The Knowledge Gradient Algorithm For Online Subset Selection

Ilya O. Ryzhov and Warren Powell

***Abstract*— We derive a one-period look-ahead policy for online subset selection problems, where learning about one subset also gives us information about other subsets. We show that the resulting decision rule is easily computable, and present experimental evidence that the policy is competitive against other online learning policies.**

I. Introduction

Suppose that we have a finite set of objects, each with an unknown effectiveness or reward, and we must choose the best possible subset of these objects. The objects may interact in unknown ways, so it may be beneficial to leave some out, or to include certain specific combinations. We are allowed to make N sequential measurements, each of which allows us to observe a sample realization of the reward of a subset, which can then be used to learn about the effectiveness of different subsets. Our objective is to allocate these measurements in order to maximize the total reward of all subsets that we measure. Crucially, we allow our beliefs about the unknown rewards to be correlated, which makes it so that one measurement may affect our beliefs about many different subsets. Correlated beliefs are not handled by the traditional bandit literature, but are central in the subset selection problem.

The problem of allocating experiments is important when we wish to find the best subset, but have relatively few measurements to work with. Furthermore, we specifically deal with online problems in this paper, where we are interested not only in finding the best subset, but also in maximizing the total reward collected over the entire time horizon. Let us give several motivating examples in which this distinction is important:

1) *Energy management*. We are installing sets of energy-saving technologies (e.g. insulation, solar panels, tinted windows) in large buildings. Different technologies interact in an unknown way that can only be measured by actually implementing portfolios of technologies and measuring their combined performance, that is, the percent reduction in heating costs. We maximize total cost reduction over all buildings.
2) *Clinical trials*. We are testing experimental drug treatments on groups of human patients. One treatment consists of multiple drugs, so our beliefs about two treatments are correlated if the treatments have at least one drug in common. We are interested in the well-being of the patients as well as in finding the best treatment, so we maximize the total benefit to all patients tested.
3) *Sports*. We have a set of players, and we can choose a different subset to play each game, with the goal of finding the most effective team. The effectiveness of a team depends on interactions between individual players, so we must play a game with one subset in order to gauge its effectiveness. We maximize the total number of games won in the season.

These problems can be viewed using the language of multi-armed bandit problems, where each subset is an individual "alternative" with an unknown reward, and we maximize the total reward collected across all measurements. In particular, [7] describes a measurement policy (which we refer to as "Gittins indices") that is asymptotically optimal as $N \to \infty$ for the case where the rewards are discounted over time. Furthermore, there are certain general heuristics (descriptions can be found e.g. in [15]) that can be applied to broad classes of optimal learning problems, including multi-armed bandits: the interval estimation policy by [11], the Boltzmann exploration policy, pure exploitation, and so on. The work by [17] presents empirical comparisons of some policies in certain settings.

However, neither the Gittins policy, nor any of the general heuristics mentioned above, is designed to handle correlated prior beliefs. The correlated case has been studied by [14], but this work relies on the assumption of binomial rewards. Yet, the ability to handle correlated beliefs is of great practical significance because it arises in any subset selection problem. In our motivating examples, the alternatives are correlated because they have common elements (common technologies, common drugs, common players). Essentially any problem where we must choose a representative set of objects poses a potential application. Economic applications can be found in [10], and applications to statistics are extensively discussed by [13].

Our analysis is motivated by the knowledge gradient (KG) concept, developed by [9] and further analyzed by [6] and [2] for the ranking and selection problem. This problem is the offline version of the multi-armed bandit problem: we must find the best of M alternatives with unknown rewards, given N chances to learn about them first, with no regard for the outcomes of the measurements beyond the information we get from them. 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. More recently, [6] extended

The authors are with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ (email: {iryzhov,powell}@princeton.edu).

This work 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.

the KG concept to the ranking and selection problem with correlated priors, and [2] extended it to the case of unknown measurement noise.

The knowledge gradient has the practical advantage of being easily computable, in contrast with the difficult calculations required for Gittins indices. The computation of Gittins indices has been discussed by [12] and [4]. An LP-based computational method is available in [1], however it is founded on a Markov decision process framework, in which the prior beliefs about the alternatives are discrete, whereas our problem has continuous, Gaussian priors. For our problem, an approximation for Gittins indices can be found in [18], but it is less accurate for small time horizons and large discount factors. However, experimental work by [16] suggests that the online KG policy is competitive against the Gittins policy even when the Gittins indices are known exactly.

In this paper, we begin with a mathematical model for an undiscounted multi-armed bandit problem, and derive the online KG policy for this setting. Then, we extend it to the correlated setting most relevant to subset selection. We compare the online KG policy against existing learning policies in an undiscounted, finite-horizon setting with correlated priors, motivated by the energy management problem described above. The experiments demonstrate that the KG policy is comparable to, or even outperforms, these other policies across many instances of the energy management problem.

II. Mathematical model for learning

We begin by considering a traditional multi-armed bandit problem. 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 observe a random reward $\hat{\mu}_x \sim \mathcal{N}\left(\mu_x, \sigma_\varepsilon^2\right)$. The measurement error σ_ε^2 is known, and we use the notation $\beta_\varepsilon = \sigma_\varepsilon^{-2}$ to denote the measurement precision. For every alternative x, the true expected reward μ_x is unknown, but we believe that $\mu_x \sim \mathcal{N}\left(\mu_x^0, \left(\sigma_x^0\right)^2\right)$. Thus, our prior beliefs about alternative x are completely characterized by μ_x^0 and σ_x^0. In the multi-armed bandit setting, our beliefs about the alternatives are independent, that is, measuring x allows us to learn only about x. We shall add correlations in Section III-C.

The random observations we make while measuring gradually improve our beliefs about the alternatives. We 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, and we use the notation $\mathbb{E}^n$ to denote the expected value given $\mathcal{F}^n$. Then, we can define $\mu_x^n = \mathbb{E}^n \mu_x$ to be our beliefs about μ_x after making exactly n measurements. Similarly, $\left(\sigma_x^n\right)^2 = Var\left(\mu^x \,|\, \mathcal{F}^n\right)$ represents the accuracy of our beliefs about μ_x at time n. We use the notation $\beta_x^n = \left(\sigma_x^n\right)^{-2}$ to denote the time-n precision of our beliefs. Thus, at time n, we believe that $\mu_x \sim \mathcal{N}\left(\mu_x^n, \left(\sigma_x^n\right)^2\right)$. If x^n is the alternative we choose to measure at time n, our beliefs after that measurement will be updated using Bayes' rule:

$$\mu_x^{n+1} = \begin{cases} \frac{\beta_x^n \mu_x^n + \beta_\varepsilon \hat{\mu}_x^{n+1}}{\beta_x^n + \beta_\varepsilon} & x = x^n \\ \mu_x^n & x \neq x^n \end{cases} \tag{1}$$

Because our beliefs about the rewards are independent, we only update one set of beliefs in each time step, corresponding to the object we have chosen. The precision of our beliefs is updated as follows:

$$\beta_x^{n+1} = \begin{cases} \beta_x^n + \beta_\varepsilon & x = x^n \\ \beta_x^n & x \neq x^n \end{cases} \tag{2}$$

We can define $\left(\tilde{\sigma}_x^n\right)^2 = Var\left(\mu_x^{n+1} \,|\, \mathcal{F}^n\right)$ to be the conditional variance of our beliefs after the next measurement, viewed from before the measurement. It can be shown that this quantity is simply the variance reduction in our beliefs about x that we achieve by measuring x, that is,

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

It is known, e.g. from [3], 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

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

where Z is standard Gaussian. Thus, viewed from the point of view of time n, the $(n+1)$st measurement is not expected to change our beliefs.

We now define the knowledge state $s^n = (\mu^n, \sigma^n)$, where $\mu^n = (\mu_1^n, ..., \mu_M^n)$ and $\sigma^n = (\sigma_1^n, ..., \sigma_M^n)$. Because μ_x^n and σ_x^n completely characterize our beliefs about alternative x at time n, it follows that s^n completely characterizes our beliefs about all the alternatives at time n. We can 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 can suppress the dependence on $\hat{\mu}_{x^n}^{n+1}$ when we write K^M, but it is important to remember that the transition function is stochastic.

In online problems, we collect rewards as we measure them. We assume that we are allowed to make N measurements, followed by one final reward at time N. Then, our objective is to choose a measurement policy π that achieves

$$\sup_\pi \mathbb{E}^\pi \sum_{n=0}^{N} \mu_{X^{\pi,n}(s^n)}. \tag{3}$$

A policy π can be thought of as a collection of decision rules $X^{\pi,n}$ for $= 0, ..., N$. The decision rule $X^{\pi,n}$ is a random function mapping an outcome ω and a knowledge state s^n to a point in the set of alternatives $\{1, ..., M\}$. We assume that $X^{\pi,n}$ is measurable with respect to $\mathcal{F}^n$, that is, the decision rule for time n is allowed to observe the outcomes of the first n measurements.

The value of following a measurement policy π, starting in knowledge state s^n at time n, is given by Bellman's

equation for dynamic programming (used in an optimal learning context by [3]):

$$V^{\pi,n}\left(s^n\right) = \mu^n_{X^{\pi,n}(s^n)} + \mathbb{E}^n V^{\pi,n+1}\left(s^{n+1}\right) \quad (4)$$

$$V^{\pi,N}\left(s^N\right) = \max_x \mu^N_x. \quad (5)$$

At time N, we can collect only one more reward, so 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 best possible measurement policy satisfies a similar equation

$$V^{*,n}\left(s^n\right) = \max_x \mu^n_x + \mathbb{E}^n V^{\pi,n+1}\left(s^{n+1}\right)$$

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

with the only difference being that it always chooses the best possible measurement, the one that maximizes the sum of the immediate and downstream rewards. Of course, viewed at time n, the knowledge state s^{n+1} evolves stochastically from s^n via the transition function K^M.

III. The knowledge gradient policy

We extend the knowledge gradient concept developed by [5] for the ranking and selection problem to the multi-armed bandit setting. We define the knowledge gradient policy to be the policy that will make an optimal allocation if we have only one measurement left. Finally, we incorporate discount factors and correlated beliefs into the policy, thus allowing us to solve subset selection problems.

A. Derivation

Suppose that we have made n measurements, reached the knowledge state s^n, and then stopped learning entirely. That is, we are still allowed to collect rewards after time n, but we are not allowed to use these rewards to update our beliefs using (1) and (2), and $s^{n'} = s^n$ for all $n' > n$. Then, the best possible policy is the empirical Bayesian policy of always choosing the alternative that looks the best based on the most recent information. Under these conditions, the expected total reward obtained by the empirical Bayesian policy after time n is

$$V^{EB,n}\left(s^n\right) = (N-n+1)\max_x \mu^n_x. \quad (6)$$

If we cannot learn any more information, 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 concept, first introduced by [8] and [9], and later developed by [5], can be stated as "choosing the measurement that would be optimal if it were our last chance to learn." Suppose now that we are at time n, with $N-n+1$ rewards left to collect, but only the $(n+1)$st reward can be used to update our beliefs. That is, $s^{n'} = s^{n+1}$ for all $n' > n+1$. Then, we need to make one decision about what to measure at time n, and we will switch to the empirical Bayesian policy starting at time $n+1$. The KG decision rule that follows from this assumption is

$$X^{KG,n}\left(s^n\right) = \arg\max_x \mu^n_x + \mathbb{E}^n V^{EB,n+1}\left(s^{n+1}\right). \quad (7)$$

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 (7) can be written as

$$\begin{aligned} & \mathbb{E}^n V^{EB,n+1}\left(s^{n+1}\right) \\ = \; & (N-n)\,\mathbb{E}^n \max_{x'} \mu^{n+1}_{x'} \\ = \; & (N-n)\,\mathbb{E}\max\left\{\max_{x'\neq x}\mu^n_{x'}, \mu^n_x + \tilde{\sigma}^n_x \cdot Z\right\} \\ = \; & (N-n)\left(\max_{x'}\mu^n_{x'}\right) + (N-n)\,\nu^{KG,n}_x \end{aligned} \quad (8)$$

where the computation of $\mathbb{E}^n \max_{x'} \mu^{n+1}_{x'}$ comes from [5]. The quantity $\nu^{KG,n}_x$ is called the *knowledge gradient* of alternative x at time n, and is defined by

$$\nu^{KG,n}_x = \mathbb{E}^n_x\left(\max_{x'}\mu^{n+1}_{x'} - \max_{x'}\mu^n_{x'}\right), \quad (9)$$

where $\mathbb{E}^n_x$ 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

$$\nu^{KG,n}_x = \tilde{\sigma}^n_x \cdot f\left(-\left|\frac{\mu^n_x - \max_{x'\neq x}\mu^n_{x'}}{\tilde{\sigma}^n_x}\right|\right) \quad (10)$$

where $f(z) = z\Phi(z) + \phi(z)$ and ϕ,Φ are the pdf and cdf of the standard Gaussian distribution. We know from [9] and [5] that (9) and (10) are equivalent in this problem, and that ν^{KG} is always positive. The term "knowledge gradient" arises from (9), where the quantity $\nu^{KG,n}_x$ is written as a difference.

It is now easy to see that (7) can be rewritten as

$$X^{KG,n}\left(s^n\right) = \arg\max_x \mu^n_x + (N-n)\,\nu^{KG,n}_x. \quad (11)$$

The term $(N-n)\max_{x'}\mu^n_{x'}$ in (8) is dropped because it does not depend on the choice of x and thus does not affect which x achieves the maximum in (7). The value of this policy follows from (4) and is given by

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

Thus, instead of choosing the alternative that appears to be the best, the KG policy adds an uncertainty bonus of $(N-n)\,\nu^{KG,n}_x$ to the most recent beliefs μ^n_x, and chooses the alternative that maximizes this sum. The knowledge gradient is given extra weight in the early stages, when n is small compared to N. 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 (11) is common in optimal learning algorithms. Many other learning policies involve similar computations, but define the uncertainty bonus in different ways. In our case, it represents the value of learning one more time.

The KG policy is optimal for $N = 1$. This follows from the fact that, for any N,

$$V^{*,N-1}\left(s^{N-1}\right) = V^{KG,N-1}\left(s^{N-1}\right),$$

which is shown in [16]. This suggests that the policy given by (11) is the correct extension of the KG concept to the multi-armed bandit problem, as all KG policies derived thus far for offline problems share the property of being optimal for $N = 1$. It is also shown in [16] that

$$b^n \leq V^{KG,n}(s) - V^{EB,n}(s) \leq c^n$$

for all times n and knowledge states s, where

$$\begin{aligned} b^n &= \max_x \left(\mu_x^n + (N-n)\,\nu_x^{KG,n}\right) - \max_x \mu_x^n \\ c^n &= \frac{1}{\sqrt{2\pi}}\left[\frac{(N-n)\,(N-n+1)}{2}\right]\max_x \tilde{\sigma}_x^n. \end{aligned}$$

Because $b^n \geq 0$, it follows that $V^{EB,n}$ is a lower bound on $V^{KG,n}$. Furthermore, because $V^{EB,n}$ represents the best value that can be obtained if learning stops at time n, the difference $V^{KG,n} - V^{EB,n}$ is precisely the value of learning, starting at time n, under the online KG policy.

B. Discount factors

We now replace (3) 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. The knowledge gradient policy for this problem is derived in the same way as in the previous section. First, in the discounted setting,

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

Repeating the calculation of (7) for the discounted case, we arrive at

$$X^{KG,n}(s^n) = \arg\max_x \mu_x^n + \gamma\frac{1-\gamma^{N-n}}{1-\gamma}\nu_x^{KG,n} \qquad (12)$$

where $\nu_x^{KG,n}$ is as in (10). If we take $N \to \infty$, we obtain the infinite-horizon discounted KG rule

$$X^{KG,n}(s^n) = \arg\max_x \mu_x^n + \frac{\gamma}{1-\gamma}\nu_x^{KG,n}. \qquad (13)$$

Both (12) and (13) look similar to (11), with a different multiplier in front of the knowledge gradient. The value of the KG policy in the discounted setting is

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

In [16], it is shown that the infinite-horizon, discounted KG policy converges almost surely. That is, if $N \to \infty$, only one alternative will be measured infinitely often by the KG policy. This alternative will not necessarily be the best alternative. However, the Gittins index policy, which is known to be optimal in the discounted infinite-horizon setting, does not necessarily converge to the best alternative either. Due to the discount factor, early 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 far in the future. The work by [16] contains an experimental comparison of the infinite-horizon, discounted KG policy to the optimal Gittins policy in a problem where the Gittins indices are known exactly, and the results suggest that KG is competitive against the optimal policy.

C. Correlated beliefs

We now return to the undiscounted setting, and the objective function from (3). 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 μ^0 is a vector of our beliefs about the true rewards, and Σ^0 is an $M \times M$ matrix representing the covariance structure of our beliefs about the true rewards. As before, if we choose to measure alternative x at time n, we observe a random reward $\hat{\mu}_x^{n+1} \sim \mathcal{N}\left(\mu_x, \sigma_\varepsilon^2\right)$. 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

$$\mu^{n+1} = \mu^n + \frac{\hat{\mu}_{x^n}^{n+1} - \mu_x^n}{\sigma_\varepsilon^2 + \Sigma^n_{x^n x^n}}\Sigma^n e_{x^n} \qquad (14)$$

$$\Sigma^{n+1} = \Sigma^n - \frac{\Sigma^n e_{x^n} e_{x^n}^T \Sigma^n}{\sigma_\varepsilon^2 + \Sigma^n_{x^n x^n}}, \qquad (15)$$

where x^n is the alternative chosen at time n, and e_{x^n} is a vector with 1 at index n and zeros everywhere else. Observe that, in the correlated setting, a single measurement leads us to update the entire vector μ^n, not just one component as before. The conditional distribution of μ^{n+1} given $\mathcal{F}^n$ is given by the vector equation

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

where Z is standard Gaussian and

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

The value of the empirical Bayesian policy, which we follow if we are unable to continue learning after time n, is still given by (6). The derivation of the online KG policy remains the same as in Section III-A, and we arrive at a decision rule

$$X^{KGC,n}(s^n) = \arg\max_x \mu_x^n + (N-n)\,\nu_x^{KGC,n} \qquad (16)$$

where

$$\nu_x^{KGC,n} = \mathbb{E}_x^n\left(\max_{x'} \mu_{x'}^{n+1} - \max_{x'} \mu_{x'}^n\right) \qquad (17)$$

as before. However, the right-hand side of (17) is more difficult to compute than the analogous expression in (9).

From the work by [6], it is known that

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

where the alternatives have been sorted in order of increasing $\tilde{\sigma}_y^{corr,n}$, f is the same as in Section III-A, and the numbers c_y are such that $y = \arg\max_{x'} \mu_{x'}^n + \tilde{\sigma}_{x'}^{corr,n}(x) \cdot z$ for $z \in [c_{y-1}, c_y)$, with ties broken by the largest-index rule. The work by [6] also gives an efficient algorithm for computing ν^{KGC} exactly, and can be used to solve the decision problem in (16). If we introduce a discount factor into the problem, the decision rule becomes as in (12) or (13), with ν^{KGC} instead of ν^{KG}.

IV. Computational experiments

The correlated-prior setting is well-suited to the problem of subset selection. We used the problem of energy portfolio selection to obtain realistic initial parameters for experiments comparing online KG to other learning policies. On average, online KG is significantly better than all the policies tested. Some policies tend to outperform KG by a small margin more frequently, however they can be very sensitive to the initial parameters, and are extremely unreliable a sizable proportion of the time.

A. Background and setup of experiments

Suppose that we are remodeling buildings to be more energy-efficient. We have a number of energy-saving technologies at our disposal, and we can combine multiple technologies into portfolios. The reward of a portfolio is the percentage by which it reduces heating costs for the building in which it is installed. The effectiveness of several technologies is not additive, so the point of interest is the effectiveness of the entire energy portfolio. Research done by the US Department of Energy suggests that a possible range for such a value is the interval $[15, 45]$ (measured in percentage points).

We consider a setting in which there are seven different energy-saving technologies (e.g. insulation, solar panels, energy-saving light bulbs, tinted windows that reduce solar heat, energy-efficient appliances, extra thermostats for more precise temperature control, sealant for air leaks), and we must choose a portfolio of three of them. Thus, there are 35 possible alternatives in our problem, and our beliefs about their rewards are correlated because the same technologies may appear in multiple portfolios. We have $N+1$ buildings which we can allocate to different portfolios. Our objective is to maximize the sum of the heating cost reductions over all groups, with no discount factor. This objective function balances the need to find the best portfolio with concern for the outcome of each installation. It is expensive to install even one portfolio, so we would like to have good results in as many experiments as possible.

In order to test a learning policy, we must first assume a truth, then evaluate the ability of the policy to find that truth. For this reason, the starting data for our experiments was randomly generated, using the context of energy management to provide realistic numbers. Because our beliefs are correlated, we used the mathematical framework in Section III-C, and the updating equations (14) and (15) in all of our experiments. The initial data for one experiment consists of a vector μ to represent the true rewards of all the portfolios, a prior $\left(\mu^0, \Sigma^0\right)$ to represent our beliefs about them, and a measurement error σ_ε^2. We generated 100 problems in all.

For each problem, we first generated the prior means μ^0 from a uniform distribution on the interval $[15, 45]$. The initial variances were set to be 7.5^2, to roughly indicate that the true values were in that same interval. These numbers represent our beliefs about the range in which the true values are likely to fall. The correlation coefficient of two portfolios was set to be $0, \frac{1}{3}$ or $\frac{2}{3}$, depending on whether the portfolios had 0, 1 or 2 technologies in common. The measurement error σ_ε^2 was chosen to be 50, to reflect a situation where the effectiveness of a treatment varies fairly widely over different buildings. The true rewards μ were generated from a multivariate Gaussian distribution with mean vector μ^0 and covariance matrix Σ^0, that is, they were drawn from the prior. This represents a situation in which we already have a reasonably good idea about the portfolios, and our prior beliefs are on average accurate.

We ran each measurement policy 10^4 times on each problem, always with a measurement budget of $N = 25$. 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,

$$C^{\pi_2} - C^{\pi_1} = \frac{1}{N+1} \sum_{n=0}^{N} \left(\mu_{X^{\pi_1,n}(s^n)} - \mu_{X^{\pi_2,n}(s^n)}\right) \quad (18)$$

is precisely the amount by which π_1 outperformed (or underperformed) π_2 on average in a single measurement. 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 those averages. The standard error of the difference in (18) 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 online KG policies are defined by the decision rules (11) and (16), respectively. The KGC policy was implemented using the algorithm from [6].

Gittins indices (Gitt). The Gittins decision rule, designed for discounted infinite-horizon problems, can be written as

$$X^{Gitt,n}(s^n) = \arg\max_x \mu_x^n + \sigma_\varepsilon \cdot \Gamma(\sigma_x^n, \gamma), \quad (19)$$

where $\Gamma(\sigma_x^n, \gamma)$ is the Gittins index based on the accuracy of our current beliefs about an alternative and the discount

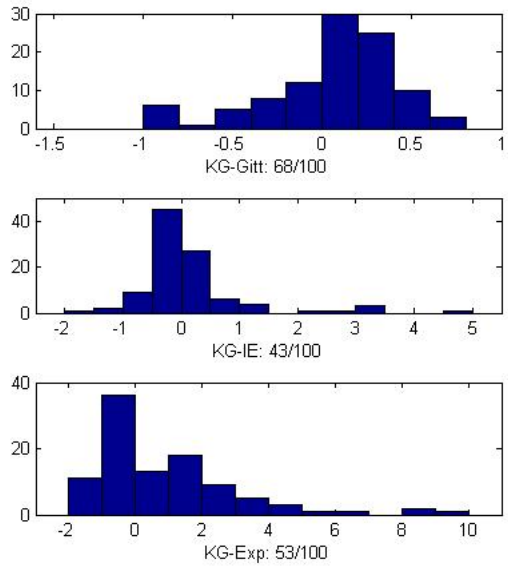

Fig. 1. Histograms of the sampled difference in opportunity cost for KG vs. other policies.

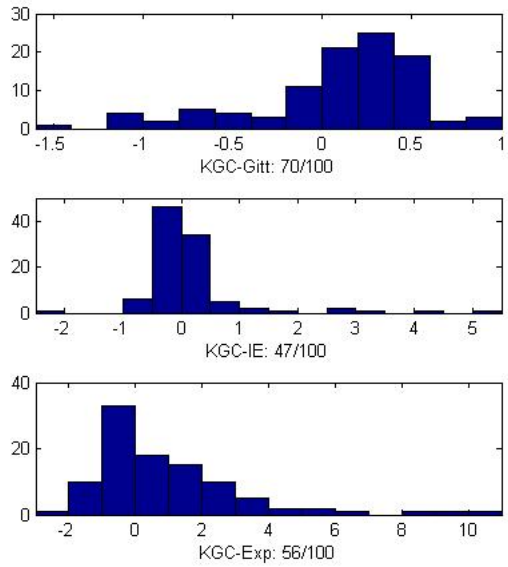

Fig. 2. Histograms of the sampled difference in opportunity cost for KGC vs. other policies.

factor γ. We can use the fact that $(\sigma_x^n)^2 \approx \frac{1}{N_x^n}$, 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. Then, (19) becomes

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

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 discount factor γ.

Even with the simplification allowed by the use of N_x^n, Gittins indices are typically very difficult to compute. Exact values are available only for a few values of γ in [7]. In order to allow us to tune the discount factor and consider values of γ for which the exact values are unknown, one can use the approximation from [18]. Define a function

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

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

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

Finally, take the Gittins index to be

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

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. In our experiments, we found that the approximation worked best for $\gamma \approx 0.9$.

Interval estimation (IE). The IE decision rule, created by [11], 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 found that the performance of IE was extremely sensitive to the choice of tuning parameter. Low values of the parameter, e.g. $z_{\alpha/2} = 0.25$, yielded good performance on many problems, but extremely bad performance on a significant proportion of problems. Larger values, e.g. $z_{\alpha/2} = 1.25$, gave better performance in these bad cases, but worse performance in the good cases. We compromised by setting $z_{\alpha/2} = 0.75$, which still gave good performance on most problems, while somewhat mitigating the worst outliers.

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.

B. Results

For each relevant comparison of two policies, we obtained 100 samples of the difference in (18). Table I gives the means and average standard errors of our estimates of (18) across the 100 problems we generated, for $N = 25$.

Figures 1 and 2 show 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: 68/100" means that the independent online KG policy outperformed the Gittins heuristic in 68/100

	Mean	Avg. SE
KG-Gitt	0.0545	0.0200
KGC-Gitt	0.0747	0.0200
KG-IE	0.1601	0.0205
KGC-IE	0.1802	0.0204
KG-Exp	0.9284	0.0220
KGC-Exp	0.9486	0.0220
KGC-KG	0.0202	0.0198

TABLE I

MEANS AND STANDARD ERRORS FOR THE EXPERIMENTS.

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 the Gittins policy about 70% of the time. Furthermore, KGC outperforms KG 63/100 times. The additional improvement brought on by correlated KG is, on average, small. KGC does not bring about a substantial improvement over KG. However, comparisons involving KGC tend to have greater positive tails than those of KG, so the correlated policy gives consistently better performance across many problems.

The IE policy actually outperforms the KG policy more than half the time, though usually by a small margin. However, there is a substantial minority of experiments where, under the KG policy, *every* measurement achieves an additional average reduction in heating cost of as much as 5% over the IE policy. This is a noticeable difference in our problem context. Pure exploitation exhibits similar behaviour. It is usually similar to the KG policy, but occasionally shows very large positive tails.

These widely varying results are due to the relationship of the truth to the prior in this experiment. Because we generate the truth from the prior, on average, the alternative that has the highest prior actually tends to be the true best alternative. When this occurs, IE and pure exploitation tend to perform extremely well, because both will choose the best alternative in the first measurement. However, when the alternative with the highest prior is not the true best alternative, such mistakes of the prior tend to cause very poor performance of IE and pure exploitation. Thus, although pure exploitation and IE behave very well in many cases, we argue that these policies are unreliable because they are prone to very bad errors.

It is also interesting to examine how much exploration is performed by each policy. For each problem we generated, we computed the number of distinct alternatives measured by each policy, averaged over 10^4 sample paths. The global averages, over 100 problems, are given in Table II. We see that, in 25 measurements, every policy tends to explore at most five different alternatives. Unsurprisingly, pure exploitation does the least exploration, because it has no uncertainty bonus and no incentive to measure anything that doesn't seem to be the best. The KG policy does more exploration than IE and pure exploitation, which helps it obtain better results in those cases where the prior is inaccurate.

Policy	No. Explored
KG	3.9370
KGC	3.6872
Gitt	5.0813
IE	3.1387
Exp	2.1295

TABLE II

NUMBER OF DISTINCT ALTERNATIVES EXPLORED BY EACH POLICY.

The KGC policy does slightly less exploration than KG. This is also unsurprising, as KGC takes correlations into account when making decisions, and thus can have a more informed picture of the future than independent KG. Finally, the Gittins policy does the most exploration, but its relatively poor performance against KG suggests that this is too much. Thus, the KG policy balances exploration and exploitation in a way that is better equipped to deal with mistakes in the prior.

V. CONCLUSION

We have proposed an easily computable decision rule for online learning problems. The KG policy proves to be versatile in a situation when we have a finite measurement budget, normally distributed priors, and normal sampling errors with known variance. Variations of the basic KG decision rule cover both undiscounted and discounted, finite- and infinite-horizon problems, and can also accommodate correlated priors. This last ability makes the KG rule attractive for subset selection problems, where the priors have an extensive correlation structure. We compared the KG policy to several other measurement policies in a realistic subset-selection setting. The KG policy is either comparable to, or better than, the other policies tested, and strikes a good balance between exploration and exploitation. We believe that the KG policy is useful for online learning applications because of its ease of use and reliability.

REFERENCES

[1] D. Bertsimas and J. Nino-Mora, "Restless bandits, linear programing relaxations, and a primal-dual index heuristic," *Operations Research,* vol. 48, no. 1, 80-90, 2000.

[2] S. Chick, J. Branke and C. Schmidt, "New myopic sequential sampling procedures," 2007, submitted for publication.

[3] M.H. DeGroot, *Optimal Statistical Decisions.* John Wiley and Sons, 1970.

[4] M.O. Duff, 'Q-learning for bandit problems," Technical report, Dept. of Comp. Sci., University of Massachusetts, Amherst, MA, 1995.

[5] P.I. Frazier, W.B. Powell and S. Dayanik, "A knowledge-gradient policy for sequential information collection," *SIAM J. on Control and Optimization*, 2008, to appear.

[6] P.I. Frazier, W.B. Powell and S. Dayanik, "The knowledge gradient policy for correlated normal rewards," 2008, submitted for publication.

[7] J. Gittins, *Multi-Armed Bandit Allocation Indices.* John Wiley and Sons, New York, 1989.

[8] S. Gupta and K. Miescke, "Bayesian look ahead one stage sampling allocations for selecting the largest normal mean," *Statistical Papers*, vol. 35, 169-177, 1994.

[9] S. Gupta and K. Miescke, "Bayesian look ahead one stage sampling allocations for selection of the best population," *Journal of Statistical Planning and Inference*, vol. 54, no. 2, 229-244, 1996.

[10] W.C. Horrace, J.T. Marchand and T.M. Smeeding, "Ranking inequality: Applications of multivariate subset selection," *Journal of Economic Inequality*, vol. 6, no. 1, 5-32, 2008.

[11] L.P. Kaelbling, *Learning in Embedded Systems.* MIT Press, Cambridge, MA, 1993.

[12] M. Katehakis and A.F. Veinott Jr., "The Multi-Armed Bandit Problem: Decomposition And Computation," *Math. of OR*, vol. 12, no. 2, pp. 262-268, 1987.
[13] A.J. Miller, *Subset Selection in Regression, 2nd ed.* CRC Press, 2002.
[14] S. Pandey, D. Chakrabarti and D. Agarwal, "Multi-armed bandit problems with dependent arms," *Proceedings of the 24th International Conference on Machine Learning*, 721-728.
[15] W.B. Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality.* John Wiley and Sons, New York, 2007.
[16] I.O. Ryzhov, W.B. Powell and P.I. Frazier, "The knowledge gradient algorithm for a general class of online learning problems," 2008, submitted for publication.
[17] J. Vermorel and M. Mohri, "Multi-armed bandit algorithms and empirical evaluation," *Proceedings of the 16th European Conference on Machine Learning*, 437-448, 2005.
[18] Y. Yao, "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, 284-294, 2006.