Submitted to *Operations Research*
manuscript (Please, provide the mansucript number!)

Sequential Bayes-Optimal Policies for Multiple Comparisons with a Control

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 known threshold. This determination is known as multiple comparisons with a control. Within a Bayesian formulation, the optimal fully sequential policy for allocating simulation effort is the solution to a dynamic program. 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 an application in ambulance positioning.

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

1. Introduction

We consider multiple comparisons with a control (MCC), in which simulation is used to determine which alternative systems under consideration have performance surpassing that of a known control. 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.

MCC problems arise in a number of applications. We give three examples in which methods for allocating simulation effort from this paper are appropriate.

- 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

calls answered in time. After this determination is made, they will choose among the acceptable methods using other criteria. We consider this application in detail.

- A firm has a number of possible projects in which it can invest its capital, and is interested in determining through simulation which of these projects have a positive net expected value.
- A logistics company is unsure about demand and weather conditions in the coming month. Executives would like to know under which conditions their current policies will be sufficient to maintain quality of service. They plan to use a simulator of their operations to answer this question.

Other applications include comparing the expected cost per period of several inventory policies, determining which allocations of buffer space in an assembly line meet performance requirements, and evaluating the improvement in the response time of a computer system if new hardware is added.

The most straightforward approach to 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, which allows us to study its solution using mathematical tools from stochastic control and dynamic programming. 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 MCC problems. Such Bayes-optimal 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 optimal policy is general. It allows two different methods for modeling the limited ability to sample: an explicit cost for each sample, and a random ceiling on the number of samples allowed. It also provides the ability to model

sampling distributions within any exponential family, which includes common families of sampling distributions like normal, Bernoulli, and Poisson.

A rich statistical literature surrounds the MCC problem, beginning with the construction of simultaneous confidence intervals in comparison to a fixed control by Dunnett (1955). For reviews of this previous literature, a number of books and survey papers are available: Hochberg and Tamhane (1987) and Hsu (1996) are general references on multiple comparisons; Goldsman and Nelson (1994) focuses on multiple comparison procedures in simulation; and Fu (1994) reviews multiple comparisons as they relate to the larger field of simulation optimization. While most of the work on multiple comparisons uses a frequentist analysis, Duncan (1965) gives the first Bayesian decision-theoretic formulation of the multiple comparison problem. Bayesian procedures for multiple comparisons are reviewed in Chapter 11 of Hochberg and Tamhane (1987). Most of this previous work, however, focuses on how the final determination should be made given a fixed sampling scheme (e.g., the construction of simultaneous confidence intervals), whereas we focus on designing fully sequential sampling schemes.

The literature on sampling schemes for MCC is smaller, but still substantial. Beyond one-stage procedures (e.g., Tukey (1953), Dunnett (1955)) that take a fixed number of samples from each alternative, and two-stage procedures (e.g., Dudewicz and Ramberg (1972), Dudewicz and Dalal (1983), Bofinger and Lewis (1992), Damerdji and Nakayama (1996), Nelson and Goldsman (2001), Yang and Nelson (1991)) that estimate nuisance parameters in a first stage and then perform a second stage that is similar in spirit to the single stage of a one-stage procedure, there are some fully sequential procedures. These include the stepwise multiple comparison significance tests proposed by Welsch (1977), the sequentially rejective type by Holm (1979) and the sequential multiple comparison with the best procedure by Hsu and Edwards (1983). Gopalan and Berry (1998) conduct sequential Bayesian multiple comparisons using Dirichlet process priors.

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

In its pursuit of Bayes-optimal fully sequential policies, the current work is related to Frazier et al. (2008), Chick and Gans (2009), and Chick and Frazier (2009), which attempt to achieve a similar feat in the related problem of sequential Bayesian ranking and selection. 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 ranking and selection policy using diffusion approximations. However, no efficient methods exist for computing the optimal sequential ranking and selection policy for more than a few alternatives. This is in contrast to the current work, in which we show that the optimal sequential MCC policy can be computed efficiently in general.

The ability shown in this paper to explicitly and efficiently compute the optimal policy also contrasts the MCC 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), sensor networks (Krause et al. 2008), and classification (Lizotte et al. 2003), 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. (2009).

We formulate the problem in Section 2. Section 3 then presents optimal policies for general sampling distributions, considering separately the case of an almost surely finite horizon with or without sampling costs (Section 3.1), and the case of an infinite horizon with sampling costs (Section 3.2). Then, in Sections 4 and 5, we specialize to two types of sampling: Bernoulli samples, and normal samples with known sampling variance. We give theoretical results particular to these more specialized cases, and provide techniques for computing the optimal policies efficiently and accurately. In Section 6 we demonstrate the resulting Bayes optimal algorithms on a collection of

illustrative example problems, and on a more realistic problem that classifies methods for positioning ambulances in a city according to whether they meet a minimum target for percentage of calls answered on time.

2. Problem Formulation

In this section we formulate the general Bayesian MCC problem, which allows both discounting of rewards in time and sampling costs. Suppose that 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, where Ξ is the parameter space. See DeGroot (1970) Chapter 9 for an in-depth treatment of exponential families and their use in Bayesian statistics. This assumption of an exponential family allows most common parametric sampling distributions, including the normal and Bernoulli distributions considered in detail in Sections 4 and 5, as well as Poisson, multinomial, and many others.

We are interested in finding 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 with parameter space Λ. (For example, in Section 4 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 $S_{01}, \ldots, S_{0k}$, each of which resides in Λ. Denote by $\mathbf{S}_0$ the vector composed of S_{0x} with x ranging from 1 to k.

Time is indexed by $n = 1, 2, \ldots$. At each time n we choose an alternative $x_n \in \{1, \ldots, k\}$ to sample from, and observe a corresponding sample y_n which has pdf or pmf $f(\cdot|\eta_{x_n})$. 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 $S_{n1}, \ldots, S_{nk}$, 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. Let τ be this number, and we require $\tau \geq 0$ to be a stopping time of the filtration, i.e., we require the event $\{\tau = n\}$ to be $\mathcal{F}_n$–measurable for each $n \geq 0$.

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 $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_n]$. For each $x = 1, \ldots, k$ define $\mu_{nx} = \mathbb{E}_n[\theta_x]$, which is fully determined by S_{nx} under our formulation. 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 MCC 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 $c_x \geq 0$ as 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 $\mathbb{E}T = \frac{1}{1-\alpha}$. 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 $1-\alpha$. 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.

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 number of alternatives correctly classified, given the available data after n samples. Formally, for all $n \geq 0$,

$$B_n = \underset{B\subseteq\{1,2,\ldots,k\},\, B\in\mathcal{F}_n}{\arg\max} \mathbb{E}_n \left[\sum_{x\in B} \mathbf{1}_{\{x\in\mathbb{B}\}} + \sum_{x\notin B} \mathbf{1}_{\{x\notin\mathbb{B}\}} \right].$$

PROPOSITION 1. *For $n \geq 0$ and $x = 1, \ldots, k$, define $P_{nx} = \mathbb{P}\{x \in \mathbb{B} \mid \mathcal{F}_n\} = \mathbb{P}\{\theta_x \geq d_x \mid \mathcal{F}_n\}$. Then $B_n = \{x : P_{nx} \geq 1/2\}$.*

By Proposition 1, our subjective belief on the objective set $\mathbb{B}$ is $B_{\tau\wedge T} = \{x : P_{\tau\wedge T, x} \geq 1/2\}$ when we stop. We then receive a reward equal to the total number of alternatives correctly classified at time $\tau \wedge T$, i.e., $\sum_{x\in B_{\tau\wedge T}} \mathbf{1}_{\{x\in\mathbb{B}\}} + \sum_{x\notin B_{\tau\wedge T}} \mathbf{1}_{\{x\notin\mathbb{B}\}}$. Our goal is to find a policy that maximizes the expected total reward, i.e., to solve the problem

$$\sup_{\pi} \mathbb{E}^{\pi} \left[\sum_{x\in B_{\tau\wedge T}} \mathbf{1}_{\{x\in\mathbb{B}\}} + \sum_{x\notin B_{\tau\wedge T}} \mathbf{1}_{\{x\notin\mathbb{B}\}} - \sum_{n=1}^{\tau\wedge T} c_{x_n} \right]. \tag{1}$$

3. The Optimal Solution

In this section we present the optimal solution to the Bayesian MCC problem (1), which allows both a geometrically distributed finite sampling horizon, and sampling costs. We first present some preliminary results, and then give solutions for a geometrically distributed horizon in Section 3.1, and for an infinite horizon with sampling costs in Section 3.2. The results in this section apply to the general sampling framework given in Section 2, and in later sections we specialize to sampling with Bernoulli observations (Section 4) and normal observations (Section 5).

We solve the problem (1) 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,

$$V(\mathbf{s}) = \sup_{\pi} \mathbb{E}^{\pi} \left[\sum_{x \in B_{\tau \wedge T}} \mathbf{1}_{\{x \in \mathbb{B}\}} + \sum_{x \notin B_{\tau \wedge T}} \mathbf{1}_{\{x \notin \mathbb{B}\}} - \sum_{n=1}^{\tau \wedge T} c_{x_n} \,\middle|\, \mathbf{S}_0 = \mathbf{s} \right]. \tag{2}$$

An optimal policy is any policy π attaining this supremum.

Before describing these optimal policies in Sections 3.1 and 3.2, 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 $h(u) = \max\{u, 1-u\}$. The value function is then simplified and transformed in the following proposition.

Proposition 2.

$$V(\mathbf{s}) = R_0 + \sup_{\pi} \mathbb{E}^{\pi} \left[\sum_{n=1}^{\tau} \alpha^{n-1} R_n \,\middle|\, \mathbf{S}_0 = \mathbf{s} \right], \tag{3}$$

where $R_0 = \sum_{x=1}^{k} h\left(P_{0x}(\mathbf{s})\right)$*;* $R_n = -c_{x_n} + h(P_{nx_n}) - h(P_{n-1,x_n})$*, for all* $n \geq 1$*.*

This proposition shows that a problem with stopping rule τ that provides a fixed initial reward R_0 and a discounted single period reward $\alpha^{n-1} R_n$ at each time $n \geq 1$ is equivalent to the original problem. Since R_0 does not affect the optimal policy, we may subtract it from the value function and instead think of V as the optimal expected incremental reward over R_0. This provides the equivalent problem,

$$V(\mathbf{s}) = \sup_{\pi} \mathbb{E}^{\pi} \left[\sum_{n=1}^{\tau} \alpha^{n-1} R_n \,\middle|\, \mathbf{S}_0 = \mathbf{s} \right], \tag{4}$$

where we have redefined V to correspond to this equivalent problem in a slight abuse of notation. A policy π that attains the supremum in (4) also attains the supremum in (3) and (2).

For later work, it is convenient to make one additional transformation in which we replace the random variables R_n in (4) with fixed functions of the alternatives' states. From (4) and the tower property,

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

For each $x = 1, \ldots, k$, define the reward function $\mathcal{R}_x : \Lambda \mapsto \mathbb{R}$ by

$$\mathcal{R}_x(s) = \mathbb{E}[R_1 \mid S_{0x} = s, x_1 = x] = -c_x + \mathbb{E}[h(P_{1x}) - h(P_{0x}) \mid S_{0x} = s, x_1 = x]. \tag{5}$$

Remark 1 in the e-companion shows that $\mathcal{R}_x(\cdot) \geq -c_x$, which is useful later. Because $\mathbb{E}^{\pi}[R_n|\mathbf{S}_{n-1}] = \mathcal{R}_{x_n}(S_{n-1,x_n})$, it follows that

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

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

$$V(\mathbf{s}) = \max \left[0, \max_x L_x(\mathbf{s}, V) \right], \ \forall \mathbf{s} \in \mathbb{S}, \tag{7}$$

where $L_x(\cdot, \cdot)$ is defined by

$$L_x(\mathbf{s}, V) = \mathcal{R}_x(s_x) + \alpha \cdot \mathbb{E}[V(\mathbf{S}_1) \mid \mathbf{S}_0 = \mathbf{s}, x_1 = x]. \tag{8}$$

Here, the value function V is not necessarily Borel-measurable, but is universally measurable, and so the expectation of V is taken in this more general sense (Dynkin and Yushkevich (1979)).

We now solve the MCC problem by solving Bellman's recursion, finding policies that attain the supremum in (6). In the following subsections we divide our assumption of a finite horizon ($\alpha < 1$) or a cost per sample ($c_x > 0$ for all x) into two distinct cases, and solve the MCC problem in each case using a distinct technique. The first (Section 3.1) assumes $\alpha < 1$ and allows $c_x = 0$ (geometric horizon with nonnegative sampling costs). The second (Section 3.2) assumes $\alpha = 1$ and $c_x > 0$ for all x (infinite horizon with strictly positive sampling costs).

3.1. Geometric Horizon With Nonnegative Sampling Costs

We first consider the MCC 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 for each possible state.

The Gittins index of alternative x at state s is defined as the maximum expected discounted reward per unit of expected discounted time, obtained by operating the single alternative x with initial state s, i.e.,

$$\nu_x(s) = \max_{\tau>0} \mathbb{E}\left[\frac{\sum_{n=1}^{\tau} \alpha^{n-1} \mathcal{R}_x(S_{n-1,x})}{\sum_{n=1}^{\tau} \alpha^{n-1}} \,\middle|\, S_{0x} = s, x_1 = \cdots = x_\tau = x \right] \tag{9}$$

Since $\mathcal{R}_x(\cdot) \geq -c_x$, it is clear that $\nu_x(\cdot) \geq -c_x$. These Gittins indices can then be computed using standard techniques (see, e.g., Varaiya et al. (1985)).

The optimal sampling rule, whose decisions we denote $(x_n^*)_{n\geq 1}$, is 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,

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

Computation of (9) is much easier than solving the full DP because the dimension of state $s \in \Lambda$ 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 (see, e.g., Powell (2007)). By using the MAB solution technique, we reduce the MCC problem from requiring the solution to a very large and often intractable DP with $|\Lambda|^k$ states to requiring the solution to a much smaller DP with only $|\Lambda|$ states. If Λ is continuous, then we must discretize or otherwise approximate the smaller DP to solve it, but the resulting approximate solution is still much easier to compute than would be a solution to the full DP over Λ^k, especially when k is large.

For the special case of $c_x = 0$ for all x, the reward functions $\mathcal{R}_x(\cdot)$ are always nonnegative by Remark 1. We may then choose τ^* to be $+\infty$.

3.2. Infinite Horizon With Strictly Positive Sampling Costs

We now consider the MCC 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 $x \in \{1, \ldots, k\}$ 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

$$V_x(s) = \sup_{\tau_x} \mathbb{E} \left[\sum_{n=1}^{\tau_x} \mathcal{R}_x(S_{n-1,x}) \,\middle|\, S_{0x} = s, x_1 = \cdots = x_{\tau_x} = x \right], \tag{10}$$

where τ_x is the stopping time for this sub-problem. We immediately have the following bounds on the value function.

PROPOSITION 3. $0 \le V_x(s) \le 1 - h\left(P_{0x}(s)\right)$.

Bellman's recursion (7) becomes

$$\begin{aligned} V_x(s) &= \max\left[0, L_x(s, V_x)\right], \quad \text{where} \\ L_x(s, V_x) &= \mathcal{R}_x(s) + \mathbb{E}[V_x(S_{1x}) \mid S_{0x} = s, x_1 = x]. \end{aligned} \tag{11}$$

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 $\Lambda \setminus \mathbb{C}_x$. These sets of states 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 (10) is the stopping time τ_x^* given by $\tau_x^* = \inf\{n \ge 0 : S_{nx} \notin \mathbb{C}_x\}$.

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, and considerable research effort has gone to creating and analyzing other policies with this property

(Siegmund 1985). However, we show later in Sections 4.1 and 5.1 that, for Bernoulli and normal sampling, τ_x^* has a known deterministic and finite upper bound.

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:*

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

For later use, we give the following proposition.

PROPOSITION 4. *Suppose that in each sub-problem with single alternative* x, τ_x^* *has a deterministic upper bound* N_x. *Then for any optimal policy with* $(x_n^*)_{n\geq 1}$ *and* τ^* *characterized in Theorem 1,* τ^* *has a deterministic upper bound* $\sum_{x=1}^{k} N_x$.

4. Specialization to Bernoulli Sampling

In this section we specialize the results of Section 3 to the specific case of Bernoulli sampling distributions. We give explicit expressions for quantities described generally in Section 3, and then present additional theoretical results and computational methods for Bernoulli sampling. Later, in Section 5, 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 sequence of sets we choose, and the reward function. 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 $n \geq 1$, $y_n \mid \theta, x_n \sim$ Bernoulli(θ_{x_n}).

We adopt a conjugate 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 $S_{nx} = (a_{nx}, b_{nx})$ 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 the vector of 0s with a single 1 at component x_n.

We know $\mu_{nx} = \mathbb{E}_n[\theta_x] = a_{nx}/(a_{nx} + b_{nx})$. Also,

$$P_{nx} = \mathbb{P}\{\theta_x \geq d_x \mid \theta_x \sim \text{Beta}(a_{nx}, b_{nx})\} = 1 - I_{d_x}(a_{nx}, b_{nx}),$$

where the regularized incomplete beta function $I_\cdot(\cdot, \cdot)$ is defined for $a, b > 0$ and $0 \leq d \leq 1$ by

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

Thus by Proposition 1, the sequence of sets $(B_n)_{n \geq 0}$ is given by $B_n = \{x : I_{d_x}(a_{nx}, b_{nx}) \leq 1/2\}$, for all $n \geq 0$. By Remark 2 in the e-companion and definition (5), we can write the reward function $\mathcal{R}_x(\cdot, \cdot)$ explicitly. For any $(a, b) \in \Lambda$,

$$\mathcal{R}_x(a, b) = -c_x - h\left(I_{d_x}(a, b)\right) + \frac{a}{a+b} \cdot h\left(I_{d_x}(a+1, b)\right) + \frac{b}{a+b} \cdot h\left(I_{d_x}(a, b+1)\right). \tag{12}$$

For later work, we give an upper bound in the following proposition.

PROPOSITION 5. $\mathcal{R}_x(a, b) \leq -c_x + 1/\sqrt{2\pi(a+b)}$.

4.1. Infinite Horizon with Sampling Costs

For the infinite horizon case with strictly positive sampling costs, we further characterize the optimal policy, which is established for general sampling distributions in Section 3.2. We show that with a Bernoulli sampling distribution, the optimal stopping rule is always finite with an explicit upper bound. This is computationally useful because it allows us to restrict the state space when solving for the optimal policy. It also contrasts our problem with other sequential information collection problems like sequential hypothesis testing (see, e.g., Siegmund (1985)) for which the optimal stopping rule has no fixed finite upper bound, even though sampling has a fixed cost.

In the sub-problem with a single alternative x, let $(a,b) \in \Lambda$ be some arbitrary state of alternative x. By Remark 2 in the e-companion,

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

Thus (11) becomes $V_x(a,b) = \max[0, L_x(a,b,V_x)]$, where

$$\begin{aligned} L_x(a,b,V_x) = -c_x - h\left(I_{d_x}(a,b)\right) &+ \frac{a}{a+b}\left[h\left(I_{d_x}(a+1,b)\right) + V_x(a+1,b)\right] \\ &+ \frac{b}{a+b}\left[h\left(I_{d_x}(a,b+1)\right) + V_x(a,b+1)\right]. \end{aligned} \tag{13}$$

Proposition 3 becomes $0 \le V_x(a,b) \le 1 - h\left(I_{d_x}(a,b)\right)$. Given $c_x > 0$, we also have

PROPOSITION 6. *If* $a+b \ge 1/(2\pi c_x^2)$, *then* $(a,b) \notin \mathbb{C}_x$.

The following proposition demonstrates that τ_x^* is always finite with a fixed upper bound.

PROPOSITION 7. $\tau_x^* \le \max\{0, \lceil 1/(2\pi c_x^2) - (a_{0x} + b_{0x})\rceil\}$.

Applying Proposition 7 and Proposition 4, the optimal stopping rule τ^* for the original problem, as characterized in Theorem 1, can then be bounded as follows.

COROLLARY 1. $\tau^* \le \sum_{x=1}^{k} N_x$, *where* $N_x := \max\{0, \lceil 1/(2\pi c_x^2) - (a_{0x} + b_{0x})\rceil\}$.

To compute the optimal policy given in Theorem 1, we evaluate the single-alternative value functions over the whole state space. Bellman's recursion enables us to obtain the values by an easy-to-compute recursion, and the upper bound on the optimal stopping time provides a reasonable initialization of the recursion.

Specifically, for each alternative x with initial state (a_{0x}, b_{0x}), we build a matrix to store the values of all the possible states in the sampling process, where the element at row n_a and column n_b is $V(a_{0x} + n_a - 1, b_{0x} + n_b - 1)$. By Propositon 6, for all elements (n_a, n_b) with $n_a + n_b \ge N_x$, we have $V(a_{0x} + n_a, b_{0x} + n_b) = 0$. Hence we only need to compute the values for $n_a + n_b = n < N_x$, which we can do using Bellman's recursion (13), starting with $n = N_x - 1$ and ending with $n = 0$.

If some N_x are large, it is reasonable to use a smaller bound N_0 instead of N_x to reduce computation. That is, we approximate $V_x(a,b)$ by 0 for $a + b \ge a_{0x} + b_{0x} + N_0$, and then approximate the values of the other states by Bellman's recursion. We use $N_0 = 1000$ in our numerical experiments.

4.2. Geometric Horizon

We now specialize the general solution for the geometric horizon case in Section 3.1 to Bernoulli sampling. We use a technique proposed by Varaiya et al. (1985) for off-line computation of the Gittins indices in a classical MAB problem, where the state-transitions of the alternatives form time-homogeneous finite-state Markov chains.

Since the state space is infinite in our problem, we use the following technique to apply this algorithm in an approximate sense. For each alternative x with initial state (a_{0x}, b_{0x}), we first assume that after we sample it N_0 (we use $N_0 = 50$ in the experiments) times, the incremental reward of continuing to sample it becomes 0. It follows that, for all $n_a + n_b > N_0$, $\nu_x(a_{0x} + n_a, b_{0x} + n_b) = 0$. We therefore only need to compute the Gittins indices for a finite set of states: $\{(a_{0x} + n_a, b_{0x} + n_b) : n_a, n_b \in \mathbb{N}, n_a + n_b \leq N_0, x = 1 \ldots, k\}$. The transition probability matrix is constructed by $\mathbb{P}[(a,b) \to (a+1,b)] = a/(a+b)$ and $\mathbb{P}[(a,b) \to (a,b+1)] = b/(a+b)$. We can then implement the procedure of Varaiya et al. (1985) and store the indices of all possible states in this finite state-space approximation before sampling begins. The optimal sampling policy is then easily implemented according to Section 3.1.

In our implementation, when an alternative is sampled more than N_0 times, we take the current state as the new initial state, and recompute the indices.

5. Specialization to Normal Sampling

We now consider normally distributed samples with known variance. As done for in Section 4 for Bernoulli samples, we give explicit expressions for the quantities described generally in Section 3 and then compute the optimal policy.

Here for each alternative x, the sampling precision is known and denoted by β_x^ϵ. The underlying mean of the sampling distribution, θ_x, is therefore the only component of η_x. For all $n \geq 1$, we have $y_n \mid \theta, x_n \sim \mathcal{N}(\theta_{x_n}, 1/\beta_{x_n}^\epsilon)$. The threshold parameters are $d_x \in (-\infty, +\infty)$.

We adopt a conjugate $\mathcal{N}(\mu_{0x}, 1/\beta_{0x})$ prior for each θ_x, under which θ_x is independent of $\theta_{x'}$ for $x \neq x'$. Our normal samples then result in a sequence of posterior distributions on θ_x, which

are again independently normally distributed with parameters $S_{nx} = (\mu_{nx}, \beta_{nx})$ 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 $n \geq 0$,

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

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 $n \geq 1$,

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

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

We know that

$$P_{nx} = \mathbb{P}\{\theta_x \geq d_x \mid \theta_x \sim \mathcal{N}(\mu_{nx}, 1/\beta_{nx})\} = 1 - \Phi\left(\sqrt{\beta_{nx}}(d_x - \mu_{nx})\right),$$

where Φ is the standard normal cdf. It is then clear that for all $n \geq 0$, $B_n = \{x : \mu_{nx} \geq d_x\}$. For any $(\mu, \beta) \in \Lambda$,

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

where Z is a standard normal random variable. Simple algebra shows that $\mathbb{E}\left[h\left(\Phi\left(\sqrt{\beta + \beta^{\epsilon}_{x}}(d_x - \mu - \tilde{\sigma}_x(\beta)Z)\right)\right)\right]$ can be computed using the cdf of the bivariate normal distribution, as described in Remark 3 in the e-companion.

For later work, we present the following property of $\mathcal{R}_x(\cdot, \cdot)$.

PROPOSITION 8. *Define* $\overline{\mathcal{R}}_x(\beta) = -c_x + (\sqrt{1 + \beta^{\epsilon}_{x}/\beta} - 1)/\sqrt{2\pi e} + \pi^{-1}\sqrt{\beta^{\epsilon}_{x}/\beta}$. *Then* $\mathcal{R}_x(\mu, \beta) \leq \overline{\mathcal{R}}_x(\beta)$, *for all* $\mu \in \mathbb{R}$. *Moreover, for any fixed* $\beta > 0$, $\mathcal{R}_x(\mu, \beta) \to -c_x$ *as* $\mu \to \pm\infty$.

5.1. Infinite Horizon with Sampling Costs

As we did for the Bernoulli sampling case in Section 4.1, we now describe methods for computing the optimal policy in the normal sampling case, and give theoretical results bounding the optimal stopping time.

Consider the single-alternative sub-problem for alternative x. For any $(\mu,\beta)\in\Lambda$, (11) becomes

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

Proposition 3 becomes

$$0 \le V_x(\mu,\beta) \le 1 - h\left(\Phi\left(\sqrt{\beta}(d_x-\mu)\right)\right). \tag{16}$$

Given $c_x > 0$, we also have the following results.

Proposition 9. *Define*

$$\overline{\beta}_x = \frac{\beta_x^\epsilon[(2\pi e)^{-1} - \pi^{-2}]}{-\pi^{-1}[(2\pi e)^{-\frac{1}{2}} + c_x] + (2\pi e)^{-\frac{1}{2}}\sqrt{\pi^{-2} + c_x^2 + 2c_x(2\pi e)^{-\frac{1}{2}}}}. \tag{17}$$

Then $\mathbb{R}\times[\overline{\beta}_x, +\infty] \subseteq \Lambda\setminus\mathbb{C}_x$.

Proposition 10. $\tau_x^* \le \max\left\{0, \lceil(\beta_x^\epsilon)^{-1}(\overline{\beta}_x - \beta_{0x})\rceil\right\}$.

Similar to the Bernoulli sampling case, we have an explicit upper bound for the optimal stopping rule of the original problem, τ^*, which is characterized in Theorem 1.

Corollary 2. $\tau^* \le \sum_{x=1}^k N_x$, *where* $N_x := \max\left\{0, \lceil(\beta_x^\epsilon)^{-1}(\overline{\beta}_x - \beta_{0x})\rceil\right\}$.

$\overline{\beta}_1,\ldots,\overline{\beta}_k$ provide fixed boundaries via Proposition 9 for the value of β within the continuation sets $\mathbb{C}_1,\ldots,\mathbb{C}_k$, which we call β-boundaries. Similarly, we have corresponding boundaries for the values of μ within the continuation sets, which we call μ-boundaries.

Proposition 11. *For each alternative x, there exist some fixed functions $\overline{\mu}_x(\cdot)$ and $\underline{\mu}_x(\cdot)$ such that for any $\beta>0$, $\mu\le\overline{\mu}_x(\beta)$ or $\mu\ge\underline{\mu}_x(\beta) \Rightarrow V_x(\mu,\beta)=0$. Thus for any given β, we have $[\overline{\mu}_x(\beta), \underline{\mu}_x(\beta)]$ as the μ-boundary of the continuation set $\mathbb{C}_x$.*

The computation of the optimal policy (given in Theorem 1) is more complicated than for Bernoulli sampling. To implement Bellman's recursion (15) directly, we need to evaluate the single-alternative value function over the whole continuous domain of μ_x, which is not possible. Thus we truncate and discretize the range of μ_x to evaluate it approximately.

Specifically, consider some alternative x with initial state (μ_{0x}, β_{0x}). Suppose $\beta_{0x} < \overline{\beta}_x$ (otherwise we should never sample from x). Then the recursion starts from $V_x(\mu, \beta_{0x} + N_x\beta_x^\epsilon) = 0$, for all $\mu \in \mathbb{R}$. Generally for any $\beta \in \{\beta_{0x} + n\beta_x^\epsilon : 0 \le n < N_x\}$, we compute the interval $[\overline{\mu}_x(\beta), \underline{\mu}_x(\beta)]$ following the proof of Proposition 11 in the e-companion, 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). We refer to these points as μ-knots. It follows that

$$
\begin{aligned}
&\mathbb{E}\left[V_x\left(\mu + \tilde{\sigma}_x(\beta)Z, \beta + \beta_x^\epsilon\right)\right] \\
\approx &\sum_i \left\{\mathbb{E}\left[V_x\left(\mu + \tilde{\sigma}_x(\beta)Z, \beta + \beta_x^\epsilon\right) \mid A^i\right] \cdot \mathbb{P}(A^i)\right\} + \mathbb{E}\left[V_x\left(\mu + \tilde{\sigma}_x(\beta)Z, \beta + \beta_x^\epsilon\right) \mid A\right] \cdot \mathbb{P}(A) \\
\approx &\sum_i \left\{V_x\left(\mu_x^i(\beta + \beta_x^\epsilon), \beta + \beta_x^\epsilon\right) \cdot \mathbb{P}(A^i)\right\} + 0 \cdot \mathbb{P}(A) \\
= &\sum_i \left\{V_x\left(\mu_x^i(\beta + \beta_x^\epsilon), \beta + \beta_x^\epsilon\right) \cdot \left[\Phi\left(\frac{\mu_x^i(\beta + \beta_x^\epsilon) + \delta/2 - \mu}{\tilde{\sigma}_x(\beta)}\right) - \Phi\left(\frac{\mu_x^i(\beta + \beta_x^\epsilon) + \delta/2 - \mu}{\tilde{\sigma}_x(\beta)}\right)\right]\right\},
\end{aligned}
$$

where Φ is the standard normal cdf, events $A^i = \{\mu + \tilde{\sigma}_x(\beta)Z \in [\mu_x^i(\beta + \beta_x^\epsilon) - \delta/2, \mu_x^i(\beta + \beta_x^\epsilon) + \delta/2]\}$, and $A = \left\{\mu + \tilde{\sigma}_x(\beta)Z \notin \left[\underline{\mu}_x(\beta + \beta_x^\epsilon), \overline{\mu}_x(\beta + \beta_x^\epsilon)\right]\right\}$. Bellman's recursion (15) can then be conducted successfully over β and these μ-knots, and we can obtain the values of these finitely many states.

Now suppose (μ, β) is some arbitrary state of alternative x, where μ is not a μ-knot. If $\mu \notin [\overline{\mu}_x(\beta), \underline{\mu}_x(\beta)]$, we set $V_x(\mu, \beta) = 0$; otherwise we set $V_x(\mu, \beta) = V_x(\mu_x^i(\beta), \beta)$, where $i = \arg\min_j\{|\mu - \mu_x^j(\beta)|\}$.

5.2. Geometric Horizon

Similar to Section 4.2, we achieve the off-line computation of the Gittins indices using the technique proposed by Varaiya et al. (1985). We first truncate the horizon for each alternative's sub-problem to N_0. With normal sampling, this is not yet enough to provide a finite set of states: although β_x is discrete, μ_x takes continuous values. Hence we truncate and discretize the range of μ_x. The following proposition gives an upper bound on the Gittins indices, that is useful for truncation.

PROPOSITION 12. *Suppose* (μ, β) *is some arbitrary state of alternative* x*. Then*

$$\nu_x(\mu,\beta) \le -c_x + 1 - h\left(\Phi\left(\sqrt{\beta}(d_x-\mu)\right)\right).$$

For each fixed $\beta \in \{\beta_{0x} + n\beta_x^\epsilon : 0 \le n \le N_0\}$, since $h\left(\Phi\left(\sqrt{\beta}(d_x-\mu)\right)\right)$ in Proposition 12 increases to 1 as $\mu \to +\infty$ or as $\mu \to -\infty$, we bound the range of μ for consideration as follows. Let $\epsilon > 0$ be small (we take $\epsilon = 0.01$ in our numerical experiments). Then there exists an interval $[\bar{\mu}_x(\beta), \underline{\mu}_x(\beta)]$, given by the bound in Proposition 12, such that for all $\mu \notin [\bar{\mu}_x(\beta), \underline{\mu}_x(\beta)]$, we have $\nu_x(\mu,\beta) < -c_x + \epsilon$. Since $\nu(\cdot,\cdot) \ge -c_x$, we approximate $\nu_x(\mu,\beta)$ as $-c_x$ for $\mu \notin [\bar{\mu}_x(\beta), \underline{\mu}_x(\beta)]$. We then discretize this interval into μ-knots with interval δ, denoted by $\{\mu_x^i(\beta)\}_i$.

We now concentrate on the computation of the Gittins indices for the finite set of states, $\{(\mu_x^i(\beta),\beta) : \beta \in \{\beta_{0x} + n\beta_x^\epsilon : 0 \le n \le N_0\}, x = 1,\dots,k\}$. To construct the transition probability matrix, we use (14) and approximate the probabilities by the density ratios:

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

where φ is the pdf of $\mathcal{N}(0,1)$. We then implement the procedure of Varaiya et al. (1985) and obtain the indices for these states. For an arbitrary state (μ,β) of alternative x, if $\mu \notin [\bar{\mu}_x(\beta), \underline{\mu}_x(\beta)]$, we set $\nu_x(\mu,\beta) = -c_x$. Otherwise we set $\nu_x(\mu,\beta) = \nu_x(\mu_x^i(\beta),\beta)$, where $i = \arg\min_j\{|\mu - \mu_x^j(\beta)|\}$.

With these modifications to our problem, we can implement the optimal policy by computing and storing the indices of all possible states before sampling begins. 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. Numerical Results

In this section, we test the performance of the Bayes-optimal policies developed in Sections 4 and 5 for Bernoulli and normal sampling with a collection of numerical experiments. We first present illustrative example problems in Section 6.1, and then present a more realistic application to ambulance quality of service in Section 6.2.

6.1. Illustrative Example Problems

We first explore the performance of the optimal policy relative to other policies in several illustrative example problems. We introduce four policies for comparison.

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

2. **Max Variance (MV):** In the max variance policy, we sample from the alternative whose θ_x has the largest variance under the posterior. Hence $x_{n+1} \in \arg\max_x \{\sigma_{nx}\}$ for all $n \geq 0$, where σ_{nx} is the standard deviation of our belief on θ_x at time n. When the prior is homogeneous, this policy is equivalent to the **equal allocation** or round robin policy that samples each alternative an equal number of times.

3. **Upper Confidence Bound (UCB):** In the UCB policy (see, e.g., Chang et al. (2007)), sampling decisions are given by $x_{n+1} \in \arg\max_x \{\mu_{nx} + z \cdot \sigma_{nx}\}$, for all $n \geq 0$. $z \geq 0$ is a tunable parameter of the policy. In our experiments, when tuning over z, we found that the best performance was obtained by setting z very large. This is functionally equivalent to the max variance policy, and so we only display results for the max variance policy.

4. **Knowledge Gradient (KG):** In the KG policy, the sampling decision is the one that would be optimal if only one measurement were to remain, i.e., $x_{n+1} \in \arg\max_x \{\mathcal{R}_x(S_{nx})\}$ 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, the KG policy decides to stop when the net single-step expected reward from sampling becomes non-positive. This stopping rule is $\tau = \inf\{n : \mathcal{R}_x(S_{nx}) \leq 0, \forall x\}$. An analogous stopping rule for ranking and selection was introduced in Frazier and Powell (2008). While the PE and MV policies are simple policies, the KG policy represents an additional level of sophistication, and we see below that it performs quite well in some situations.

In each of the first three policies (PE, MV, UCB), only a sampling rule is specified. For these policies, we use a deterministic stopping rule, in which τ is a fixed parameter of the policy. In

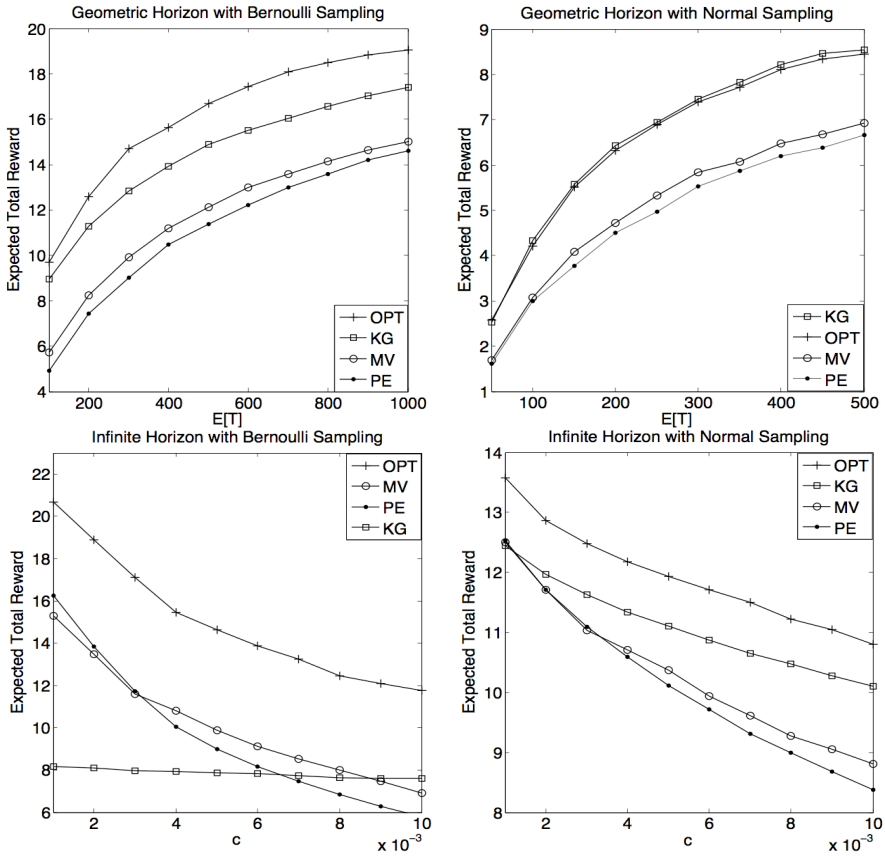

Figure 1 Performance of four policies: pure exploration (PE), max variance (MV), knowledge gradient (KG), and optimal (OPT). The maximum length of the 95%-confidence intervals for the values in each plot is respectively 0.21, 0.16, 0.11 and 0.10. In the normal sampling case, OPT is approximately optimal because of the discretization and truncation used to compute the policy.

problems with strictly positive sampling costs we then optimize over the value of τ using a simple grid search, and report performance with this best deterministic value of τ. We find that the best deterministic value of τ is usually near the expected stopping time under the optimal policy. We set $\tau = \infty$ in problems with no sampling costs. In the KG policy, both sampling and stopping rules are specified above. In all of these policies, ties in the arg max are broken uniformly at random.

Figure 1 shows the relative performance of these policies and the optimal policy on four different problem settings: geometric horizon with Bernoulli sampling; geometric horizon with normal sampling; infinite horizon with Bernoulli sampling; and infinite horizon with normal sampling. The parameters used are described as follows. For Bernoulli sampling, we test $k = 100$ alternatives with θ_x and d_x randomly generated according to the uniform $[0, 1]$ distribution ($a_{0x} = b_{0x} = 1$ for all x).

For normal sampling, we test $k = 50$ alternatives with θ_x and d_x randomly generated according to prior distribution $\mathcal{N}(0, 100)$ ($(\mu_{0x}, \beta_{0x}) = (0, 0.01)$ for all x); the sampling precisions β_x^ϵ are chosen uniformly from $[0.5, 2]$. For the geometric horizon, we adopt zero sampling cost ($c_x = 0$ for all x). We vary the discounting factor α such that the expected stopping time $\mathbb{E}[T] = \frac{1}{1-\alpha}$ varies within $[100, 1500]$ in the Bernoulli sampling case, and $[50, 500]$ in the normal sampling case. The actual number of samples T in the experiments are randomly generated according to the corresponding geometric distribution with parameter $1 - \alpha$. For the infinite horizon, we adopt an identical sampling cost c for each alternative and vary c within $[0.001, 0.01]$.

In all four problem settings, we see that the optimal policy significantly outperforms the PE and MV policies. Because these alternative strategies are the ones most commonly used in practice (recall that MV is the same as equal allocation when the prior is homogeneous), we see that practitioners can make substantial gains in performance by using the optimal policy over other suboptimal naive strategies.

The performance of the KG policy depends greatly on the problem. It is almost optimal in the geometric-horizon normal-sampling setting, while in the other three 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.

To understand the behavior of KG, first consider the two problem settings with normal sampling. KG makes its decisions using the one-step approximation $\mathcal{R}_x(S_{nx})$ 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 often leaves the performance of allocation decisions relatively unaffected because the level of underestimation is relatively constant

across alternatives, and the alternative with the largest one-step VOI tends to also have a 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 problem setting, where there are no sampling costs and the only decisions concern allocation. Indeed, we find that KG outperforms our (approximate) implementation of the optimal policy by a small but statistically significant margin for large $\mathbb{E}[T]$, because of numerical inaccuracies when discretizing the continuous state-space.

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 about whether to place the 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 strictly positive sampling costs, or allocates its sample randomly (an inefficient strategy) among the alternatives if sampling costs are 0. For this reason, KG performs poorly in both settings with Bernoulli sampling.

6.2. Ambulance Quality of Service Application

To demonstrate the fully sequential Bayes-optimal policy 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 MCC problem.

Each alternative x is a pair of the per-hour call arrival rate λ_x and the ambulance allocation plan $\mathcal{A}_x$. A sample from alternative x is the number of calls answered on time during a two-week period given $(\lambda_x, \mathcal{A}_x)$, and θ_x is the underlying mean of this sampling distribution, which is unknown. The emergency calls are generated according to a Poisson process with hourly rate λ_x, so the expected total number of calls during two weeks for each alternative $x = (\lambda_x, \mathcal{A}_x)$ is known analytically as $m_x = 24 \times 14 \times \lambda_x$. The long-term percentage of calls answered on time is then θ_x/m_x. 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 $d_x = 0.7 \times m_x = 0.7 \times 24 \times 14 \times \lambda_x$. To select the alternatives for our experiment, we chose 25 allocation plans and 25 values for the hourly call arrival rate from $[3, 6.6]$ with equal distance 0.15 between them. 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 ambulance allocation plans and call arrival rates. 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). Calculating the optimal policy under this new statistical model would then require further work.

We use independent normal priors for 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 requiring the policy to begin by taking a

single sample from each alternative. We then followed one of several different sampling policies. Their resulting performance is measured by the similarity between the set $\mathbb{B}$ and its estimates $(B_n)_{n\geq 0}$, where as before $B_n = \{x : \mu_{nx} \geq d_x\}$.

To support this measure of similarity, we independently estimated $\mathbb{B}$ through exhaustive simulation. We sampled each alternative 1000 times and used the sample mean as the estimate of θ_x. These estimates provide us the underlying boundary of $\mathbb{B}$ used below. We also sorted the 25 ambulance allocation used to construct the alternatives, in order of decreasing θ_x (at a fixed value of the call rate) to make the set $\mathbb{B}$ easier to visualize.

Figure 2 compares the optimal policy against three other policies: PE, MV, and KG. We assume a geometric horizon with no sampling costs. (We set $\alpha = 0.999$ in the optimal policy.) For each policy and after each of 500, 1000, 2000, and 5000 samples we plot the current estimate B_n as the light region. Each panel also shows a black line, which is the independently obtained high-accuracy estimate of the boundary between $\mathbb{B}$ and its complement. As previously described, the ambulance allocation plans have been sorted so that this boundary is a decreasing line.

Figure 2 shows that PE and MV behave poorly in distinguishing among the alternatives. Under these policies, 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. In contrast, the KG and optimal policies are much more efficient. The excellent performance of the KG policy relative to optimal should not be surprising: samples are normally distributed and there are no sampling costs, which is the setting from Section 6.1 in which KG was nearly optimal. Had the problem used Bernoulli sampling or sampling costs, then KG likely would not have performed as well.

As shown in Figure 2, after 5000 samples under KG or the optimal policy, we have estimated the set $\mathbb{B}$ 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 MV. 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 MCC.

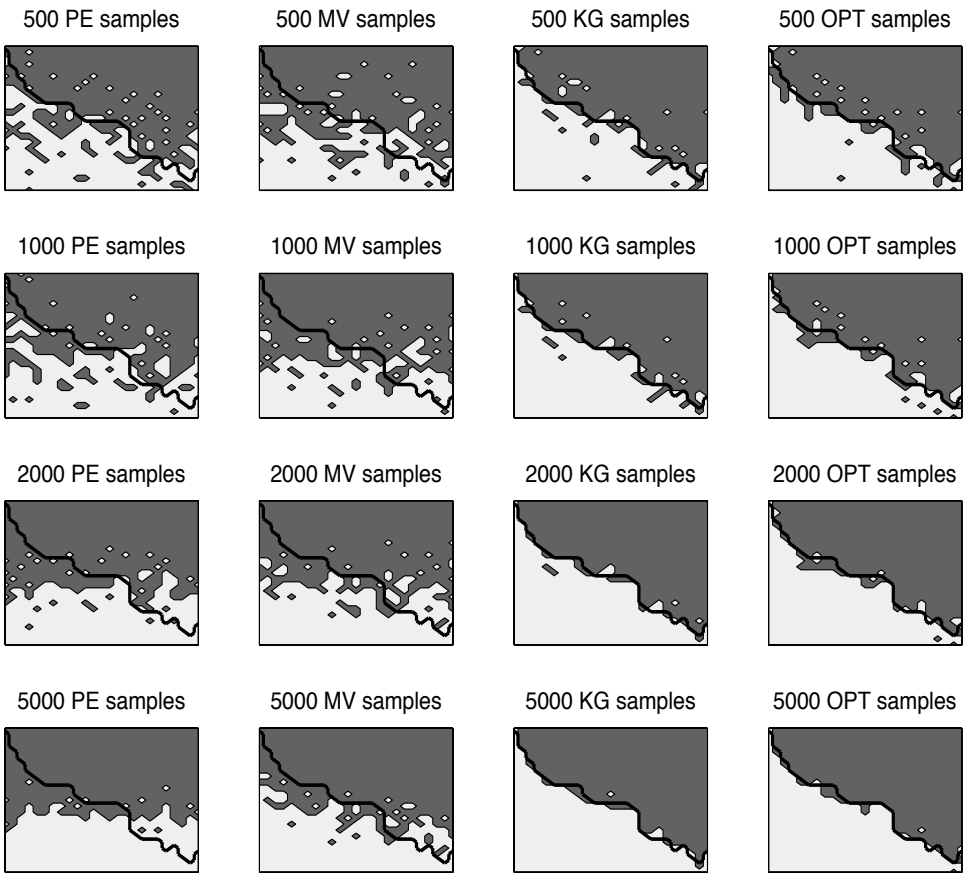

Figure 2 Performance of four policies in the ambulance service quality application: pure exploration (PE), max variance (MV), knowledge gradient (KG), and optimal (OPT). In each plot, the black curve is the boundary of the set $\mathbb{B}$ (the set of alternatives answering at least 70% of the emergency calls on time); the light region is the estimate of the set $\mathbb{B}$ under the corresponding sampling policy given the stated number of samples; and the dark region is the complement of the light region. The hourly call arrival rates $(3, 3.15, 3.3, \ldots, 6.6)$ are distributed along the y-axis. Ambulance allocation plans are distributed along the x-axis. There are 25 allocation plans, and they are sorted to make the black line decreasing.

7. Conclusions

By applying methods from multi-armed bandits and optimal stopping, we are able to efficiently solve the dynamic program for the sequential Bayesian MCC 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 Bayesian experimental design and effort allocation for simulation, this goal has remained elusive in many problems, and so the results in this paper place the MCC problem together with a select group of problems in sequential experimental design

for which the fully sequential Bayes-optimal policy 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 stopping time or sampling costs or both, and allowing sampling distributions from any exponential family. 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. For practitioners facing MCC problems of this type, the KG policy is extremely easy to use and can be used as a high-performance alternative to the Bayes-optimal policy.

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

References

Abramowitz, M., I.A. Stegun. 1964. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications.

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

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.

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.

Chang, H.S., M.C. Fu, J. Hu, S.I. Marcus. 2007. *Simulation-Based Algorithms for Markov Decision Processes*. Springer, Berlin.

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.

Duncan, D.B. 1965. A Bayesian approach to multiple comparisons. *Technometrics* **7**(2) 171–222.

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.

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*.

Gopalan, R., D.A. Berry. 1998. Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association* 1130–1139.

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.

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

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**(2) 65–70.

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

Hsu, J.C., D.G. Edwards. 1983. Sequential multiple comparisons with the best. *Journal of the American Statistical Association* **78**(384) 958–964.

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.

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.

Lizotte, D.J., O. Madani, R. Greiner. 2003. Budgeted learning of naıve-bayes classifiers. *Uncertainty in Artificial Intelligence*, vol. 3. 378–385.

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.

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. 2009. The knowledge gradient algorithm for a general class of online learning problems. Submitted.

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

Sloane, N. 2007. The on-line encyclopedia of integer sequences. *Towards Mechanized Mathematical Assistants* 130–130.

Tukey, J.W. 1953. Multiple comparisons. *Journal of the American Statistical Association* **48** 624–5.

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.

Welsch, R.E. 1977. Stepwise multiple comparison procedures. *Journal of the American Statistical Association* **72**(359) 566–575.

Yang, W.N., B.L. Nelson. 1991. Using common random numbers and control variates in multiple-comparison procedures. *Operations Research* **39**(4) 583–591.

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Mathematical Proofs.

EC.1. Proof of Proposition 1.

PROPOSITION 1. *For $n \geq 0$ and $x = 1, \ldots, k$, define $P_{nx} = \mathbb{P}\{x \in \mathbb{B} \mid \mathcal{F}_n\} = \mathbb{P}\{\theta_x \geq d_x \mid \mathcal{F}_n\}$. Then $B_n = \{x : P_{nx} \geq 1/2\}$.*

Proof of Proposition 1. For any $B \in \mathcal{F}_n$, we write

$$\mathbb{E}_n\left[\sum_{x\in B} \mathbf{1}_{\{x\in\mathbb{B}\}} + \sum_{x\notin B} \mathbf{1}_{\{x\notin\mathbb{B}\}}\right] = \sum_{x\in B} \mathbb{E}_n\left[\mathbf{1}_{\{x\in\mathbb{B}\}}\right] + \sum_{x\notin B} \mathbb{E}_n\left[\mathbf{1}_{\{x\notin\mathbb{B}\}}\right] = \sum_{x\in B} P_{nx} + \sum_{x\notin B}(1 - P_{nx}).$$

It follows that

$$B_n = \arg\max_{B\in\mathcal{F}_n} \mathbb{E}_n\left[\sum_{x\in B} \mathbf{1}_{\{x\in\mathbb{B}\}} + \sum_{x\notin B} \mathbf{1}_{\{x\notin\mathbb{B}\}}\right] = \{x : P_{nx} \geq 1/2\}.$$

EC.2. Proof of Proposition 2.

PROPOSITION 2.

$$V(\mathbf{s}) = R_0 + \sup_{\pi} \mathbb{E}^{\pi}\left[\sum_{n=1}^{\tau} \alpha^{n-1} R_n \,\middle|\, \mathbf{S}_0 = \mathbf{s}\right], \tag{EC.1}$$

where $R_0 = \sum_{x=1}^{k} h\left(P_{0x}(\mathbf{s})\right)$; $R_n = -c_{x_n} + h(P_{nx_n}) - h(P_{n-1,x_n})$, for all $n \geq 1$.

Proof of Proposition 2. Using the proof of Proposition 1, we know for all $n \geq 0$,

$$\mathbb{E}_n\left[\sum_{x\in B_n} \mathbf{1}_{\{x\in\mathbb{B}\}} + \sum_{x\notin B_n} \mathbf{1}_{\{x\notin\mathbb{B}\}}\right] = \sum_{x=1}^{k} h(P_{nx}).$$

Hence by the tower property of conditional expectation,

$$\mathbb{E}^{\pi}\left[\sum_{x\in B_{\tau\wedge T}} \mathbf{1}_{\{x\in\mathbb{B}\}} + \sum_{x\notin B_{\tau\wedge T}} \mathbf{1}_{\{x\notin\mathbb{B}\}}\right] = \mathbb{E}^{\pi}\left[\mathbb{E}^{\pi}_{\tau\wedge T}\left[\sum_{x\in B_{\tau\wedge T}} \mathbf{1}_{\{x\in\mathbb{B}\}} + \sum_{x\notin B_{\tau\wedge T}} \mathbf{1}_{\{x\notin\mathbb{B}\}}\right]\right] = \mathbb{E}^{\pi}\left[\sum_{x=1}^{k} h(P_{\tau\wedge T,x})\right].$$

We can therefore write (2) as follows,

$$V(\mathbf{s}) = \sup_{\pi} \mathbb{E}^{\pi}\left[\sum_{x=1}^{k} h(P_{\tau\wedge T,x}) - \sum_{n=1}^{\tau\wedge T} c_{x_n} \,\middle|\, \mathbf{S}_0 = \mathbf{s}\right].$$

We restructure this into a sequence of single period rewards. For a fixed policy π and hence a stopping rule τ, since T is independent of the sampling filtration,

$$
\begin{aligned}
&\mathbb{E}^\pi\left[\sum_{x=1}^{k} h(P_{\tau\wedge T,x}) - \sum_{n=1}^{\tau\wedge T} c_{x_n}\right] \\
&= \mathbb{E}^\pi\left[\sum_{t=1}^{\tau}\left[(1-\alpha)\alpha^{t-1}\left(\sum_{x=1}^{k} h(P_{tx}) - \sum_{n=1}^{t} c_{x_n}\right)\right] + \alpha^\tau\left(\sum_{x=1}^{k} h(P_{\tau x}) - \sum_{n=1}^{\tau} c_{x_n}\right)\right] \\
&= \mathbb{E}^\pi\left[\sum_{x=1}^{k}\left[\sum_{t=1}^{\tau}[(1-\alpha)\alpha^{t-1}h(P_{tx})] + \alpha^\tau h(P_{\tau x})\right] - \sum_{t=1}^{\tau}\left[(1-\alpha)\alpha^{t-1}\sum_{n=1}^{t} c_{x_n}\right] - \alpha^\tau\sum_{n=1}^{\tau} c_{x_n}\right] \\
&= \mathbb{E}^\pi\left[\sum_{x=1}^{k}\left[h(P_{0x}) + \sum_{t=1}^{\tau}\alpha^{t-1}[h(P_{tx}) - h(P_{t-1,x})]\right] - (1-\alpha)\sum_{n=1}^{\tau}\left[c_{x_n}\sum_{t=n}^{\tau}\alpha^{t-1}\right] - \alpha^\tau\sum_{n=1}^{\tau} c_{x_n}\right] \\
&= \mathbb{E}^\pi\left[\sum_{x=1}^{k} h(P_{0x}) + \sum_{t=1}^{\tau}\alpha^{t-1}[h(P_{tx_t}) - h(P_{t-1,x_t})] - \sum_{t=1}^{\tau}\alpha^{t-1}c_{x_t}\right] \\
&= \mathbb{E}^\pi\left[\sum_{x=1}^{k} h(P_{0x}) + \sum_{t=1}^{\tau}\alpha^{t-1}[-c_{x_t} + h(P_{tx_t}) - h(P_{t-1,x_t})]\right].
\end{aligned}
$$

The second to last equation follows from simple computation and the fact that at each time $n = 1, 2, \ldots, \tau$, for all non-selected alternatives $x \neq x_n$, we have $s_{nx} = s_{n-1,x}$ and hence $P_{nx} = P_{n-1,x}$.

Finally, for each x, P_{0x} is a function of the initial state $\mathbf{s}$ and is fully determined by $\mathbf{s}$, thus

$$
V(\mathbf{s}) = \sum_{x=1}^{k} h\left(P_{0x}(\mathbf{s})\right) + \sup_{\pi}\mathbb{E}^\pi\left[\sum_{n=1}^{\tau}\alpha^{n-1}[-c_{x_n} + h(P_{nx_n}) - h(P_{n-1,x_n})] \,\middle|\, \mathbf{S}_0 = \mathbf{s}\right],
$$

which is (EC.1).

EC.3. Remark 1.

We show that for each x, $\mathcal{R}_x(\cdot) \geq -c_x$. Note that $h(\cdot)$ is a convex function and that $\mathbb{E}_0\left[P_{1x_1}\right] = \mathbb{E}_0\left[\mathbb{E}_1\left[\mathbf{1}_{\{x_1\in\mathbb{B}\}}\right]\right] = \mathbb{E}_0\left[\mathbf{1}_{\{x_1\in\mathbb{B}\}}\right] = P_{0x_1}$. Hence by Jensen's inequality, we have $\mathbb{E}_0[R_1] = -c_{x_1} + \mathbb{E}_0[h(P_{1x_1})] - h(P_{0x_1}) \geq -c_{x_1}$.

EC.4. Proof of Proposition 3.

Proposition 3. $0 \leq V_x(s) \leq 1 - h\left(P_{0x}(s)\right)$.

Proof of Proposition 3. We receive a zero reward if we take $\tau_x = 0$. Thus $V_x(\cdot) \geq 0$. By (5), we can write (10) as $V_x(s) = \sup_{\tau_x}\left\{-c_x\tau_x - h\left(P_{0x}(s)\right) + \mathbb{E}[h(P_{\tau_x x}) \mid S_{0x} = s, x_1 = \cdots = x_{\tau_x} = x]\right\}$. Note that $\tau_x \geq 0$ and $h(\cdot) \leq 1$. It follows that $V_x(s) \leq 1 - h\left(P_{0x}(s)\right)$.

EC.5. Proof of Theorem 1.

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:*

$$x_{n+1}^* \in \{x : S_{nx} \in \mathbb{C}_x\},\ \forall n \geq 0; \qquad \tau^* = \inf\{n \geq 0 : S_{nx} \notin \mathbb{C}_x, \forall x\}. \tag{EC.2}$$

Proof of Theorem 1. For any arbitrary policy π with stopping time τ, we denote the number of times we sample from each alternative x by m_x. Then $\tau = \sum_{x=1}^{k} m_x$. Denote the collection of times when we sample from x, $\{1 \leq n \leq \tau : x_n = x\}$, by $\{n_i^x\}_{1\leq i\leq m_x}$.

Since the reward for each period only depends on the alternative being sampled during that period, and the states of all the other alternatives remain frozen, we know that the order of the sequence of sampling decisions does not affect the expected total reward. Hence the original problem can be naturally decomposed into k sub-problems as follows.

$$\begin{aligned}
&\mathbb{E}^\pi\left[\sum_{n=1}^{\tau}\mathcal{R}_{x_n}(S_{n-1,x_n})\,\middle|\,\mathbf{S}_0=\mathbf{s}\right] = \sum_{x=1}^{k}\left\{\mathbb{E}^\pi\left[\sum_{i=1}^{m_x}\mathcal{R}_x(S_{n_i^x-1,x})\,\middle|\,S_{0x}=s_x\right]\right\} \\
&= \sum_{x=1}^{k}\left\{\mathbb{E}^\pi\left[\sum_{n=1}^{m_x}\mathcal{R}_x(S_{n-1,x})\,\middle|\,S_{0x}=s_x, x_1=\cdots=x_{m_x}=x\right]\right\} \leq \sum_{x=1}^{k} V_x(s_x),
\end{aligned} \tag{EC.3}$$

where the last inequality follows from (10). Thus

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

On the other hand, if we adopt a policy satisfying $x_{n+1} \in \{x : S_{nx} \in \mathbb{C}_x\}$ for all $n \geq 0$ and $\tau = \inf\{n \geq 0 : S_{nx} \notin \mathbb{C}_x, \forall x\}$, then for each x, m_x is exactly the number of samples from alternative x needed for the state of x to leave $\mathbb{C}_x$ for the first time. Hence in each decomposed sub-problem with single alternative x, m_x is an optimal solution equivalent to τ_x^*. As a result,

$$\mathbb{E}^\pi\left[\sum_{n=1}^{m_x}\mathcal{R}_x(S_{n-1,x})\,\middle|\,S_{0x}=s_x, x_1=\cdots=x_{m_x}=x\right] = V_x(s_x),$$

the inequality in (EC.3) becomes equality, and $V(\mathbf{s}) = \sum_{x=1}^{k} V_x(s_x)$. This also shows that any policy satisfying the conditions (EC.2) is optimal.

EC.6. Proof of Proposition 4.

PROPOSITION 4. *Suppose that in each sub-problem with single alternative x, τ_x^* has a deterministic upper bound N_x. Then for any optimal policy with $(x_n^*)_{n\geq 1}$ and τ^* characterized in Theorem 1, τ^* has a deterministic upper bound $\sum_{x=1}^k N_x$.*

Proof of Proposition 4. We apply ideas similar to those in the proof of Theorem 1. First, $\tau^* = \sum_{x=1}^k m_x$. Under the optimal policy, since m_x is the number of samples from alternative x needed for the state of x to leave $\mathbb{C}_x$ for the first time, its distribution is the same as the distribution of τ_x in the decomposed sub-problem with single alternative x. It follows that $m_x \leq N_x$, for each x. Thus $\tau^* \leq \sum_{x=1}^k N_x$.

EC.7. Remark 2.

Consider the distribution of $\mathbf{S}_1$ given $\mathbf{S}_0 = (\mathbf{a}, \mathbf{b})$ and $x_1 = x$. Since $\mathbb{P}_0\{y_1 = 1\} = \mathbb{E}_0[y_1] = \mathbb{E}_0[\mathbb{E}_0[y_1|\theta]] = \mathbb{E}_0[\theta_{x_1}] = \mu_{0x_1}$, we immediately have the following expressions:

$$\mathbb{P}\{\mathbf{S}_1 = (\mathbf{a}+\mathbf{e}_x, \mathbf{b}) \mid \mathbf{S}_0 = (\mathbf{a}, \mathbf{b}), x_1 = x\} = \mathbb{P}\{y_1 = 1 \mid \mathbf{S}_0 = (\mathbf{a}, \mathbf{b}), x_1 = x\} = a_x/(a_x + b_x),$$

$$\mathbb{P}\{\mathbf{S}_1 = (\mathbf{a}, \mathbf{b}+\mathbf{e}_x) \mid \mathbf{S}_0 = (\mathbf{a}, \mathbf{b}), x_1 = x\} = \mathbb{P}\{y_1 = 0 \mid \mathbf{S}_0 = (\mathbf{a}, \mathbf{b}), x_1 = x\} = b_x/(a_x + b_x).$$

EC.8. Proof of Proposition 5.

PROPOSITION 5. $\mathcal{R}_x(a, b) \leq -c_x + 1/\sqrt{2\pi(a+b)}$.

EC.8.1. Preparatory Material

Use Stirling's approximation, for large a and b,

$$B(a, b) \sim \sqrt{2\pi}\frac{a^{a-\frac{1}{2}}b^{b-\frac{1}{2}}}{(a+b)^{a+b-\frac{1}{2}}}.$$

More generally, we have the following lemma.

LEMMA EC.1. *For $a, b \geq 1$,*

$$B(a, b) \geq \sqrt{2\pi}\frac{a^{a-\frac{1}{2}}b^{b-\frac{1}{2}}}{(a+b)^{a+b-\frac{1}{2}}}.$$

Proof of Lemma EC.1. By Stirling's asymptotic series (see, e.g., Abramowitz and Stegun (1964) and Sloane (2007)), we write

$$\Gamma(z) = e^{-z} z^{z-\frac{1}{2}} \sqrt{2\pi} e^{\lambda_z}, \quad \text{with } \frac{1}{12z+1} < \lambda_z < \frac{1}{12z}.$$

Hence

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \sqrt{2\pi} \exp\{\lambda_a + \lambda_b - \lambda_{a+b}\} \frac{a^{a-\frac{1}{2}} b^{b-\frac{1}{2}}}{(a+b)^{a+b-\frac{1}{2}}}.$$

The result then follows since for $a, b \geq 1$,

$$\lambda_a + \lambda_b - \lambda_{a+b} > \frac{1}{12a+1} + \frac{1}{12b+1} - \frac{1}{12(a+b)} = \frac{(12a+\frac{1}{2})^2 + (12b+\frac{1}{2})^2 + 144ab - \frac{3}{2}}{12(a+b)(12a+1)(12b+1)} > 0.$$

LEMMA EC.2. *For $a, b \geq 1$ and $d_x \in (0,1)$,*

$$\frac{d_x^a (1-d_x)^b}{B(a,b)} \leq \frac{1}{2} \sqrt{\frac{a+b}{2\pi}}.$$

Proof of Lemma EC.2. Denote $t = a+b$ and $\mu = \frac{a}{a+b}$. Then by Lemma EC.1,

$$\frac{d_x^a (1-d_x)^b}{B(a,b)} \leq \frac{d_x^{\mu t} (1-d_x)^{(1-\mu)t} t^{t-\frac{1}{2}}}{\sqrt{2\pi} (\mu t)^{\mu t - \frac{1}{2}} [(1-\mu)t]^{(1-\mu)t - \frac{1}{2}}} = \sqrt{\frac{\mu(1-\mu)t}{2\pi}} \left[\left(\frac{d_x}{\mu} \right)^{\mu} \left(\frac{1-d_x}{1-\mu} \right)^{1-\mu} \right]^t.$$

Define function $g(\cdot)$ on $(0,1)$ by

$$g(u) = \left(\frac{d_x}{u} \right)^u \left(\frac{1-d_x}{1-u} \right)^{1-u}.$$

Then since $\frac{d}{du}[\log g(u)] = \log \frac{d_x(1-u)}{u(1-d_x)}$, it is easy to see that $g(\cdot)$ is unimodal on $(0,1)$ with peak at $u = d_x$. Finally, since $\mu \in (0,1)$, we know $\sqrt{\mu(1-\mu)} \leq \frac{1}{2}$, and hence

$$\frac{d_x^a (1-d_x)^b}{B(a,b)} \leq \frac{1}{2} \sqrt{\frac{t}{2\pi}} [g(\mu)]^t \leq \frac{1}{2} \sqrt{\frac{t}{2\pi}} [g(d_x)]^t = \frac{1}{2} \sqrt{\frac{t}{2\pi}} = \frac{1}{2} \sqrt{\frac{a+b}{2\pi}}.$$

EC.8.2. Main Proof

Proof of Proposition 5. We first state the following property of the regularized incomplete beta function $I_d(\cdot,\cdot)$,

$$I_d(a+1,b) = I_d(a,b) - \frac{d^a (1-d)^b}{aB(a,b)}; \quad I_d(a,b+1) = I_d(a,b) + \frac{d^a (1-d)^b}{bB(a,b)}.$$

Hence by (12), if

$$I_{d_x}(a,b) \geq \frac{1}{2} + \frac{d_x^a(1-d_x)^b}{aB(a,b)},$$

then

$$\begin{aligned}\mathcal{R}_x(a,b) &= -c_x - I_{d_x}(a,b) + \frac{a}{a+b} I_{d_x}(a+1,b) + \frac{b}{a+b} I_{d_x}(a,b+1) \\ &= -c_x - \frac{a}{a+b} \cdot \frac{d_x^a(1-d_x)^b}{aB(a,b)} + \frac{b}{a+b} \cdot \frac{d_x^a(1-d_x)^b}{bB(a,b)} \\ &= -c_x.\end{aligned}$$

Similarly, if

$$I_{d_x}(a,b) \leq \frac{1}{2} - \frac{d_x^a(1-d_x)^b}{bB(a,b)},$$

we have $\mathcal{R}_x(a,b) = -c_x$.

Now, if

$$\frac{1}{2} \leq I_{d_x}(a,b) < \frac{1}{2} + \frac{d_x^a(1-d_x)^b}{aB(a,b)},$$

then

$$\begin{aligned}\mathcal{R}_x(a,b) &= -c_x - I_{d_x}(a,b) + \frac{a}{a+b}[1 - I_{d_x}(a+1,b)] + \frac{b}{a+b} I_{d_x}(a,b+1) \\ &= -c_x + \frac{a}{a+b}[1 - 2I_{d_x}(a,b)] + \frac{2d_x^a(1-d_x)^b}{(a+b)B(a,b)} \\ &\leq -c_x + \frac{2d_x^a(1-d_x)^b}{(a+b)B(a,b)}.\end{aligned}$$

Similarly, if

$$\frac{1}{2} - \frac{d_x^a(1-d_x)^b}{bB(a,b)} < I_{d_x}(a,b) \leq \frac{1}{2},$$

we still have

$$\mathcal{R}_x(a,b) \leq -c_x + \frac{2d_x^a(1-d_x)^b}{(a+b)B(a,b)}. \tag{EC.4}$$

Thus (EC.4) holds for all $(a,b) \in \Lambda = [1,+\infty) \times [1,+\infty)$. Now applying Lemma EC.2, we know

$$\mathcal{R}_x(a,b) \leq -c_x + 1/\sqrt{2\pi(a+b)}.$$

EC.9. Proof of Proposition 6.

Proposition 6. *If* $a+b \geq 1/(2\pi c_x^2)$, *then* $(a,b) \notin \mathbb{C}_x$.

Proof of Proposition 6. In the sub-problem with single alternative x,

$$V_x(a,b) = \sup_{\tau_x} \mathbb{E}\left[\sum_{n=0}^{\tau_x-1} \mathcal{R}_x(a_n,b_n) \,\middle|\, (a_0,b_0)=(a,b)\right].$$

For any $\tau_x \geq 1$ and $0 \leq n \leq \tau_x - 1$, $a_n + b_n = a + b + n \geq 1/(2\pi c_x^2)$, and hence by Proposition 5, $\mathcal{R}_x(a_n,b_n) \leq 0$. Thus $\tau_x = 0$ is optimal and $V_x(a,b) = 0$, i.e., $(a,b) \notin \mathbb{C}_x$.

EC.10. Proof of Proposition 7.

PROPOSITION 7. $\tau_x^* \leq \max\{0, \lceil 1/(2\pi c_x^2) - (a_{0x} + b_{0x})\rceil\}$.

Proof of Proposition 7. Let $n = \max\{0, \lceil -(a_{0x}+b_{0x}) + 1/(2\pi c_x^2)\rceil\}$. Then $a_{nx} + b_{nx} = a_{0x} + b_{0x} + n \geq 1/(2\pi c_x^2)$. By Proposition 6, $(a_{nx}, b_{nx}) \notin \mathbb{C}_x$, and thus $\tau_x^* \leq n$.

EC.11. Remark 3.

Let $\rho = d_x - \mu$, then

$$\begin{aligned}
&\mathbb{E}\left[h\left(\Phi\left(\sqrt{\beta+\beta_x^\epsilon}(d_x - \mu - \tilde{\sigma}_x(\beta)Z)\right)\right)\right] \\
&= \int_{-\infty}^{\rho/\tilde{\sigma}_x(\beta)} \Phi\left(\sqrt{\beta+\beta_x^\epsilon}(\rho - \tilde{\sigma}_x(\beta)z)\right)\varphi(z)dz + \int_{\rho/\tilde{\sigma}_x(\beta)}^{+\infty}\left[1 - \Phi\left(\sqrt{\beta+\beta_x^\epsilon}(\rho - \tilde{\sigma}_x(\beta)z)\right)\right]\varphi(z)dz \\
&= X + Y,
\end{aligned}$$

where X and Y are defined to be the first and second integral respectively in the second line. Let Z_1 and Z_2 be two independent standard normal random variables. Then

$$X = \mathbb{P}\left[Z_1 \leq \sqrt{\beta+\beta_x^\epsilon}(\rho - \tilde{\sigma}_x(\beta)Z_2), Z_2 \leq \rho/\tilde{\sigma}_x(\beta)\right] = \mathbb{P}\left[(Z_2, Z_3) \leq (\rho/\tilde{\sigma}_x(\beta), 0)\right],$$

where $Z_3 := Z_1 - \sqrt{\beta+\beta_x^\epsilon}(\rho - \tilde{\sigma}_x(\beta)Z_2) \sim \mathcal{N}(-\rho\sqrt{\beta+\beta_x^\epsilon}, 1 + \beta_x^\epsilon/\beta)$ and $\mathrm{Cov}(Z_2, Z_3) = \sqrt{\beta_x^\epsilon/\beta}$. It follows that X can be evaluated from the cdf of a bivariate normal distribution. A similar argument can be applied to Y.

EC.12. Proof of Proposition 8.

PROPOSITION 8. *Define* $\overline{\mathcal{R}}_x(\beta) = -c_x + (\sqrt{1+\beta_x^\epsilon/\beta} - 1)/\sqrt{2\pi e} + \pi^{-1}\sqrt{\beta_x^\epsilon/\beta}$, *then* $\mathcal{R}_x(\mu,\beta) \leq \overline{\mathcal{R}}_x(\beta)$, *for all* $\mu \in \mathbb{R}$. *Moreover, for any fixed* $\beta > 0$, $\mathcal{R}_x(\mu,\beta) \to -c_x$ *as* $\mu \to \pm\infty$.

EC.12.1. Preparatory Material

LEMMA EC.3. *Define* $g\colon (-\infty,+\infty)^2 \mapsto [1/2,+\infty)$ *by*

$$g(u,v) = h\left(\Phi(u)\right) + |v-u| \cdot \varphi(u).$$

Then for any fixed u *and all* v, $h\left(\Phi(v)\right) \le g(u,v)$, *and the equation holds iff* $v = u$.

Proof of Lemma EC.3. It is clear that

$$h\left(\Phi(u)\right) = \begin{cases} \Phi(u), & \text{if } u \ge 0; \\ 1-\Phi(u) = \Phi(-u), & \text{if } u < 0. \end{cases}$$

Also notice that $h\left(\Phi(u)\right) = h\left(1-\Phi(-u)\right) = h\left(\Phi(-u)\right)$ and $\varphi(u) = \varphi(-u)$.

First assume that $u \ge 0$. Then $g(u,u) = h\left(\Phi(u)\right)$.

If $v > u$, by the Mean Value Theorem, $h\left(\Phi(v)\right) = \Phi(v) = \Phi(u) + (v-u)\cdot\varphi(w)$, where $w \in (u,v)$. Since $\varphi(w) < \varphi(u)$, it follows that $h\left(\Phi(v)\right) < g(u,v)$.

Otherwise if $v < u$, we claim that $h\left(\Phi(v)\right) < h\left(\Phi(2u-v)\right)$. The reason is as follows: if $v \ge 0$, then $v < u < 2u-v$, and hence $h\left(\Phi(v)\right) = \Phi(v) < \Phi(2u-v) = h\left(\Phi(2u-v)\right)$; if $v < 0$, then $0 < -v \le 2u-v$, and hence $h\left(\Phi(v)\right) = \Phi(-v) \le \Phi(2u-v) = h\left(\Phi(2u-v)\right)$. Now since $2u-v > u$, we know $h\left(\Phi(2u-v)\right) < g(u,2u-v) = g(u,v)$, it follows that $h\left(\Phi(v)\right) < g(u,v)$.

So the result holds for $u \ge 0$.

Now if $u \le 0$, then $-u \ge 0$, and hence for all v, we have $h\left(\Phi(v)\right) = h\left(\Phi(-v)\right) \le g(-u,-v)$. Moreover, $g(-u,-v) = h\left(\Phi(-u)\right) + |-v+u| \cdot \varphi(-u) = h\left(\Phi(u)\right) + |v-u|\cdot\varphi(u) = g(u,v)$. Thus we still have $h\left(\Phi(v)\right) \le g(u,v)$.

EC.12.2. Main Proof

Proof of Proposition 8. We first notice that the function $s \mapsto |s|\varphi(s)$ is maximized at $s = \pm 1$ and is strictly decreasing on $[1,+\infty)$ and strictly increasing on $(-\infty,-1]$, with $|s|\varphi(s) \to 0$ as $s \to \pm\infty$.

Now denote $\rho = d_x - \mu$. Then

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

where

$$
\begin{aligned}
0 &\le -h\left(\Phi\left(\sqrt{\beta}\rho\right)\right) + \mathbb{E}\left[h\left(\Phi\left(\sqrt{\beta+\beta_x^\epsilon}(\rho-\tilde{\sigma}_x(\beta)Z)\right)\right)\right] \\
&\le -h\left(\Phi\left(\sqrt{\beta}\rho\right)\right) + \mathbb{E}\left[g\left(\sqrt{\beta}\rho, \sqrt{\beta+\beta_x^\epsilon}(\rho-\tilde{\sigma}_x(\beta)Z)\right)\right] \\
&= -h\left(\Phi\left(\sqrt{\beta}\rho\right)\right) + \mathbb{E}\left[h\left(\Phi\left(\sqrt{\beta}\rho\right)\right) + \left|\sqrt{\beta+\beta_x^\epsilon}(\rho-\tilde{\sigma}_x(\beta)Z) - \sqrt{\beta}\rho\right| \cdot \varphi(\sqrt{\beta}\rho)\right] \\
&= \varphi\left(\sqrt{\beta}\rho\right) \cdot \mathbb{E}\left[\left|\sqrt{\beta+\beta_x^\epsilon}(\rho-\tilde{\sigma}_x(\beta)Z) - \sqrt{\beta}\rho\right|\right] \\
&= \varphi\left(\sqrt{\beta}\rho\right) \cdot \mathbb{E}\left[\left|(\sqrt{\beta+\beta_x^\epsilon} - \sqrt{\beta})\rho - \sqrt{\beta+\beta_x^\epsilon}\cdot\tilde{\sigma}_x(\beta)Z\right|\right] \\
&\le \varphi\left(\sqrt{\beta}\rho\right) \cdot \left[\left(\sqrt{\beta+\beta_x^\epsilon} - \sqrt{\beta}\right)|\rho| + \sqrt{\beta_x^\epsilon/\beta}\cdot\mathbb{E}|Z|\right] \\
&= \left(\sqrt{1+\beta_x^\epsilon/\beta} - 1\right)\cdot\varphi\left(\sqrt{\beta}\rho\right)\sqrt{\beta}|\rho| + \sqrt{\beta_x^\epsilon/\beta}\cdot\varphi\left(\sqrt{\beta}\rho\right)\cdot\mathbb{E}|Z|.
\end{aligned}
$$

If we fix β and let $\mu \to \pm\infty$, then $\rho \to \mp\infty$, and the above expression goes to 0. Hence $\mathcal{R}_x(\mu,\beta) \to -c_x$.

$\mathcal{R}_x(\mu,\beta) \le \overline{\mathcal{R}}_x(\beta)$ follows from $\varphi\left(\sqrt{\beta}\rho\right)\mathbb{E}|Z| \le \varphi(0)\mathbb{E}|Z| = 1/\pi$ and $\sup_s\{|s|\varphi(s)\} = \varphi(1) = 1/\sqrt{2\pi e}$.

EC.13. Proof of Proposition 9.

PROPOSITION 9. *Define*

$$
\overline{\beta}_x = \frac{\beta_x^\epsilon[(2\pi e)^{-1} - \pi^{-2}]}{-\pi^{-1}[(2\pi e)^{-\frac{1}{2}} + c_x] + (2\pi e)^{-\frac{1}{2}}\sqrt{\pi^{-2} + c_x^2 + 2c_x(2\pi e)^{-\frac{1}{2}}}}, \tag{EC.5}
$$

then $\mathbb{R} \times [\overline{\beta}_x, +\infty] \subseteq \Lambda \setminus \mathbb{C}_x$.

Proof of Proposition 9. It is clear that $\overline{\mathcal{R}}_x(\cdot)$ is a continuous, strictly decreasing function with $\lim_{\beta\to 0^+}\overline{\mathcal{R}}_x(\beta) = +\infty$ and $\lim_{\beta\to+\infty}\overline{\mathcal{R}}_x(\beta) = -c_x < 0$. Hence there exits a unique root, which is given by $\overline{\beta}_x$ in (EC.5). For any arbitrary μ and $\beta \ge \overline{\beta}_x$, $\mathcal{R}_x(\mu,\beta) \le \overline{\mathcal{R}}_x(\beta) \le 0$. In the sub-problem with single alternative x,

$$
\begin{aligned}
V_x(\mu,\beta) &= \sup_{\tau_x}\mathbb{E}\left[\sum_{n=0}^{\tau_x-1}\mathcal{R}_x(\mu_n,\beta_n)\,\middle|\,(\mu_0,\beta_0) = (\mu,\beta)\right] \\
&= \sup_{\tau_x}\mathbb{E}\left[\sum_{n=0}^{\tau_x-1}\mathcal{R}_x(\mu_n,\beta_0+n\beta_x^\epsilon)\,\middle|\,(\mu_0,\beta_0) = (\mu,\beta)\right] \\
&\le 0.
\end{aligned}
$$

It follows that $\tau_x = 0$ is optimal and $V_x(\mu,\beta) = 0$, i.e. $(\mu,\beta) \notin \mathbb{C}_x$.

EC.14. Proof of Proposition 10.

PROPOSITION 10. $\tau_x^* \le \max\left\{0, \lceil (\beta_x^\epsilon)^{-1}(\overline{\beta}_x - \beta_{0x})\rceil\right\}$.

Proof of Proposition 10. Let $n = \max\left\{0, \lceil (\beta_x^\epsilon)^{-1}(\overline{\beta}_x - \beta_{0x})\rceil\right\}$. Then $\beta_{nx} = \beta_{0x} + n \cdot \beta_x^\epsilon \ge \overline{\beta}_x$. By Proposition 9, $(\mu_{nx}, \beta_{nx}) \notin \mathbb{C}_x$, and thus $\tau_x^* \le n$.

EC.15. Proof of Proposition 11.

PROPOSITION 11. *For each alternative x, there exist some fixed functions $\overline{\mu}_x(\cdot)$ and $\underline{\mu}_x(\cdot)$ such that for any $\beta > 0$, $\mu \le \overline{\mu}_x(\beta)$ or $\mu \ge \underline{\mu}_x(\beta) \Rightarrow V_x(\mu, \beta) = 0$. Thus for any given β, we have $[\overline{\mu}_x(\beta), \underline{\mu}_x(\beta)]$ as the μ-boundary of the continuation set $\mathbb{C}_x$.*

Proof of Proposition 11. Pick some α such that $0 < \alpha < \min\{1, 2c_x\}$. Let $I_\alpha = [-z_\alpha, z_\alpha] = [-\Phi^{-1}(1-\alpha/2), \Phi^{-1}(1-\alpha/2)]$ be the $100(1-\alpha)\%$ confidence interval for the standard normal distribution.

For any fixed $\beta > 0$, by (16) and the fact that $h(\cdot) \ge 1/2$, we know

$$\begin{aligned}
&\mathbb{E}[V_x(\mu + \tilde{\sigma}_x(\beta)Z, \beta + \beta_x^\epsilon)] \\
&\le \mathbb{E}\left[1 - h\left(\Phi\left(\sqrt{\beta + \beta_x^\epsilon}(\rho - \tilde{\sigma}_x(\beta)Z)\right)\right)\right] \\
&\le \alpha/2 + (1-\alpha) \cdot \mathbb{E}\left[1 - h\left(\Phi\left(\sqrt{\beta + \beta_x^\epsilon}(\rho - \tilde{\sigma}_x(\beta)Z)\right)\right) \,\middle|\, Z \in I_\alpha\right],
\end{aligned}$$

where $\rho := d_x - \mu$. When $\rho \ge \tilde{\sigma}_x(\beta) z_\alpha$, since $h\left(\Phi(\cdot)\right) = \Phi(\cdot)$ on $[0, +\infty)$,

$$\mathbb{E}\left[h\left(\Phi\left(\sqrt{\beta + \beta_x^\epsilon}(\rho - \tilde{\sigma}_x(\beta)Z)\right)\right) \,\middle|\, Z \in I_\alpha\right] \ge \Phi\left(\sqrt{\beta + \beta_x^\epsilon}(\rho - \tilde{\sigma}_x(\beta) z_\alpha)\right).$$

From the proof of Proposition 8,

$$\mathcal{R}_x(\mu, \beta) \le -c_x + \left(\sqrt{1 + \beta_x^\epsilon/\beta} - 1\right) \cdot \varphi\left(\sqrt{\beta}\rho\right) \sqrt{\beta}|\rho| + \sqrt{\beta_x^\epsilon/\beta} \cdot \varphi\left(\sqrt{\beta}\rho\right) \cdot \mathbb{E}|Z|.$$

It follows that

$$L_x(\mu, \beta, V_x) = \mathcal{R}_x(\mu, \beta) + \mathbb{E}[V_x(\mu + \tilde{\sigma}_x(\beta)Z, \beta + \beta_x^\epsilon)] \le -c_x + \alpha/2 + \delta(\rho),$$

where

$$\begin{aligned}
\delta(\rho) := &\left(\sqrt{1 + \beta_x^\epsilon/\beta} - 1\right) \cdot \varphi\left(\sqrt{\beta}\rho\right) \sqrt{\beta}|\rho| + \sqrt{\beta_x^\epsilon/\beta} \cdot \varphi\left(\sqrt{\beta}\rho\right) \cdot \mathbb{E}|Z| \\
&+ (1-\alpha)\left[1 - \Phi\left(\sqrt{\beta + \beta_x^\epsilon}(\rho - \tilde{\sigma}_x(\beta) z_\alpha\right)\right]
\end{aligned}$$

is strictly decreasing on $\left[1/\sqrt{\beta}, +\infty\right)$ and goes to 0 as $\rho \to +\infty$ (see the proof of Proposition 8). Since $c_x - \alpha/2 > 0$, there exists some $\overline{\rho} \geq \max\{1/\sqrt{\beta}, \tilde{\sigma}_x(\beta) z_\alpha\}$ such that for all $\rho \geq \overline{\rho}$, $\delta(\rho) \leq c_x - \alpha/2$. Let $\underline{\mu}_x(\beta) = d_x - \overline{\rho}$, then equivalently for all $\mu \leq \underline{\mu}_x(\beta)$, we have $L_x(\mu, \beta, V_x) \leq 0$ and hence $V_x(\mu, \beta) = 0$.

Symmetrically, we can find some $\overline{\mu}_x(\beta)$ such that for all $\mu \geq \overline{\mu}_x(\beta)$, $V_x(\mu, \beta) = 0$.

EC.16. Proof of Proposition 12.

PROPOSITION 12. *Suppose (μ, β) is some arbitrary state of alternative x. Then*

$$\nu_x(\mu, \beta) \leq -c_x + 1 - h\left(\Phi\left(\sqrt{\beta}(d_x - \mu)\right)\right).$$

Proof of Proposition 12. The following expectations are taken with respect to the sub-problem with a single alternative x.

$$\begin{aligned}
\nu_x(\mu, \beta) + c_x &= \max_{\tau > 0} \frac{\mathbb{E}\left[\sum_{n=1}^{\tau} \alpha^{n-1} [\mathcal{R}_x(S_{n-1,x}) + c_x] | S_{0x} = (\mu, \beta)\right]}{\sum_{n=1}^{\tau} \alpha^{n-1}} \\
&\leq \max_{\tau > 0} \mathbb{E}\left[\sum_{n=1}^{\tau} [\mathcal{R}_x(S_{n-1,x}) + c_x] \,\middle|\, S_{0x} = (\mu, \beta)\right] \qquad \text{since } \mathcal{R}_x(S_{n-1,x}) + c_x \geq 0 \text{ and } 0 < \alpha < 1 \\
&= \mathbb{E}\left[\sum_{n=0}^{\infty} [\mathcal{R}_x(S_{nx}) + c_x] \,\middle|\, S_{0x} = (\mu, \beta)\right] \\
&= \mathbb{E}\left[\sum_{n=1}^{\infty} [R_n + c_x] \,\middle|\, S_{0x} = (\mu, \beta)\right] \qquad \text{use } \mathcal{R}_x(S_{nx}) = \mathbb{E}[R_{n+1} | S_n] \text{ and the tower property} \\
&= \mathbb{E}\left[\sum_{n=1}^{\infty} [h(P_{nx}) - h(P_{n-1,x})] \,\middle|\, S_{0x} = (\mu, \beta)\right] \\
&= \mathbb{E}\left[\lim_{n \to \infty} h(P_{nx}) \,\middle|\, S_{0x} = (\mu, \beta)\right] - h\left(\Phi\left(\sqrt{\beta}(d_x - \mu)\right)\right) \\
&= 1 - h\left(\Phi\left(\sqrt{\beta}(d_x - \mu)\right)\right),
\end{aligned}$$

where the last equation follows because $\lim_{n \to \infty} P_{nx} = \mathbf{1}_{\{x \in \mathbb{B}\}}$ almost surely, by the strong law of large numbers.

Acknowledgments

This work was partially supported by the Air Force Office of Scientific Research. The authors would also like to thank Matthew S. Maxwell and Shane G. Henderson for the use of their ambulance simulation software, and their help in using it.

References

See references list in the main paper.

Abramowitz, M., I.A. Stegun. 1964. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover Publications.

Sloane, N. 2007. The on-line encyclopedia of integer sequences. *Towards Mechanized Mathematical Assistants* 130–130.