Submitted to *Operations Research*
manuscript OPRE-2011-04-182

Sequential Bayes-Optimal Policies for Multiple Comparisons with a Known Standard

Jing Xie
School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853, jx66@cornell.edu

Peter I. Frazier
School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853, pf98@cornell.edu

We consider the problem of efficiently allocating simulation effort to determine which of several simulated systems have mean performance exceeding a threshold of known value. Within a Bayesian formulation of this problem, the optimal fully sequential policy for allocating simulation effort is the solution to a dynamic program. When sampling is limited by probabilistic termination or sampling costs, we show that this dynamic program can be solved efficiently, providing a tractable way to compute the Bayes-optimal policy. The solution uses techniques from optimal stopping and multi-armed bandits. We then present further theoretical results characterizing this Bayes-optimal policy, compare it numerically to several approximate policies, and apply it to applications in emergency services and manufacturing.

Key words: multiple comparisons with a standard; sequential experimental design; dynamic programming; Bayesian statistics; value of information.

1. Introduction

We consider multiple comparisons with a (known) standard (MCS), in which simulation is used to determine which alternative systems under consideration have performance surpassing that of a standard with known value. We focus on how simulation effort should be allocated among the alternative systems to best support comparison of alternatives when sampling stops. We formulate the problem of allocating effort as a problem in sequential decision-making under uncertainty, and solve the resulting dynamic program, providing Bayes-optimal procedures.

MCS problems arise in many applications. We consider the following two examples in detail:

• Administrators of a city's emergency medical services would like to know which of several methods under consideration for positioning ambulances satisfy mandated minimums for percentage of emergency calls answered on time.

• A manufacturing firm is unsure about market and operations conditions in the coming month. Executives would like to know under which conditions their production line can maintain a positive net expected revenue. They plan to use a simulator of their operations to answer this question.

The most straightforward approach for allocating sampling effort, and the approach most commonly employed by practitioners, is to simulate each system an equal number of times. This is inefficient, because some alternatives have performance far from the control and can be immediately established as being substantially better or substantially worse after only a few samples. Other alternatives need many more samples before an accurate determination can be made.

To design a strategy that samples more efficiently, we first formulate the problem in a Bayesian framework. Using methods from multi-armed bandits and optimal stopping (see, e.g., Gittins and Jones (1974) and DeGroot (1970) respectively), we explicitly characterize and then efficiently compute Bayes-optimal sequential sampling policies for MCS problems. Such policies provide optimal average case performance, where the average is taken under a prior distribution that we choose.

Our framework and the resulting ability to compute the Bayes-optimal policies is general. First, it allows two different methods for modeling the limited ability to sample: an explicit cost for each sample (appropriate, e.g., when using on-demand cloud-computing services, in which fees are proportional to the number of CPU hours consumed), and/or a random ceiling on the number of samples allowed. Second, it allows a broad class of terminal payoff functions for modeling the consequences of correct and incorrect comparisons. Third, it provides the ability to model sampling distributions within any exponential family, which includes common families of sampling distributions like normal with known sampling variance, normal with unknown sampling variance, Bernoulli, and Poisson. We present in detail two such cases, normal with known variance and Bernoulli, and derive additional theoretical results for these special settings.

While our results allow computing the Bayes-optimal policies in a broad class of problem settings, these results have two important limitations. First, they cannot incorporate information about a *known* sampling budget. In this situation, we recommend a heuristic policy based on the optimal policy for a stochastic budget. Second, they require that the value of the standard against which systems are compared is known.

We review the related literature in Section 2, and formulate the problem in Section 3. In Section 4 we present Bayes-optimal policies for general sampling distributions, considering separately the case of an almost surely finite horizon with or without sampling costs (Section 4.1), and the case of an infinite horizon with sampling costs (Section 4.2). Then, in Sections 5 and 6, we specialize to two types of sampling: Bernoulli samples, and normal samples with known variance. We give theoretical results particular to these more specialized cases, and provide techniques for computing the optimal policies efficiently. In Section 7 we demonstrate the resulting Bayes-optimal algorithms on illustrative problems, and on two examples in emergency services and manufacturing.

2. Literature Review

The MCS problem is frequently studied as a special case of a larger class of problems called multiple comparisons with a control (MCC). In MCC problems, the standard (also called the "control") against which we compare is itself the mean performance of a stochastic system. When this mean is modeled as known, we recover the MCS problem. This can be appropriate when the standard system has been in use for a long period of time, providing sufficient data to estimate its mean performance with high accuracy (Nelson and Goldsman 2001). MCS problems also arise when the standard is some known performance requirement, rather than the mean of a stochastic system.

Several books and survey papers review the previous literature on MCC: Hochberg and Tamhane (1987) and Hsu (1996) are general references on multiple comparisons; Goldsman and Nelson (1994) focuses on multiple comparisons in simulation; and Fu (1994) reviews multiple comparisons as they relate to simulation optimization. Within this MCC literature, work on designing sampling procedures focuses on creating simultaneous confidence intervals for the differences of the mean of each system with that of an unknown control: Paulson (1952), Dunnett (1955) create one-stage procedures assuming independent normal sampling with common known variance; and Dudewicz and Ramberg (1972), Dudewicz and Dalal (1983), Bofinger and Lewis (1992), Damerdji and Nakayama (1996) create two-stage procedures allowing more general sampling distributions. This previous work focuses on difficulties introduced by an unknown control, while we consider a known standard. Also, these procedures have only one or two stages, while we focus on fully sequential procedures, whose ability to adapt the sampling scheme to previous samples offers better sampling efficiency.

Procedures for variants of the MCS problem have also been developed in the indifference-zone ranking and selection (R&S) literature (see Kim and Nelson (2006) for a survey of R&S). Paulson (1962) and Kim (2005) provide fully sequential procedures for determining if any of the systems is better than the standard, and if so, selecting the one with the best mean. Bechhofer and Turnbull (1978), Nelson and Goldsman (2001) provide two-stage procedures for the same task. Andradóttir et al. (2005) and Andradóttir and Kim (2010) consider the problem of finding the system with the best expected primary performance measure, subject to a constraint on the expected value of a secondary performance measure. They present a procedure that determines systems' feasibility in the presence of a stochastic constraint, and combines this procedure with a finding-the-best procedure to identify the best feasible system. Batur and Kim (2010) provides fully sequential procedures for identifying a set of feasible or near-feasible systems in the presence of multiple stochastic constraints. Healey et al. (2012) extends this procedures to select the best feasible system with multiple constraints and possible correlation across systems. These papers all consider problems in a frequentist context, where the goal is to create a procedure with a worst-case statistical guarantee

on solution quality. In contrast, we work within a Bayesian context, where the goal is to provide good performance in the average case.

Within a large deviations framework, which was introduced for R&S by Glynn and Juneja (2004), Szechtman and Yücesan (2008) considers MCS (also called feasibility determination). That paper characterizes the fractional allocation that maximizes the asymptotic rate of decay of the expected number of incorrect determinations, and presents a stochastic approximation algorithm that yields a budget allocation that provably converges to the optimal one. While this is appealing, Glynn and Juneja (2011) argues that such methods, which calculate optimal allocations assuming large deviations rate functions are known, and then plug in estimates of the rate functions, may have rates of convergence that are significantly worse than desired because of the difficulty of estimating rate functions. This difficulty can be avoided when sampling distributions are bounded, but the difficulty remains when sampling distributions are unbounded. Hunter et al. (2011), on the other hand, considers selecting an optimal system from among a finite set of competing systems, based on a stochastic objective function and subject to multiple stochastic constraints. Hunter and Pasupathy (2012) then assumes a bivariate normal distribution for the objective and (single) constraint performance measures, and gives explicit and asymptotically exact characterizations.

In other related work, Picheny et al. (2010) proposes adaptive experimental designs with kriging metamodels for approximating a continuous function accurately around a particular level-set.

The current work differs from all previous work by finding optimal fully sequential procedures in a Bayesian formulation of the MCS problem that explicitly models a limited ability to sample.

In its pursuit of Bayes-optimal fully sequential policies, the current work is related to the value of information approach, exemplified by Frazier et al. (2008), Chick and Gans (2009) and Chick and Frazier (2009), which attempt to achieve a similar goal in sequential Bayesian R&S by managing the trade-off between the consequences of an immediate decision and the cost of additional sampling. In the special case of a single unknown alternative, Chick and Gans (2009) and Chick and Frazier (2009) compute a close approximation to the optimal R&S policy using diffusion approximations. However, no efficient methods exist for computing the optimal sequential R&S policy for more than a few alternatives. This is in contrast to the current work, in which we show that the optimal sequential MCS policy can be computed efficiently in general.

The ability shown in this paper to explicitly and efficiently compute the optimal policy also contrasts the MCS problem with other problems in Bayesian experimental design and Bayesian optimal learning, including global optimization (Mockus 1989), dynamic pricing (Araman and Caldenty 2009), inventory control (Ding et al. 2002), and sensor networks (Krause et al. 2008), where finding the optimal policy is usually considered intractable. In such problems, a common suboptimal approach is to compute a myopic one-step lookahead policy (Gupta and Miescke 1996,

Chick et al. 2010, Jones et al. 1998, Lizotte et al. 2007). Policies of this type are also called knowledge-gradient (KG) policies (Frazier 2009). In our numerical experiments, we derive the KG policy and compare it against the optimal policy. We find that in some cases the KG policy performs extremely well (see also Frazier et al. (2008, 2009)), while in other cases it performs poorly. This variability in performance is similar to results in Frazier and Powell (2010), Ryzhov et al. (2012).

3. Problem Formulation

In this section we formulate the general Bayesian MCS problem, which allows both a random ceiling on the number of samples, and sampling costs. We have k alternative systems that we can simulate, and samples from each alternative are independent and from distributions that do not change over time. For each $x = 1, 2 \ldots, k$, let $f(\cdot|\eta_x)$ be the probability density function (pdf) or probability mass function (pmf) for samples from alternative x, where η_x is an unknown parameter or vector of parameters residing in a parameter space Ξ. We further assume that the space of possible sampling distributions $\{f(\cdot|\eta) : \eta \in \Xi\}$ form an exponential family. See DeGroot (1970) Chapter 9 for an in-depth treatment of exponential families. This assumption of an exponential family allows most common parametric sampling distributions, including the normal (with known variance) and Bernoulli distributions considered in detail in Sections 5 and 6, as well as normal with unknown variance, Poisson, multinomial, and many others.

We wish to find the set of alternatives whose underlying performance is above a corresponding threshold or control. The underlying performance of each alternative x is characterized by the mean of its sampling distribution, θ_x, which is a known function of η_x. The corresponding threshold is d_x. Hence we want to determine the set $\mathbb{B} = \{x : \theta_x \geq d_x\}$.

We take a Bayesian approach, placing a prior probability distribution on each unknown η_x. This prior distribution represents our subjective beliefs about this sampling distribution. To facilitate computation, we adopt independent conjugate priors. Specifically, we suppose the independent prior distributions on $\eta_1, \ldots, \eta_k$ come from a common conjugate exponential family $\mathcal{D}$ with parameter space Λ. For example, in Section 5 where samples are Bernoulli-distributed, the prior is beta-distributed and Λ is the space of parameters of the beta distribution. Let the corresponding parameters of these prior distributions be $S_{0,1}, \ldots, S_{0,k}$, each of which resides in Λ. Denote by $\mathbf{S}_0$ the vector composed of $S_{0,x}$ with x ranging from 1 to k.

Time is indexed by $n = 1, 2, \ldots$. At each time n we choose an alternative $x_n \in \{1, \ldots, k\}$ from which to sample, and observe a corresponding sample y_n which has pdf or pmf $f(\cdot \mid \eta_{x_n})$. We refer to the decision x_n as our "sampling decision", and our focus in this paper is on how to best make these sampling decisions, and a related stopping decision discussed below.

As our prior is conjugate to our sampling distribution, our samples result in a sequence of posterior distributions on $\eta_1, \ldots, \eta_k$, each of which resides in the same conjugate family parameterized by Λ. We denote the parameters of these posteriors at time $n \geq 1$ by $S_{n,1}, \ldots, S_{n,k}$, and the vector composed of them by $\mathbf{S}_n$. Then for $n \geq 1$, $\mathbf{S}_n = G(\mathbf{S}_{n-1}, x_n, y_n)$, where $G(\cdot,\cdot,\cdot)$ is some known and fixed function determined by the exponential and conjugate families. Moreover, for all x, the posterior remains independent across x under this update. Define $\mathbb{S} = \Lambda^k$, which is the state space of the stochastic process $(\mathbf{S}_n)_{n\geq 0}$. We will sometimes refer to a generic element of $\mathbb{S}$ as $\mathbf{s} = (s_1, \ldots, s_k)$, and a generic element of Λ as s. In this paper we use boldfaced parameters to refer to multiple alternatives and regular font to refer to a single alternative.

We allow decisions to depend only upon the data available from previous samples. To make this requirement more formal, we define a filtration $(\mathcal{F}_n)_{n\geq 0}$, where $\mathcal{F}_n$ is the sigma-algebra generated by $x_1, y_1, \ldots, x_n, y_n$. We require that $x_{n+1} \in \mathcal{F}_n$ for $n \geq 0$. In addition to the sampling decisions $(x_n)_{n\geq 1}$, we also choose the total number of samples we take, denoted by τ. We require τ to be a stopping time of the filtration, i.e., we require the event $\{\tau = n\}$ to be $\mathcal{F}_n$–measurable for all $n \geq 0$.

We refer to a collection of rules for making all of the required decisions in a decision-making problem as a policy. Thus, in this problem a policy π is composed of a sampling rule for choosing the sequence of sampling decisions $(x_n)_{n\geq 1}$, and a stopping rule for choosing τ.

For each $n \geq 0$, let $\mathbb{E}_n$ denote the conditional expectation with respect to the information available after n samples, so $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_n]$. When the expectation depends on the policy π, we write $\mathbb{E}^\pi$ for the unconditional expectation, and $\mathbb{E}^\pi_n$ for the conditional expectation with respect to $\mathcal{F}_n$.

In the general formulation of the MCS problem that we consider here, we model the need to sample efficiently in two complementary ways. First, we suppose that each sample incurs a nonnegative cost. For $x = 1, \ldots, k$, denote by $c_x \geq 0$ the sampling cost for alternative x. Second, we suppose that there is some random time horizon T beyond which we will be unable to sample, so that we stop sampling at time $\tau \wedge T$, where $\wedge$ is the minimum operator.

Most frequently this horizon T is imposed because the results of the simulation are needed by the simulation analyst. For analytical convenience, we assume that T is geometrically distributed, and independent of the sampling process. Let $1-\alpha$ be the parameter of this geometric distribution, with $0 < \alpha < 1$, so that $\mathbb{E}[T] = 1/(1-\alpha)$. We also allow T to be a random variable that is infinite with probability 1, in which case we take $\alpha = 1$. In either case, we can equivalently model this random time horizon by supposing that external circumstances may require us to stop after each sample independently with probability $1-\alpha$. While our model does not allow a deterministic horizon T, one can apply the Bayes-optimal procedures we develop in such situations as heuristics by choosing α so that $T = 1/(1-\alpha)$. We use this method in Section 7.1 and 7.2.

We assume that sampling is penalized, either through a finite horizon ($\alpha < 1$), or a cost per sample ($c_x > 0$ for all x), or both. That is, we disallow the combination of $\alpha = 1$ and $c_x = 0$, which prevents the unrealistic situation of sampling from x forever at no cost.

Define a terminal payoff function r with the following decomposition:

$$r(B;\theta,d) = \sum_{x\notin B} r_0(x;\theta_x) + \sum_{x\in B} r_1(x;\theta_x),$$

where r_0 and r_1 are known real-valued functions. This decomposition is necessary for our analysis, and without it many of our results would not hold.

Adapted to the information filtration $(\mathcal{F}_n)_{n\geq 0}$, we choose a sequence of sets $(B_n)_{n\geq 0}$ to approximate the objective set $\mathbb{B}$. We require that each $B_n \subseteq \{1,2,\ldots,k\}$ is chosen to maximize the expected terminal payoff given the available data after n samples. Formally, for all $n \geq 0$,

$$B_n = \operatorname*{arg\,max}_{B\subseteq\{1,2,\ldots,k\},\, B\in\mathcal{F}_n} \mathbb{E}_n\left[r(B;\theta,d)\right] = \operatorname*{arg\,max}_{B\subseteq\{1,2,\ldots,k\},\, B\in\mathcal{F}_n} \left\{\sum_{x\notin B} \mathbb{E}_n\left[r_0(x;\theta_x)\right] + \sum_{x\in B} \mathbb{E}_n\left[r_1(x;\theta_x)\right]\right\}.$$

Define $h_{ix}(s) = \mathbb{E}\left[r_i(x;\theta_x) \mid \eta_x \sim \mathcal{D}(s)\right]$, $i = 0,1$, and $h_x(s) = \max\{h_{0x}(s), h_{1x}(s)\}$ for $s \in \Lambda$ and $x = 1,\ldots,k$. Simple algebra then yields

$$B_n = \{x : h_{0x}(S_{n,x}) \leq h_{1x}(S_{n,x})\} \quad \text{and} \quad \mathbb{E}_n\left[r(B_n;\theta,d)\right] = \sum_{x=1}^{k} h_x(S_{n,x}). \tag{1}$$

Our estimate of the set $\mathbb{B}$ is $B_{\tau\wedge T}$ when sampling stops. Our goal is to find a policy that maximizes the expected total reward, i.e., to solve the problem

$$\sup_{\pi} \mathbb{E}^{\pi}\left[r(B_{\tau\wedge T};\theta,d) - \sum_{n=1}^{\tau\wedge T} c_{x_n}\right]. \tag{2}$$

3.1. Terminal Payoff Functions: Conditions and Examples

While our results apply to general terminal payoff functions, some of our theoretical results require additional conditions defined below. Payoff Condition 1 states that, for each x, $\{h_x(S_{n,x})\}_{n\geq 0}$ is a sub-martingale, i.e., that the expected terminal payoff (not including the cost of sampling) improves as we collect more sample information. Payoff Condition 2 states that this improvement has an upper bound that may depend upon the starting posterior, but not on the number of samples taken. These conditions together provide necessary bounds for the reward functions, the Gittins indices, and the value functions introduced later in Section 4. Payoff Condition 3 states, roughly speaking, that the improvement in expected terminal payoff from an arbitrarily large amount of additional sampling vanishes as we sample more and more. Payoff 3 provides additional results that further facilitate computation of the optimal policy. When a theoretical result requires one or more of these conditions, we state this explicitly in the result.

Payoff Condition 1 *For any* $x=1,\ldots,k$ *and* $s\in\Lambda$, $h_x(s)\leq\mathbb{E}\left[h_x\left(S_{1,x}\right)\mid S_{0,x}=s,x_1=x\right]$.

Payoff Condition 2 *There exist deterministic non-negative functions* $H_1,\ldots,H_k$ *on* Λ *such that, for any* x, $n\geq 0$ *and* $s\in\Lambda$, $\mathbb{E}\left[h_x\left(S_{n,x}\right)\mid S_{0,x}=s,x_1=\cdots=x_n=x\right]-h_x(s)\leq H_x\left(s\right)$.

Payoff Condition 3 *There exist deterministic non-negative functions* $\widetilde{H}_1,\ldots,\widetilde{H}_k$ *on* Λ *such that, for any* x *and* $s\in\Lambda$,

$$\mathbb{E}\left[h_x\left(S_{1,x}\right)\mid S_{0,x}=s,x_1=x\right]-h_x(s)\leq\widetilde{H}_x(s), \tag{3}$$

$$\lim_{n\to\infty}\left[\sup_{s\in PS(x;n)}\widetilde{H}_x(s)\right]=0, \tag{4}$$

where $PS(x;n):=\left\{s\in\Lambda\colon\exists s'\in\Lambda\ s.t.\ \mathbb{P}\left[S_{n,x}=s\mid S_{0,x}=s',x_1=\cdots=x_n=x\right]>0\right\}$.

We will consider the following two terminal payoff functions in detail throughout the paper. Let m_{ix} for $i=0,1$ and $x=1,\ldots,k$ be non-negative constants.

Example 1: (**0-1 Terminal Payoff**) $r_0\left(x;\theta_x\right)=m_{0x}\cdot\mathbf{1}_{\{x\notin\mathbb{B}\}}$, $r_1\left(x;\theta_x\right)=m_{1x}\cdot\mathbf{1}_{\{x\in\mathbb{B}\}}$.

Example 2: (**Linear Terminal Payoff**) $r_0\left(x;\theta_x\right)=m_{0x}\left(d_x-\theta_x\right)$, $r_1\left(x;\theta_x\right)=m_{1x}\left(\theta_x-d_x\right)$.

We further characterize these example terminal payoff functions in Table 1, giving explicit expressions for $h_{ix}(s)$, B_n, and $H_x(s)$ under these payoff functions, and showing that both payoff functions satisfy Payoff Conditions 1 and 2. The table uses the following additional notation:

$$p_x(s):=\mathbb{P}\left\{\theta_x\geq d_x\mid\eta_x\sim\mathcal{D}(s)\right\},\quad\mu(s):=\mathbb{E}\left[\theta_x\mid\eta_x\sim\mathcal{D}(s)\right],\quad\mu_{nx}:=\mathbb{E}_n\left[\theta_x\right]=\mu\left(S_{n,x}\right),$$
$$B(x):=\left\{s\colon p_x(s)\geq\frac{m_{0x}}{m_{0x}+m_{1x}}\right\},\quad A_0(s):=\mathbb{E}\left[m_{0x}(\theta_x-d_x)^-+m_{1x}(\theta_x-d_x)^+\,\middle|\,\eta_x\sim\mathcal{D}(s)\right],$$

where $(z)^+=\max(z,0)$ and $(z)^-=\max(-z,0)$ denote the positive part and the negative part respectively. The proof of the statements in this table can be found in the e-companion.

Table 1 Example Terminal Payoff Functions and Their Properties

	0-1 Terminal Payoff	Linear Terminal Payoff
$h_{0x}(s)$	$m_{0x}\left[1-p_x(s)\right]$	$m_{0x}\left[d_x-\mu(s)\right]$
$h_{1x}(s)$	$m_{1x}\cdot p_x(s)$	$m_{1x}\left[\mu(s)-d_x\right]$
B_n	$\left\{x:p_x\left(S_{n,x}\right)\geq\frac{m_{0x}}{m_{0x}+m_{1x}}\right\}$	$\left\{x:\mu_{nx}\geq d_x\right\}$
Payoff Condition 1	Yes	Yes
Payoff Condition 2	Yes	Yes
$H_x(s)$	$m_{0x}\cdot\mathbf{1}_{\{s\notin B(x)\}}+m_{1x}\cdot\mathbf{1}_{\{s\in B(x)\}}-h_x(s)$	$A_0(s)-h_x(s)$

4. The Optimal Solution

In this section we present the optimal solution to the Bayesian MCS problem (2), which allows both a geometrically distributed sampling horizon, and sampling costs. We first present some preliminary results, and then give solutions for a geometrically distributed horizon in Section 4.1, and for an infinite horizon with sampling costs in Section 4.2. The results in this section apply to the general sampling framework given in Section 3, and in later sections we specialize to sampling with Bernoulli observations (Section 5) and normal observations (Section 6).

We solve the problem (2) using dynamic programming (DP) (Bellman (1954), and see references Dynkin and Yushkevich (1979), Bertsekas (2005, 2007), Powell (2007)). In the DP approach, we define a value function $V : \mathbb{S} \mapsto \mathbb{R}$. For each state $\mathbf{s} \in \mathbb{S}$, $V(\mathbf{s})$ is the optimal expected total reward attainable when the initial state is $\mathbf{s}$. That is,

$$V(\mathbf{s}) = \sup_{\pi} \mathbb{E}^{\pi} \left[r\left(B_{\tau\wedge T}; \theta, d\right) - \sum_{n=1}^{\tau\wedge T} c_{x_n} \,\middle|\, \mathbf{S}_0 = \mathbf{s} \right]. \tag{5}$$

An optimal policy is any policy π attaining this supremum.

Before describing these optimal policies in Sections 4.1 and 4.2, we transform the value function to a form that supports later theoretical development. Define a function $h\colon [0,1] \mapsto [1/2, 1]$ by $h(u) = \max\{u, 1-u\}$ and let $R_0(\mathbf{s}) := \sum_{x=1}^{k} h_x(s_x)$. Since R_0 is a deterministic function of the initial state, we can subtract it from the value function, and redefine V as the optimal expected incremental reward over R_0. We then have the following proposition.

Proposition 1.

$$V(\mathbf{s}) = \sup_{\pi} \mathbb{E}^{\pi} \left[\sum_{n=1}^{\tau} \alpha^{n-1} \mathcal{R}_{x_n}(S_{n-1,x_n}) \,\middle|\, \mathbf{S}_0 = \mathbf{s} \right], \tag{6}$$

where the reward functions $\mathcal{R}_x : \Lambda \mapsto \mathbb{R}$ *for* $x = 1, \ldots, k$ *are defined by*

$$\mathcal{R}_x(s) = \mathbb{E}\left[h_x(S_{1,x}) \mid S_{0,x} = s, x_1 = x\right] - h_x(s) - c_x. \tag{7}$$

It is clear that $\mathcal{R}_x$ is bounded below by $-c_x$ under Payoff Payoff Condition 1. It is bounded above by $-c_x + H_x$ under Payoff Condition 2, and by $-c_x + \widetilde{H}_x$ under Payoff Condition 3.

In the following subsections we divide our assumption of an almost surely finite horizon ($\alpha < 1$) or a cost per sample ($c_x > 0$ for all x) into two distinct cases, and solve the MCS problem in each case using a distinct technique. The first (Section 4.1) assumes $\alpha < 1$ (geometric horizon), and the second (Section 4.2) assumes $\alpha = 1$ and $c_x > 0$ for all x (infinite horizon with sampling costs).

4.1. Geometric Horizon

We first consider the MCS problem with T almost surely finite, i.e., with $0<\alpha<1$. We make no restrictions on the sampling costs c_x, allowing them to be 0 or strictly positive. In this case, (6) is a multi-armed bandit (MAB) problem (see, e.g., Mahajan and Teneketzis (2008)).

To solve a Bayesian MAB problem, Gittins and Jones (1974) showed that it is sufficient to compute Gittins indices $\nu_x(s)$ for each possible state s, which can be written here as

$$\nu_x(s)=\max_{\tau>0}\mathbb{E}\left[\frac{\sum_{n=1}^{\tau}\alpha^{n-1}\mathcal{R}_x(S_{n-1,x})}{\sum_{n=1}^{\tau}\alpha^{n-1}}\,\middle|\,S_{0,x}=s,x_1=\cdots=x_\tau=x\right]. \tag{8}$$

The optimal sampling rule, whose decisions we denote $(x_n^*)_{n\geq 1}$, is then to select at each time the alternative with a corresponding state that has the largest Gittins index. The optimal stopping time, which we write as τ^*, is the first time when all the k indices are non-positive. Formally,

$$x_{n+1}^*=\arg\max_x\{\nu_x(S_{n,x})\},\ \forall n\geq 0;\qquad \tau^*=\inf\{n:\nu_x(S_{n,x})\leq 0,\forall x\}.$$

Computation of (8) is much easier than solving the full DP because the dimension of $s\in\Lambda$ is smaller than that of $\mathbf{s}\in\mathbb{S}=\Lambda^k$, and the computational complexity of solving a DP scales poorly with the dimension of the state space, due to the so-called curse of dimensionality (Powell 2007).

We introduce the following bounds on the Gittins indices to serve approximate computation of the optimal policy when the state space is infinite (see Section 6.1).

PROPOSITION 2. *Under Payoff Condition 1, $\nu_x\geq -c_x$. Under Payoff Conditions 1 and 2, $-c_x\leq \nu_x\leq -c_x+H_x$.*

When $c_x=0$ for all x and Payoff Condition 1 holds, the Gittins indices are always nonnegative by Proposition 2. We may then choose τ^* to be $+\infty$.

4.2. Infinite Horizon With Sampling Costs

We now consider the MCS problem with $T=\infty$ almost surely and positive sampling costs, i.e., $\alpha=1$ and $c_x>0$ for all x. With these values for α and c_x, (6) becomes

$$V(\mathbf{s})=\sup_{\pi}\mathbb{E}^{\pi}\left[\sum_{n=1}^{\tau}\mathcal{R}_{x_n}(S_{n-1,x_n})\,\middle|\,\mathbf{S}_0=\mathbf{s}\right].$$

Now fix some x and consider a sub-problem in which only alternative x can be sampled. The optimal expected reward for this single-alternative problem with initial state s is then

$$V_x(s)=\sup_{\tau_x}\mathbb{E}\left[\sum_{n=1}^{\tau_x}\mathcal{R}_x(S_{n-1,x})\,\middle|\,S_{0,x}=s,x_1=\cdots=x_{\tau_x}=x\right], \tag{9}$$

where τ_x is the stopping time. We immediately have the following bounds on V_x.

PROPOSITION 3. *$V_x \geq 0$. Under Payoff Condition 2, $V_x \leq H_x$.*

Standard results from the DP literature (see, e.g., Dynkin and Yushkevich (1979)) show that V_x satisfies Bellman's recursion,

$$\begin{aligned} V_x(s) &= \max\left[0, L_x(s, V_x)\right], \quad \text{where} \\ L_x(s, V_x) &= \mathcal{R}_x(s) + \mathbb{E}[V_x(S_{1,x}) \mid S_{0,x} = s, x_1 = x]. \end{aligned} \tag{10}$$

Here, the value function V_x is not necessarily Borel-measurable, but is universally measurable, and so the expectation of V_x is taken in this more general sense.

This problem is a standard optimal stopping problem (for details see Bertsekas (2007) Section 3.4) that can be solved by specifying the set of states $\mathbb{C}_x$ on which we should continue sampling (also called the continuation set), which implicitly specifies the set on which we should stop (the stopping set) as $\Lambda \setminus \mathbb{C}_x$. They are optimally specified as $\mathbb{C}_x = \{s \in \Lambda : V_x(s) > 0\}$ and $\Lambda \setminus \mathbb{C}_x = \{s \in \Lambda : V_x(s) = 0\}$. Then, an optimal solution to (9) is the stopping time τ_x^* given by $\tau_x^* = \inf\{n \geq 0 : S_{n,x} \notin \mathbb{C}_x\}$.

We allow τ_x^* to be ∞, in which case the state of alternative x never leaves $\mathbb{C}_x$. Even if τ_x^* is almost surely finite, there may be no deterministic upper bound. For example, in the Bayesian formulation of the sequential hypothesis testing problem (Wald and Wolfowitz 1948), there is no fixed almost sure upper bound on the number of samples taken by the Bayes-optimal policy even though sampling has a fixed cost, and considerable research effort has gone to creating good sub-optimal policies that stop within a fixed amount of time (Siegmund 1985). In our problem, however, the set of possible states for x after n samples, i.e., $PS(x; n)$, shrinks under Payoff Condition 3 so that it is contained by $\Lambda \setminus \mathbb{C}_x$ when n exceeds a deterministic value. This gives a deterministic upper bound on τ_x^*, as demonstrated by the following proposition.

PROPOSITION 4. *Under Payoff Condition 3, τ_x^* has a deterministic upper bound N_x, where*

$$N_x := \min\left\{ n : \left[\sup_{s \in PS(x;n')} \widetilde{H}_x(s) \right] \leq c_x, \ \forall n' \geq n \right\}. \tag{11}$$

This bound is also computationally useful because it allows us to restrict the state space when solving Bellman's recursion for the optimal policy (see Sections 5.2 and 6.2).

Given the independence among the alternatives, our original problem can be decomposed into k sub-problems. This decomposition is used in the proof of the following theorem, Theorem 1, which relates the value functions of these sub-problems to the original problem and gives the optimal policy for the original problem.

THEOREM 1. *The value function is given by, $V(\mathbf{s}) = \sum_{x=1}^{k} V_x(s_x)$. Furthermore, any policy with sampling decisions $(x_n^*)_{n \geq 1}$ and stopping time τ^* satisfying the following conditions is optimal:*

$$x_{n+1}^* \in \{x : S_{n,x} \in \mathbb{C}_x\}, \ \forall n \geq 0; \qquad \tau^* = \inf\{n \geq 0 : S_{n,x} \notin \mathbb{C}_x, \forall x\}.$$

The following proposition shows that if each τ_x^* is bounded above, then the optimal stopping time for the whole problem is also bounded above.

PROPOSITION 5. *Suppose that each τ_x^* has a deterministic upper bound N_x. Then the optimal stopping rule τ^*, as characterized in Theorem 1, has a deterministic upper bound $\sum_{x=1}^k N_x$.*

5. Specialization to Bernoulli Sampling

In this section we specialize the results of Section 4 to the specific case of Bernoulli samples. We give explicit expressions for quantities described generally in Section 4, and then present additional theoretical results and computational methods. Later, in Section 6, we pursue the same agenda for another commonly considered type of sampling: normal samples with known sampling variance.

We first give explicit expressions for the statistical model, the reward function, and Bellman's recursion. Here for each x, the underlying performance parameter $\theta_x \in (0,1)$ is the only component of η_x, and the corresponding threshold is $d_x \in (0,1)$. At each time $n \geq 1$, $y_n \mid \theta, x_n \sim \text{Bernoulli}(\theta_{x_n})$.

We adopt a conjugate $\text{Beta}(a_{0x}, b_{0x})$ prior for each θ_x with $a_{0x}, b_{0x} \geq 1$, under which θ_x is independent of $\theta_{x'}$ for $x \neq x'$. Our Bernoulli samples then result in a sequence of posterior distributions on θ_x which are again independently beta-distributed with parameters $S_{n,x} = (a_{nx}, b_{nx})$ in parameter space $\Lambda = [1,+\infty) \times [1,+\infty)$. We take $\mathbf{S}_n = (\mathbf{a}_n, \mathbf{b}_n)$ as the state of the DP, where the state space is $\mathbb{S} = [1,+\infty)^k \times [1,+\infty)^k$. The state update function G is given by, for $n \geq 1$,

$$(\mathbf{a}_n, \mathbf{b}_n) = \mathbf{1}_{\{y_n=1\}} \cdot (\mathbf{a}_{n-1} + \mathbf{e}_{x_n}, \mathbf{b}_{n-1}) + \mathbf{1}_{\{y_n=0\}} \cdot (\mathbf{a}_{n-1}, \mathbf{b}_{n-1} + \mathbf{e}_{x_n}),$$

where $\mathbf{e}_{x_n}$ denotes a length-k vector with 1 at element x_n and 0 elsewhere.

Now for any $(a,b) \in \Lambda$, $\mathcal{D}(a,b) = \text{Beta}(a,b)$, $\mu(a,b) = a/(a+b)$ and $p_x(a,b) = 1 - I_{d_x}(a,b)$, where the regularized incomplete beta function $I_\cdot(\cdot,\cdot)$ is defined for $a,b>0$ and $0 \leq d \leq 1$ by

$$I_d(a,b) = \frac{B(d;a,b)}{B(a,b)} = \frac{\int_0^d t^{a-1}(1-t)^{b-1}dt}{\int_0^1 t^{a-1}(1-t)^{b-1}dt}.$$

By Remark 1 in the e-companion and definition (7), we also have

$$\mathcal{R}_x(a,b) = -c_x - h_x(a,b) + \frac{a}{a+b} \cdot h_x(a+1,b) + \frac{b}{a+b} \cdot h_x(a,b+1). \tag{12}$$

When $\alpha = 1$ and $c_x > 0$ for all x, in each sub-problem with alternative x,

$$\mathbb{E}[V_x(S_{1,x}) \mid S_{0,x} = (a,b), x_1 = x] = \frac{a}{a+b} V(a+1,b) + \frac{b}{a+b} V(a,b+1),$$

by Remark 1 in the e-companion. Thus (10) becomes $V_x(a,b) = \max[0, L_x(a,b,V_x)]$, where

$$L_x(a,b,V_x) = -c_x - h_x(a,b) + \frac{a}{a+b}\left[h_x(a+1,b) + V_x(a+1,b)\right] + \frac{b}{a+b}\left[h_x(a,b+1) + V_x(a,b+1)\right]. \tag{13}$$

5.1. Geometric Horizon

To compute the optimal policy for a geometric horizon in Section 4.1 with respect to Bernoulli sampling, we use a technique from Varaiya et al. (1985) for off-line computation of Gittins indices, which assumes a finite state space.

Since the state space is infinite in our problem, we apply this technique in an approximate sense. For each alternative x with initial state (a_{0x}, b_{0x}), we first truncate the horizon for its sub-problem to N_0 (we use $N_0 = 50$ in our experiments). We then apply Varaiya et al. (1985) to pre-compute the Gittins indices for a finite set of states: $\{(a_{0x}+n_a, b_{0x}+n_b) : n_a, n_b \in \mathbb{N}, n_a+n_b \leq N_0, x = 1\ldots, k\}$. When an alternative is sampled more than N_0 times, we take the current state as the new (a_{0x}, b_{0x}), and recompute the indices for a new set of states.

5.2. Infinite Horizon with Sampling Costs

For the infinite horizon case with sampling costs, we explicitly compute the optimal policy for Bernoulli sampling, which is characterized for general sampling distributions in Section 4.2.

By Theorem 1, it suffices to evaluate for each alternative x all possible states in the sampling process. If Payoff Condition 3 holds, then $V_x(a_{0x}+n_a, b_{0x}+n_b) = 0$ for $n_a+n_b \geq N_x$ by Propositions 3 and 4; and for $n_a+n_b = n < N_x$, $V_x(a_{0x}+n_a, b_{0x}+n_b)$ can be computed recursively from $n = N_x - 1$ to $n = 0$ using (13). If Payoff Condition 3 does not hold, or some N_x are large, a pre-specified N_0 can be used instead of N_x to reduce computation. That is, we approximate $V_x(a_{0x}+n_a, b_{0x}+n_b)$ by 0 for $n_a+n_b \geq N_0$ (we use $N_0 = 1000$ in our experiments).

5.3. Example Terminal Payoff Functions

We state in Table 2 that 0-1 and linear terminal payoff functions satisfy Payoff Condition 3 with Bernoulli sampling, and we give explicit expressions for N_x and $\widetilde{H}_x(a,b)$. The proof of the statements in the table can be found in the e-companion.

Table 2 Example Terminal Payoff Functions with Bernoulli Sampling

	0-1 Terminal Payoff	Linear Terminal Payoff
Payoff Condition 3	Yes	Yes
$\widetilde{H}_x(a,b)$	$\frac{m_{0x}+m_{1x}}{2\sqrt{2\pi(a+b)}}$	$\frac{\max\{m_{0x},m_{1x}\}+m_{0x}}{4(a+b+1)}$
N_x	$\left(\lceil \frac{(m_{0x}+m_{1x})^2}{8\pi c_x^2} \rceil - 2\right)^+$	$\left(\lceil \frac{\max\{m_{0x},m_{1x}\}+m_{0x}}{4c_x} \rceil - 3\right)^+$

6. Specialization to Normal Sampling

We now consider normally distributed samples with known variance. As done in Section 5 for Bernoulli samples, we give explicit expressions for the quantities described generally in Section 4, and then present additional theoretical results to compute the optimal policy.

Here the sampling precision is known for each alternative x and denoted by β_x^ϵ. Hence η_x only consists of θ_x, and $d_x \in (-\infty, +\infty)$ for all x. We have $y_n \mid \theta, x_n \sim \mathcal{N}(\theta_{x_n}, 1/\beta_{x_n}^\epsilon)$ for all $n \geq 1$. We adopt an independent conjugate $\mathcal{N}(\mu_{0x}, 1/\beta_{0x})$ prior for each θ_x, and our normal samples result in a sequence of normal posterior distributions on θ_x with parameters $S_{n,x} = (\mu_{nx}, \beta_{nx})$ in parameter space $\Lambda = (-\infty, +\infty) \times [0, +\infty)$. We take $\mathbf{S}_n = (\boldsymbol{\mu}_n, \boldsymbol{\beta}_n)$ as the state of the DP, where the state space is $\mathbb{S} = (-\infty, +\infty)^k \times [0, +\infty)^k$.

Using Bayes rule, we write the state update function G as follows. For all $n \geq 0$,

$$\mu_{n+1,x} = \begin{cases} [\beta_{nx}\mu_{nx} + \beta_x^\epsilon y_{n+1}]/\beta_{n+1,x} & \text{if } x = x_{n+1} \\ \mu_{nx} & \text{otherwise} \end{cases}, \qquad \beta_{n+1,x} = \begin{cases} \beta_{nx} + \beta_x^\epsilon & \text{if } x = x_{n+1} \\ \beta_{nx} & \text{otherwise} \end{cases}.$$

Frazier et al. (2008) gives a probabilistically equivalent form of this update in terms of an $\mathcal{F}_n$-adapted sequence of standard normal random variables $Z_1, Z_2, \ldots$. More specifically, for all $n \geq 1$,

$$(\boldsymbol{\mu}_n, \boldsymbol{\beta}_n) = \left(\boldsymbol{\mu}_{n-1} + \tilde{\sigma}_{x_n}(\beta_{n-1,x_n}) Z_n \mathbf{e}_{x_n}, \boldsymbol{\beta}_{n-1} + \beta_{x_n}^\epsilon \mathbf{e}_{x_n}\right), \tag{14}$$

where $\tilde{\sigma}_x \colon (0, \infty] \mapsto [0, \infty)$ for each x is defined by $\tilde{\sigma}_x(\gamma) = \sqrt{(\gamma)^{-1} - (\gamma + \beta_x^\epsilon)^{-1}} = \sqrt{\beta_x^\epsilon / [\gamma(\gamma + \beta_x^\epsilon)]}$.

It follows that for any $(\mu, \beta) \in \Lambda$, $\mathcal{D}(\mu, \beta) = \mathcal{N}(\mu, 1/\beta)$ and

$$p_x(\mu, \beta) = 1 - \Phi\left(\sqrt{\beta}(d_x - \mu)\right), \ \mathcal{R}_x(\mu, \beta) = -c_x + \mathbb{E}\left[h_x\left(\mu + \tilde{\sigma}_x(\beta)Z, \beta + \beta_x^\epsilon\right)\right] - h_x(\mu, \beta),$$

where Φ is the standard normal cdf and Z is a standard normal random variable.

When $\alpha = 1$ and $c_x > 0$ for all x, (10) becomes

$$V_x(\mu, \beta) = \max\left[0, L_x(\mu, \beta, V_x)\right], \text{ where } L_x(\mu, \beta, V_x) = \mathcal{R}_x(\mu, \beta) + \mathbb{E}[V_x(\mu + \tilde{\sigma}_x(\beta)Z, \beta + \beta_x^\epsilon)]. \tag{15}$$

6.1. Geometric Horizon

Similar to the Bernoulli sampling case in Section 5.1, we first truncate the horizon for each subproblem to N_0. With normal sampling, this is not yet enough to provide a finite set of states: although β is discrete, μ takes continuous values. We thus need the following condition on H_x.

Special Condition 1 *For any fixed x and β, $H_x(\mu, \beta) \to 0$ as $\mu \to +\infty$ or $\mu \to -\infty$.*

Under Payoff Conditions 1, 2 and Special Condition 1, we know from (7), Propositions 2 and 3 that for any fixed β, $\mathcal{R}_x(\mu, \beta) \to -c_x$, $\nu_x(\mu, \beta) \to -c_x$, and $V_x(\mu, \beta) \to 0$ as $\mu \to +\infty$ or $\mu \to -\infty$. We can then truncate and discretize the range of μ_x as follows.

Let $\epsilon, \delta > 0$ be small (we take $\epsilon = \delta = 0.01$ in our experiments). For each fixed $\beta \in \{\beta_{0x} + n\beta_x^\epsilon : 0 \leq n \leq N_0\}$, we compute an interval $\left[\overline{\mu}_x(\beta), \underline{\mu}_x(\beta)\right]$ (guaranteed to exist under Special Condition 1) such that for all $\mu \notin \left[\overline{\mu}_x(\beta), \underline{\mu}_x(\beta)\right]$, we have $0 \leq H_x(\mu, \beta) \leq \epsilon$, and hence $-c_x \leq \nu_x(\mu, \beta) \leq -c_x + \epsilon$ by Proposition 2. We then discretize $\left[\overline{\mu}_x(\beta), \underline{\mu}_x(\beta)\right]$ into points with interval δ, denoted by $\{\mu_x^i(\beta)\}_i$.

We now use the technique from Varaiya et al. (1985) to pre-compute the Gittins indices for a finite set of states, $\{(\mu_x^i(\beta), \beta) : \beta \in \{\beta_{0x} + n\beta_x^\epsilon : 0 \leq n \leq N_0\}, x = 1, \ldots, k\}$, with the transition probability matrix approximated by the density ratios using (14):

$$\mathbb{P}\left[\left(\mu_x^i(\beta), \beta\right) \to \left(\mu_x^j(\beta + \beta_x^\epsilon), \beta + \beta_x^\epsilon\right)\right] = \varphi\left(\frac{\mu_x^j(\beta + \beta_x^\epsilon) - \mu_x^i(\beta)}{\tilde{\sigma}_x(\beta)}\right) \Big/ \sum_k \varphi\left(\frac{\mu_x^k(\beta + \beta_x^\epsilon) - \mu_x^i(\beta)}{\tilde{\sigma}_x(\beta)}\right),$$

where φ is the standard normal pdf. For an arbitrary state (μ, β) of alternative x, we set

$$\nu_x(\mu, \beta) = \begin{cases} -c_x & \text{if } \mu \notin \left[\overline{\mu}_x(\beta), \underline{\mu}_x(\beta)\right], \\ \nu_x\left(\mu_x^i(\beta), \beta\right) & \text{otherwise, where } i = \arg\min_j \{|\mu - \mu_x^j(\beta)|\}. \end{cases}$$

As we did in the Bernoulli sampling case, we also track the number of samples taken from each alternative and recompute the indices when an alternative is sampled more than N_0 times.

6.2. Infinite Horizon with Sampling Costs

Computation of the optimal policy in Section 4.2 is not trivial for normal sampling. To implement Bellman's recursion (15) directly, we need to evaluate each single-alternative value function over the whole continuous domain of μ, which is not possible. Instead, we truncate and discretize the range of μ to evaluate it approximately. The following condition is required for applying truncation.

Special Condition 2 *For any fixed x and β, we can compute an interval $\left[\overline{\mu}_x(\beta), \underline{\mu}_x(\beta)\right]$ such that $\mu \notin \left[\overline{\mu}_x(\beta), \underline{\mu}_x(\beta)\right] \Rightarrow V_x(\mu, \beta) = 0$.*

Under Payoff Condition 3 and Special Condition 2, we can evaluate each V_x as follows. By Propositions 3 and 4, $V_x\left(\mu, \beta_{0x} + n\beta_x^\epsilon\right) = 0$ for all $n \geq N_x$ and $\mu \in \mathbb{R}$. For each $\beta \in \{\beta_{0x} + n\beta_x^\epsilon : 0 \leq n < N_x\}$, we compute $\left[\overline{\mu}_x(\beta), \underline{\mu}_x(\beta)\right]$ (given in Special Condition 2) as a boundary of the value of μ within $\mathbb{C}_x$, and discretize it into points $\{\mu_x^i(\beta)\}_i$ with an interval of δ between them (we set $\delta = 0.01$ in our experiments). Using Remark 2 in the e-companion and (15), we know that each $V_x\left(\mu_x^i(\beta), \beta\right)$ can be computed recursively for $\beta \in \{\beta_{0x} + n\beta_x^\epsilon : 0 \leq n < N_x\}$. For any arbitrary state (μ, β), we set

$$V_x(\mu, \beta) = \begin{cases} 0 & \text{if } \mu \notin \left[\overline{\mu}_x(\beta), \underline{\mu}_x(\beta)\right], \\ V_x\left(\mu_x^i(\beta), \beta\right) & \text{otherwise, where } i = \arg\min_j \{|\mu - \mu_x^j(\beta)|\}. \end{cases}$$

6.3. Example Terminal Payoff Functions

Remark 3 in the e-companion describes the explicit computation of $\mathcal{R}_x(\mu,\beta)$ for $(\mu,\beta)\in\Lambda$ under 0-1 and linear terminal payoff. Table 3 shows that with normal sampling, these payoff functions satisfy Payoff Condition 3, Special Conditions 1 and 2. It uses the following additional notation:

$$A_1 = A_4^2 + 2A_4A_5, \qquad A_2 = 2(A_4+A_5)/\pi, \qquad A_3 = \pi^{-2} - A_4^2,$$
$$\text{where} \qquad A_4 = 1 + 1/\sqrt{2\pi e} \quad \text{and} \quad A_5 = c_x/\max\{m_{0x}, m_{1x}\}.$$

The proof of the statements in this table can be found in the e-companion.

Table 3 Example Terminal Payoff Functions with Normal Sampling

	0-1 Terminal Payoff	Linear Terminal Payoff
Payoff Condition 3	Yes	Yes
$\widetilde{H}_x(\mu,\beta)$	$\max\{m_{0x},m_{1x}\}\left[\left(\sqrt{1+\beta_x^\epsilon/\beta}-1\right)A_3+\pi^{-1}\sqrt{\beta_x^\epsilon/\beta}\right]$	$\frac{m_{0x}+m_{1x}}{\sqrt{2\pi\beta}}$
N_x	$\left\lceil\frac{A_2+\sqrt{A_2^2-4A_1A_3}}{2A_1}\right\rceil$	$\left\lceil\frac{(m_{0x}+m_{1x})^2}{2\pi c_x^2\beta_x^\epsilon}\right\rceil$
Special Condition 1	Yes	Yes
Special Condition 2	Yes	Yes

7. Numerical Results

In this section, we test the performance of the Bayes-optimal policies with a collection of numerical experiments. We first present illustrative example problems in Section 7.1, and then present applications to ambulance positioning in Section 7.2, and the revenue of a production line in Section 7.3.

We introduce the following sampling policies for comparison. In the implementation of all of these policies, ties in the arg max are broken uniformly at random.

1. **Pure Exploration (PE):** In this policy, we choose the next alternative to sample uniformly and independently at random, i.e., $x_n \sim \text{Uniform}(1,\ldots,k)$ for all $n \geq 1$.

2. **Large Deviations (LD):** In this policy, we sample according to the asymptotically optimal fractional allocations $p_1^*,\ldots,p_k^*$ with the highest exponential rate of decay for the expected number of incorrect determinations. Szechtman and Yücesan (2008) provides explicit characterizations for Bernoulli and normal sampling. Our implementation of LD is idealized, as it uses the exact values of p^*, which require knowing θ. These exact values are unavailable in practice, and so practical algorithms, such as those proposed in Szechtman and Yücesan (2008), use estimates instead.

3. **Andradóttir and Kim (AK):** AK is a feasibility check procedure for normally distributed systems with a single constraint, described in Algorithm $\mathcal{F}$ in Andradóttir and Kim (2010), and Procedure $\mathcal{F}_B^I$ in Batur and Kim (2010). It provides a statistical guarantee on performance with pre-specified tolerance level ϵ^{AK} and lower bound $1-\alpha^{AK}$ on the probability of correct decision (PCD). It uses an initial stage with $n_0^{AK} \geq 2$ samples from each alternative. Our implementation uses known variance instead of an estimate. Batur and Kim (2010) states that AK is robust to non-normality, so we apply it to both Bernoulli and normal sampling.

4. **Knowledge Gradient (KG):** In this policy, the sampling decision is the one that would be optimal if only one measurement were to remain, i.e., $x_{n+1} \in \arg\max_x \{\mathcal{R}_x(S_{n,x})\}$ for all $n \geq 0$. Such policies have also been called myopic or one-step-lookahead policies, and are discussed in detail in Frazier (2009). When sampling costs are strictly positive, KG stops when the one-step expected reward becomes non-positive. This stopping rule is $\tau = \inf\{n \geq 0 : \mathcal{R}_x(S_{n,x}) \leq 0, \forall x\}$. An analogous stopping rule for ranking and selection was introduced in Frazier and Powell (2008).

7.1. Illustrative Example Problems

We first explore the relative performance of these policies and the Bayes-optimal one on six different problem settings: geometric horizon with Bernoulli sampling; geometric horizon with normal sampling; deterministic horizon with Bernoulli sampling; deterministic horizon with normal sampling; infinite horizon with Bernoulli sampling; and infinite horizon with normal sampling. We adopt the 0-1 terminal payoff function with $m_{0x} = m_{1x} = 1$ for all x.

For Bernoulli sampling, we test $k = 100$ alternatives with each d_x picked from Uniform $[0,1]$ and each θ_x randomly generated from prior distribution Beta$(1,1)$. For normal sampling, we test $k = 50$ alternatives with d_x picked from $\mathcal{N}(0,100)$, θ_x generated from prior distribution $\mathcal{N}(0,100)$, and β_x^ϵ picked from Uniform $[0.5,2]$. For the geometric horizon, we set $c_x = 0$ for all x and vary α so that $\mathbb{E}[T] = \frac{1}{1-\alpha}$ varies within $[100,1000]$ in the Bernoulli sampling case, and within $[50,500]$ in the normal sampling case. The actual values of T are randomly generated according to the corresponding geometric distribution. For the deterministic horizon, we apply the Bayes-optimal policy as a heuristic, by choosing α so that $\frac{1}{1-\alpha}$ is equal to the horizon. For the infinite horizon, we set $c_x = c$ for all x and vary c within $[0.001, 0.01]$.

Since LD does not have a stopping rule, we only implement it for the geometric horizon and the deterministic horizon. For AK, we set $n_0^{AK} = 2$. If AK has not yet classified an alternative as in or out of the level set by the end of the horizon, we classify it according to the optimal level set estimate B_n. For the finite (geometric or deterministic) horizon, we optimize AK's performance over the values of ϵ^{AK} and α^{AK} using a simple grid search, and report performance with the best values of these parameters. For the infinite horizon with sampling costs, we found that AK's expected

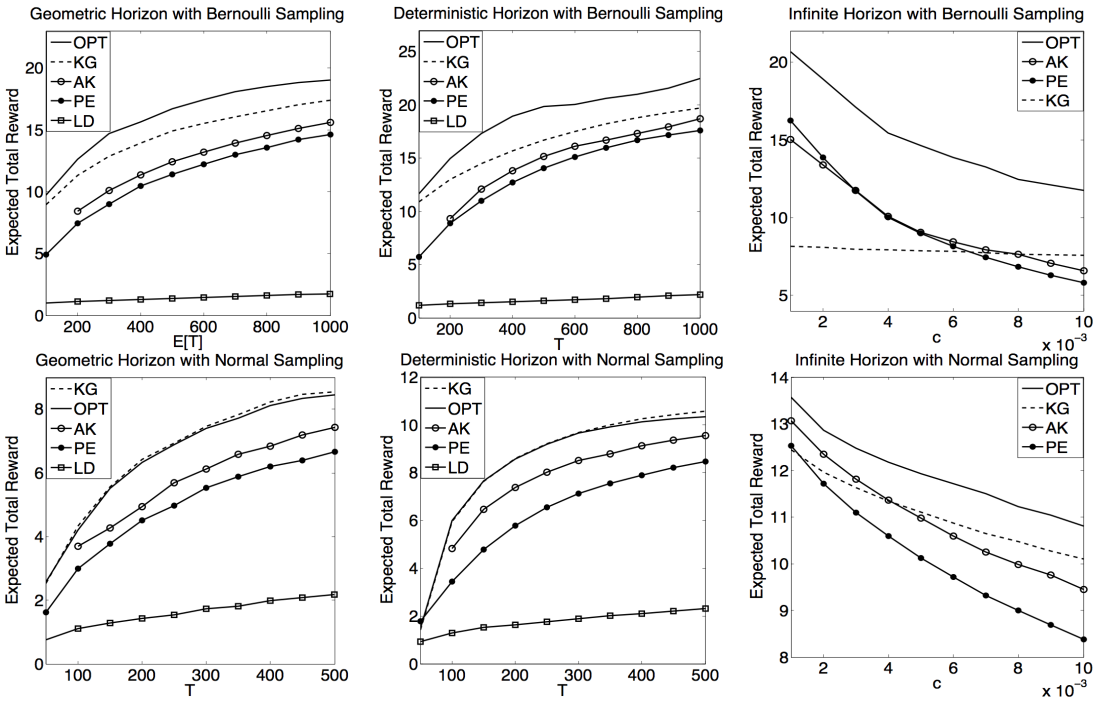

Figure 1 Performance of the following policies: Pure Exploration (PE), Large Deviations (LD), Andradóttir and Kim (AK), Knowledge Gradient (KG), and the approximate implementation of Bayes-Optimal (OPT) using necessary truncation and discretization. The maximum length of the 95% confidence intervals for the values in the plots is 0.24.

total reward was significantly lower than that of the other policies, for both normal and Bernoulli sampling and across all values of ϵ^{AK}, α^{AK} and c. This is due to AK's tendency to occasionally stop too late in the presence of sampling costs. To improve its performance in the infinite horizon setting, we introduce an additional parameter $\overline{T}$ to the procedure and prevent AK from starting new sampling stages after a deterministic time $\overline{T}$. We then optimize over the values of $\epsilon^{AK}, \alpha^{AK}$ and $\overline{T}$ and use the best values found. For PE, we also use a deterministic stopping rule $\tau = \overline{T}$ for the infinite horizon, and optimize over $\overline{T}$ to report its performance. We find that the best deterministic value of $\overline{T}$ for PE and AK is usually near the Bayes-optimal policy's expected stopping time.

Figure 1 shows that in all six problem settings, the Bayes-optimal policies significantly outperform PE. Because naive strategies like PE are the ones most commonly used in practice, we see that practitioners can substantially improve performance by using optimal policies instead.

LD's poor performance seems surprising, but can be understood as follows. In most sample paths, LD allocates all its samples to one or two alternatives that have significantly smaller $|\theta_x - d_x|$ and hence significantly larger p^*_x (see Szechtman and Yücesan (2008)). As a result, the total reward earned by sampling, which is the increment in the number of correct determinations, is always below 3. Indeed, given a fixed or expected sample size N, the expected number of samples allocated to any

alternative x with $p^*_x < \frac{1}{N}$ is less than 1. This can be the case for most of the k alternatives, especially when N/k is small. While LD has optimal asymptotic performance, this does not necessarily lead to good finite time performance.

We now look at the performance of AK. With normal sampling, it performs significantly better than PE, but is outperformed uniformly by the Bayes-optimal policy, and by KG when c is relatively large. The main reason is that AK looks for a feasible set with the target tolerance level and PCD, but does not consider that the payoff of a high-quality determination may not be worth the sampling effort required to achieve it given a limited budget or positive sampling costs. With Bernoulli sampling, AK performs closer to PE. While Batur and Kim (2010) demonstrates that batching can be helpful in achieving approximate normality and satisfying statistical guarantees for difficult Bernoulli cases, where θ_x is close to 0 or 1, we implement AK directly without batching because our focus is on expected total reward, rather than satisfying statistical guarantees.

The performance of KG depends greatly on the problem. It is almost optimal in the geometric-horizon normal-sampling and the deterministic-horizon normal-sampling settings, while in the other four settings it is significantly suboptimal. Its worst performance comes in the infinite-horizon Bernoulli-sampling setting, where it is even worse than PE and MV with small values of c.

To understand the behavior of KG, first consider the two problem settings with normal sampling. KG makes its decisions using the one-step approximation $\mathcal{R}_x(S_{n,x})$ to the true value of sampling alternative x. This approximation is the sum of a one-step value of information (VOI) and the cost of sampling. As observed in Frazier and Powell (2010), Chick and Frazier (2011), the one-step VOI for normal sampling can significantly underestimate the true VOI when more samples will be taken later. This causes KG stopping rules to stop too soon (Chick and Frazier 2011), hurting their performance in problems with strictly positive sampling costs. This is likely to be the largest contributor to KG's suboptimality in the infinite-horizon normal-sampling problem.

Unlike the stopping decision, this underestimation of the true VOI has little effect on allocation decisions, because the level of underestimation is relatively constant across alternatives, and the alternative with the largest one-step VOI tends to also have near-maximal true VOI (Chick and Frazier 2011, Frazier et al. 2008, 2009). This is the reason that KG does so well in the geometric-horizon normal-sampling and deterministic-horizon normal-sampling problem settings, where there are no sampling costs and the only decisions concern allocation. In this case, KG's performance is comparable with that of the Bayes-optimal policy. Although KG actually outperforms our implementation of the Bayes-optimal policy by a slight margin for large $\mathbb{E}[T]$ in the geometric horizon and large T in the deterministic horizon, this small gap is an artifact due to numerical inaccuracies introduced by discretizing the state-space when solving the DP that defines the Bayes-optimal

policy, and would vanish with finer discretization. For discussions of discretization error in dynamic programming, see Powell (2007) and Bertsekas and Tsitsiklis (1996).

In problems with Bernoulli sampling, the discrete nature of the samples often causes the one-step VOI to be 0. This occurs when a single sample x_n, y_n is not enough to alter our decision on whether to place an alternative x in B_n, even when significant uncertainty about θ_x remains and more than one sample could alter our decision. In these situations, KG stops sampling immediately if there are positive sampling costs, or otherwise allocates its sample randomly (an inefficient strategy) among the alternatives. For this reason, KG performs poorly in both settings with Bernoulli sampling.

7.2. Ambulance Quality of Service Application

To demonstrate the Bayes-optimal policies in a more realistic application setting, we use it to analyze methods for positioning ambulances in a large city. We use the ambulance simulation introduced by Maxwell et al. (2010), which simulates ambulances responding to emergency calls in a city. The simulation model is very loosely based on the city of Edmonton, but is sufficiently modified in call arrival rates, etc., that the results have no bearing on actual ambulance performance in that city. The city considers an emergency call to be answered on time if an ambulance arrives within 8 minutes, and otherwise it considers the call to be missed. We suppose that the city is considering several different static allocations of their fleet of 16 ambulances across the 11 bases in the city, and would like to know for each candidate allocation and each of several different call arrival rates whether it meets the minimum requirement of 70% of calls answered on time.

This is an MCS problem. Each alternative x corresponds to an hourly call arrival rate λ_x and an ambulance positioning plan. Based on these, each sample from x gives the number of calls answered on time during a two-week simulation. Since the emergency calls are generated according to a Poisson process, the expected total number of calls during two weeks for alternative x is known analytically as $M_x = 24 \times 14 \times \lambda_x$. Instead of directly measuring the fraction of calls answered on time in each simulation and estimating its expectation, we take θ_x/M_x as the long-term percentage of calls answered on time (see Henderson (2000)). The set of alternatives meeting or exceeding 70% of calls answered on time is therefore $\mathbb{B} = \{x : \theta_x/M_x \geq 0.7\} = \{x : \theta_x \geq d_x\}$, where $d_x = 0.7 \times M_x$. We chose 25 ambulance positioning plans and 25 values for the hourly call arrival rate from $[3, 6.6]$ for our experiment. This provides a collection of $25 \times 25 = 625$ alternatives.

The number of calls answered on time in a two-week simulation is approximately normally distributed. This was confirmed by visual examination of the empirical distribution for several randomly picked alternatives. We also assume a common sampling precision for all the alternatives. We confirmed that this assumption is reasonable by calculating and comparing the sampling precisions of several different alternatives chosen at random. To estimate the common sampling

precision, we randomly chose 5 alternatives, sampled 20 times from each of them to estimate their individual sampling precisions, and used the average of the 5 sampling precisions as the estimate of the common sampling precision. This estimate was 1.4×10^{-3}. In problems with a high degree of variation in the sampling variances, one might instead estimate the sampling precisions separately for each alternative, or assume normal samples with unknown mean and unknown variance, with an inverse-gamma prior on the unknown sampling variance (DeGroot 1970). In this second case, the optimal policy could be computed by applying the theoretical results in this paper to the exponential family of normal distributions with unknown mean and unknown variance, but computing solutions to these dynamic programs would require further work.

We use independent normal priors for each θ_x. We take a single sample from each alternative, set the prior mean μ_{0x} to this sampled value, and the prior precision β_{0x} to the common sampling precision. This is equivalent to using a non-informative prior and starting sampling by taking a single sample from each alternative. We then follow one of several different sampling policies for a deterministic horizon, where we investigate performance at a few fixed times. For the heuristic Bayes-optimal policy, we set $\alpha = 0.999$ (corresponding to a horizon of $1/(1-\alpha) = 1000$) and $c_x = 0$ for all x. We assume a 0-1 terminal payoff function with $m_{0x} = m_{1x} = 1$ for x, hence our estimate of set $\mathbb{B}$ at each time n is $B_n = \{x : \mu_{nx} \geq d_x\}$ by Table 1 and (15). For AK, we report its best performance over multiple sample paths with $n_0^{AK} = 2$, ϵ^{AK} chosen from $\{5, 10, 25, 50, 100, 110, 120, 130, 140, 150\}$, and α^{AK} chosen from $\{0.1, 0.3, 0.5, 0.7, 0.9, 0.99, 0.999, 0.9999, 0.99999, 0.999999\}$. As $1 - \alpha^{AK}$ gives a lower bound on AK's probability of correct decision, one would typically choose α^{AK} closer to 0, e.g., to 0.1 or 0.3. The set over which we optimize includes both these smaller values, as well as values closer to 1, because we found that increasing α^{AK} improved AK's performance.

Figure 2 compares the Bayes-optimal policy against PE, LD, AK, and KG. The performance of each policy is measured by the similarity between $\mathbb{B}$ and $(B_n)_{n \geq 0}$, where we independently estimated $\mathbb{B}$ through exhaustive simulation of each θ_x. We also sorted the ambulance positioning plans used to construct the alternatives, in order of decreasing θ_x (at a fixed value of the call arrival rate), to make the set $\mathbb{B}$ easier to visualize. Each panel shows a black line, which is the independently obtained high-accuracy estimate of the boundary between $\mathbb{B}$ and its complement. For each policy and after each of 500, 1000, 2000, and 5000 samples we plot the current estimate B_n as the light region, and the complement of B_n as the dark region.

Figure 2 shows that PE behaves poorly in distinguishing among the alternatives. Under this policy, after 5000 samples total, approximately $5000/625 = 8$ samples have been taken from each alternative. For those alternatives x with θ_x close to d_x, this number of samples is much too small to accurately estimate whether x is in $\mathbb{B}$ or not. Moreover, the estimates given by LD barely change as the number of samples increases. The reason is that it only samples from 2 alternatives out of 625 in

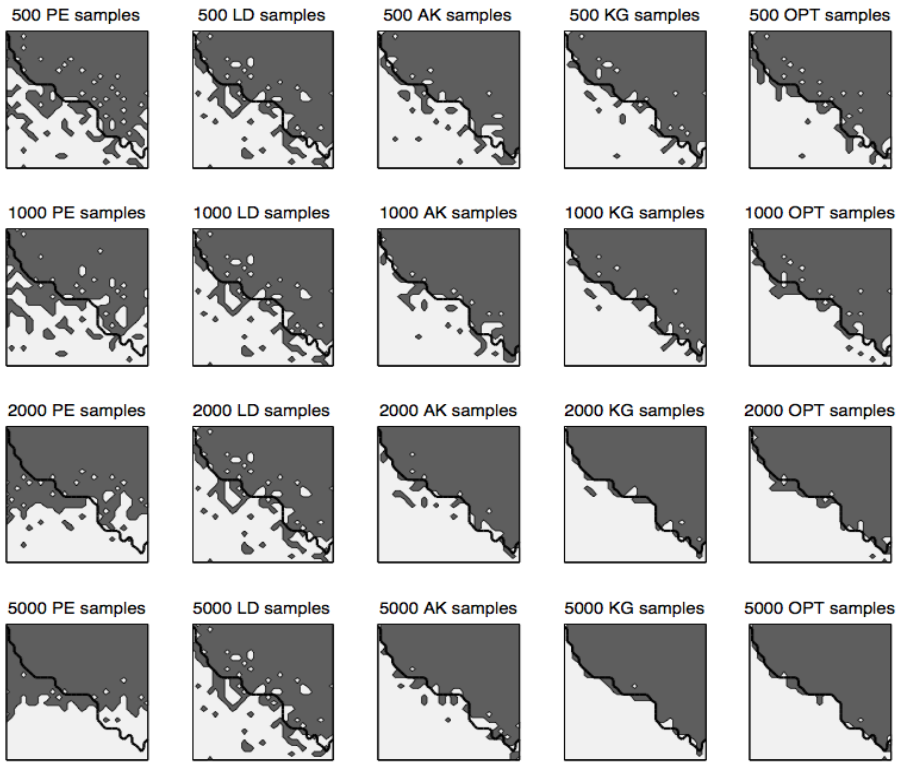

Figure 2 Performance of the sampling policies in the ambulance quality of service application: Pure Exploration (PE), Large Deviations (LD), Andradóttir and Kim (AK), Knowledge Gradient (KG), and Bayes-Optimal (OPT). In each plot, the black curve is the boundary of $\mathbb{B}$ (the set of alternatives answering at least 70% of the emergency calls on time); the light region is the estimate of the set $\mathbb{B}$ under the corresponding policy given the marked number of samples; and the dark region is the complement of the light region. Hourly call arrival rates $3, 3.15, 3.3, \ldots, 6.6$ are distributed along the vertical axis. Ambulance positioning plans are distributed along the horizontal axis, and are sorted to make the black line decreasing.

the 5000 samples. In contrast, AK, KG and the Bayes-optimal policy are much more efficient, while the latter two significantly outperform AK. AK's best performance occurs at a very large tolerance level, $\epsilon^{AK} = 130$, and a very small lower bound on the PCD, $1 - \alpha^{AK} = 10^{-4}$. We believe this is because the budget we are considering is much smaller than AK would typically require to provide a meaningful bound on PCD with a large number of alternatives ($k = 625$). The excellent performance of KG relative to optimal should not be surprising: samples are approximately normally distributed and there are no sampling costs, which is the setting from Section 7.1 in which KG was nearly optimal. Had the problem used Bernoulli sampling or strictly positive sampling costs, then KG likely would not have performed as well.

As shown in Figure 2, after 5000 samples under KG or the Bayes-optimal policy, we have estimated $\mathbb{B}$ with a high degree of accuracy. Indeed, with only 500 samples from either of these policies,

we estimate $\mathbb{B}$ with greater accuracy than is possible with 5000 samples under PE or LD. This factor of 10 in sampling effort represents a dramatic savings of simulation time, and demonstrates the value of using an optimal or near-optimal sampling policy when performing MCS.

7.3. Revenue of a Production Line Application

We consider a manufacturing firm that can choose, on each day, whether or not to operate a production line. If it operates the line, it earns a stochastic net revenue that can be positive or negative. If it chooses not to operate the line, its net revenue is 0.

At the beginning of each day, the firm observes market and operational conditions X. Each X takes one of $k < \infty$ values. The firm has a simulator that, given X, can simulate a net daily revenue $Y(X)$. We suppose that this simulator is cumbersome, taking a long time to run, and hence the plant manager does not want to run it at the beginning of each day. Instead, at time 0, the manager would like to use simulation to estimate $\theta_x = \mathbb{E}[Y(X) \mid X = x]$ for a wide variety of conditions x, and then estimate the set $\mathbb{B} = \{x : \theta_x \geq 0\}$ in which it is profitable to operate the production line. The manager can then make decisions when each X becomes known with a simple look up table based on this estimate of $\mathbb{B}$.

Suppose that the plant manager recreates the look up table every month. From historical data and forecasts, she has estimated the long-run percentage of time each condition x occurs as q_x. Thus the expected monthly net revenue she earns from an estimate B of $\mathbb{B}$ is a one-sided linear terminal payoff with $d_x = 0$ and $m_x = 30q_x$ for all x, i.e., $r(B;\theta,d) = \sum_{x\in B} m_x\theta_x$.

In performing her simulations, the plant manager has no time limit, but she is using a third-party computing service such as Amazon EC2 to perform her computation. This service charges her a monetary cost of c_x for each simulation of condition x, which is determined by the length of time that this simulation requires to run and the unit cost per CPU hour determined from the computing service's pricing structure (Amazon.com 2012).

The simulation of the production line we consider is based on Buchholz and Thummler (2005) and "Optimization of a Production Line" in SimOpt.org (2011). It has 4 service queues arranged serially, each of which has a fixed finite capacity 5. Parts (customer orders) arrive to queue 1 according to a Poisson process with rate λ. Each queue sends the parts into a corresponding single server with first-come-first-serve discipline, and the service time is exponentially distributed with rate γ. Parts leaving queue i after service are immediately transferred to the next queue $i+1$ (if possible). Whenever the service queue $i+1$ is full, the server i is said to be blocked and the part in it cannot leave even if it is completed, since there is no room in the next queue.

We have $k = 500$ alternatives in our experiment. Each alternative condition x corresponds to a customer arrival rate λ_x in $\{5.1, 5.2, \ldots, 7.5\}$ and a queue service rate γ_x in $\{5.1, 5.15, \ldots, 6.05\}$.

Table 4 Performance of the Sampling Policies in the Revenue of a Production Line Application

Sampling Policy	PE	AK	KG	OPT
expected terminal payoff: $\mathbb{E}\left[\sum_{x\in B_\tau} m_x\theta_x\right]$	2709	2748	2667	2859
expected sampling cost: $\mathbb{E}\left[\sum_{n=1}^{\tau} c_{x_n}\right]$	198	195	84	198
expected total reward: $\mathbb{E}\left[\sum_{x\in B_\tau} m_x\theta_x - \sum_{n=1}^{\tau} c_{x_n}\right]$	2511	2553	2583	2661

Denote by Γ_x the expected number of parts leaving the last queue (expected number of completed orders) in an 8-hour period under condition x. We assume that the net revenue per order filled is \$50, and operation of the production line over an 8-hour period has a fixed basic cost of \$2800. The expected revenue of the production line under condition x is then $\theta_x = 50\Gamma_x - 2800$. The level set we wish to estimate is thus $\mathbb{B} = \{x : \Gamma_x \geq 56\}$.

The number of parts leaving the last queue in an 8-hour simulation is approximately normally distributed, which was confirmed by visual examination of the empirical distribution for several randomly picked alternatives. As in the ambulance quality of service application, we assume a common sampling precision for all the alternatives and estimate it by calculating and averaging over the sampling precisions of several different alternatives chosen at random. We also use independent normal priors for each θ_x, where we set μ_{0x} to be the value of an initial sample, and set β_{0x} to be the common sampling precision. We then follow one of several different sampling policies, assuming an infinite horizon with $q_x = 1/500$, representing homogeneous long-run percentages for each condition, $m_x = 30q_x$ as described above, $c_x = 0.06$, and $\beta_x^\epsilon = 1.2 \times 10^{-3}$ for all x. This sampling cost c_x of \$0.06 per simulation corresponds to a computing service that charges \$0.12 per CPU hour, and a simulation that takes 30 minutes for each replication. Our estimate of the set $\mathbb{B}$ when we stop at time τ is then $B_\tau = \{x : \mu_{\tau x} \geq 0\}$ by Table 1. To examine the expected total reward under each policy, we independently estimate each θ_x through exhaustive simulation.

Table 4 compares the Bayes-optimal policy against three other policies: PE, AK, and KG, showing their expected terminal payoff, expected sampling cost, and expected total reward by averaging over 2000 independent sample paths. The maximum length of the 95% confidence intervals for the values in the table is 3. Similar to Section 7.1, we report the performance of PE with a best deterministic stopping rule $\tau = \overline{T}$, and report the performance of AK with best pre-specified values of $\epsilon^{AK}, \alpha^{AK}$ and $\overline{T}$. Our results show that the best deterministic value of $\overline{T}$ for PE and AK is very close to the expected stopping time under the Bayes-optimal policy. While PE, AK and KG are all sub-optimal, KG is the best among the three sub-optimal policies. Though it stops too soon, its advanced efficiency in allocating samples still results in a relatively high expected total reward.

We also performed an additional experiment to assess the degradation of the Bayes-optimal policy caused by approximating the sampling precisions as known. We find that an optimal policy

with a high-fidelity estimate of each alternative's individual sampling precision obtained through exhaustive simulation has an expected total reward of 2664, which offers a 0.1% improvement over our previous implementation. We see that in this particular example, the degradation is very small.

8. Conclusions

By applying methods from multi-armed bandits and optimal stopping, we are able to efficiently solve the dynamic program for the Bayesian MCS problem and find Bayes-optimal fully sequential sampling and stopping policies. While researchers have searched for Bayes-optimal policies for other related problems in sequential experimental design and effort allocation for simulation, tractable computation of the optimal policies has remained elusive in many problems, and so the results in this paper place the MCS problem together with a select group of problems in sequential experimental design for which the sequential Bayes-optimal policies can be computed efficiently.

The Bayes-optimal policies presented are flexible, allowing limitations on the ability to sample to be modeled with either a random horizon or sampling costs or both, allowing sampling distributions from any exponential family, and allowing a broad class of terminal payoff functions. While they do not allow for problems with fixed horizon, the optimal policy for random horizons can be used heuristically in such situations. We provide explicit computations for Bernoulli sampling and normal sampling with known variance. We also provide expressions for the KG policy and show that it works extremely well for normal sampling with a geometrically distributed horizon and no sampling costs. Although the KG policy is not Bayes optimal, and results in some performance loss, its ease of use may make it attractive to practitioners facing MCS problems of this type.

In conclusion, the results in this paper provide new tools for simulation analysts facing MCS problems. These new tools dramatically improve efficiency over naive sampling methods, and make it possible to efficiently and accurately solve previously intractable MCS problems.

Acknowledgments

This work was supported by the Air Force Office of Scientific Research under FA9550-11-1-0083 and FA9550-12-1-0200, and by the National Science Foundation under CMMI-1254298, IIS-1142251, and IIS-1247696.

The authors would like to thank Matthew S. Maxwell and Shane G. Henderson for the use of their ambulance simulation software, and their help in using it. The authors would also like to thank the Associate Editor, and several anonymous referees, whose comments greatly improved the manuscript.

Author Biographies

JING XIE is a PhD student in the School of Operations Research and Information Engineering at Cornell University. She received a B.S. in mathematics from Fudan University at Shanghai, China in 2009. She is the recipient of the McMullen Fellowship in 2009-2010 and the Robert Kaplan PhD

Fellowship in 2013. She has a research interest in sequential design of simulation experiments. Her e-mail is jx66@cornell.edu and her web page is http://people.orie.cornell.edu/jx66/.

PETER I. FRAZIER is an assistant professor in the School of Operations Research and Information Engineering at Cornell University. He received a Ph.D. in Operations Research and Financial Engineering from Princeton University in 2009. He is the recipient of an AFOSR Young Investigator Award, and an NSF CAREER Award. His research interest is in dynamic programming and Bayesian statistics, focusing on the optimal acquisition of information. He works on applications in simulation, optimization, operations management, and medicine. His email address is pf98@cornell.edu and his web page is http://people.orie.cornell.edu/pfrazier/.

References

Amazon.com. 2012. http://aws.amazon.com/ec2/pricing/.

Andradóttir, S., D. Goldsman, S.H. Kim. 2005. Finding the best in the presence of a stochastic constraint. *Simulation Conference, 2005 Proceedings of the Winter*. IEEE, 7–pp.

Andradóttir, S., S.H. Kim. 2010. Fully sequential procedures for comparing constrained systems via simulation. *Naval Research Logistics (NRL)* **57**(5) 403–421.

Araman, V.F., R. Caldenty. 2009. Dynamic pricing for perishable products with demand learning. *Operations Research* **57** 1169 – 1188.

Batur, D., S.H. Kim. 2010. Finding feasible systems in the presence of constraints on multiple performance measures. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* **20**(3) 13.

Bechhofer, R.E., B.W. Turnbull. 1978. Two (k+1)-decision selection procedures for comparing k normal means with a specified standard. *Journal of the American Statistical Association* 385–392.

Bellman, R. 1954. The theory of dynamic programming. *Bulletin of the American Mathematical Society* **60** 503–516.

Bertsekas, D.P. 2005. *Dynamic Programming and Optimal Control, vol. I*. 3rd ed. Athena Scientific.

Bertsekas, D.P. 2007. *Dynamic Programming and Optimal Control, vol. II*. 3rd ed. Athena Scientific.

Bertsekas, D.P., J.N. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

Bofinger, E., G.J. Lewis. 1992. Two-stage procedures for multiple comparisons with a control. *American Journal of Mathematical and Management Sciences* **12** 253–253.

Buchholz, P., A. Thummler. 2005. Enhancing evolutionary algorithms with statistical selection procedures for simulation optimization. *Proceedings of the 2005 Winter Simulation Conference* 11–pp.

Chick, S.E., J. Branke, C. Schmidt. 2010. Sequential sampling to myopically maximize the expected value of information. *INFORMS Journal on Computing* **22**(1) 71–80.

Chick, S.E., P.I. Frazier. 2009. The conjunction of the knowledge gradient and economic approach to simulation selection. *Winter Simulation Conference Proceedings, 2009* 528–539.

Chick, S.E., P.I. Frazier. 2011. Sequential sampling for selection with ESP. In review.

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

Damerdji, H., M.K. Nakayama. 1996. Two-stage procedures for multiple comparisons with a control in steady-state simulations. *Proceedings of the 28th Winter Simulation Conference* 372–375.

DeGroot, M.H. 1970. *Optimal Statistical Decisions*. McGraw Hill, New York.

Ding, X., M.L. Puterman, A. Bisi. 2002. The censored newsvendor and the optimal acquisition of information. *Operations Research* **50**(3) 517–527.

Dudewicz, E.J., S.R. Dalal. 1983. Multiple comparisons with a control when variances are unknown and unequal. *American Journal of Mathematics and Management Sciences* **4** 275–295.

Dudewicz, E.J., J.S. Ramberg. 1972. Multiple comparisons with a control: Unknown variances. *The Annual Technical Conference Transactions of the American Society of Quality Control*, vol. 26.

Dunnett, C.W. 1955. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**(272) 1096–1121.

Dynkin, E.B., A.A. Yushkevich. 1979. *Controlled Markov Processes*. Springer, New York.

Frazier, P.I. 2009. Knowledge-gradient methods for statistical learning. Ph.D. thesis, Princeton University.

Frazier, P.I., W.B. Powell. 2008. The knowledge-gradient stopping rule for ranking and selection. *Winter Simulation Conference Proceedings, 2008* .

Frazier, P.I., W.B. Powell. 2010. Paradoxes in learning and the marginal value of information. *Decision Analysis* **7**(4).

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

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

Fu, M. 1994. Optimization via simulation: A review. *Annals of Operations Research* **53**(1) 199–248.

Gittins, J.C., D.M. Jones. 1974. A dynamic allocation index for the sequential design of experiments. J. Gani, ed., *Progress in Statistics*. North-Holland, Amsterdam, 241–266.

Glynn, P., S. Juneja. 2004. A large deviations perspective on ordinal optimization. *Proceedings of the 36th conference on Winter simulation*. Winter Simulation Conference, 577–585.

Glynn, P., S. Juneja. 2011. Ordinal optimization: a nonparametric framework. *Proceedings of the 2011 Winter Simulation Conference*. Winter Simulation Conference, 4062–4069.

Goldsman, D., B. Nelson. 1994. Ranking, selection and multiple comparisons in computer simulation. J. D. Tew, S. Manivannan, D. A. Sadowski, A. F. Seila, eds., *Proceedings of the 1994 Winter Simulation Conference*.

Gupta, S.S., K.J. 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.

Healey, C., S. Andradóttir, S.H. Kim. 2012. Selection procedures for simulations with multiple constraints under independent and correlated sampling. In review.

Henderson, S.G. 2000. Mathematics for simulation. *Simulation Conference Proceedings, 2000. Winter*, vol. 1. IEEE, 137–146.

Hochberg, Y., A.C. Tamhane. 1987. *Multiple Comparison Procedures*. Wiley New York.

Hsu, J.C. 1996. *Multiple Comparisons: theory and methods*. CRC Press, Boca Raton.

Hunter, SR, NA Pujowidianto, Chun-Hung Chen, Loo Hay Lee, R Pasupathy, Chee Meng Yap. 2011. Optimal sampling laws for constrained simulation optimization on finite sets: The bivariate normal case. *Simulation Conference (WSC), Proceedings of the 2011 Winter*. IEEE, 4289–4297.

Hunter, Susan R, Raghu Pasupathy. 2012. Optimal sampling laws for stochastically constrained simulation optimization on finite sets. *INFORMS Journal on Computing* .

Jones, D.R., M. Schonlau, W.J. Welch. 1998. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* **13**(4) 455–492.

Kim, S.H. 2005. Comparison with a standard via fully sequential procedures. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* **15**(2) 155–174.

Kim, S.H., B.L. Nelson. 2006. Selecting the best system. *Handbooks in operations research and management science: simulation* **13** 501–534.

Krause, A., J. Leskovec, C. Guestrin, J. VanBriesen, C. Faloutsos. 2008. Efficient sensor placement optimization for securing large water distribution networks. *Journal of Water Resources Planning and Management* **134** 516.

Lizotte, D., T. Wang, M. Bowling, D. Schuurmans. 2007. Automatic gait optimization with gaussian process regression. *Proceedings of International Joint Conferences on Artificial Intelligence*. 944–949.

Mahajan, A., D. Teneketzis. 2008. Multi-armed bandit problems. *Foundations and Applications of Sensor Management* 121–151.

Maxwell, M.S., M. Restrepo, S.G. Henderson, H. Topaloglu. 2010. Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing* **22** 266–281.

Mockus, J. 1989. *Bayesian Approach to Global Optimization: Theory and applications*. Kluwer Academic, Dordrecht.

Nelson, B.L., D. Goldsman. 2001. Comparisons with a standard in simulation experiments. *Management Science* **47**(3) 449–463.

Paulson, E. 1952. On the comparison of several experimental categories with a control. *The Annals of Mathematical Statistics* **23**(2) 239–246.

Paulson, E. 1962. A sequential procedure for comparing several experimental categories with a standard or control. *The Annals of mathematical statistics* 438–443.

Picheny, V., D. Ginsbourger, O. Roustant, R.T. Haftka, N.H. Kim, et al. 2010. Adaptive designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design* **132** 071008.

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

Ryzhov, I., W.B. Powell, P.I. Frazier. 2012. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research* **60**(1).

Siegmund, D. 1985. *Sequential Analysis: Tests and confidence intervals*. Springer Series in Statistics, Springer-Verlag, New York.

SimOpt.org. 2011. http://simopt.org/.

Szechtman, R., E. Yücesan. 2008. A new perspective on feasibility determination. *Proceedings of the 40th Conference on Winter Simulation*. Winter Simulation Conference, 273–280.

Varaiya, P.P., J.C. Walrand, C. Buyukkoc. 1985. Extensions of the multiarmed bandit problem: The discounted case. *IEEE Transactions on Automatic Control* **30**(5) 426–439.

Wald, A., J. Wolfowitz. 1948. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics* **19**(3) 326–339.