shows the effect of λ on the performance of different measurement policies. Pure exploitation and UCB are omitted from the figure, because they were found to significantly underperform all other policies for each value of λ considered. We see that the relative performance of the remaining policies stays roughly the same as before as λ is varied. The 2Step policy continues to yield virtually the same performance as KG. The tuned IE policy performs comparably to KG overall, yielding slightly better results for low λ.

Generally, all policies tend to do better when all the alternatives are heavily correlated (this occurs for low values of λ). In this case, a single measurement will reveal a great deal of information about all the alternatives, which means that we can quickly get a sense of the best value by measuring almost any alternative. However, even in this setting, KG and IE are able to learn more efficiently than approximate Gittins or UCB1.

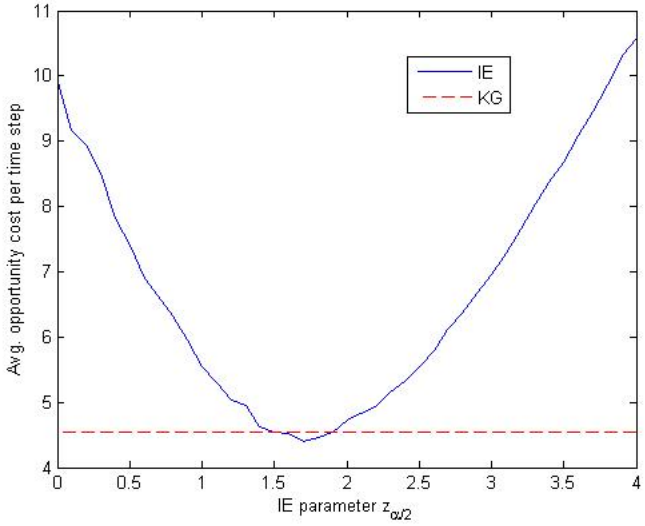


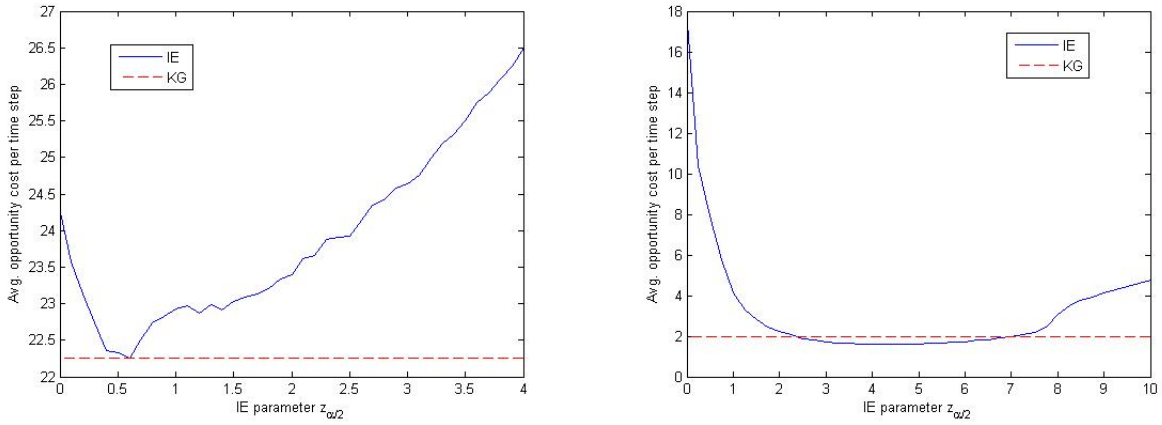
Figure 6: Sensitivity of IE to the tuning parameter $z_{\alpha/2}$ in a truth-from-prior problem.

7.3 Sensitivity of interval estimation

Our experiments show that a properly tuned interval estimation policy can work quite well, even for problems with correlated beliefs. Since IE is particularly easy to implement, it is worth addressing the robustness of the tuning parameter $z_{\alpha/2}$. We find that the process of tuning seems to capture quite a bit of information about the function.

Figure 6 shows how the performance of IE varies for different values of $z_{\alpha/2}$ in the same truth-from-prior problem that we examined in Section 7.2. The best value of $z_{\alpha/2}$ is about 1.7, but the performance is quite sensitive to this parameter and deteriorates quickly as we move away from the optimal value. Furthermore, the best value of $z_{\alpha/2}$ is highly problem-dependent. Figure 7 gives two examples of problems where the best value of the tuning parameter is very different from the truth-from-prior example. Figure 7(a) shows the sensitivity of IE on one of the equal-prior problems from Section 6, with the addition of a power-exponential covariance structure. We see that the best value of $z_{\alpha/2}$ is 0.6; a value of 1.7 yields much worse performance.

Figure 7(b) shows the sensitivity of IE in the *sine-truth* problem, where $\mu_x = -35 \sin\left(\frac{0.3}{\pi}x\right)$ for $x \in \{1, 2, ..., 100\}$. In this problem, the true values have a single peak with a value of 35 around $x = 50$, two smaller local maxima at 0 and 100, and two minima with values of -35. The prior means are set to 0 for all alternatives, halfway between the smallest and largest truths, and Σ^0 is given by a power-exponential structure with $\lambda = 0.01$ and $\Sigma^0_{xx} = 25$ for all x. The measurement



(a) Equal-prior problem.

(b) Sine-truth problem.

Figure 7: Sensitivity of IE to the tuning parameter $z_{\alpha/2}$ in other problems.

noise is set to $\sigma_\varepsilon^2 = 25$. The smallest and largest truth are very far apart in this problem, and the prior does not provide any information about the structure of the truth. In this situation, we see that there is a range of values of $z_{\alpha/2}$ that provide comparable performance to KG, with the best value being 4.5. However, the value 1.7 is not in this range, and causes IE to be outperformed by KG. Furthermore, the range of $z_{\alpha/2}$ values for which IE performs well can be shifted to the right by increasing the amplitude of the sine wave, making it even more difficult to tune IE when the truth is unknown.

The literature contains other examples of problems with different optimal values of $z_{\alpha/2}$. In the offline problem studied by Frazier et al. (2008), the best value is 3.1. In the particular online problems considered by Ryzhov & Powell (2009*a*) and Ryzhov & Powell (2009*b*), the best values of $z_{\alpha/2}$ are 1 and 0.75, respectively. Clearly, there is no one value of $z_{\alpha/2}$ that always works well.

The sensitivity of IE to the choice of $z_{\alpha/2}$ is a weakness of the IE policy. Although it can be tuned to perform equivalently to correlated KG, the range of values of $z_{\alpha/2}$ that yield good performance is relatively small. Furthermore, the optimal range may change drastically depending on the problem. We have presented examples of problems where the best values of $z_{\alpha/2}$ are 0.6, 1.7, and 4.5 respectively. Each of these values yields good performance in one problem and poor performance in the other two. In light of this issue, we can conclude that correlated KG has one attractive advantage over IE: it requires no tuning at all, while yielding comparable performance to a finely-tuned IE policy.

	Truth-from-prior	Equal-prior	Sine-truth	Quadratic-truth
Mean	0.0517	0.6124	1.2161	258.5264
Avg. SE	0.0914	0.0719	0.0233	4.3547

Table 2: Difference between correlated and independent KG decision rules for different problem settings.

7.4 Comparison to independent KG decision rule

Finally, we consider the question of whether the correlated KG decision rule given in (22) is able to offer substantial improvement, in a correlated problem, over the independent KG decision rule given in (14), with (13) used to compute the KG factor. In other words, this is the issue of how much KG gains by incorporating covariances directly into the decision rule. Table 2 shows the average difference in opportunity cost between the correlated and independent KG policies for four distinct problem settings: the truth-from-prior problem considered in Section 7.2, the equal-prior and sine-truth problems considered in Section 7.3, and the *quadratic-truth problem*, where $\mu_x = -x^2 + 101x - 100$ for $x \in \{1, 2, ..., 100\}$. Like the other three problems, the quadratic-truth problem uses a power-exponential covariance structure with $\lambda = 0.01$ and $\Sigma^n_{xx} = 500^2$. The measurement noise is $\sigma^2_\varepsilon = 300^2$.

We see that correlated KG outperforms independent KG in all four settings. However, the margin of victory in the truth-from-prior problem is not statistically significant (meaning that correlated and independent KG yield similar performance). When the priors start out equal, however, correlated KG offers significant improvement. When the truth has a specific structure, as in the sine-truth and quadratic-truth problems, the improvement offered by correlated KG becomes even more dramatic.

We conclude that the value added by the correlated KG decision rule over regular KG is problem-dependent. Recall that the independent KG rule is itself a non-index policy that considers the estimates μ^n_x relative to each other when making a decision. In a truth-from-prior setting, where the prior is fairly accurate from the very beginning, this examination of the relative magnitudes of μ^n_x can capture enough information about the relationships between the alternatives to allow us to obtain reasonably good performance with the independent KG policy (at a lower computational cost than correlated KG). However, if the prior contains less information about the truth, as in the equal-prior setting, it becomes more important to consider covariances when making a decision. Furthermore, if the truth happens to have more structure (e.g. if we are trying to find the maximum

of a continuous function), it is worth paying the additional computational cost required to use the correlated KG rule.

8 Conclusion

We have proposed a new type of decision rule for online learning problems, which can be used for finite or infinite horizons. In contrast with the Gittins index policy, which looks at one alternative at a time over an infinite horizon, the knowledge gradient considers all alternatives at once, but only looks one time period into the future. There is an explicit expression for the value of information gained in a single time step, resulting in an easily computable decision rule for the KG policy. In the classic bandit setting (infinite-horizon discounted), the probability that KG finds the best alternative converges to 1 as the discount factor approaches 1. Experiments show that KG performs competitively against the best known approximation to the optimal Gittins index policy.

One major advantage of the KG method is its ability to handle problems with correlated beliefs. Index policies are inherently unable to do this, because they depend on the ability to consider each alternative separately from the others. The non-index nature of KG allows it to incorporate the effects of correlation into the computation of the KG factor. Experiments show that KG outperforms or is competitive against a number of index policies from the traditional bandit literature on problems with correlations. To our knowledge, KG is the first learning policy that is able to consider a multivariate Gaussian prior while making decisions. We believe that KG represents an important step in the study of problems with correlated beliefs, while remaining a worthwhile alternative to the index policy approach in the traditional multi-armed bandit setting.

The empirical conclusions regarding the performance of different policies reflect, of course, the specific experiments we chose to run. It is not possible to generate every variation, and further research is needed to compare these policies in the context of different problems. However, we feel that the experiments reported here are encouraging, and suggest that other researchers should consider using the knowledge gradient as a potential alternative to Gittins indices and other index policies.

Acknowledgments

The authors thank Lawrence Manning for his assistance with the computational aspect of this study. We are also grateful to the Area Editor, Associate Editor, and two reviewers for their thorough reading of the paper and helpful comments. This research was supported in part by AFOSR contract FA9550-08-1-0195 and ONR contract N00014-07-1-0150 through the Center for Dynamic Data Analysis.

References

Auer, P., Cesa-Bianchi, N. & Fischer, P. (2002), 'Finite-time analysis of the multiarmed bandit problem', *Machine Learning* **47**(2-3), 235–256.

Bertsimas, D. & Nino-Mora, J. (2000), 'Restless bandits, linear programming relaxations, and a primal-dual index heuristic', *Operations Research* **48**(1), 80–90.

Brezzi, M. & Lai, T. (2000), 'Incomplete learning from endogenous data in dynamic allocation', *Econometrica* **68**(6), 1511–1516.

Brezzi, M. & Lai, T. (2002), 'Optimal learning and experimentation in bandit problems', *Journal of Economic Dynamics and Control* **27**(1), 87–108.

Chick, S. & Gans, N. (2009), 'Economic analysis of simulation selection problems', *Management Science* **55**(3), 421–437.

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

DeGroot, M. H. (1970), *Optimal Statistical Decisions*, John Wiley and Sons.

Duff, M. (1995), 'Q-learning for bandit problems', *Proceedings of the 12th International Conference on Machine Learning* pp. 209–217.

Feldman, D. (1962), 'Contributions to the Two-Armed Bandit Problem', *The Annals of Mathematical Statistics* **33**(3), 847–856.

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

Frazier, P. I., Powell, W. B. & Dayanik, S. (2009), 'The knowledge-gradient policy for correlated normal rewards', *INFORMS J. on Computing* **21**(4), 599–613.

Ginebra, J. & Clayton, M. (1995), 'Response surface bandits', *Journal of the Royal Statistical Society* **B57**(4), 771–784.

Gittins, J. (1989), *Multi-Armed Bandit Allocation Indices*, John Wiley and Sons, New York.

Gittins, J. C. & Jones, D. M. (1974), A dynamic allocation index for the sequential design of experiments, *in* J. Gani, ed., 'Progress in Statistics', pp. 241–266.

Gittins, J. C. & Jones, D. M. (1979), 'A dynamic allocation index for the discounted multiarmed bandit problem', *Biometrika* **66**(3), 561–565.

Goel, A., Khanna, S. & Null, B. (2009), The ratio index for budgeted learning, with applications, *in* 'Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms', pp. 18–27.

Gupta, S. & Miescke, K. (1994), 'Bayesian look ahead one stage sampling allocations for selecting the largest normal mean', *Statistical Papers* **35**, 169–177.

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

Kaelbling, L. P. (1993), *Learning in Embedded Systems*, MIT Press, Cambridge, MA.

Katehakis, M. & Veinott, A. (1987), 'The Multi-Armed Bandit Problem: Decomposition and Computation', *Mathematics of Operations Research* **12**(2), 262–268.

Keener, R. (1985), 'Further contributions to the "two-armed bandit" problem', *The Annals of Statistics* **13**(1), 418–422.

Lai, T. (1987), 'Adaptive treatment allocation and the multi-armed bandit problem', *The Annals of Statistics* **15**(3), 1091–1114.

Lai, T. L. & Robbins, H. (1985), 'Asymptotically efficient adaptive allocation rules', *Advances in Applied Mathematics* **6**, 4–22.

Mahajan, A. & Teneketzis, D. (2008), Multi-armed bandit problems, *in* A. Hero, D. Castanon, D. Cochran & K. Kastella, eds, 'Foundations and Applications of Sensor Management', Springer, pp. 121–152.

Mersereau, A., Rusmevichientong, P. & Tsitsiklis, J. (2008), A Structured Multiarmed Bandit Problem and the Greedy Policy, *in* 'Proceedings of the 47th IEEE Conference on Decision and Control', pp. 4945–4950.

Mersereau, A., Rusmevichientong, P. & Tsitsiklis, J. (2009), 'A Structured Multiarmed Bandit Problem and the Greedy Policy', *IEEE Transactions on Automatic Control* **54**(12), 2787–2802.

Pandey, S., Chakrabarti, D. & Agarwal, D. (2007), 'Multi-armed bandit problems with dependent arms', *Proceedings of the 24th International Conference on Machine Learning* pp. 721–728.

Powell, W. B. (2007), *Approximate Dynamic Programming: Solving the curses of dimensionality*, John Wiley and Sons, New York.

Ryzhov, I. O. & Powell, W. B. (2009*a*), A Monte Carlo Knowledge Gradient Method For Learning Abatement Potential Of Emissions Reduction Technologies, *in* M. Rosetti, R. Hill, B. Johansson, A. Dunkin & R. Ingalls, eds, 'Proceedings of the 2009 Winter Simulation Conference', pp. 1492–1502.

Ryzhov, I. O. & Powell, W. B. (2009*b*), The knowledge gradient algorithm for online subset selection, *in* 'Proceedings of the 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning', pp. 137–144.

Ryzhov, I. O. & Powell, W. B. (2010), Approximate Dynamic Programming With Correlated Bayesian Beliefs, *in* 'Proceedings of the 48th Allerton Conference on Communication, Control and Computing', pp. 1360–1367.

Ryzhov, I. O. & Powell, W. B. (2011), 'Information collection on a graph', *Operations Research* **59**(1), 188–201.

Ryzhov, I. O., Valdez-Vivas, M. R. & Powell, W. B. (2010), Optimal Learning of Transition Probabilities in the Two-Agent Newsvendor Problem, *in* B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan & E. Yücesan, eds, 'Proceedings of the 2010 Winter Simulation Conference', pp. 1088–1098.

Steele, M. J. (2000), *Stochastic Calculus and Financial Applications*, Springer, New York.

Sutton, R. & Barto, A. (1998), *Reinforcement Learning*, The MIT Press, Cambridge, Massachusetts.

Tesauro, G. & Galperin, G. (1996), On-Line Policy Improvement using Monte Carlo Search, *in* M. Mozer, M. Jordan & T. Pesche, eds, 'Advances in Neural Information Processing Systems', Vol. 9, Cambridge, MA: MIT Press, pp. 1068–1074.

Tewari, A. & Bartlett, P. (2007), Optimistic Linear Programming gives Logarithmic Regret for Irreducible MDPs, *in* J. Platt, D. Koller, Y. Singer & S. Roweis, eds, 'Advances in Neural Information Processing Systems', Vol. 20, Cambridge, MA: MIT Press, pp. 1505–1512.

Vermorel, J. & Mohri, M. (2005), 'Multi-armed bandit algorithms and empirical evaluation', *Proceedings of the 16th European Conference on Machine Learning* pp. 437–448.

Washburn, R. (2008), Applications of multi-armed bandits to sensor management, *in* A. Hero, D. Castanon, D. Cochran & K. Kastella, eds, 'Foundations and Applications of Sensor Management', Springer, pp. 153–176.

Whittle, P. (1980), 'Multi-armed bandits and the Gittins index', *Journal of the Royal Statistical Society* **B42**(2), 143–149.

Yao, Y. (2006), Some results on the Gittins index for a normal reward process, *in* H. Ho, C. Ing & T. Lai, eds, 'Time Series and Related Topics: In Memory of Ching-Zong Wei', Institute of Mathematical Statistics, Beachwood, OH, USA, pp. 284–294.

9 Appendix: Proofs

9.1 Proof of Proposition 3.1

For any n and any alternative x',

$$\begin{aligned} \mu_{x'}^n &\leq \mu_{x'}^n + (N-n)\,\nu_{x'}^{KG,n} \\ &\leq \max_x \left(\mu_x^n + (N-n)\,\nu_x^{KG,n}\right) \\ &= \mu_{X^{KG,n}(s)}^n + (N-n)\,\nu_{X^{KG,n}(s)}^{KG,n}. \end{aligned} \tag{30}$$

The first inequality holds because $\nu_{x'}^{KG,n} \geq 0$ for any n and any x', and the last line follows from (14). In particular, we can let $n = N-1$ and $x' = \arg\max_x \mu_x^{N-1}$. Combined with (9), this yields

$$\begin{aligned} V^{SL,N-1}(s) &= 2\max_x \mu_x^{N-1} \\ &\leq \left(\max_x \mu_x^{N-1}\right) + \mu_{X^{KG,N-1}(s)}^{N-1} + \nu_{X^{KG,N-1}(s)}^{KG,N-1} \\ &= V^{KG,N-1}(s). \end{aligned}$$

Suppose now that $V^{KG,n'}(s) \geq V^{SL,n'}(s)$ for all s and all $n' > n$. Then,

$$\begin{aligned} V^{KG,n}(s) &= \mu_{X^{KG,n}(s)}^n + \mathbb{E}^n V^{KG,n+1}\left(K^M\left(s, X^{KG,n}(s)\right)\right) \\ &\geq \mu_{X^{KG,n}(s)}^n + \mathbb{E}^n V^{SL,n+1}\left(K^M\left(s, X^{KG,n}(s)\right)\right) \\ &= \mu_{X^{KG,n}(s)}^n + (N-n)\left(\max_{x'} \mu_x^n\right) + (N-n)\,\nu_{X^{KG,n}(s)}^{KG,n} \\ &\geq (N-n+1)\max_{x'} \mu_{x'}^n \\ &= V^{SL,n}(s). \end{aligned}$$

The first inequality is due to the monotonicity of conditional expectation and the inductive hypothesis for $n' = n+1$. The second inequality follows from (30). □

9.2 Proof of Proposition 4.1

Let A be the set of all sample paths ω for which the KG policy measures at least two distinct alternatives infinitely often. By the strong law of large numbers, if we measure an alternative x infinitely often, we have $\mu_x^n \to \mu_x$ almost surely. Furthermore, $\tilde{\sigma}_x^n \to 0$ and $\nu_x^{KG,n} \to 0$ in n almost surely. Lastly, under the normal prior, we have $\mu_x \neq \mu_y$ for any $x \neq y$, almost surely. If we let A' be the subset of A for which all of these properties hold, we have $P(A') = P(A)$.

Let $\omega \in A'$, and suppose that alternatives x and y are measured infinitely often by the KG policy on ω. Then, if we define

$$Q_{x'}^n(\omega) = \mu_{x'}^n(\omega) + \frac{\gamma}{1-\gamma}\nu_{x'}^{KG,n}(\omega)$$

to be the quantity computed by the KG policy for alternative x' at time n on this sample path, it follows that $Q_x^n(\omega) \to \mu_x(\omega)$ and $Q_y^n(\omega) \to \mu_y(\omega)$ in n. Then, letting $\varepsilon = |\mu_x(\omega) - \mu_y(\omega)|$, we can find an integer K_ω such that, for all $n > K_\omega$,

$$\left|Q_x^n(\omega) - \mu_x(\omega)\right|, \left|Q_y^n(\omega) - \mu_y(\omega)\right| < \frac{\varepsilon}{2}.$$

Consequently, at all times after time K_ω, the KG policy will prefer one of these alternatives to the other, namely the one with the higher true reward. This contradicts the assumption that both x and y are measured infinitely often on the sample path ω. It follows that $A' = \emptyset$, whence $P(A') = P(A) = 0$, meaning that the KG policy will measure only one alternative infinitely often on almost every sample path. □

9.3 Proof of Proposition 4.2

The proof has two parts. First, we fix γ and $K > 0$, and choose an outcome ω such that $N_\gamma(\omega) > K$. Now, we show that $N_{\gamma'}(\omega) > K$ for all $\gamma' > \gamma$. If this is not the case, then there is some $n \leq K$ for which we can find $\gamma' > \gamma$ and some alternative y such that, for $x = \arg\max_x \nu_x^{KG,n}(\omega)$,

$$\mu_x^n(\omega) + \frac{1}{1-\gamma'}\nu_x^{KG,n}(\omega) < \mu_y^n(\omega) + \frac{1}{1-\gamma'}\nu_y^{KG,n}(\omega)$$

which means that

$$\frac{1}{1-\gamma'} < \frac{\mu_y^n(\omega) - \mu_x^n(\omega)}{\nu_x^{KG,n}(\omega) - \nu_y^{KG,n}(\omega)}$$

where the sign of the inequality does not change because $\nu_x^{KG,n}(\omega) > \nu_y^{KG,n}(\omega)$ by assumption. However, we know that $\frac{1}{1-\gamma} < \frac{1}{1-\gamma'}$ because $\gamma < \gamma'$, hence it follows that

$$\frac{1}{1-\gamma} < \frac{\mu_y^n(\omega) - \mu_x^n(\omega)}{\nu_x^{KG,n}(\omega) - \nu_y^{KG,n}(\omega)},$$

implying that y is preferred to x by online KG under the discount factor γ, which is not the case because $n < N_\gamma(\omega)$ by assumption. Thus, if offline and online KG agree under γ, they also agree under any $\gamma' > \gamma$. It follows that N_γ is increasing in γ.

We now show that for any ω and any γ, we can find γ' such that $N_\gamma(\omega) < N_{\gamma'}(\omega)$. Let $x = \arg\max_{x'} \nu_{x'}^{KG,N_\gamma(\omega)}(\omega)$. Then, taking

$$\gamma' > 1 - \left(\max_y \frac{\mu_y^{N_\gamma(\omega)}(\omega) - \mu_x^{N_\gamma(\omega)}(\omega)}{\nu_x^{KG,N_\gamma(\omega)}(\omega) - \nu_y^{KG,N_\gamma(\omega)}(\omega)} \right)^{-1}$$

we obtain

$$\frac{1}{1-\gamma'} > \max_y \frac{\mu_y^{N_\gamma(\omega)}(\omega) - \mu_x^{N_\gamma(\omega)}(\omega)}{\nu_x^{KG,N_\gamma(\omega)}(\omega) - \nu_y^{KG,N_\gamma(\omega)}(\omega)}.$$

We can assume that the maximum is strictly positive, so $\gamma' < 1$. The denominator of each term of the maximum is strictly positive because x has the highest KG factor by assumption. At least

one term must have a positive numerator because there must be at least one alternative y that is preferred to x by online KG under γ, hence $\mu_y^{N_\gamma(\omega)}(\omega) > \mu_x^{N_\gamma(\omega)}(\omega)$. It follows that

$$\mu_x^{N_\gamma(\omega)}(\omega) + \frac{1}{1-\gamma'}\nu_x^{KG,N_\gamma(\omega)}(\omega) > \mu_y^{N_\gamma(\omega)}(\omega) + \frac{1}{1-\gamma'}\nu_y^{KG,N_\gamma(\omega)}(\omega).$$

We have already found that increasing γ' will not change the alternatives chosen by online KG before time $N_\gamma(\omega)$, so it follows that $N_{\gamma'}(\omega) > N_\gamma(\omega)$. Consequently, $N_\gamma(\omega) \nearrow \infty$ as $\gamma \nearrow 1$. Since N_γ is integer-valued, it follows that $N_\gamma \nearrow \infty$ almost surely for $\gamma \nearrow 1$. □

9.4 Proof of Proposition 4.3

From Frazier et al. (2008), we know that the process $M_n = \max_x \mu_x^n$ is uniformly integrable. Because $(\mu_x^n)_{n\in\mathbb{N}}$ is a martingale, it follows that M_n is a submartingale. Therefore, M_n converges almost surely and in L^1, and

$$\lim_{n\to\infty} \mathbb{E}M_n = \mathbb{E}\left(\lim_{n\to\infty} M_n\right) \geq \mathbb{E}M_T$$

for any stopping time T. Thus, we can write

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}^{KG(\gamma)}\left(\max_x \mu_x^n\right) &= \mathbb{E}^{KG(\gamma)} \lim_{n\to\infty}\left(\max_x \mu_x^n\right) \\ &\geq \mathbb{E}^{KG(\gamma)}\left(\max_x \mu_x^{N_\gamma}\right)\end{aligned}$$

because N_γ is a stopping time. □

9.5 Proof of Proposition 4.4

Because offline and online KG agree on all measurements before N_γ, we have

$$\begin{aligned}\lim_{\gamma\nearrow 1} \mathbb{E}^{KG(\gamma)}\left(\max_x \mu_x^{N_\gamma}\right) &= \lim_{\gamma\nearrow 1} \mathbb{E}^{Off}\left(\max_x \mu_x^{N_\gamma}\right) \\ &= \lim_{n\to\infty} \mathbb{E}^{Off}\left(\max_x \mu_x^n\right) \\ &= \mathbb{E}\left(\max_x \mu_x\right).\end{aligned}$$

The second line is due to Proposition 4.2 together with the uniform integrability of $(\max_x \mu_x^n)_{n\in\mathbb{N}}$. The last line is due to the asymptotic optimality of offline KG, shown in Frazier et al. (2008). □

9.6 Proof of Theorem 4.1

We can obtain one direction of the inequality by writing

$$\begin{aligned}\lim_{\gamma\nearrow 1}\lim_{n\to\infty}\mathbb{E}^{KG(\gamma)}\left(\max_x \mu_x^n\right) &= \lim_{\gamma\nearrow 1}\lim_{n\to\infty}\mathbb{E}^{KG(\gamma)}\left(\max_x \mathbb{E}^n\mu_x\right)\\ &\leq \lim_{\gamma\nearrow 1}\lim_{n\to\infty}\mathbb{E}^{KG(\gamma)}\mathbb{E}^n\left(\max_x \mu_x\right)\\ &= \mathbb{E}\left(\max_x \mu_x\right).\end{aligned}$$

The first line follows by the tower property of conditional expectation. The second line is due to Jensen's inequality, and the last line is due to the tower property and the fact that μ does not depend on the policy chosen.

For the other direction, we write

$$\begin{aligned}\lim_{\gamma\nearrow 1}\lim_{n\to\infty}\mathbb{E}^{KG(\gamma)}\left(\max_x \mu_x^n\right) &\geq \lim_{\gamma\nearrow 1}\mathbb{E}^{KG(\gamma)}\left(\max_x \mu_x^{N_\gamma}\right)\\ &= \mathbb{E}\left(\max_x \mu_x\right).\end{aligned}$$

The inequality follows by Proposition 4.3 and the second line follows by Proposition 4.4. □

9.7 Proof of Proposition 4.5

Suppose that, under the $KG(\gamma)$ policy, for some ω we have $\arg\max_x \mu_x^n(\omega) = x_*^\gamma(\omega) = x_*(\omega)$ for all $n \geq N_\gamma(\omega)$, and yet $x_*(\omega)$ is not measured infinitely often. By Proposition 4.1, the $KG(\gamma)$ policy must converge to some alternative almost surely, so we can assume without loss of generality that the policy converges to some $y(\omega) \neq x_*(\omega)$ on the sample path ω. As in the proof of Proposition 4.1, $\lim_{n\to\infty}\mu_{y(\omega)}^n + \frac{\gamma}{1-\gamma}\nu_{y(\omega)}^{KG,n} = \mu_{y(\omega)}$.

Let $T(\omega)$ be the last time when the $KG(\gamma)$ policy measures any alternative $z \neq y(\omega)$ on the sample path ω. Because $y(\omega)$ is measured infinitely often, it must follow that

$$\mu_{y(\omega)}(\omega) \geq \mu_{x_*(\omega)}^{T(\omega)}(\omega) + \frac{\gamma}{1-\gamma}\nu_{x_*(\omega)}^{KG,T(\omega)}(\omega) \geq \mu_{x_*(\omega)}^{T(\omega)}(\omega). \tag{31}$$

The first inequality is due to the fact that the $KG(\gamma)$ policy prefers $y(\omega)$ to $x_*(\omega)$ for all $n > T(\omega)$. The second inequality is due to the positivity of the KG factor.

At the same time, we have by assumption that $\mu^n_{x_*(\omega)}(\omega) \geq \mu^n_{y(\omega)}(\omega)$ for all $n \geq N_\gamma(\omega)$. Letting $n \to \infty$ on both sides, we obtain

$$\mu^{T(\omega)}_{x_*(\omega)}(\omega) \geq \mu_{y(\omega)}(\omega). \tag{32}$$

Combining (31) and (32) yields $\mu_{y(\omega)}(\omega) = \mu^{T(\omega)}_{x_*(\omega)}(\omega)$. Since μ is continuous, the set of ω for which this holds has measure zero. We conclude that, for almost every ω for which $\arg\max_x \mu^n_x(\omega) = x^\gamma_*(\omega) = x_*(\omega)$ for all $n \geq N_\gamma(\omega)$, the alternative $x_*(\omega)$ is measured infinitely often by the $KG(\gamma)$ policy. □

9.8 Proof of Proposition 4.6

Let A_1 denote the event that, for any $y \neq x^\gamma_*$, we have $B^{x^\gamma_*}_0 > B^y_0$ and $B^{x^\gamma_*}_1 > B^y_1$. Also let A_2 denote the event that, for any $y \neq x^\gamma_0$, we have $B^{x^\gamma_*}_{t^n} > B^y_{t^{n'}}$ for all n, n'. Clearly, the event that $L_{x^\gamma_*} > \max_{x \neq x^\gamma_*} U_x$ implies both A_1 and A_2. However,

$$P(A_1 \cap A_2) \leq P^{KG(\gamma)}\left(x_* = x^\gamma_*,\ \arg\max_x \mu^{N_\gamma+n}_x = x^\gamma_*\ \forall n \geq 0\right)$$

because event A_1 is analogous to having $x_* = x^\gamma_*$, and A_2 ensures (due to the independence of the alternatives) that $\mu^{N_\gamma+n}_{x^\gamma_*} > \mu^{N_\gamma+n}_y$ for all $y \neq x^\gamma_*$ and all $n \geq 0$, regardless of how measurements are allocated after time N_γ. □

9.9 Proof of Proposition 4.7

First, observe that

$$\begin{aligned} h\left(\mu^{N_\gamma}, \sigma^{N_\gamma}\right) &= P^{KG(\gamma)}\left(\mu^{N_\gamma}_{x^\gamma_*} - \sigma^{N_\gamma}_{x^\gamma_*}\left|W^{x^\gamma_*}_1\right| > \max_{x \neq x^\gamma_*}\left(\mu^{N_\gamma}_x + \sigma^{N_\gamma}_x |W^x_1|\right) \mid \mathcal{F}^{N_\gamma}\right) \\ &= P^{KG(\gamma)}\left(\mu^{N_\gamma}_{x^\gamma_*} - \sigma^{N_\gamma}_{x^\gamma_*}\left|W^{x^\gamma_*}_1\right| > \mu^{N_\gamma}_x + \sigma^{N_\gamma}_x |W^x_1|\ \forall x \neq x^\gamma_* \mid \mathcal{F}^{N_\gamma}\right). \end{aligned} \tag{33}$$

Now, if

$$\sigma^{N_\gamma}_x |W^x_1| < \frac{1}{2}\left(\mu^{N_\gamma}_{x^\gamma_*} - \max_{x' \neq x^\gamma_*} \mu^{N_\gamma}_{x'}\right) \tag{34}$$

for all x (including x^γ_*), then

$$\sigma^{N_\gamma}_x |W^x_1| + \sigma^{N_\gamma}_{x^\gamma_*}\left|W^{x^\gamma_*}_1\right| < \mu^{N_\gamma}_{x^\gamma_*} - \max_{x' \neq x^\gamma_*} \mu^{N_\gamma}_{x'} \leq \mu^{N_\gamma}_{x^\gamma_*} - \mu^{N_\gamma}_x$$

for any $x \neq x_*^\gamma$ and the event in (33) occurs. The probability of the event in (34) is given by

$$P^{KG(\gamma)}\left(|W_1^x| < \frac{\mu_{x_*^\gamma}^{N_\gamma} - \max_{x \neq x_*^\gamma} \mu_x^{N_\gamma}}{2\sigma_x^{N_\gamma}} \; \forall x \,|\, \mathcal{F}^{N_\gamma}\right) = g\left(\frac{\mu_{x_*^\gamma}^{N_\gamma} - \max_{x' \neq x_*^\gamma} \mu_{x'}^{N_\gamma}}{2}, \sigma^{N_\gamma}\right). \tag{35}$$

The quantity $\mu_{x_*^\gamma}^{N_\gamma} - \max_{x' \neq x_*^\gamma} \mu_{x'}^{N_\gamma}$ is a.s. strictly positive by the definition of x_*^γ, so the right-hand side of (35) is well-defined. □

9.10 Proof of Theorem 4.2

Combining Propositions 4.5 and 4.6 yields

$$P^{KG(\gamma)}(B) \geq P^{KG(\gamma)}\left(L_{x_*^\gamma} > \max_{x \neq x_*^\gamma} U_x\right)$$

Next, we write

$$\begin{aligned} P^{KG(\gamma)}\left(L_{x_*^\gamma} > \max_{x \neq x_*^\gamma} U_x\right) &= \mathbb{E}^{KG(\gamma)} h\left(\mu^{N_\gamma}, \sigma^{N_\gamma}\right) \\ &= \mathbb{E}^{Off} h\left(\mu^{N_\gamma}, \sigma^{N_\gamma}\right) \\ &\geq \mathbb{E}^{Off} g\left(\frac{\mu_{x_*^\gamma}^{N_\gamma} - \max_{x \neq x_*^\gamma} \mu_x^{N_\gamma}}{2}, \sigma^{N_\gamma}\right). \end{aligned}$$

The first line is due to the tower property of conditional expectation. The second line follows because the offline KG policy agrees with the $KG(\gamma)$ policy up to the stopping time N_γ, and the quantity $f\left(\mu^{N_\gamma}, \sigma^{N_\gamma}\right)$ depends only on our beliefs at that time. The last line comes from Proposition 4.7.

It follows that

$$\begin{aligned} \lim_{\gamma \nearrow 1} P^{KG(\gamma)}(B) &\geq \lim_{\gamma \nearrow 1} \mathbb{E}^{Off} g\left(\frac{\mu_{x_*^\gamma}^{N_\gamma} - \max_{x \neq x_*^\gamma} \mu_x^{N_\gamma}}{2}, \sigma^{N_\gamma}\right) \\ &= \mathbb{E}^{Off} \lim_{\gamma \nearrow 1} g\left(\frac{\mu_{x_*^\gamma}^{N_\gamma} - \max_{x \neq x_*^\gamma} \mu_x^{N_\gamma}}{2}, \sigma^{N_\gamma}\right), \end{aligned}$$

where we can pass the limit inside the expectation because $|g| \leq 1$, and so the dominated convergence theorem applies. Observe now that g is a continuous function due to the continuity of Φ and

the fact that $\Phi\left(\frac{a}{b}\right) \to 1$ as $b \to 0$. Therefore,

$$\begin{aligned}
\mathbb{E}^{Off} \lim_{\gamma \nearrow 1} g\left(\frac{\mu_{x_*^\gamma}^{N_\gamma} - \max_{x \neq x_*^\gamma} \mu_x^{N_\gamma}}{2}, \sigma^{N_\gamma}\right) &= \mathbb{E}^{Off} g\left(\lim_{\gamma \nearrow 1} \frac{\mu_{x_*^\gamma}^{N_\gamma} - \max_{x \neq x_*^\gamma} \mu_x^{N_\gamma}}{2}, \lim_{\gamma \nearrow 1} \sigma^{N_\gamma}\right) \\
&= \mathbb{E}^{Off} g\left(\frac{\mu_{x_*} - \max_{x \neq x_*} \mu_x}{2}, 0\right) \\
&= 1,
\end{aligned}$$

because $\lim_{\gamma \nearrow 1} N_\gamma = \infty$ a.s. by Proposition 4.2, and therefore, under the offline KG policy, $\mu^{N_\gamma} \to \mu$ and $\sigma^{N_\gamma} \to 0$ almost surely as $\gamma \nearrow 1$. The quantity $\mu_{x_*} - \max_{x \neq x_*} \mu_x$ is almost surely strictly positive, so g evaluated at the limit is almost surely equal to 1. Since probabilities are bounded above by 1, we conclude that

$$\lim_{\gamma \nearrow 1} P^{KG(\gamma)}(B) = 1,$$

as required. □

9.11 Proof of Proposition 5.1

Because f is increasing, (21) yields

$$\mu_x^n + (N-n)\,\nu_x^{KGC,n} \leq \mu_x^n + (N-n)\,\frac{1}{2\pi} \max_{x'} \tilde{\sigma}_{x'}^{corr,n}(x)\,.$$

Since $\max_{x'} \mu_{x'}^n \leq \max_{x'} \mu_x^n + (N-n)\,\nu_{x'}^{KGC,n}$ by the positivity of the KG factor, (23) implies that x will not be chosen by the KG decision rule. □

10 Appendix: a discussion of a non-Gaussian model

Throughout our paper, we have focused on a Gaussian belief model with Gaussian measurements. This framework provides us with a powerful and general way to handle correlated beliefs using multivariate Gaussian priors. However, the knowledge gradient concept is not limited to the Gaussian model, and can be used for many other types of belief structures. In this section, we briefly discuss one such problem, in which the measurements have 0/1 outcomes. We argue that KG is not limited to a single method for Gaussian beliefs, but rather is a general methodology that can be applied to many broad classes of learning problems.

Consider an online learning problem with M alternatives. Alternative x has an unknown success probability $\rho_x \sim Beta\left(\alpha_x^0, \beta_x^0\right)$, where the parameters α_x^0 and β_x^0 represent our beliefs. For example, ρ_x may represent the probability of success for a particular medical treatment in curing a disease. This application has motivated many classic studies of bandit problems, such as Gittins & Jones (1979).

If we choose to measure alternative x at time n, we make an observation $W_x^{n+1} \sim Bernoulli\left(\rho_x\right)$ denoting success or failure. The posterior distribution of ρ_x is also beta, and the updating equations are given by (DeGroot (1970))

$$\begin{aligned}
\alpha_x^{n+1} &= \alpha_x^n + W_x^{n+1} \\
\beta_x^{n+1} &= \beta_x^n + \left(1 - W_x^{n+1}\right).
\end{aligned}$$

We assume that the alternatives are independent in this example. If we measure x at time n, we update our beliefs about ρ_x, but not about the other success probabilities. As before, $\mathcal{F}^n$ denotes the sigma-algebra generated by the first n measurement decisions as well as the outcomes of those decisions.

Our objective is to choose a measurement policy that maximizes the total number of successes across all trials. In an infinite-horizon setting, this can be stated as

$$\sup_{\pi} \mathbb{E}^{\pi} \sum_{n=0}^{\infty} \gamma^n \rho_{X^{\pi,n}(s^n)},$$

where $s^n = (\alpha^n, \beta^n)$. The KG logic from Section 3.2 still applies. The KG policy yields the decision rule

$$X^{KG,n}\left(s^n\right) = \arg\max_x \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n} + \frac{\gamma}{1-\gamma} \nu_x^{KG,n},$$

where

$$\nu_x^{KG,n} = \mathbb{E}_x^n \left[\left(\max_{x'} \frac{\alpha_{x'}^{n+1}}{\alpha_{x'}^{n+1} + \beta_{x'}^{n+1}} \right) - \left(\max_{x'} \frac{\alpha_{x'}^n}{\alpha_{x'}^n + \beta_{x'}^n} \right) \right].$$

Just as in the Gaussian case, the KG factor can be computed explicitly. We now present this computational result.

Proposition 10.1. *The KG factor in the beta-Bernoulli model can be computed as*

$$\nu_x^{KG,n} = \begin{cases} \frac{\alpha_x^n}{\alpha_x^n+\beta_x^n} \left(\frac{\alpha_x^n+1}{\alpha_x^n+\beta_x^n+1} - C_x^n \right) & \text{if } \frac{\alpha_x^n}{\alpha_x^n+\beta_x^n} \le C_x^n < \frac{\alpha_x^n+1}{\alpha_x^n+\beta_x^n+1} \\ \frac{\beta_x^n}{\alpha_x^n+\beta_x^n} \left(C_x^n - \frac{\alpha_x^n}{\alpha_x^n+\beta_x^n+1} \right) & \text{if } \frac{\alpha_x^n}{\alpha_x^n+\beta_x^n+1} \le C_x^n < \frac{\alpha_x^n}{\alpha_x^n+\beta_x^n} \\ 0 & \text{otherwise,} \end{cases}$$

where $C_x^n = \max_{x' \neq x} \frac{\alpha_{x'}^n}{\alpha_{x'}^n + \beta_{x'}^n}$.

Proof: The first step is to compute the predictive distribution of W_x^{n+1} given $\mathcal{F}^n$. Since there are only two possible values that W_x^{n+1} can take on, it suffices to consider each outcome individually. We can write

$$\begin{aligned} P^n\left(W_x^{n+1} = 1\right) &= \mathbb{E}^n P^n\left(W_x^{n+1} = 1 \,|\, \rho_x\right) \\ &= \mathbb{E}^n \rho_x \\ &= \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n}. \end{aligned}$$

Therefore, $P^n\left(W_x^{n+1} = 0\right) = \frac{\beta_x^n}{\alpha_x^n + \beta_x^n}$. Furthermore, $\alpha_x^{n+1} + \beta_x^{n+1} = \alpha_x^n + \beta_x^n + 1$ regardless of which outcome occurs. It follows that

$$\begin{aligned} \mathbb{E}_x^n\left(\max_{x'} \frac{\alpha_{x'}^{n+1}}{\alpha_{x'}^{n+1} + \beta_{x'}^{n+1}}\right) &= \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n} \max\left\{C_x^n, \frac{\alpha_x^n + 1}{\alpha_x^n + \beta_x^n + 1}\right\} \\ &+ \frac{\beta_x^n}{\alpha_x^n + \beta_x^n} \max\left\{C_x^n, \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n + 1}\right\}. \end{aligned} \tag{36}$$

Observe that

$$\frac{\alpha_x^n}{\alpha_x^n + \beta_x^n + 1} < \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n} < \frac{\alpha_x^n + 1}{\alpha_x^n + \beta_x^n + 1}$$

for any $\alpha_x^n, \beta_x^n > 0$. Therefore, if $C_x^n < \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n + 1}$, it follows that $\max_{x'} \frac{\alpha_{x'}^n}{\alpha_{x'}^n + \beta_{x'}^n} = \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n}$ and (36) becomes

$$\begin{aligned} \mathbb{E}_x^n\left(\max_{x'} \frac{\alpha_{x'}^{n+1}}{\alpha_{x'}^{n+1} + \beta_{x'}^{n+1}}\right) &= \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n} \frac{\alpha_x^n + 1}{\alpha_x^n + \beta_x^n + 1} + \frac{\beta_x^n}{\alpha_x^n + \beta_x^n} \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n + 1} \\ &= \frac{\alpha_x^n\left(\alpha_x^n + \beta_x^n + 1\right)}{\left(\alpha_x^n + \beta_x^n\right)\left(\alpha_x^n + \beta_x^n + 1\right)} \\ &= \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n} \end{aligned}$$

and $\nu_x^{KG,n} = 0$.

In the second case, when $\frac{\alpha_x^n}{\alpha_x^n + \beta_x^n + 1} \leq C_x^n < \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n}$, we compute (36) as

$$\mathbb{E}_x^n\left(\max_{x'} \frac{\alpha_{x'}^{n+1}}{\alpha_{x'}^{n+1} + \beta_{x'}^{n+1}}\right) = \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n} \frac{\alpha_x^n + 1}{\alpha_x^n + \beta_x^n + 1} + \frac{\beta_x^n}{\alpha_x^n + \beta_x^n} C_x^n. \tag{37}$$

Subtracting $\frac{\alpha_x^n}{\alpha_x^n+\beta_x^n}$ yields

$$\begin{aligned}\nu_x^{KG,n} &= \frac{\alpha_x^n}{\alpha_x^n+\beta_x^n}\left(\frac{\alpha_x^n+1}{\alpha_x^n+\beta_x^n+1}-1\right)+\frac{\beta_x^n}{\alpha_x^n+\beta_x^n}C_x^n \\ &= \frac{\beta_x^n}{\alpha_x^n+\beta_x^n}\left(C_x^n-\frac{\alpha_x^n}{\alpha_x^n+\beta_x^n+1}\right),\end{aligned}$$

as required.

In the third case, when $\frac{\alpha_x^n}{\alpha_x^n+\beta_x^n} \le C_x^n < \frac{\alpha_x^n+1}{\alpha_x^n+\beta_x^n+1}$, we have $\max_{x'} \frac{\alpha_{x'}^n}{\alpha_{x'}^n+\beta_{x'}^n} = C_x^n$, whence (37) still holds, and

$$\begin{aligned}\nu_x^{KG,n} &= \frac{\alpha_x^n}{\alpha_x^n+\beta_x^n}\frac{\alpha_x^n+1}{\alpha_x^n+\beta_x^n+1}+C_x^n\left(\frac{\beta_x^n}{\alpha_x^n+\beta_x^n}-1\right) \\ &= \frac{\alpha_x^n}{\alpha_x^n+\beta_x^n}\left(\frac{\alpha_x^n+1}{\alpha_x^n+\beta_x^n+1}-C_x^n\right).\end{aligned}$$

Finally, when $C_x^n \ge \frac{\alpha_x^n+1}{\alpha_x^n+\beta_x^n+1}$, the right-hand side of (36) is equal to C_x^n, and $\nu_x^{KG,n} = 0$. □

The KG policy has an intuitive interpretation. In the beta-Bernoulli model, a single measurement of x can only change our beliefs about the best alternative if $\frac{\alpha_x^n}{\alpha_x^n+\beta_x^n}$ and C_x^n are sufficiently close together. If these quantities are far enough apart, that is, $C_x^n < \frac{\alpha_x^n}{\alpha_x^n+\beta_x^n+1}$ or $C_x^n \ge \frac{\alpha_x^n+1}{\alpha_x^n+\beta_x^n+1}$, then neither outcome of the measurement will bring about any improvement in our understanding of the best alternative, and the measurement has no value. We did not see this property in the Gaussian model because Gaussian measurements are continuous, and there is always a possibility that we will observe a sufficiently large or small outcome to change our beliefs.

As in Section 6, we tested the KG policy on 100 randomly generated truth-from-prior problems where the initial prior parameters α_x^0, β_x^0 were generated from uniform distributions on $[0, 5]$. We ran 10^4 sample paths on each problem, with a new set of truths $\rho_x \sim Beta\left(\alpha_x^0, \beta_x^0\right)$ for every sample path. All problems were infinite-horizon problems with a discount factor of 0.9.

In the beta-Bernoulli model, Gittins indices can be approximated using a computationally expensive dynamic programming procedure. Analytic approximations of the Gittins index policy are not as advanced as those available for the Gaussian model. However, the seminal work by Brezzi & Lai (2002) uses a central limit argument to recommend using the Gaussian approximation

(the one used throughout this paper), with

$$\begin{aligned}
\mu_x^n &\approx \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n}, \\
(\sigma_x^n)^2 &\approx \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n} \frac{\beta_x^n}{\alpha_x^n + \beta_x^n} \frac{1}{\alpha_x^n + \beta_x^n + 1}, \\
\sigma_\varepsilon^2 &\approx \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n} \frac{\beta_x^n}{\alpha_x^n + \beta_x^n}
\end{aligned}$$

used to stand in for the Gaussian parameters at time n. We followed this recommendation, but again used the most advanced Gaussian approximation derived by Chick & Gans (2009).

Furthermore, we also compared against the pure exploitation policy

$$X^{Exp,n}(s^n) = \arg\max_x \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n}$$

and the UCB1 policy of Auer et al. (2002) for rewards in $[0, 1]$, given by

$$X^{UCB1,n}(s^n) = \arg\max_x \frac{\alpha_x^n}{\alpha_x^n + \beta_x^n} + \sqrt{\frac{2 \log n}{N_x^n}}.$$

As before, we can estimate the difference in opportunity cost for KG and other policies, averaged over 10^4 sample paths, in each problem. The means and standard errors for these estimated differences are given in Table 3.

We see that KG significantly outperforms both the Gittins and UCB1 policies, and is competitive against pure exploitation (beating it by a small margin). Furthermore, KG outperformed Gittins and UCB1 on all 100 problems. KG outperformed pure exploitation on 58/100 problems. These results suggest that the online KG policy is also competitive in the beta-Bernoulli setting.

Our goal in this discussion has been to illustrate the applicability of knowledge gradient methods to a variety of learning problems. In fact, the idea of computing the expected value of information may apply to more complex resource allocation problems where Gittins indices are no longer applicable. Such problems are outside the scope of this paper, but work in this area is ongoing. For example, Ryzhov & Powell (2011) considers a stochastic shortest-path problem on a graph where

	KG-Gitt	KG-UCB1	KGC-Exp
Mean	0.5250	1.3463	0.0045
Avg. SE	0.0052	0.0036	0.0031

Table 3: Means and standard errors for the beta-Bernoulli experiments.

the mean arc lengths are unknown, and sequential sampling can be used to adaptively estimate them. This problem moves beyond the multi-armed bandit setting by creating a distinction between measurement and implementation decisions: while we measure individual arcs, we are not interested in finding the shortest arc. Rather, an arc is only valuable as long as it provides information about the shortest path. While no analogue of Gittins indices (or interval estimation or other index policies) exists for this problem, it is possible to apply value of information concepts and derive a KG-type policy that computes the expected value of a single measurement exactly. Other work along these lines includes Ryzhov et al. (2010), which considers the problem of learning uncertain transition probabilities in a Markov decision process, and Ryzhov & Powell (2010), where the online KG concept put forth in this paper is used to create a decision-making strategy for approximate dynamic programming. We believe that, while the bandit problem with multivariate Gaussian priors is itself a general class of online learning problems, the KG concept reaches beyond this setting.