The knowledge gradient algorithm for online learning

Ilya O. Ryzhov   Warren Powell   Peter I. Frazier

July 22, 2008

Abstract

We derive a one-period look-ahead policy for finite- and infinite-horizon online optimal learning problems with Gaussian rewards. The resulting decision rule easily extends to a variety of settings, including the case where our prior beliefs about the rewards are correlated. Experiments show that the KG policy performs competitively against other learning policies in diverse situations. In the case where the optimal policy is known, the KG policy performs comparably well, while being substantially more convenient to use.

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 random rewards are independent of each other and follow a Gaussian distribution with fixed, known variance. Our objective is to maximize the total expected reward collected over a horizon of N measurements. We allow several variations of this basic setup: the rewards may be discounted over time, the time horizon could be finite or infinite, and our beliefs about the unknown rewards may be correlated.

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, the problem has correlated rewards.

2. *Call centers.* We are assigning calls arriving in series to technicians. The reward is the time needed by a technician to resolve the issue. The objective is to minimize the total time needed to finish all jobs. This problem is online, with a potentially large time horizon and no discount factor.

3. *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.

4. *Sensor management.* In this area (see Mahajan & Teneketzis (2008) and Washburn (2008) for more on applications), 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.

Variations of this problem have been widely studied under the name "multi-armed bandit problems." In particular, Gittins (1989) describes a measurement policy (referred to as "Gittins indices") that is asymptotically optimal as $N \to \infty$ in the discounted case. 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: the interval estimation policy by Kaelbling (1993), the Boltzmann exploration policy, pure exploitation, and the equal-allocation policy. Some general bounds on the performance of learning policies were obtained by Lai & Robbins (1985). Empirical comparisons of some policies in certain settings are available in Vermorel & Mohri (2005).

We consider several generalizations of online learning problems, including finite-horizon problems, both discounted and undiscounted, and problems with correlated prior beliefs. Many applications have a natural finite time horizon, such as the problem of finding the best type of diabetes medication within a particular budget, or the problem of finding the best pricing strategy for a product during its lifetime. Correlated problems, where our prior belief about the mean value of one alternative is correlated with our beliefs about mean values of other alternatives, also arise frequently in applications. However, the Gittins policy and the other heuristics listed above do not handle correlations. The correlated case has been studied by Pandey et al. (2007), but this work relies on the assumption of binomial rewards.

Our analysis is motivated by the knowledge gradient (KG) principle, developed by Gupta & Miescke (1994) and further analyzed by Frazier et al. (2008*a*) and Chick et al. (2007) 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. It is optimal for $N = 1$ and $N \to \infty$, and performs well in practice for other values of N, while providing an easily computable decision rule. More recently, Frazier et al. (2008*b*) extended the KG principle to the ranking and selection problem with correlated priors, and Chick et al. (2007) extended it 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. The computation of Gittins indices in the general case is discussed by Katehakis & Veinott Jr (1987) and Duff (1995). An LP-based computational method was developed by Bertsimas & Nino-Mora (2000), however it is founded on a Markov decision process framework, in which the prior beliefs about the alternatives are limited to a finite set of values, whereas our problem has continuous, Gaussian priors and rewards. For our problem, an approximation for Gittins indices can be found in Yao (2006), but it is less accurate for small time horizons and large discount factors. However, we show that our KG policy closely matches the performance of the Gittins policy even for problems where the Gittins indices are known and optimal. Furthermore, the knowledge gradient is a general methodology that can be applied to other distributions, although these require the development of different computational formulas.

We proceed as follows. In Section 2, we lay out a dynamic-programming-based mathematical model of the problem. In Section 3, we derive the KG measurement policy for undiscounted, finite-horizon problems, and provide lower and upper bounds on the value of learning under this policy. We also show that the marginal value of a measurement does not always decrease with the amount of measurements already performed. In Section 4, we extend the KG policy to the discounted and correlated cases. Finally, we present numerical results comparing our online KG rules to existing learning policies. We focus on an undiscounted, finite-horizon setting with correlated rewards, motivated by the medical application described above, but we also test the KG policy against the Gittins policy in a discounted, infinite-horizon problem. 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 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 $\mu^n = \left(\mu_1^n, ..., \mu_M^n\right)$ and $\beta^n = \left(\beta_1^n, ..., \beta_M^n\right)$. We also let

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

be the reduction in the variance of our beliefs about x that we achieve by measuring x at time n. 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 $\mathcal{N}\left(\mu_x^n, \left(\tilde{\sigma}_x^n\right)^2\right)$. 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*

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

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

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

where the transition function K^M is described by (1) and (2). For notational convenience, we suppress the dependence on $\hat{\mu}^{n+1}_{x^n}$ when we write K^M, but it is important to remember that the transition function is stochastic.

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

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

where $X^{\pi,i}\left(s^i\right)$ is the alternative chosen by policy π at time i given a knowledge state s^i. Then 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 (used in an optimal learning context by DeGroot (1970)):

$$\begin{aligned} V^{\pi,n}\left(s^n\right) &= \mu^n_{X^{\pi,n}(s^n)} + \mathbb{E}^n V^{\pi,n+1}\left(K^M\left(s^n, X^{\pi,n}\left(s^n\right)\right)\right) &(5)\\ V^{\pi,N}\left(s^N\right) &= \max_x \mu^N_x. &(6)\end{aligned}$$

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

$$\begin{aligned} V^{*,n}\left(s^n\right) &= \max_x \mu^n_x + \mathbb{E}^n V^{*,n+1}\left(K^M\left(s^n, x\right)\right) &(7)\\ V^{*,N}\left(s^N\right) &= \max_x \mu^N_x &(8)\end{aligned}$$

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.

3 The online knowledge gradient policy

We derive an easily computable online decision rule using the KG principle. We then define the value of learning under the online KG policy, and prove lower and upper bounds on this quantity. Finally, we discuss the general behaviour of the value of learning, and present an example in which the marginal value of information is non-concave.

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 *empirical Bayesian* policy of 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

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

If we cannot make any more measurements, but we can still collect $N-n+1$ more rewards, we should always choose the alternative that looks the best given everything that we were able to learn up to time n.

The *knowledge gradient principle*, first described by Gupta & Miescke (1994) and later developed by Frazier et al. (2008*a*), 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 empirical Bayesian policy starting at time $n+1$. The KG decision rule that follows from this assumption is

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

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

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

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

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

where $\mathbb{E}_x^n$ observes all the information known at time n, as well as the choice to measure x at time n. The knowledge gradient can be computed exactly using the formula

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

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*a*) that (13) and (12) are equivalent in this problem, and that ν^{KG} is always positive. The origin of the term "knowledge gradient" arises from (12), where the quantity $\nu_x^{KG,n}$ is written as a difference.

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

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

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

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

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. The form of the decision rule in (14) is common in optimal learning algorithms. Many other policies, such as interval estimation and Gittins indices, do the same thing, but define the value of exploration in different ways. In our case, it represents the value of learning one more time.

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*

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

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

Remark 3.1 suggests that the policy given by (14) is the correct extension of the KG principle to the online bandit problem. The decision rule itself is different from the one for ranking and selection. In particular, it is not stationary (as it is for offline problems), because the right-hand side of (14) depends on n as well as on s^n. If we collect a reward in each time period, knowledge is more useful early on, while there are still many rewards left to collect. For this reason, if $n_1 < n_2$, the uncertainty bonus for an alternative at time n_1 is greater than the uncertainty bonus for the same alternative at time n_2, even if the available information is the same at both times.

3.2 Properties of the online KG policy

We continue our discussion of the undiscounted case with several results on the value of information. First, we show that it is better to measure under the KG policy than to not measure at all. Second, we derive a lower bound on (15). Although the KG policy is suboptimal for $N > 1$, the lower bound tells us that we are guaranteed to achieve a certain expected reward by following it. Finally, we derive an upper bound on (15). Together, these results enable us to narrow the value of information under the KG policy to an interval.

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

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

Proof: For any n and any alternative x',

$$\begin{aligned} \mu^n_{x'} &\leq \mu^n_{x'} + (N-n)\,\nu^{KG,n}_{x'} \\ &\leq \max_x \mu^n_x + (N-n)\,\nu^{KG,n}_x \\ &= \mu^n_{X^{KG,n}(s)} + (N-n)\,\nu^{KG,n}_{X^{KG,n}(s)} \end{aligned} \tag{16}$$

where the first inequality is due to the fact that $\nu^{KG,n}_{x'} \geq 0$ for any n and any x', and the last line follows from (14). In particular, we can let $n = N-1$ and $x' = \arg\max_x \mu^{N-1}_x$. Combined with (9), this yields

$$\begin{aligned} V^{EB,N-1}(s) &= 2\max_x \mu^{N-1}_x \\ &\leq \max_x \mu^{N-1}_x + \mu^{N-1}_{X^{KG,N-1}(s)} + \nu^{KG,N-1}_{X^{KG,N-1}(s)} \\ &= V^{KG,N-1}(s). \end{aligned}$$

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

$$\begin{aligned} V^{KG,n}(s) &= \mu^n_{X^{KG,n}(s)} + \mathbf{E}^n V^{KG,n+1}\left(K^M\left(s, X^{KG,n}(s)\right)\right) \\ &\geq \mu^n_{X^{KG,n}(s)} + \mathbf{E}^n V^{EB,n+1}\left(K^M\left(s, X^{KG,n}(s)\right)\right) \\ &= \mu^n_{X^{KG,n}(s)} + (N-n)\,\mathbf{E}^n \max_x \mu^{n+1}_x \\ &= \mu^n_{X^{KG,n}(s)} + (N-n)\left(\max_{x'} \mu^n_x\right) + (N-n)\,\nu^{KG,n}_{X^{KG,n}(s)} \\ &\geq (N-n+1)\max_{x'} \mu^n_{x'} \\ &= V^{EB,n}(s). \end{aligned}$$

The first inequality is due to the monotonicity of conditional expectation and the inductive hypothesis for $n' = n+1$. The second inequality follows from (16). □

Proposition 3.2. *For any s and any n,*

$$V^{KG,n}(s) \geq (N-n)\left(\max_{x'} \mu^n_{x'}\right) + \max_x \left(\mu^n_x + (N-n)\,\nu^{KG,n}_x\right).$$

Proof: Observe that

$$\begin{aligned} V^{KG,n}(s) &= \mu^n_{X^{KG,n}(s)} + \mathbb{E}^n V^{KG,n+1}\left(K^M\left(s, X^{KG,n}(s)\right)\right) \\ &\geq \mu^n_{X^{KG,n}(s)} + \mathbb{E}^n V^{EB,n+1}\left(K^M\left(s, X^{KG,n}(s)\right)\right) \\ &= \mu^n_{X^{KG,n}(s)} + (N-n)\left(\max_{x'} \mu^n_{x'}\right) + (N-n)\,\nu^{KG,n}_{X^{KG,n}(s)} \\ &= (N-n)\left(\max_{x'} \mu^n_{x'}\right) + \max_x \left(\mu^n_x + (N-n)\,\nu^{KG,n}_x\right) \end{aligned}$$

where the inequality is due to Proposition 3.1 and the last line follows from (14). □

Proposition 3.3. *For any s and any n,*

$$V^{KG,n}(s) \leq (N-n+1)\max_x \mu^n_x + \left[\frac{(N-n)(N-n+1)}{2}\right] c^n \tag{17}$$

where

$$c^n = \frac{1}{\sqrt{2\pi}} \max_x \tilde{\sigma}^n_x.$$

Proof: From (12), we have

$$\begin{aligned} \nu^{KG,n}_x &= \mathbb{E}^n_x\left(\max_{x'} \mu^{n+1}_{x'} - \max_{x'} \mu^n_{x'}\right) \\ &\leq \mathbb{E}^n_x\left(\max_{x'} \mu^{n+1}_{x'} - \mu^n_{x'}\right) \\ &= \mathbb{E}\left(\max\left(0, \tilde{\sigma}^n_x \cdot Z\right)\right) \\ &= \frac{1}{\sqrt{2\pi}}\tilde{\sigma}^n_x \end{aligned}$$

for any x, whence it follows that $\nu^{KG,n}_x \leq c^n$ for any n and any x. Furthermore, we have $c^{n+1} \leq c^n$ for any n, because the variance reduction for the alternative measured at time n

is smaller at time $n+1$ (after the measurement) than at time n, and the variance reduction for the other alternatives stays the same from time n to time $n+1$.

Then, for any s^{N-1}, it follows from (11) and Remark 3.1 that

$$\begin{aligned} V^{KG,N-1}\left(s^{N-1}\right) &= \mu^{N-1}_{X^{KG,N-1}\left(s^{N-1}\right)} + \mathbb{E}^{N-1} V^{EB,N}\left(K^M\left(s^{N-1}, X^{KG,N-1}\left(s^{N-1}\right)\right)\right) \\ &= \mu^{N-1}_{X^{KG,N-1}\left(s^{N-1}\right)} + \max_x \mu^{N-1}_x + \nu^{KG,N-1}_{X^{KG,N-1}\left(s^{N-1}\right)} \\ &\leq 2\max_x \mu^{N-1}_x + c^{N-1} \end{aligned}$$

which is exactly (17) for $n = N-1$. Suppose now that (17) holds for all s and all $n' > n$. Then, for n, we have

$$\begin{aligned} V^{KG,n}\left(s^n\right) &= \mu^n_{X^{KG,n}\left(s^n\right)} + \mathbb{E}^n V^{KG,n+1}\left(K^M\left(s^n, X^{KG,n}\left(s^n\right)\right)\right) \\ &\leq \max_x \mu^n_x + \mathbb{E}^n V^{KG,n+1}\left(K^M\left(s^n, X^{KG,n}\left(s^n\right)\right)\right) \\ &\leq \max_x \mu^n_x + \mathbb{E}^n \left\{(N-n)\max_x \mu^{n+1}_x + \left[\frac{(N-n-1)(N-n)}{2}\right] c^{n+1}\right\} \\ &= (N-n+1)\max_x \mu^n_x + (N-n)\,\nu^{KG,n}_{X^{KG,n}\left(s^n\right)} + \left[\frac{(N-n-1)(N-n)}{2}\right] c^{n+1} \\ &\leq (N-n+1)\max_x \mu^n_x + (N-n)\,c^n + \left[\frac{(N-n-1)(N-n)}{2}\right] c^{n+1} \\ &\leq (N-n+1)\max_x \mu^n_x + \left[\frac{(N-n)(N-n+1)}{2}\right] c^n \end{aligned}$$

which completes the proof. The second inequality is due to the inductive hypothesis for $n' = n+1$, and the third and fourth inequalities are due to the fact that $\nu^{KG,n}_x \leq c^n$ and $c^{n+1} \leq c^n$. □

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

$$0 \leq b^n \leq V^{KG,n}(s) - V^{EB,n}(s) \leq \left[\frac{(N-n)(N-n+1)}{2}\right] c^n$$

where c^n is as in Proposition 3.3, and

$$b^n = \max_x \left(\mu^n_x + (N-n)\,\nu^{KG,n}_x\right) - \max_x \mu^n_x.$$

Corollary 3.1 follows directly from Propositions 3.2 and 3.3, and gives us a range for the value of information under the KG policy. In our discussion, $V^{EB,n}$ represents the best value that can be obtained if learning stops at time n. Therefore, $V^{KG,n} - V^{EB,n}$ represents exactly the value of learning, starting at time n, under the online KG policy. Note that the interval given by Corollary 3.1 becomes tighter as n increases. This reflects the intuitive idea that earlier measurements (corresponding to smaller n) are somehow more unpredictable, which is discussed below in greater detail.

3.3 The marginal value of information

The quantity $V^{KG,n} - V^{EB,n}$ is difficult to compute. However, an analogous expression $V^{\pi,n} - V^{EB,n}$, representing the value of learning under some policy π, can be computed for certain simple choices of π. Suppose that we have an undiscounted, finite-horizon problem in which there are only two alternatives, A and B. Alternative A is known to yield zero reward, that is, $\mu_A^0 = 0$ and $\beta_A^0 = \infty$. Let $\mu_B^0 = 10$ and $\beta_B^0 = 1$ represent our beliefs about alternative B. In addition, let $\beta_\varepsilon = 1$.

Now, define π_B to be the deterministic policy that always measures alternative B, in every time step. From (5), we have

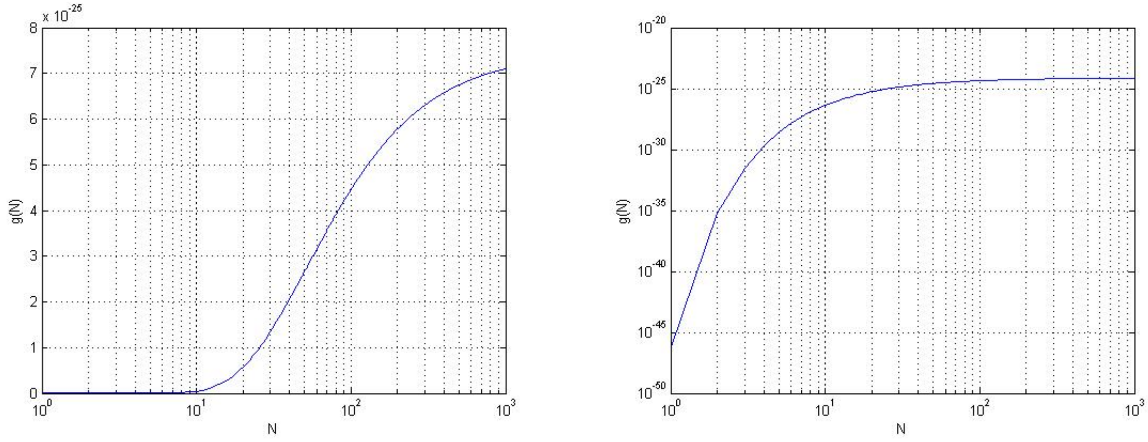
$$V^{\pi_B,n}\left(s^n\right) = \mu_B^n + \mathbb{E}^n V^{\pi_B,n+1}\left(K^M\left(s^n, B\right)\right).$$

Since, for any n, we have $\mathbb{E}^n V^{\pi_B,n+1}\left(K^M\left(s^n, B\right)\right) = \mu_B^n + \mathbb{E}^n V^{\pi_B,n+2}\left(K^M\left(s^n, B\right)\right)$ by the tower property of conditional expectation, it follows that

$$\begin{aligned} V^{\pi_B,0}\left(s^0\right) &= N \cdot \mu_B^0 + \mathbb{E}^0 \max\left\{0, \mu_B^N\right\} \\ &= N \cdot \mu_B^0 + \mathbb{E} \max\left\{0, \mu_B^0 + \tilde{\sigma}_B^0\left(N\right) \cdot Z\right\} \end{aligned} \tag{18}$$

where Z is a standard Gaussian random variable, and

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



(a) Non-concavity.

(b) Log-concavity.

Figure 1: Properties of the value of learning under policy π_B over different time horizons.

is the total reduction in the variance of our beliefs about x achieved by k consecutive measurements of x starting at time n. In this setting,

$$\tilde{\sigma}_B^0(N) = \sqrt{1 - \frac{1}{N+1}} = \sqrt{\frac{N}{N+1}}.$$

Then, (18) becomes

$$V^{\pi_B,0}\left(s^0\right) = N \cdot \mu_B^0 + \max\left\{0, \mu_B^0\right\} + g(N) = (N+1) \cdot \mu_B^0 + g(N)$$

where

$$g(N) = \tilde{\sigma}_B^0(N) \cdot f\left(-\left|\frac{\mu_B^0}{\tilde{\sigma}_B^0(N)}\right|\right) \tag{19}$$

and the function f is as in (13). Because $\max\left\{\mu_A^0, \mu_B^0\right\} = \max\left\{0, \mu_B^0\right\} = \mu_B^0$, it follows that

$$V^{\pi_B,0}\left(s^0\right) - V^{EB,0}\left(s^0\right) = (N+1) \cdot \mu_B^0 + g(N) - (N+1) \cdot \mu_B^0 = g(N)$$

so $g(N)$ is precisely the total value of learning under the policy π_B. Equation (19) allows us to easily compute and plot $g(N)$ versus N. Figure 1 shows the shape of g in both semi-log and log-scale.

We discover that $g(N)$ is not concave in N. The shape of g looks convex for small N and concave for large N. This means that the marginal value of one more measurement of B does not necessarily diminish with the total number of measurements of B performed up to that point. In fact, for small time horizons, later measurements of B are actually more valuable than earlier ones. At the same time, $g(N)$ appears to be log-concave.

This behaviour has an intuitive explanation. The early beliefs are based on few observations and heavily rely on the prior. Our beliefs still have high variance and it is difficult to tell whether the prior means are accurate. Successive measurements should be increasingly more valuable in the early stages, because they play a key role in revealing the structure of the problem. Later on, additional measurements serve to refine an already accurate estimate, and we observe typical diminishing returns.

Not all learning problems have these properties. If we let $\mu_B^0 = 0$ in the preceding example, we do not observe the non-concave behaviour at all. On the other hand, if N is too small, we may never observe the eventual concavity. Many applications have small measurement budgets, and in discounted problems, early iterations are more important than later ones. This example suggests that the quality of an optimal learning algorithm is largely determined by its performance in a relatively small number of iterations early on.

The S-curve in Figure 1(a) can be said to reflect the problem of having too many choices. In a situation where there are multiple choices, but the measurement budget is limited, we will tend to simply ignore some of the choices, and spend our time measuring a subset of the choices multiple times. Of course, these will be the choices where the marginal value of information is non-concave.

4 Extensions of the KG policy

The definition of the KG policy given in (10) is quite versatile. We can extend it to online problems with a discount factor, as well as online problems with correlated normal priors. In both cases, the KG policy resembles the one given by (14), with some differences in the computation of the uncertainty bonus.

4.1 Discounted problems

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

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

where $\gamma \in (0,1)$ is a given parameter. The knowledge gradient policy for this problem is derived the same way as in Section 3. First, in the discounted setting,

$$V^{EB,n}(s^{n}) = \frac{1-\gamma^{N-n+1}}{1-\gamma} \max_{x} \mu_{x}^{n}.$$

Then, (10) is computed as

$$\begin{aligned} X^{KG,n}(s^{n}) &= \arg\max_{x} \mu_{x}^{n} + \gamma \cdot \mathbb{E}^{n} V^{EB,n+1}\left(K^{M}(s^{n},x)\right) \\ &= \arg\max_{x} \mu_{x}^{n} + \gamma \frac{1-\gamma^{N-n}}{1-\gamma} \nu_{x}^{KG,n} \end{aligned} \tag{20}$$

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

$$X^{KG,n}(s^{n}) = \arg\max_{x} \mu_{x}^{n} + \frac{\gamma}{1-\gamma} \nu_{x}^{KG,n}. \tag{21}$$

Both (20) and (21) look similar to (14), with a different multiplier in front of the knowledge gradient. For both finite- and infinite-horizon problems, the value of the discounted KG policy is given by

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

The next result shows that, in the limit as $N \to \infty$, the discounted KG policy will converge. In other words, only one alternative will be measured infinitely often by the policy. Like Corollary 3.1, this suggests that information becomes less valuable over time. In the limit, the value of additional information converges to zero, and the KG policy settles on an alternative that it believes to be the best.

Proposition 4.1. *Suppose that $\mu_x \neq \mu_y$ for any $x \neq y$. Then, for almost every sample path, only one alternative will be measured infinitely often by the infinite-horizon discounted KG policy.*

Proof: 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. Therefore, if we let A' be the subset of A for which 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 $n > K_\omega$,

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

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. $\square$

We see that the asymptotic behaviour of the online KG policy is exactly the opposite of that of the offline KG policy for ranking and selection. We know from Frazier et al. (2008*a*) that the offline KG policy will measure every alternative infinitely often, if there are infinitely many opportunities to measure. In the online setting, however, the KG policy converges almost surely to one alternative. Note that this alternative will not necessarily be the best one. The KG policy may miss the alternative that is truly the best if the Q-factor of that alternative is low (for example, due to an inaccurate prior). However, the optimal Gittins policy does not necessarily converge to the best alternative either. In the discounted setting, earlier measurements are more important than later ones, so it is more important for a policy to learn well early on than to converge to the optimal alternative in the future.

4.2 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 $\mu^0 = (\mu_1^0, ..., \mu_M^0)$ 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

$$\mu^{n+1} = \mu^n + \frac{\hat{\mu}_{x^n}^n - \mu_x^n}{\sigma_\varepsilon^2 + \Sigma_{x^n x^n}^n} \Sigma^n e_{x^n} \tag{22}$$

$$\Sigma^{n+1} = \Sigma^n - \frac{\Sigma^n e_{x^n} e_{x^n}^T \Sigma^n}{\sigma_\varepsilon^2 + \Sigma_{x^n x^n}^n} \tag{23}$$

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

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

where Z is standard Gaussian and

$$\tilde{\sigma}^{corr,n}\left(x^n\right) = \frac{\Sigma^n e_{x^n}}{\sqrt{\sigma_\varepsilon^2 + \Sigma_{x^n x^n}^n}}.$$

The empirical Bayesian 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

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

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

$$\nu_x^{KGC,n} = \mathbb{E}_x^n \left(\max_{x'} \mu_{x'}^{n+1} - \max_{x'} \mu_{x'}^n \right)$$

be the analog of (12) in the correlated setting. From the work by Frazier et al. (2008*b*), it is known that

$$\nu_x^{KGC,n} = \sum_{y=1}^{M-1} \left(\tilde{\sigma}_{y+1}^{corr,n} - \tilde{\sigma}_y^{corr,n} \right) f\left(-\left|c_y\right|\right)$$

where the alternatives have been sorted in order of increasing $\tilde{\sigma}_y^{corr,n}$, f is as in (13), and the numbers c_y are such that $y = \arg\max_{x'} \mu_{x'}^n + \tilde{\sigma}_{x'}^{corr,n}(x) \cdot z$ for $z \in [c_{y-1}, c_y)$. The online KG decision rule for the correlated case is given by

$$X^{KGC,n}(s^n) = \arg\max_x \mu_x^n + (N-n)\, \nu_x^{KGC,n}. \tag{24}$$

An efficient algorithm for computing ν^{KGC} exactly is presented in Frazier et al. (2008*b*), and can be used to solve this decision problem. If we introduce a discount factor into the problem, the decision rule becomes as in (20) or (21), using ν^{KGC} instead of ν^{KG}.

5 Computational experiments

We used the problem of clinical trials of experimental diabetes treatments to obtain realistic initial parameters for experiments comparing online KG to other learning policies. In a situation where our prior beliefs give us a lot of information about the rewards, online KG is comparable to the competition. When the prior beliefs tell us nothing about which alternative is the best, online KG consistently outperforms the other policies. Also, in a discounted infinite-horizon setting, online KG is comparable to the known optimal policy, while being easier to compute.

5.1 Background and setup of experiments

In the final stages of clinical drug trials, the most promising treatments are chosen for testing on several thousand human patients. Each treatment is an alternative, in our terminology, and the reward is the effectiveness of the treatment. The effectiveness of a new diabetes drug can be expressed in terms of the resulting reduction in fasting plasma glucose (FPG), the blood sugar level after the patient has not eaten for eight hours. The FPG level is measured in millimoles per liter (mmol/L). A healthy subject typically has an FPG level of $4 - 6$ mmol/L, whereas diabetes patients can show FPG levels of $10 - 15$ mmol/L. A single drug can reduce FPG level by as little as 0.5 (Figure 5 in DREAM Trial Investigators (2006)) or as much as 2 (Figure 3 in UKPDS Group (1998)) mmol/L over a period of several months.

Often, the treatment is actually a combination of several drugs. The effectiveness of the drugs is not additive, so the point of interest is the effectiveness of the entire treatment. A possible range for such a value might be the interval $[2, 5]$. We consider a setting in which there are six drugs (e.g. metformin, insulin, glibenclamide, chlorpropamide, rosiglitazone, and conventional treatment), and each treatment consists of three drugs. Thus, there are 20 possible alternatives in our problem, and rewards are correlated because the same drugs can appear in multiple treatments. In practice, it would take too long to test all patients one by one in series, so we suppose that the patients are split up into $N + 1$ large groups, and every patient within a group is assigned to the same treatment. One measurement is the average FPG reduction across one group, so our normality assumptions will be plausible if the groups are large enough. Our objective is to maximize the sum of the average FPG reductions over all groups, with no discount factor. This objective balances the need to find the best treatment with concern for the well-being of the patients.

In order to test a learning policy, we must first assume a truth, then evaluate the ability of the policy to find that truth. Furthermore, the truths should represent a wide variety of situations. For this reason, the starting data for our experiments was randomly generated, using the context of diabetes treatments to provide realistic numbers. Because the rewards are correlated, we used the mathematical framework in Section 4.2, and the updating equations (22) and (23), in all of our experiments. The initial data for one experiment consists of a vector μ to represent the true rewards, a prior (μ^0, Σ^0) to represent our initial beliefs about the rewards, and a measurement error σ_ε^2. We generated two sets of 100 experiments.

In the first set, referred to as the *heterogeneous-prior* experiments, we first generated the prior means μ^0 from a uniform distribution on the interval $[2, 5]$. The variances were generated from a uniform distribution on $[0.25, 0.75]$. These numbers represent our beliefs about the range in which the true values are likely to fall. The correlation coefficient of two treatments was set to be 0, $\frac{1}{3}$ or $\frac{2}{3}$, depending on whether the two treatments had 0, 1 or 2 drugs in common. The measurement error σ_ε^2 was chosen to be 0.5 $\mathrm{mmol}^2/\mathrm{L}^2$, to reflect a situation where the effectiveness of a drug varies fairly widely over groups of patients. The true rewards μ were then generated from a multivariate Gaussian distribution with mean vector μ^0 and covariance matrix Σ^0. That is, the truths were drawn from the prior. This represents a situation in which we already have a reasonably good idea about the treatments, and our prior beliefs are on average accurate.

In the second set of experiments, referred to as the *equal-prior* experiments, we first generated the true rewards μ from a triangular distribution on $[2, 5]$ with mode 3.5. The prior means μ^0 were all set to 3.5, and the prior variances were all set to 0.75^2. The covariances and the measurement error were chosen the same way as in the heterogeneous-prior experiments. This set of experiments represents a situation where we only know a general range of values for the true rewards, but we know nothing at all about which treatment is better. The truths are mostly concentrated around the prior, but there are a few very effective treatments.

For each experiment, in either set, we ran each measurement policy 10^4 times starting from the same initial data, for each of four different measurement budgets $N = 5, 10, 15, 20$. For each policy, we observed the average opportunity cost per reward collected, defined as

$$C^\pi = \max_x \mu_x - \frac{1}{N+1} \sum_{n=0}^{N} \mu_{X^{\pi,n}(s^n)}$$

for a generic policy π. The policies were compared by taking the difference of their opportunity costs. For policies π_1, π_2,

$$C^{\pi_2} - C^{\pi_1} = \frac{1}{N+1} \sum_{n=0}^{N} \left(\mu_{X^{\pi_1,n}(s^n)} - \mu_{X^{\pi_2,n}(s^n)} \right) \qquad (25)$$

is precisely the amount by which π_1 outperformed (or underperformed) π_2. The 10^4 sample paths were divided into groups of 500 in order to obtain approximately normal samples of average opportunity cost and the standard errors of these averages. The standard error of the difference in (25) is the square root of the sum of the squared standard errors of C^{π_1}, C^{π_2}. Five policies were tested overall; we briefly describe the implementation of each.

Independent and correlated online KG (KG/KGC). The independent and correlated KG policies are defined by the decision rules (14) and (24), respectively. The KGC policy was implemented using the algorithm from Frazier et al. (2008*b*), discussed in Section 4.2.

Gittins indices (Gitt). The Gittins decision rule, designed for discounted infinite-horizon problems, is given by

$$X^{Gitt,n}\left(s^n\right) = \arg\max_x \Gamma\left(\mu_x^n, \sigma_x^n, \sigma_\varepsilon, \gamma\right), \tag{26}$$

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 can use the identity

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

known from Gittins (1989). Furthermore, once the measurement error has been normalized in this manner, we can use the fact that $\left(\sigma_x^n\right)^2 \approx \frac{1}{N_x^n}$, where N_x^n is the number of times alternative x has been visited up to and including time n, to avoid having to compute Gittins indices for arbitrary σ_x^n. Thus, we can rewrite (26) as

$$X^{Gitt,n}\left(s^n\right) = \arg\max_x \mu_x^n + \sigma_\varepsilon \cdot \Gamma\left(N_x^n, \gamma\right),$$

where Γ is now a function only of the discount factor and the number of times an alternative has been measured.

The Gittins policy is not designed for undiscounted, finite-horizon problems. Therefore, we view it as a heuristic, with a tunable parameter in the form of the Gittins discount factor γ. It is reasonable to choose γ to satisfy $\sum_{n=0}^{\infty} \gamma^n = N$, so the effective time horizon under the Gittins policy is equal to the actual time horizon in our problem.

However, even with the above simplifications, Gittins indices are still hard to compute. An LP-based method was developed by Bertsimas & Nino-Mora (2000), but it assumes that the knowledge state is discrete, with only finitely many possible values of μ^n, σ^n. Discretization is not practical for our problem, where the priors and rewards are Gaussian.

Furthermore, the method is computationally complex, requiring the size of the LP to grow exponentially in order to obtain an exact solution.

In the setting of Gaussian priors and rewards, exact values are only available for a few choices of γ in Gittins (1989). In order to allow us to tune the discount factor and consider values of γ for which the exact Gittins indices are unknown, one can use the approximation from Yao (2006). Define a function

$$\Psi(s) = \begin{cases} \sqrt{\frac{s}{2}} & s \leq 0.2 \\ 0.49 - 0.11 s^{-\frac{1}{2}} & 0.2 < s \leq 1 \\ 0.63 - 0.26 s^{-\frac{1}{2}} & 1 < s \leq 5 \\ 0.77 - 0.57 s^{-\frac{1}{2}} & 5 < s \leq 15 \\ (2\log s - \log\log s - \log 16\pi)^{-\frac{1}{2}} & s > 15 \end{cases}$$

Now let $s = -\frac{1}{n \log \gamma}$ and define

$$\begin{aligned} \Gamma^{LB}(n) &= \frac{1}{\sqrt{n}} \Psi(s) - \frac{0.583 n^{-1}}{\sqrt{1+n^{-1}}} \\ \Gamma^{UB}(n) &= \frac{1}{\sqrt{n}} \sqrt{\frac{s}{2}} - \frac{0.583 n^{-1}}{\sqrt{1+n^{-1}}}. \end{aligned}$$

Finally, take the Gittins index to be

$$\Gamma(n, \gamma) \approx \frac{1}{2}\left(\Gamma^{LB} + \Gamma^{UB}\right).$$

This approximation will perform very well for any value of γ, as long as n is high enough. However, it can be inaccurate for low values of n and high values of γ, which is important for our application, where the number of measurements is relatively small.

In our study, we found that the approximation worked best for $\gamma \in [0.85, 0.9]$ across all the time horizons. For $N = 15$, the exact Gittins indices for $\gamma = 0.95$, obtained from Gittins (1989), yielded very similar results to the approximation. Our figures in the subsequent discussion were obtained using these exact values. However, the approximation is a viable alternative for any setting requiring a smaller discount factor.

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

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

where $z_{\alpha/2}$ is a tunable parameter. We obtained the best performance for relatively low values of $z_{\alpha/2}$, e.g. $z_{\alpha/2} = 1.25$. In general, interval estimation worked better than the Gittins policy, but was sensitive to the choice of $z_{\alpha/2}$.

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.

5.2 Results: heterogeneous-prior experiments

The results for all four time horizons were similar. For this reason, we focus on one time horizon $N = 15$ throughout this discussion, and explain the minor differences between time horizons at the end. For each relevant comparison of two policies, we obtained 100 samples of the difference in (25). Table 1 gives the means and average standard errors of our estimates of (25) across the 100 problems in the heterogeneous-prior set, for $N = 15$.

Figure 2 shows the distribution of the sampled differences. The label on each histogram names the two policies that were compared and gives the number of times the first policy outperformed the second. Bars to the right of zero indicate that the first policy outperformed the second policy, and bars to the left of zero indicate the converse. For example, "KG-Gitt: 73/100" means that the independent online KG policy outperformed the Gittins heuristic in 73/100 experiments, and bars to the right of zero in this histogram represent those experiments where KG performed better.

We see that KG and KGC outperform Gittins and IE about 70% of the time. Furthermore, KGC outperforms KG 73/100 times. The additional improvement brought about by correlated KG can be observed in Table 1. Comparisons involving KGC tend to have smaller negative tails and greater positive tails than KG. This is most evident in the comparison with interval estimation.

	KG-Gitt	KG-IE	KG-Exp	KGC-Gitt	KGC-IE	KGC-Exp	KGC-KG
Mean	0.0217	0.0008	0.0241	0.0251	0.0042	0.0275	0.0034
Avg. SE	0.0018	0.0017	0.0018	0.0019	0.0018	0.0018	0.0018

Table 1: Means and standard errors for the heterogeneous-prior experiments with $N = 15$.

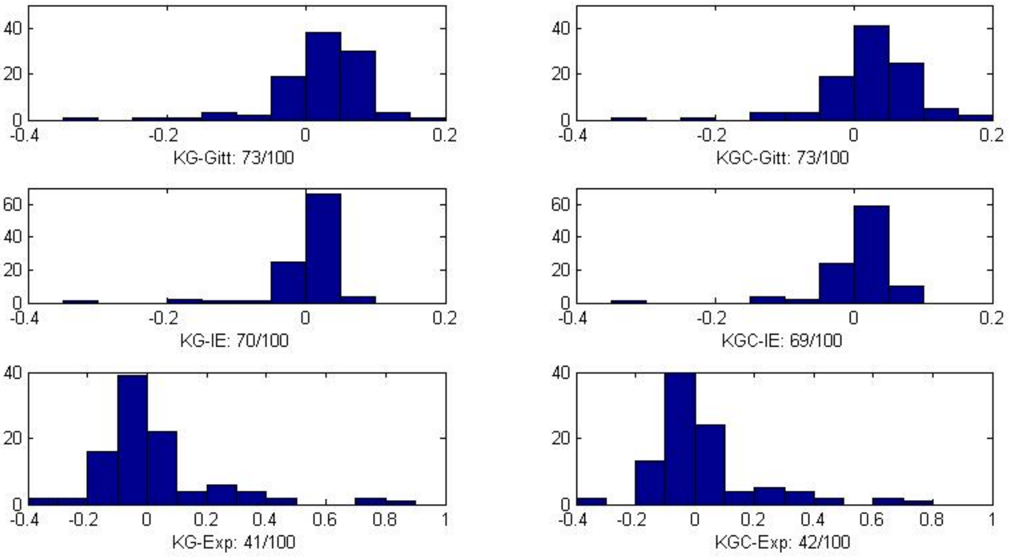
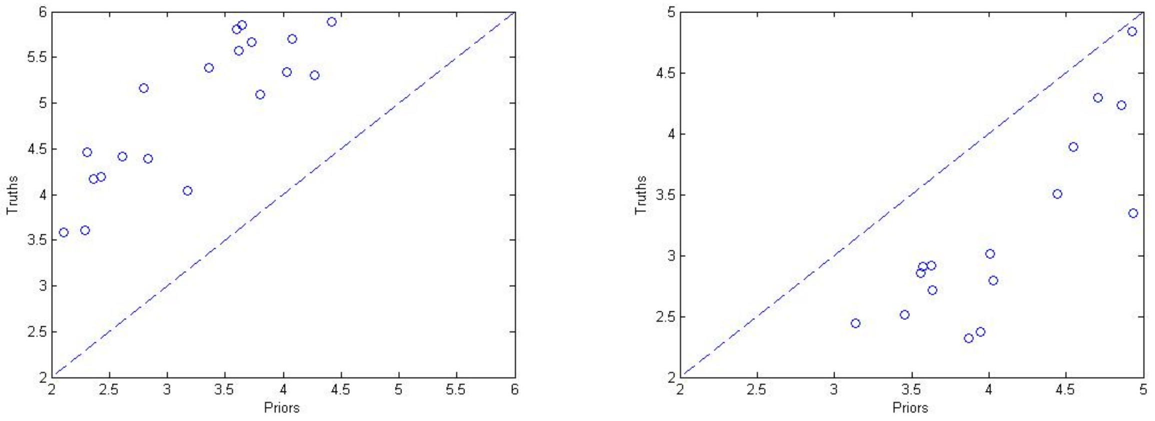


Figure 2: Histograms of the sampled difference in opportunity cost for competing policies across 100 heterogeneous-prior experiments, with $N = 15$.

The comparisons exhibit noticeable negative tails. In a minority of experiments, KG and KGC perform worse than IE and Gittins. We examined these outliers and found that they were mostly coming from the same experiments. For example, the same experiment produces the single worst outlier in all of the comparisons in Figure 2. A different experiment produces the second best outlier in KG-Gitt and the best in KGC-Gitt, and gives good (positive) results in KG-IE and KGC-IE. Figure 3 shows the relationship between the true rewards and prior beliefs in both of these experiments.

In Figure 3(a), the prior consistently underestimates the truth, for every alternative. On the other hand, in Figure 3(b), the prior consistently overestimates the truth, also for every alternative. We believe that the first case leads KG to stop exploring sooner than the other policies, and settle in on a suboptimal alternative that appears to be better than the others. The second case, however, leads KG to explore more, as every measurement appears to be worse than was previously believed. Thus, it appears that online KG is more susceptible to being misled by a bad prior. However, a consistently, excessively pessimistic prior such as in Figure 3(a) is unlikely in practice. Furthermore, in cases when the truths are more evenly spread around the priors, KG outperforms Gittins and IE.



(a)

(b)

Figure 3: True rewards and prior beliefs for two experiments: (a) one producing the worst outlier for all comparisons, and (b) one producing the best outlier for KGC-Gitt and second-best for KG-Gitt.

Pure exploitation outperforms *all* of the other policies 60% of the time, though usually by a small margin. However, there is a substantial minority of experiments where *every* measurement achieves an additional FPG reduction of as much as 0.8 mmol/L under the KG policy, the equivalent of adding a whole new drug to the treatment. These widely varying results are also due to the relationship of the truth to the prior in each experiment. The experiment in Figure 3(a) again contributes the single worst outlier for KG-Exp and KGC-Exp, this time because the alternative with the highest prior also has the highest truth, far ahead of the alternative with the second-highest prior. Pure exploitation happens to find the optimal alternative in the first measurement, and usually stays there, while KG and other policies are still exploring. In Figure 3(b), we see the opposite situation: the alternative with the highest prior has a much lower truth than the alternative with the second-highest prior. Thus, pure exploitation makes a mistake in the first measurement, while other policies are more likely to use it to obtain more useful information.

Thus, although pure exploitation performs very well on most of the heterogeneous-prior experiments, we argue that it is unreliable because it is prone to worse errors than any other policy, and these errors are particularly bad for this application. In general, the good performance of pure exploitation is due to the fact that the prior is heterogeneous. The alternative with the highest prior is likely to have the highest truth also, allowing pure exploitation to quickly discover it. Pure exploitation is less effective in the equal-prior experiments.

The choice of N does not change the overall portrait of the results. We found that smaller values of N led to larger tails, both positive and negative, in the comparisons KG-Gitt, KGC-Gitt, KG-IE, KGC-IE. The majority of experiments still fell into the range $0 - 0.1$, but this majority increased with N. The KG-Exp and KGC-Exp comparisons showed the opposite behaviour: larger values of N led to larger positive tails. Thus, pure exploitation becomes less reliable as the time horizon gets larger. We conclude that, in a situation where we have a reasonably accurate estimate of the true rewards, KG and KGC are more reliable than pure exploitation, and comparable to Gittins and IE.

5.3 Results: equal-prior experiments

In the equal-prior setting, the KG policy yields consistently superior performance compared to the other policies. The means and average standard errors of our estimates of (25) are given in Table 2. It is notable that the Gittins heuristic does particularly poorly on this set of experiments. In fact, as we see in Figure 4, both KG and KGC outperform the Gittins heuristic a full 100% of the time. The Gittins policy makes more severe mistakes in the correlated setting when our prior reveals nothing about which alternative is best.

By contrast, interval estimation performs quite well, usually losing to the KG policy by a very small margin. However, this comparison also shows the most dramatic improvement achieved by using KGC instead of KG. Where KG outperforms IE in 78/100 experiments, for KGC this proportion is 93/100. The pure exploitation policy has the worst performance of all the policies tested. As one might expect, it is less effective when it has a smaller chance of immediately picking a good alternative.

	KG-Gitt	KG-IE	KG-Exp	KGC-Gitt	KGC-IE	KGC-Exp	KGC-KG
Mean	0.1167	0.0079	0.1426	0.1258	0.0170	0.1516	0.0090
Avg. SE	0.0038	0.0044	0.0059	0.0039	0.0045	0.0060	0.0045

Table 2: Means and standard errors for the equal-prior experiments with $N = 15$.

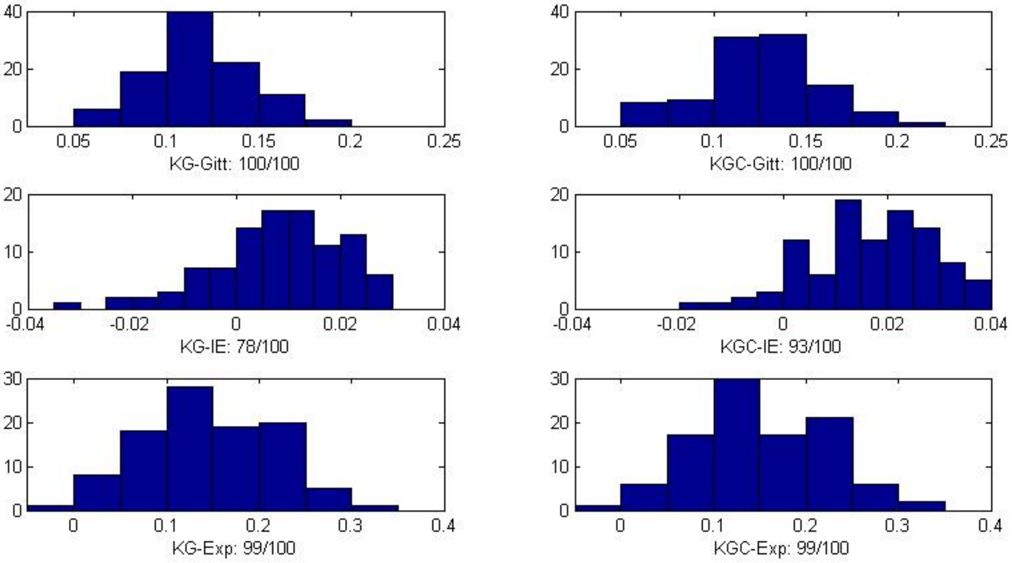


Figure 4: Histograms of the sampled difference in opportunity cost for competing policies across 100 equal-prior experiments, with $N = 15$.

The choice of N has the same effect as in the heterogeneous-prior experiments. Larger values of N lead to smaller tails, although the KG policy still consistently outperforms the other policies. We conclude that, in a situation where we know nothing about how the alternatives are ranked, KG and KGC offer reliable ways to distinguish between the alternatives quickly. KGC is especially well-suited to this situation, because the correlations between rewards become more important when we have little prior knowledge of the rewards.

5.4 Long-run comparison to Gittins policy

We end our computational study with a comparison of the independent KG policy to the Gittins index policy in a situation where the latter is optimal: a discounted problem with a large measurement budget and independent rewards. We use the knowledge transition function given by (1) and (2), and the discounted variant of the KG decision rule given by (20). Again, we measure the opportunity cost of each policy, but now we average over the effective time horizon:

$$C^{\pi} = \max_x \mu_x - \frac{1}{\sum_{n=0}^{N} \gamma^n} \sum_{n=0}^{N} \gamma^n \mu_{X^{\pi,n}(s^n)}$$

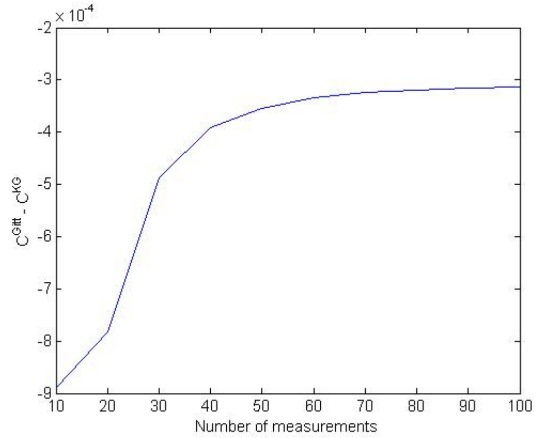


Figure 5: KG-Gitt comparison over $N = 100$.

Then, (25) becomes

$$C^{\pi_2} - C^{\pi_1} = \frac{1}{\sum_{n=0}^{N} \gamma^n} \sum_{n=0}^{N} \gamma^n \left(\mu_{X^{\pi_1,n}(s^n)} - \mu_{X^{\pi_2,n}(s^n)} \right).$$

We are interested in the difference $C^{Gitt} - C^{KG}$. As before, a positive difference would mean that KG outperforms Gittins, and a negative difference means the converse.

We tested the policies using all the initial data from the heterogeneous-prior experiments, except without the covariance structure. Earlier we observed that the KG policy has more difficulty learning in this setting than in the equal-prior case. We chose a discount factor of $\gamma = 0.95$, in order to use the exact Gittins indices from Gittins (1989), and simulated each policy over a time horizon of $N = 100$, reporting the difference $C^{Gitt} - C^{KG}$ every 10 measurements. Our estimates of $C^{Gitt} - C^{KG}$, averaged over 100 heterogeneous-prior experiments and 10^4 sample paths in each experiment, are shown in Figure 5.

As expected, the KG policy is outperformed by the Gittins policy. However, even for $N = 10$, the difference between them is very small, on the order of 10^{-4}. In the context of diabetes treatments, this is negligible. Furthermore, the difference shrinks as N increases. In other words, the Gittins policy does choose better alternatives than the KG policy in the early iterations, though this difference is not large to begin with. However, eventually, the KG policy tends to converge to the same alternative as the Gittins policy, and the difference between them shrinks.

This suggests that the KG policy is competitive against the Gittins policy even when the Gittins policy is known to be optimal. However, unlike Gittins indices, the KG decision rule is easy to compute, and does not require us to rely on approximations or limit ourselves to a few choices of γ. Thus, the KG policy emerges as a viable alternative to the Gittins policy.

6 Conclusion

We have proposed an easily computable decision rule for online learning problems. Within the class of problems with a finite measurement budget, normally distributed priors, and normal sampling errors with known variance, the KG policy proves to be versatile. Variations of the basic KG decision rule cover both undiscounted and discounted, finite- and infinite-horizon problems, and can also accommodate correlated priors. We compared the KG policy to several other measurement policies in a realistic setting. In our experiments, we considered different situations that might arise in practice: one where we already have some information about the rewards of the alternatives, and one where we only know that they fall in a certain range, not how they compare to each other. In the correlated setting, the KG policy is either comparable to or better than the other policies tested, and performs especially well when the prior provides little information.

We also showed that, in the uncorrelated, discounted setting, the KG policy is comparable to the known optimal policy (Gittins indices), while being far easier to implement and compute for any discount factor. For this reason, the KG policy emerges as a worthwhile alternative even in problems where the optimal policy is known, with the additional advantage of extending to the correlated setting. We have constructed our experimental study to illustrate the potential of the KG policy for application to problems with a finite measurement budget.

Acknowledgements

The authors thank Patrick Williams at McGill University for his explanation of clinical drug trials, and his help in constructing the initial parameters for our experiments.

References

Bertsimas, D. & Nino-Mora, J. (2000), 'Restless bandits, linear programming relaxations, and a primal-dual index heuristic', *Operations Research* **48**(1), 80–90. 3, 22

Chick, S., Branke, J. & Schmidt, C. (2007), 'New myopic sequential sampling procedures', *Submitted for publication.* 3

DeGroot, M. H. (1970), *Optimal statistical decisions*, John Wiley and Sons. 5, 6

DREAM Trial Investigators (2006), 'Effect of rosiglitazone on the frequency of diabetes in patients with impaired glucose tolerance or impaired fasting glucose: a randomised controlled trial', *The Lancet* **368**(9549), 1096–1105. 20

Duff, M. O. (1995), Q-learning for bandit problems, Technical report, Department of Computer Science, University of Massachusetts, Amherst, MA. 3

Frazier, P., Powell, W. & Dayanik, S. (2008*a*), 'A knowledge-gradient policy for sequential information collection', *SIAM Journal on Control and Optimization (to appear).* 3, 7, 8, 17

Frazier, P., Powell, W. & Dayanik, S. (2008*b*), 'The knowledge-gradient policy for correlated normal rewards', *Submitted for publication.* 3, 19, 22

Gittins, J. (1989), *Multi-Armed Bandit Allocation Indices*, John Wiley and Sons, New York. 2, 22, 23, 29

Gupta, S. & Miescke, K. (1994), 'Bayesian look ahead one-stage sampling allocations for selecting the largest normal mean', *Statistical Papers* **35**, 169–177. 3, 7, 8

Kaelbling, L. P. (1993), *Learning in embedded systems*, MIT Press, Cambridge, MA. 2, 23

Katehakis, M. & Veinott Jr, A. (1987), 'The Multi-Armed Bandit Problem: Decomposition and Computation', *Mathematics of Operations Research* **12**(2), 262–268. 3

Lai, T. L. & Robbins, H. (1985), 'Asymptotically efficient adaptive allocation rules', *Advances in Applied Mathematics* **6**, 4–22. 3

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. 2

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. 3

Powell, W. B. (2007), *Approximate Dynamic Programming: Solving the curses of dimensionality*, John Wiley and Sons. 2

Sutton, R. & Barto, A. (1998), *Reinforcement Learning*, The MIT Press, Cambridge, Massachusetts. 2

UKPDS Group (1998), 'UK Prospective Diabetes Study 34: Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes', *The Lancet* **352**(9131), 854–865. 20

Vermorel, J. & Mohri, M. (2005), 'Multi-armed bandit algorithms and empirical evaluation', *Proceedings of the 16th European Conference on Machine Learning* pp. 437–448. 3

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. 2

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. 3, 23