Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

A Monte-Carlo Knowledge Gradient Method For Learning Abatement Potential Of Emissions Reduction Technologies

Ilya O. Ryzhov
Warren Powell

Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08540, U.S.A.

ABSTRACT

Suppose that we have ten emissions reduction technologies whose greenhouse gas abatement potential is unknown, and we wish to select an optimal portfolio of five of these technologies. This problem can be viewed as an online optimal learning problem with correlated prior beliefs. We propose a learning policy which uses Monte Carlo sampling to narrow down the choice set to a relatively small number of promising portfolios, and then applies a one-period look-ahead approach using knowledge gradients to choose among this reduced set. We present experimental evidence that this policy is competitive against other online learning policies that consider the entire choice set.

1 INTRODUCTION

We consider the problem of reducing greenhouse gas emissions by refitting residential households with various existing and emerging technologies. A study by McKinsey & Company (2007) finds potential for emissions abatement in some of the following areas:

- Residential lighting
- Energy-efficient water heaters
- Energy-efficient appliances
- Improved ventilation systems
- Efficient air conditioning equipment
- Heating systems and furnaces
- Increased use of solar energy (solar panels)
- Fuel economy packages for cars
- New shell improvements for residential buildings
- Shell retrofits for residential buildings

Suppose that we wish to select a portfolio of five of the emissions-reducing technologies on this list. We do not know the true emissions abatement that can be achieved by choosing any given portfolio, but we do have certain prior information about the abatement potential. Different technologies may interact in unknown ways, so the overall abatement potential of a portfolio is best measured by installing that portfolio into one or more residential buildings. By doing so, we can obtain more accurate information about the effectiveness of that portfolio, as well as all other portfolios that contain one or more of the same technologies.

Suppose that we are allowed to make N sequential trials. Each trial allows us to observe a sample realization of the abatement potential of a portfolio by observing the effectiveness of that portfolio in a group of buildings. Because each such trial is expensive, our objective is to allocate the trials in such a way as to maximize the total effectiveness of all portfolios chosen for testing. Thus, we have an online subset selection problem in which we must choose one out of $\binom{10}{5} = 252$ subsets.

Viewed in this way, the problem can be treated as a multi-armed bandit problem, where each subset is a distinct "alternative" with an unknown reward (abatement potential), and we maximize the total reward collected across all measurements. The traditional multi-armed bandit problem, which assumes that all alternatives have independent rewards, has been widely studied. A major breakthrough in this area was the development of index policies. In an index policy, every alternative is assigned an index which depends only on our most recent beliefs about that alternative, considered separately from all other alternatives. The policy then measures the alternative with the largest index. The index itself can be defined in different ways; the work by Lai and Robbins (1985) provides bounds on the number of times suboptimal alternatives are measured by certain classes of index policies. The culmination of the index policy approach is the Gittins index policy (Gittins 1989), which is asymptotically optimal as $N \to \infty$ in the presence of a discount factor. Additionally, there are many other general heuristics (described e.g. by Powell 2007) that have been proposed for online learning problems. These include interval estimation (Kaelbling 1993), the Boltzmann exploration policy, pure exploitation, and so on. An empirical comparison of some heuristics is available in Vermorel and Mohri (2005).

However, the portfolio selection problem described above has one crucial feature that is not considered by the traditional bandit literature, namely the presence of correlated beliefs. If two energy portfolios have one or more technologies in common, then their rewards are correlated. By observing the effectiveness of one portfolio, we also learn something about other portfolios that we believe to be similar. The correlated case has been studied by Pandey et al. (2007), but only in the case of binomial rewards. The ability to handle correlated beliefs is important in the context of energy portfolio selection, because our measurement budget N may only allow us to look at a small portion of the set of alternatives. In the example given above, we have 252 distinct alternatives, but we may only be able to test five or ten percent of the total number of possible portfolios. Thus, it becomes especially important to be able to use the correlation structure of the problem to intelligently choose those portfolios that provide the most useful information about the entire set of choices.

Our analysis builds on the knowledge gradient (KG) principle, originally developed by Gupta and Miescke (1996) and later studied by Chick et al. (2009) and Frazier et al. (2008) for the ranking and selection problem. This problem is the offline version of the multi-armed bandit problem: the objective is solely to find the alternative with the highest reward, with no additional regard for the outcome of each trial. The KG policy for ranking and selection always chooses the alternative that would be optimal if it were the last alternative we were allowed to measure. Thus, the policy 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. The KG concept was extended by Frazier et al. (2009) to the ranking and selection problem with correlated beliefs, and by Chick et al. (2009) to a setting with unknown measurement noise. The first KG policy for online problems was given by Ryzhov et al. (2008), and studied by Ryzhov and Powell (2009) in the specific context of subset selection. A summary of this work is given in Section 3.1.

An important advantage of KG in the previously studied problems has been its usability. In many settings, knowledge gradients are easy to calculate, unlike Gittins indices, which are well-known to be difficult to compute (see Katehakis and Veinott Jr 1987 and Duff 1995 for approaches to this issue). Furthermore, there is experimental evidence (Ryzhov et al. 2008) that suggests that the online KG policy is competitive against the Gittins policy even when the Gittins indices are known exactly. However, knowledge gradient methods require significantly more computational effort than simpler heuristics such as interval estimation or approximated Gittins indices (Yao 2006). When the number of alternatives become large, it becomes expensive to consider the correlation structure of the problem when making decisions.

In this paper, we propose a modification of the online KG policy that uses Monte Carlo sampling to reduce the set of choices, and then runs existing knowledge gradient algorithms on the reduced set. We compare the Monte Carlo KG policy (MCKG) against three other learning policies in the context of energy portfolio selection: an undiscounted, finite-horizon setting where the number of choices makes it costly to compute knowledge gradients for every possible portfolio, and the time horizon N is much smaller than the total number of alternatives. Our experiments demonstrate that MCKG retains the KG policy's ability to consider correlations when making a decision, while incurring little cost in performance relative to other policies.

2 MATHEMATICAL MODEL

We present a mathematical model for an online learning problem with multivariate normal rewards and a general covariance structure. This model has been previously used by Frazier et al. (2009) for the correlated ranking and selection problem. Suppose that there are M distinct energy portfolios. In every time step, we can perform a trial on any one portfolio of our choice. By measuring portfolio x, we make a random observation $\hat{\mu}_x \sim \mathcal{N}\left(\mu_x, \sigma_\varepsilon^2\right)$ of the effectiveness of x. The quantity μ_x is the true abatement potential of portfolio x, and is unknown to us. Letting $\mu = (\mu_1, ..., \mu_M)^T$ be the vector of true values for all portfolios, we can express our initial beliefs about the true values with a multivariate normal distribution,

$$\mu \sim \mathcal{N}\left(\mu^0, \Sigma^0\right),$$

where $\mu^0 = \left(\mu_1^0, ..., \mu_M^0\right)$ is a vector containing our beliefs about each portfolio, and Σ^0 is an $M \times M$ matrix representing the uncertainty of our beliefs, as well as the covariance structure of the true values. The initial parameters μ^0 and Σ^0 thus completely characterize our prior beliefs about all portfolios.

Let $\mathscr{F}^n$ be the sigma-algebra generated by our choices of portfolios for the first n trials, as well as the observations we made of their true values. We use the notation $\mathbb{E}^n$ to mean the expected value given $\mathscr{F}^n$, and we say that something happens "at time n" if it happens after we have made exactly n observations, and before we have made our decision about the $(n+1)$st trial. Then, we can define $\mu^n = \mathbb{E}^n \mu$ to be our beliefs about μ after making exactly n measurements. Similarly, $\Sigma^n = Var\left(\mu \,|\, \mathscr{F}^n\right)$ represents the time-n uncertainty and correlation structure of our beliefs. Thus, at time n, we believe that $\mu \sim \mathscr{N}\left(\mu^n, \Sigma^n\right)$.

If we measure portfolio x^n at time n, the evolution of our beliefs is given by the equations

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

$$\Sigma^{n+1} = \Sigma^n - \frac{\Sigma^n e_{x^n} e_{x^n}^T \Sigma^n}{\sigma_\varepsilon^2 + \Sigma_{x^n x^n}^n} \quad (2)$$

where $\hat{\mu}_{x^n}^{n+1}$ is the random observation we make as a result of the measurement, and e_{x^n} is a column vector with component x^n equal to 1 and all other components equal to zero. Thus, one measurement has the potential to change our entire vector of beliefs, and the new beliefs will have less uncertainty because the components of Σ^{n+1} are smaller than the components of Σ^n. We assume that the random rewards $\hat{\mu}_{x^n}^n$ are independent of each other, conditionally on x^n.

We define the knowledge state $s^n = \left(\mu^n, \Sigma^n\right)$. The knowledge state completely characterizes our beliefs about all the alternatives at time n. Then, 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).

The energy portfolio problem is online, because we wish to ensure good outcomes for all trials, in addition to learning from them. We assume that we are allowed to make N trials in all, followed by one final reward at time N. When we allocate a trial at time n, we are allowed to use the information encoded in the knowledge state s^n to make our decision. Our goal is to arrive at a set of rules for allocating each trial efficiently based on the most recent information available at a given time. A measurement policy π represents a collection of decision rules $X^{\pi,n}$ for $n = 0, ..., N$. The decision rule $X^{\pi,n}$ maps a knowledge state s^n to a point in $\{1, ..., M\}$ representing an alternative. We require that $X^{\pi,n}$ be measurable with respect to $\mathscr{F}^n$, that is, the decision rule at time n is allowed to know the past leading up to time n, but not the future. Our objective is to choose a measurement policy π that achieves

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

The value of following a policy π, starting in knowledge state s^n at time n, is given by Bellman's equation for dynamic programming, first adapted to the context of optimal learning problems by DeGroot (1970):

$$V^{\pi,n}\left(s^n\right) = \mu_{X^{\pi,n}(s^n)}^n + \mathbb{E}^n V^{\pi,n+1}\left(K^M\left(s^n, X^{\pi,n}\left(s^n\right), \hat{\mu}_{X^{\pi,n}(s^n)}^{n+1}\right)\right) \quad (3)$$

$$V^{\pi,N}\left(s^N\right) = \max_x \mu_x^N. \quad (4)$$

At time N, we can collect only one more reward, and the information we collect from this reward will no longer be useful, so we choose the alternative that seems to be the best based on the most recent information. At time $n < N$, we expect to collect a reward of $\mu_{X^{\pi,n}(s^n)}^n$ from the portfolio we measure, as well as an expected downstream reward for future measurements. The best possible measurement policy satisfies a similar equation

$$V^{*,n}\left(s^n\right) = \max_x \mu_x^n + \mathbb{E}^n V^{\pi,n+1}\left(K^M\left(s^n, x, \hat{\mu}_x^{n+1}\right)\right) \quad (5)$$

$$V^{*,N}\left(s^N\right) = \max_x \mu_x^N. \quad (6)$$

We take a time-n expectation of the downstream reward because s^{n+1} evolves stochastically from s^n according to the transition function K^M.

3 THE MONTE-CARLO KNOWLEDGE GRADIENT POLICY

We review the existing work on knowledge gradients, including their extension to correlated online problems. We then propose a measurement policy that uses Monte Carlo simulation to reduce the set of alternatives, and then applies online KG methods to that reduced set. This methodology can be used to reduce the computational effort needed to calculate knowledge gradients in the energy portfolio problem, where the number of alternatives can be relatively large.

3.1 Overview of Knowledge Gradients

We briefly summarize the derivation of the online KG policy given by Ryzhov et al. (2008) and Ryzhov and Powell (2009). Suppose that it is time n, and we have reached state s^n. Now, suppose that the knowledge state will remain fixed for the rest of the time horizon, that is, $s^{n'} = s^n$ for all $n' > n$. We will continue to select portfolios and collect rewards at each time step, but these rewards will no longer impact our beliefs through (1) and (2). Then the best possible policy is the one that always chooses the alternative that we believe to be the best given the most current information. The expected total reward obtained by this "stop-learning" policy is

$$V^{SL,n}\left(s^n\right) = \left(N-n+1\right)\max_x \mu_x^n \tag{7}$$

because we will collect exactly $N-n+1$ more rewards until the end of the time horizon.

The knowledge gradient concept, first introduced by Gupta and Miescke (1994), can be stated as, "choosing the measurement that would be optimal if it were our last chance to learn." Suppose that the time-n measurement is now the last measurement that will change our beliefs. That is, $s^{n+1} = K^M\left(s^n, x^n, \hat{\mu}_{x^n}^{n+1}\right)$ as usual, but $s^{n'} = s^{n+1}$ for all $n' > n+1$. Then, after we decide what to do at time n, we will switch to the SL policy starting at time $n+1$. The KG decision rule following from this assumption is found using Bellman's equation:

$$\begin{aligned} X^{KG,n}\left(s^n\right) &= \arg\max_x \mu_x^n + \mathbb{E}^n V^{SL,n+1}\left(K^M\left(s^n, x, \hat{\mu}_x^{n+1}\right)\right) \\ &= \arg\max_x \mu_x^n + \left(N-n\right)\mathbb{E}^n \max_{x'} \mu_{x'}^{n+1} \\ &= \arg\max_x \mu_x^n + \left(N-n\right)\mathbb{E}^n \left(\max_{x'} \mu_{x'}^{n+1} - \max_{x'} \mu_{x'}^n\right). \end{aligned} \tag{8}$$

The last equality holds because the quantity $\max_{x'} \mu_{x'}^n$ does not affect the argmax. The quantity

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

is called the knowledge gradient of portfolio x at time n. Thus, the KG policy yields the decision rule

$$X^{KG,n}\left(s^n\right) = \arg\max_x \mu_x^n + \left(N-n\right)\nu_x^{KG,n}. \tag{9}$$

It remains to compute the quantity $\nu^{KG,n}$. In the special case where Σ^0 is diagonal (all portfolios are assumed to be independent), the knowledge gradient can be found using the explicit formula (Gupta and Miescke 1996, Frazier et al. 2008)

$$\nu_x^{KG,n} = \tilde{\sigma}_x^n \cdot f\left(-\left|\frac{\mu_x^n - \max_{x' \neq x} \mu_{x'}^n}{\tilde{\sigma}_x^n}\right|\right) \tag{10}$$

where $\tilde{\sigma}_x^n = \sqrt{\left(\sigma_x^n\right)^2 - \left(\sigma_x^{n+1}\right)^2}$ is the variance reduction achieved by measuring x and $f\left(z\right) = z\Phi\left(z\right) + \phi\left(z\right)$ with Φ, ϕ being the standard normal cdf and pdf, respectively. For arbitrary covariance matrices, the knowledge gradient is given by (Frazier et al. 2009)

$$\nu_x^{KG,n} = \sum_{y=1}^{M-1} \frac{b_{y+1}^n - b_y^n}{\sqrt{\sigma_\varepsilon^2 + \Sigma_{xx}^n}} f\left(-\left|c_y\right|\right) \tag{11}$$

where f is as before, the vector b^n is equal to the xth column of Σ^n with the components sorted in increasing order, and the numbers c_y are such that

$$y = \arg\max_{x'} \mu^n_{x'} + \frac{\Sigma^n_{x',x}}{\sqrt{\sigma^2_\varepsilon + \Sigma^n_{xx}}} \cdot z \qquad z \in [c_{y-1}, c_y)$$

with ties broken by the largest-index rule. The work by Frazier et al. (2009) also gives an algorithm for computing (11) exactly.

Much of the research done on online learning problems has been devoted to the development of index policies of the form $X^{\pi,n}(s^n) = \arg\max_x \mu^n_x + I_x(\mu^n_x, \Sigma^n_{xx})$, where the index I_x of choice x depends only on our beliefs about choice x. Index policies can be shown to have desirable asymptotic properties in the case where Σ^0 is diagonal. In particular, the work by Lai and Robbins (1985) establishes that the expected number of times suboptimal choices are measured under a certain class of index policies is $O(\log N)$. Various index policies for which this property holds were put forth by Auer et al. (2002), although these policies rely on restrictive assumptions about the range of the true values. Furthermore, the Gittins index policy by Gittins (1989) is known to collect information optimally if the choices are independent.

The knowledge gradient policy is not an index policy. It is clear from (10) and (11) that the knowledge gradient depends on our beliefs about all choices, not just the one under consideration. However, it is precisely this feature that makes the KG policy suitable for correlated problems, such as the problem of energy portfolio selection. Index policies are inherently not designed for this setting, because they rely on the ability to look at each choice separately from the others, ignoring the covariance structure of the problem. The work by Ryzhov et al. (2008) derives a number of theoretical properties of the KG policy, such as a bound on $V^{KG,n}(s^n)$, and presents experimental evidence showing that online KG is competitive against Gittins indices even when the latter are known exactly.

3.2 Action Elimination Using Monte Carlo Sampling

For a problem with arbitrary Σ^0, the KG decision rule given by (9) requires us to compute $\nu^{KG,n}_x$ for all x and all n. The algorithm given by Frazier et al. (2009) for computing the knowledge gradients in a single time step has complexity $O\left(M^2 \log M\right)$. In the energy problem under consideration, where there are $\binom{10}{5} = 252$ portfolios to choose from, it can be expensive to simulate the performance of the KG policy over many sample paths. We propose a method that uses Monte Carlo sampling to reduce the set of choices.

The intuition behind this procedure is as follows. Experimental work (see e.g. Ryzhov and Powell 2009) suggests that, in many online optimal learning problems, measurement policies such as KG and Gittins indices do not explore the entire choice set, instead preferring to vacillate between a small number of choices that "look good." The problem reduces to refining the distinctions between these top choices. We can use Monte Carlo sampling to find those top choices, and limit the KG computations to those choices only.

Suppose that we are at time n and knowledge state s^n. We can generate K sample realizations of the random variable $\bar{\mu}^n \sim \mathcal{N}(\mu^n, \Sigma^n)$. Let $\bar{\mu}^n(\omega_k)$ be the kth sample realization. Then, $\bar{\mu}^n(\omega_k)$ is an M-vector for each k. Let $x_k = \arg\max_x \bar{\mu}^n_x(\omega_k)$ be the energy portfolio that appears to be the best, based on this sample realization. Because some of our Monte Carlo samples may have the same argmax, let K_0 be the number of distinct portfolios obtained from this procedure, and let $x_1, ..., x_{K_0}$ represent those distinct portfolios. Now define a matrix A^n of size $M \times K_0$ by

$$A^n = \left[e_{x_1} \quad \ldots \quad e_{x_0}\right]$$

. Then the equations

$$\mu^{MC,n} = (A^n)^T \mu^n \tag{12}$$

$$\Sigma^{MC,n} = (A^n)^T \Sigma^n A^n \tag{13}$$

give the parameters of the time-n marginal distribution of $\left(\mu_{x_1}, ..., \mu_{x_{K_0}}\right)$. We can then apply the decision rule

$$X^{MCKG,n}(s^n) = \arg\max_k \mu^n_{x_k} + (N-n)\, \nu^{MCKG,n}_{x_k} \tag{14}$$

```
1. Begin with an initial knowledge state $s^0 = \left(\mu^0, \sigma^0\right)$, and set $n = 0$.
2. Let $R = \emptyset$. For $k = 1, ..., K$, do the following:
    (a) Generate a Monte Carlo sample $\bar{\mu}^n(\omega_k)$ from the distribution $\mathcal{N}(\mu^n, \Sigma^n)$.
    (b) Let $y = \arg\max_x \bar{\mu}^n_x(\omega_k)$.
    (c) If $y \notin R$, then add $y$ to $R$.
3. Letting $K_0 = |R|$ and $x_1, ..., x_{K_0}$ be the elements of $R$, compute $\mu^{MC,n}$ and $\Sigma^{MC,n}$ using (12) and (13).
4. If $K_0 = 1$, let $x^n = x_1$. Otherwise, compute
       $\max_k \mu^n_{x_k} + (N-n)\, \nu^{MCKG,n}_{x_k}$,
   where $\nu^{MCKG,n}_{x_k}$ is as in (11), using $K_0$ instead of $M$, and $\mu^{MC,n}$ and $\Sigma^{MC,n}$ instead of $\mu^n$ and $\Sigma^n$. Let $x^n$ be the choice that achieves this maximum.
5. Observe a random measurement $\hat{\mu}^{n+1}_{x^n} \sim \mathcal{N}\left(\mu_{x^n}, \sigma^2_\varepsilon\right)$.
6. Compute $s^{n+1}$ using (1) and (2).
7. Let $n = n+1$. If $n < N$, go to step 2.
8. Compute $\max_x \mu^N_x$ and measure the alternative that achieves this maximum.
```

Figure 1: Summary of the online MCKG policy.

where the knowledge gradients $\nu^{MCKG,n}$ are computed by running existing KG algorithms on the reduced choice set $\left\{x_1, ..., x_{K_0}\right\}$ with time-n distribution characterized by $\mu^{MC,n}$ and $\Sigma^{MC,n}$. This policy, which we call the Monte Carlo KG policy, is summarized in Figure 1.

To put it in words, we first find a set of K_0 different portfolios by generating K Monte Carlo samples of the time-n distribution of our beliefs, and taking the portfolio that seems to be the best from each sample. Then, given $\mathscr{F}^n$, the vector of true values for those portfolios has a multivariate normal distribution with mean $\mu^{MC,n}$ and covariance matrix $\Sigma^{MC,n}$. We can now compute knowledge gradients for the K_0 portfolios generated, and repeat the computation (11) using $\mu^{MC,n}$ and $\Sigma^{MC,n}$ instead of μ^n and Σ^n. This procedure returns a portfolio $X^{MCKG,n}(s^n)$.

The KG policy prefers to visit either those portfolios that we believe to be very good (the ones with high μ^n_x), or those portfolios that seem to have a lot of uncertainty and correlation with other portfolios (those with high $\nu^{KG,n}_x$). Intuitively, these are precisely the choices that will emerge from the Monte Carlo sampling. When we draw from the distribution $\mathcal{N}(\mu^n, \Sigma^n)$, the most likely winners are either those portfolios with the highest μ^n_x, or those portfolios with slightly lower μ^n_x but larger uncertainty.

It should be noted that the cost of computing knowledge gradients for every portfolio in a single time step is $O\left(M^2 \log M\right)$, whereas the cost of generating a single Monte Carlo sample from $\mathcal{N}(\mu^n, \Sigma^n)$ using the Cholesky factorization of Σ^n is $O\left(M^3\right)$. In practice, however, the computation of Cholesky factorizations in modern linear algebra packages (e.g. in MATLAB) has been extensively optimized, and the Monte Carlo samples can be generated as quickly as, or faster than, the knowledge gradients.

4 EXPERIMENTAL RESULTS

We compared the Monte Carlo KG policy to other learning policies in the context of the energy portfolio selection problem. On average, MCKG is significantly better than the other policies tested across 100 randomly generated problems. Furthermore, MCKG can achieve good performance with a small sample size, effectively eliminating the majority of the choice set.

4.1 Setup of Experiments

The most reliable way to validate a measurement policy is to run it on a problem where the true values of the choices are known, and evaluate its ability to find the true best choice. In practice, the true values are rarely known, but we can test a policy on randomly generated problems where the true values are fixed. We do not allow the policy to see the true values while it is making a decision, but we can use the true values to evaluate its performance at the end.

We compared MCKG to several measurement policies by running them on 100 randomly generated problems, using our motivating example of energy portfolio selection to provide context. The initial data for one problem consists of a vector μ containing the true values of each portfolio, a prior $\left(\mu^0, \Sigma^0\right)$ to represent our initial beliefs, and a measurement error σ_ε^2. Each problem has $M = 252$ distinct choices. We set the time horizon to $N = 25$, allowing ourselves to visit at most ten percent of the choice set. The study by McKinsey & Company (2007) finds that individual emissions reduction technologies can have abatement potential ranging from 0.2 to 3.0 gigatons of CO_2 equivalent per year. We generated our prior beliefs μ^0 about the abatement potential of a portfolio of five technologies from a uniform distribution on the interval $[3, 10]$. The initial variances were set to be 4, to roughly indicate that the true values were in a similar interval. The correlation coefficient of portfolios x and y was chosen to be $\frac{1}{5}c$, where $c \in \{0, 1, 2, 3, 4\}$ is the number of technologies that x and y had in common. The measurement error was chosen to be $\sigma_\varepsilon^2 = 3$, to reflect a situation where the observed effectiveness of a portfolio can vary fairly widely between groups of buildings. Finally, the true rewards μ were generated from a multivariate normal distribution with mean μ^0 and covariance matrix Σ^0. This represents a situation where our prior provides us with reasonably accurate information about the portfolios.

We ran each measurement policy 10^4 times on each problem. Our performance measure for each policy is the average opportunity cost per true reward collected, which we define 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. We can then compare two policies by taking the difference of their opportunity costs. Thus,

$$C^{\pi_2} - C^{\pi_1} = \frac{1}{N+1} \sum_{n=0}^{N} \mu_{X^{\pi_1,n}(s^n)} - \mu_{X^{\pi_2,n}(s^n)} \tag{15}$$

is the amount by which policy π_1 outperforms (or underperforms) policy π_2 on average in a single trial. For each policy, the 10^4 runs were divided into groups of 500 to obtain approximately normal samples of average opportunity cost and the standard errors of these averages. The standard error of (15) is the square root of the sum of the squared standard errors of C^{π_1} and C^{π_2}. In all, we tested five policies, which are briefly described below.

Monte Carlo KG (MCKG). The online MCKG policy is defined by the decision rule (14). The computation of knowledge gradients was implemented using the algorithm from Frazier et al. (2009). The number of samples was chosen to be $K = 25$. Thus, we only compute knowledge gradients for at most 10% of the choice set in every time step.

Independent KG (IndKG). The independent KG policy is defined by (9) with (10) used to compute knowledge gradients. This policy ignores the covariance structure and takes what would be the one-period look-ahead action if the portfolios were independent. The complexity of this policy is $O(M)$, and thus we do not use Monte Carlo sampling to reduce the choice set.

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

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

where γ is the discount factor and $\Gamma\left(\Sigma_{xx}^n, \gamma\right)$ is the Gittins index of portfolio x based on the information available at time n. To simplify this computation, we can approximate $\left(\sigma_x^n\right)^2 \approx \frac{1}{N_x^n}$ where N_x^n is the number of times portfolio x has been visited up to and including time n, which yields the decision rule

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

Our problem is not discounted, so we view γ as a tunable parameter. Gittins indices are typically difficult to compute; in order to allow ourselves to tune the discount factor and consider values of γ for which the exact Gittins indices are unknown, we use the following approximation from Yao (2006). Define a function

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

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 by Kaelbling (1993) is given by

$$X^{IE,n}(s^n) = \arg\max_x \mu^n_x + \sqrt{\Sigma^n_{xx}} \cdot z_{\alpha/2}$$

where $z_{\alpha/2}$ is a tunable parameter. The performance of IE is highly sensitive to the choice of tuning parameter: low values of $z_{\alpha/2}$ yield very good performance on many problems, but very bad performance on a significant proportion of problems, whereas high values reduce the extreme cases, but cause a drop in overall performance. We set $z_{\alpha/2} = 1$, which somewhat reduced the outliers while giving good performance on most problems.

Pure exploitation (Exp). The pure exploitation decision rule is given by $X^{Exp,n}(s^n) = \arg\max_x \mu^n_x$. It requires no tuning, and does not take the uncertainty of our beliefs into account.

4.2 Main Results

For each comparison of MCKG to another policy, we obtained 100 samples of the difference in (15), one for each problem we generated. Table 1 gives the means and average standard errors of our estimates of (15) across the 100 problems, for $N = 25$. On average, the MCKG outperformed all competing policies by a statistically significant amount.

Figure 2 shows the distribution of the sampled differences. Each histogram is labeled with 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 opposite. Thus, for instance, we see that the MCKG policy outperformed the Gittins heuristic in $76/100$ problems.

We see that the MCKG policy outperformed all competitors more than 60% of the time. In many cases, the difference in performance was small. However, all comparisons exhibit significant positive tails. For example, there are problems where the abatement potential of *every* portfolio chosen by MCKG is (on average) greater than the abatement potential of the portfolio chosen by pure exploitation by as much as 3 gigatons of CO_2 equivalent per year.

Table 1: Means and standard errors for the experiments.

	MCKG-IndKG	MCKG-Gitt	MCKG-IE	MCKG-Exp
Mean	0.0116	0.0930	0.0395	0.4915
Avg. SE	0.0071	0.0069	0.0071	0.0075

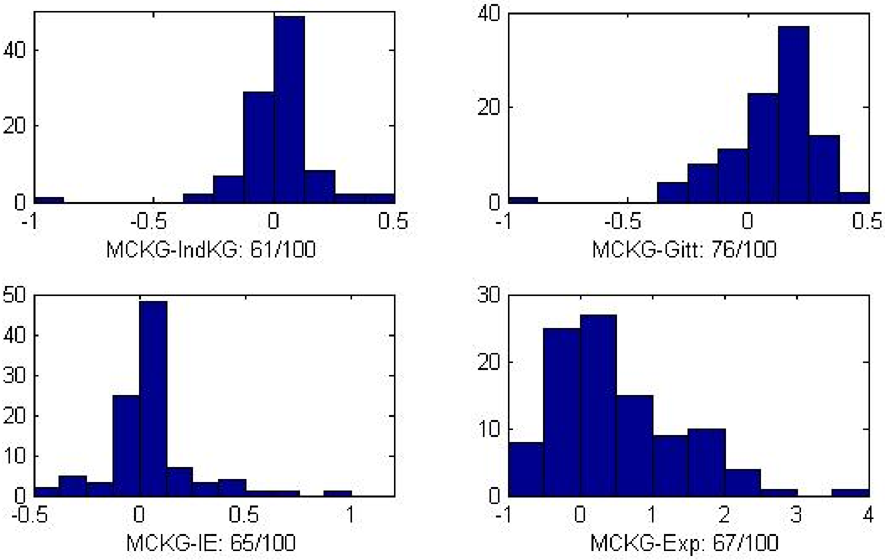

Figure 2: Histograms of the sampled difference in opportunity cost for MCKG vs. other policies.

In the case of the pure exploitation policy, we see notable negative tails. About one third of the time, the abatement potential of the pure exploitation portfolio is greater than the abatement potential of the MCKG portfolio by up to 1 gigaton of CO_2 equivalent per year. This behaviour arises from the fact that, in these experiments, the true values are generated from the prior values. Because the prior is accurate on average, it is often the case that the portfolio that seems to be the best at the very start of the trials is actually the true best portfolio, and pure exploitation immediately finds it. However, when the alternative with the highest prior value is not the true best alternative, we start to observe severe positive tails.

We also examined the number of distinct portfolios examined by each policy. The average values across 100 problems are given in Table 2. Although our time horizon allows us to look at up to 10% of the choice set, all five policies explore less than 4%. As expected, the pure exploitation policy does the least exploration. The MCKG policy is in the middle; because it takes correlations between portfolios into consideration when making decisions, it can afford to do less exploration than independent KG and the Gittins heuristic, which were designed for uncorrelated problems. We see that the MCKG policy, although it only looks at the learning potential of a reduced choice set, is still able to explore enough alternatives to achieve good performance.

4.3 Effect of Sample Size on KG Performance

Finally, we consider the effect of the sample size K on the performance of the MCKG policy. We chose the problem on which MCKG achieved its median performance (out of the full set of 100), and ran MCKG on that problem for $K = 5, 10, ..., 100$. Figure 3(a) shows the effect of K on C^{MCKG} for this problem. It is easily seen that larger values of K result in noticeably improved performance up to around $K = 30$. After that, the improvement is less than 0.01 for each increment of K, and begins to fluctuate. We conclude that $K = 30$ is sufficient to generate all the portfolios which MCKG would be interested in measuring, and larger values of K increase computational time without resulting in more exploration.

Figure 3(b) shows the effect of K on the number of distinct portfolios examined by MCKG. We see that MCKG does much more exploration for small values of K. The number of distinct alternatives steadily decreases until around $K = 30$, then levels off. For small sample sizes, the reduced choice set does not include all the portfolios that the knowledge gradient

Table 2: Number of distinct portfolios examined by each policy.

Policy	MCKG	IndKG	Gitt	IE	Exp
No. Explored	5.8325	6.2111	10.0880	5.5957	2.1988

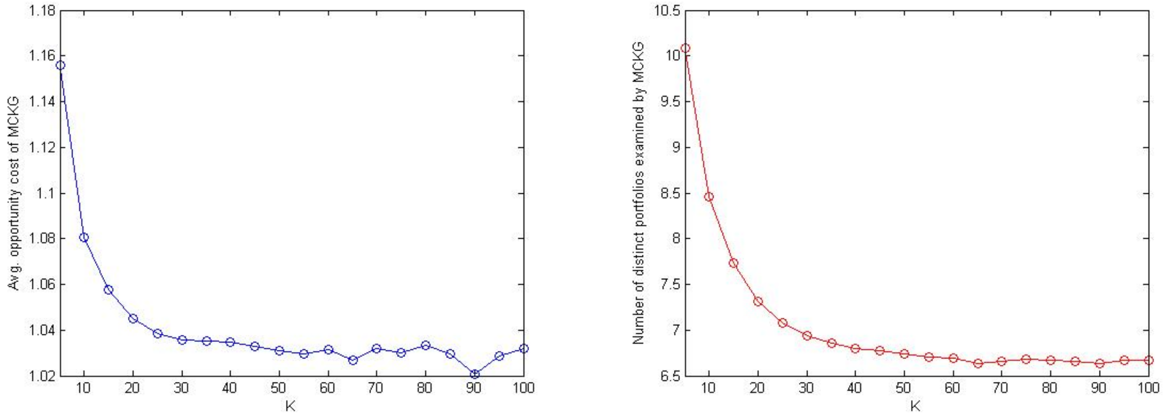

Figure 3: Effect of sample size K on performance of MCKG policy.

policy should look at in order to make a decision, causing MCKG to explore more in search of those portfolios. When K is large enough for all such portfolios to be included, MCKG limits itself to them. To obtain satisfactory performance, it is sufficient to limit K to 10% of the full choice set.

5 CONCLUSION

We have proposed a decision rule for online learning problems. The MCKG policy is based on the logic of knowledge gradients, but uses Monte Carlo sampling to compute knowledge gradients for a small subset of the choice set. In a problem where the measurement budget is smaller than the number of choices by an order of magnitude, MCKG is effective at finding the most interesting choices. We compared the MCKG policy to several other learning policies in the context of energy portfolio selection, and found that it was competitive against the other policies tested. We believe that the Monte Carlo-based methodology we have proposed is a reliable approach to learning problems. In particular, the usefulness of MCKG would increase in a situation where we could obtain a sample realization of the best alternative, but where it would be difficult to enumerate every alternative.

ACKNOWLEDGMENTS

This research was supported in part by AFOSR contract FA9550-08-1-0195 and ONR contract N00014-07-1-0150 through the Center for Dynamic Data Analysis.

REFERENCES

Auer, P., N. Cesa-Bianchi, and P. Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47:235–256.

Chick, S., J. Branke, and C. Schmidt. 2009. Sequential sampling to myopically maximize the expected value of information.

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

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

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

Frazier, P., W. Powell, and S. Dayanik. 2009. The knowledge-gradient policy for correlated normal rewards. *INFORMS J. on Computing (to appear)*.

Gittins, J. 1989. *Multi-armed bandit allocation indices*. New York: John Wiley and Sons.

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

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

Kaelbling, L. P. 1993. *Learning in embedded systems*. Cambridge, MA: MIT Press.

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

Lai, T. L., and H. Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6:4–22.

McKinsey & Company 2007. Reducing U.S. Greenhouse Gas Emissions: How Much at What Cost? U.S. Greenhouse Gas Abatement Mapping Initiative, Executive Report.

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

Powell, W. B. 2007. *Approximate dynamic programming: Solving the curses of dimensionality*. New York: John Wiley and Sons.

Ryzhov, I., and W. Powell. 2009. The knowledge gradient algorithm for online subset selection. In *Proceedings of the 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, Nashville, TN*, 137–144.

Ryzhov, I., W. Powell, and P. Frazier. 2008. The knowledge gradient algorithm for a general class of online learning problems. *Submitted for publication*.

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

Yao, Y. 2006. Some results on the Gittins index for a normal reward process. In *Time Series and Related Topics: In Memory of Ching-Zong Wei*, ed. H. Ho, C. Ing, and T. Lai, 284–294. Institute of Mathematical Statistics, Beachwood, OH, USA.

AUTHOR BIOGRAPHIES

ILYA O. RYZHOV is a Ph.D. student in the Department of Operations Research and Industrial Engineering at Princeton University. His research is concerned with Bayesian information collection in stochastic optimization problems, e.g. shortest path problems with unknown arc costs. His email address for these proceedings is <iryzhov@princeton.edu>.

WARREN B. POWELL is a Professor in the Department of Operations Research and Financial Engineering at Princeton University. ... He is the author of the book *Approximate Dynamic Programming: Solving the curses of dimensionality*, published by John Wiley & Sons. His email address for these proceedings is <powell@princeton.edu>.