MANAGEMENT SCIENCE
Vol. 58, No. 3, March 2012, pp. 550–569
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1110.1425

Sequential Sampling with Economics of Selection Procedures

Stephen E. Chick
Technology and Operations Management Area, INSEAD, 77305 Fontainebleau, France,
stephen.chick@insead.edu

Peter Frazier
Department of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853,
pf98@cornell.edu

Sequential sampling problems arise in stochastic simulation and many other applications. Sampling is used to infer the unknown performance of several alternatives before one alternative is selected as best. This paper presents new economically motivated fully sequential sampling procedures to solve such problems, called economics of selection procedures. The optimal procedure is derived for comparing a known standard with one alternative whose unknown reward is inferred with sampling. That result motivates heuristics when multiple alternatives have unknown rewards. The resulting procedures are more effective in numerical experiments than any previously proposed procedure of which we are aware and are easily implemented. The key driver of the improvement is the use of dynamic programming to model sequential sampling as an option to learn before selecting an alternative. It accounts for the expected benefit of adaptive stopping policies for sampling, rather than of one-stage policies, as is common in the literature.

Key words: simulation; statistical analysis; probability; diffusion; decision analysis; dynamic programming; Bayesian
History: Received December 2, 2009; accepted June 27, 2011, by Assaf Zeevi, stochastic models and simulation. Published online in *Articles in Advance* October 7, 2011.

This paper focuses on the use of sampling to select the best of a finite set of alternatives, where the "best" alternative is the one with maximum expected value, and the expectation is to be inferred with statistical sampling. Such problems arise in stochastic simulation projects for process design decisions in business, industrial, and service applications, agricultural and pharmaceutical tests, and a variety of other applications.

In simulation experiments, for example, an alternative might correspond to a particular choice of design parameters for a manufacturing process (Law 2007). Simulation replications of alternatives can be simulated sequentially until sampling stops and an alternative is selected for implementation. We are interested in the case where the implemented alternative has economic value, such as the net profit from implementing a manufacturing process with design parameters that are chosen on the basis of simulation experiments.

An extensive statistical literature addresses this problem (see, e.g., Gupta et al. 1979, Bechhofer et al. 1995, Kim and Nelson 2006). The majority of work seeks to minimize the expected number of samples required to provide statistical guarantees for the probability of correct selection. A typical guarantee is of the form "the alternative that is selected is within some prespecified error tolerance of the true best, with at least some prespecified probability." Although there is some economic value gained in minimizing the expected number of samples subject to such a worst-case constraint, this approach is statistically conservative and typically results in excessive sampling. Furthermore, this approach does not typically consider and react to the cost of sampling (Chan and Lai 2006 is a notable exception), even though this cost varies widely across applications. For example, at a recent Dagstuhl workshop[1] entitled "Sampling-based Optimization in the Presence of Uncertainty," participants noted that a single replication of a simulation can require anywhere from a fraction of a second up to a full day, depending on the application. In sequential sampling with interactive questionnaires, biological tests, or medical screening, the marginal cost of a sample is usually much larger than in simulation.

This paper describes how sampling costs and the expected benefits from implementing an alternative,

[1] See http://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=09181.

rather than statistical criteria that ignore sampling costs, can be used as the driver for dynamically and sequentially deciding which alternatives to sample and when to stop sampling. Insufficient sampling reduces sampling costs but decreases the potential benefit of selecting a good alternative with high probability. Excessive sampling increases this potential benefit because it allows the mean performance of each alternative to be better estimated, but the cost of excessive sampling can overwhelm that benefit. This paper studies the balance between these costs and benefits.

We frame the problem using dynamic programming (DP) techniques in §1 for the special case of normally distributed samples with unknown means and known variances. That section also describes the optimal sequential sampling policy. Uncertainty about the unknown mean performance of the alternatives is described with a Bayesian formulation, because this has been found to be effective in related work (Branke et al. 2007).

Section 2 analyzes the problem when a single alternative with unknown mean reward is being compared to the expected value of a known standard. This analysis uses a diffusion approximation, which enables us to write the solution in terms of a single standardized problem that is independent of parameters characterizing the sampling costs and variances. This makes the solution much easier to use. The diffusion approximation is motivated by an approach developed by Chernoff (1961), which has general appeal in its own right for optimal stopping problems in Bayesian statistics and has led to a variety of related work in hypothesis testing, sequential sampling, and the multiarmed bandit problem (Chernoff and Ray 1965, Lai 1987, Brezzi and Lai 2002, Chick and Gans 2009). The analysis below provides a continuation set, within which it is optimal to continue sampling and outside of which it is optimal to stop and to select the better of the alternative with an unknown mean or the known standard. The continuation set identifies whether there exists a dynamic and nonanticipative sampling plan whose expected value of information (EVI) from sampling exceeds its expected cost of sampling.

Section 3 demonstrates how the solution to the problem with $k = 1$ unknown alternative from §2 can be used to define effective heuristics to handle the problem with $k > 1$ unknown alternatives. We call the resulting procedures economics of selection procedures (ESPs). This approach, which determines the shape of the optimal stopping boundary from economic considerations, differs in spirit and structure from the frequentist literature, which uses stopping boundaries of a fixed shape (e.g., triangular or parabolic) and scales or resizes that shape to achieve a specific frequentist probability of correct selection guarantee (e.g., see Kim and Nelson 2006).

We recall several previously proposed sequential sampling procedures and bounds on the optimal performance in §4 to assess the new sequential sampling procedure derived in §§2 and 3. The comparators, given in §§4.2 and 4.3, include the procedures that performed the best in a recent and very large-scale assessment of selection procedures (Branke et al. 2007). Numerical experiments in §5 demonstrate that our new approach performs more effectively than the best of those earlier procedures, as measured by the expected net benefit of selecting an alternative less the cost of sampling.

Section 6 extends the analysis for normally distributed samples with known variances to a broader class of sampling distributions and develops that analysis specifically for normally distributed samples with unknown variances. The resulting procedure is simple to implement (it does not require the calculation of statistical constants, uses only simple algebraic functions, and has no free parameters) and is shown in a numerical experiment to outperform all other procedures for this problem of which we are aware.

In spite of the preponderance of previous work focusing on statistical criteria, this paper is not the first to consider the cost and EVI from sampling. Early work on Bayesian selection procedures examined the EVI of one-stage sampling procedures and noted that those procedures might be repeated in sequential fashion to obtain a dynamic procedure (Gupta and Miescke 1996). Chick and Inoue (2001) extended that analysis and incorporated sampling cost criteria into such a one-stage procedure. Frazier et al. (2008) also provided one-step lookahead allocations and used a DP framework to provide interesting characterizations for policies that repeatedly use one-step lookahead allocations in sequential fashion. The resulting policies can perform highly efficiently relative to some other procedures that have been proposed (Frazier and Powell 2008).

This paper extends those papers by providing a more thorough analysis of a range of nonanticipative sampling policies, not just one-stage allocations that stand alone or are put in sequence. The analytical methodology for doing so is commensurately more challenging, and our characterization of the optimal solution to the diffusion approximation for the special case of $k = 1$ unknown alternative differs from all of that prior work. In spite of the challenging derivation, the resulting procedure is easy to implement.

The prior work that is closest in nature to this paper is Chick and Gans (2009), which also uses DP techniques to assess the value of the option to continue sampling before selecting an alternative as

best. The key difference between our work and that paper is that the latter considers discounted sampling costs and discounted rewards for the alternative that is selected, whereas this paper focuses on sampling costs and rewards that are not discounted. We show below that the optimal stopping regions for the two cases of discounted versus undiscounted costs have very different structural properties.

In summary, this paper shows how to dynamically decide which alternatives to sample and for how long to sample before selecting an alternative for implementation by using economic criteria rather than statistical criteria. We consider a broader class of adaptive sampling policies than has been considered previously. Doing so provides the most effective sequential selection procedure to date. The paper also demonstrates another application of useful DP tools from Chernoff (1961) for optimal stopping problems that involve Bayesian inference. This paper extends initial work reported in Chick and Frazier (2009).

1. The Sampling Selection Problem

An analyst seeks to implement one of k alternatives whose rewards X_i are random with unknown means (for $i = 1, 2, \ldots, k$) or to implement a known standard (labeled $i = 0$) whose expected reward is the known value $\mathrm{E}[X_0] = m$. For example, if the analyst has the option to implement one of k alternatives or to reject them all and "do nothing," then $m = 0$.

Before selecting an alternative to implement ($i \in 0, 1, \ldots, k$), the analyst can choose to sequentially sample one or more of the k alternatives to infer the unknown means. In simulation experiments, for example, the sample $X_{i,j}$ is a random variable whose realization $x_{i,j}$ is the output of the jth simulation replication for alternative i, for $i = 0, 1, \ldots, k$ and $j = 1, 2, \ldots$.

At each stage of this sequential process, the choice of which alternative to sample or to select for implementation can depend upon all of the data observed before making that choice. The analyst should maximize the expected net reward, which sums the cost of sampling and the (single) reward received after sampling ceases, and we implement the one alternative that appears best. This section formalizes this "sampling selection problem" for the special case of normally distributed samples with known sampling variances and unknown sampling means. It also presents theoretical results regarding the optimal solution to that problem.

Let U_i be the sampling mean of alternative i, and let σ_i^2 be the known sampling variance. Let $\mathbf{U} = (U_0, U_1, \ldots, U_k)$ be the vector of unknown means and $\mathbf{u} = (u_0, u_1, \ldots, u_k)$ be the corresponding vector of realizations. We assume that samples $X_{i,j}$ are conditionally independent given the means U_i,

$$\{X_{i,j}\colon j = 1, 2, \ldots\} \mid U_i \overset{\text{iid}}{\sim} \texttt{Normal}(U_i, \sigma_i^2) \quad \text{for } i = 0, 1, \ldots, k. \tag{1}$$

To describe the analyst's initial uncertainty about the mean rewards, we assume a conjugate prior distribution for the unknown means,

$$U_i \sim \texttt{Normal}(\mu_{i,0}, \sigma_i^2/n_{i,0}) \quad \text{for } i = 1, 2, \ldots, k, \tag{2}$$

with the U_i independent for $i > 0$ and with $U_0 = m$ for the known alternative. The real-valued $n_{i,0} > 0$ can be interpreted as the effective number of samples that is embodied by the prior distribution for the unknown mean of alternative i. It will be convenient to refer to the vector $(\mu_{i,0}, n_{i,0})$, the so-called hyperparameter for the unknown mean U_i, by a single term. We therefore define $\Theta_{i,0} = (\mu_{i,0}, n_{i,0})$.

We now turn to the sampling process for inferring the unknown means. We model sampling as occurring sequentially in stages indexed by $t = 0, 1, 2, \ldots$. Suppose that t samples have been observed in total, of which $l_{i,t}$ of them have been for alternative $i > 0$, so that $t = \sum_{i=1}^{k} l_{i,t}$. By Bayes' rule, the posterior distribution of U_i given the data through time t is $\texttt{Normal}(\mu_{i,t}, \sigma_i^2/n_{i,t})$, where $n_{i,t} = n_{i,0} + l_{i,t}$, $\bar{x}_{i,t}$ is the sample average of the $l_{i,t}$ observations for alternative i, and $\mu_{i,t} = (n_{i,0}\mu_{i,0} + l_{i,t}\bar{x}_{i,t})/n_{i,t}$. We set $\Theta_{i,t} = (\mu_{i,t}, n_{i,t})$ to denote the hyperparameters for the unknown U_i given data to time t, for $i > 0$. The known standard is described by $\Theta_{0,t} = (m, \infty)$, because a known mean can result from an infinite number of samples. Let $\vec{\Theta}_t = (\Theta_{0,t}, \Theta_{1,t}, \ldots, \Theta_{k,t})$ for $t = 0, 1, \ldots$.

A choice to sample alternative $i(t) = i > 0$ at time t causes a Markovian state transition from a known state $\Theta_{i,t}$ to a random state $\Theta_{i,t+1}$. This transition is determined by the probability distribution $P_{X_{i,t+1}\mid\Theta_{i,t}}$ for the next sample given information up to time t and Bayes' rule, which implies $\mu_{i,t+1} = (n_{i,t}\mu_{i,t} + X_{i,t+1})/(n_{i,t}+1)$ for normally distributed samples. For $j \neq i(t)$, we have $\Theta_{j,t+1} = \Theta_{j,t}$.

The analyst must choose a sequence of alternatives to sample from, and then ultimately select an alternative, so that the stream of costs and terminal reward together maximize the expected net reward. We use the notion of a policy to model those choices. Informally, a policy π is a dynamic method of choosing, at each time t, whether to sample an alternative or to select an alternative for implementation (which delivers the ultimate reward and stops the process). More precisely, a policy π defines a mapping $i(t, \vec{\Theta}_t)$ at each

time t to one of $2k+1$ possible decisions.[2] We write $i(t)$ for $i(t, \vec{\Theta}_t)$ to simplify notation. The $2k+1$ possible decisions are to sample alternative $i(t) = i$ for $i = 1, 2, \ldots, k$ or to stop sampling to implement alternative i (denoted by $i(t) = k+1+i$ for $i = 0, 1, \ldots, k$). We define T to be the first time that an alternative is not sampled (the smallest t such that $i(t) > k$) and define $I(T) = i(T) - k - 1$ to be the alternative that is selected for implementation. We need not define $i(t)$ for $t > T$ because only one alternative can be implemented.

By construction, π is nonanticipative. Both $i(t)$ and the event $\{T = t\}$ depend only on the prior distribution and the data observed to time t, so that T is a stopping time. Let Π be the set of such policies. We write E_π to indicate the expectation with respect to the measure that π induces on the sequence of observations and decisions, and E to indicate the expectation when it does not depend on π.

We assume that the incremental cost of each sample of alternative i is $c_i > 0$. Under this assumption, the expected value of any policy with a strictly positive probability of sampling forever (choosing $T = \infty$) is $-\infty$. Thus, we restrict Π to policies with $T < \infty$ almost surely. Given a generic prior distribution $\vec{\Theta} = (\Theta_0, \Theta_1, \ldots, \Theta_k)$ and a policy $\pi \in \Pi$, the expected value of the future stream of rewards is

$$V^\pi(\vec{\Theta}) \;=\; \mathrm{E}_\pi\left[\sum_{t=0}^{T-1}(-c_{i(t)}) \;+\; X_{I(T),\,T+1}\,\middle|\,\vec{\Theta}_0 = \vec{\Theta}\right]. \quad (3)$$

Formally, we define the analyst's undiscounted *sampling selection problem* to be the choice of a selection policy that maximizes this undiscounted expected reward:

$$V^*(\vec{\Theta}_0) = \sup_{\pi\in\Pi} V^\pi(\vec{\Theta}_0). \quad (4)$$

The value function V^* and optimal policies can be characterized by standard results from dynamic programming for infinite horizon undiscounted problems. To do so, we first write the problem as one with nonnegative costs for sampling and selecting.

Proposition 1. *For policies $\pi \in \Pi$,*

$$V^\pi(\vec{\Theta}_0) = \mathrm{E}\Big[\max_{i=0,1,\ldots,k} U_i \mid \vec{\Theta}_0\Big] - \mathrm{E}_\pi\left[\sum_{t=0}^{T-1} c_{i(t)} + L_{I(T)}\,\middle|\,\vec{\Theta}_0\right],$$

where $L_i = (\max_{j=0,1,\ldots,k} U_j) - U_i$ is the loss associated with selecting alternative i, for $i = 0, 1, \ldots, k$.

The appendix provides mathematical proofs of all claims that are not justified in the main text.

[2] It is sufficient to consider dependence only on $(t, \vec{\Theta}_t)$ because $\{(t, \vec{\Theta}_t)\colon t = 0, 1, \ldots\}$ is a Markov process, so an additional dependence on the past cannot bring additional expected reward (Bertsekas and Shreve 1978, Proposition 9.1).

The term $\mathrm{E}[\max_{i=0,1,\ldots,k} U_i \mid \vec{\Theta}_0]$ is the expected reward of having perfect information about the means, with no sampling cost, before selecting the best alternative for implementation. That term does not depend on π. A policy π therefore maximizes $V^\pi(\vec{\Theta}_0)$ if and only if (iff) it minimizes the sum of the expected total sampling cost $\mathrm{E}_\pi[\sum_{t=0}^{T-1} c_{i(t)} \mid \vec{\Theta}_0]$ and the expected opportunity cost ($\mathrm{E}[OC]$) of potentially selecting an alternative that is not best, $\mathrm{E}[OC] = \mathrm{E}_\pi[L_{I(T)} \mid \vec{\Theta}_0]$.

Because both c_i and L_i are nonnegative, this equivalent problem satisfies the (P) assumption of Chap. 9 of Bertsekas and Shreve (1978). Proposition 9.8 of Bertsekas and Shreve (1978) then shows that the value function satisfies Bellman's recursion,

$$V^*(\vec{\Theta}_t) = \max\Big(\max_{i=1,2,\ldots,k} \mathrm{E}[-c_i + V^*(\vec{\Theta}_{t+1}) \mid \vec{\Theta}_t, i(t) = i], \max_{i=0,1,\ldots,k} \mathrm{E}\big[U_i \mid \vec{\Theta}_t\big]\Big). \quad (5)$$

Here, $\mathrm{E}[-c_i + V^*(\vec{\Theta}_{t+1}) \mid \vec{\Theta}_t, i(t) = i]$ is the expected reward of sampling from alternative $i(t) = i$ at time t and acting optimally afterward, and $\mathrm{E}[U_i \mid \vec{\Theta}_t]$ is the expected reward of stopping at time $T = t$ and selecting alternative $I(T) = i$. The presence of the term $X_{I(T),\,T+1}$ in (3) as compared with the term $\mathrm{E}[U_i \mid \vec{\Theta}_t]$ in (5) is explained by noting that selecting alternative $I(T) = \arg\max_{i\in 0,1,\ldots,k} \mathrm{E}[U_i \mid \vec{\Theta}_t]$ at time $T = t$ results in a reward $\mathrm{E}[X_{I(T),\,T+1} \mid \vec{\Theta}_T] = \mathrm{E}[\mathrm{E}[X_{I(T),\,T+1} \mid U_{I(T)}, \vec{\Theta}_T] \mid \vec{\Theta}_T] = \mathrm{E}[U_{I(T)} \mid \vec{\Theta}_T]$.

Proposition 2. *Any policy π whose decisions attain the maximum in Bellman's recursion in* (5) *is optimal, i.e., $V^\pi(\vec{\Theta}_0) = V^*(\vec{\Theta}_0)$.*

Proof. The proof follows directly from Proposition 9.12 of Bertsekas and Shreve (1978).

The proofs of Propositions 1 and 2 hold in some interesting cases beyond normal distributions with known sampling variances. For example, see §6 below for normal distributions with unknown sampling variances.

The following proposition shows that there is a deterministic upper bound on the number of samples taken by the optimal policy, even though there is no a priori constraint to stop in a finite time. This is in contrast to other sequential information collection problems in which there is no deterministic upper bound on the number of samples taken by the optimal policy. For example, consider sequential hypothesis testing and the discussion of the truncated sequential probability ratio test in Siegmund (1985).

Proposition 3. *Assume* (1) *and* (2). *Under any optimal policy, the stopping time T is bounded above by a deterministic quantity, $T \le k + \sum_{i=1}^{k}[\sigma_i^2/(2\pi c_i^2) - n_{i,0}]$.*

The proof of Proposition 3 relies on specific properties of the normal distribution and the assumption that the sampling variances are known.

2. Comparing One Alternative with an Unknown Mean to a Known Standard

This section examines the special case of sequential sampling to compare $k=1$ alternative whose mean is unknown with a known mean reward, m. In this section we assume that samples are normally distributed with an unknown mean whose prior distribution is in accordance with (1) and (2) above. The results of the analysis here for $k=1$ will also be used in §3 where we study $k>1$.

We drop i, $i(t)$, and $I(T)$ in subscripts for notational convenience in this section because $k=1$. Terms like U, Θ_t, and X_t refer to alternative 1 in this section, and we rewrite Θ_t as (μ_t, n_t) here to describe the distribution for the unknown mean U of the lone alternative. Thus, (4) becomes the optimal stopping problem

$$V^*(m,\mu_0,n_0) = \sup_{\pi} \mathrm{E}_\pi[-cT+\max\{m,\mathrm{E}[U\mid\Theta_T]\}\mid\Theta_0=(\mu_0,n_0)], \quad (6)$$

and Bellman's recursion in (5) becomes

$$V^*(m,\mu_t,n_t) = \max\big\{m, -c+\mathrm{E}[V^*(m, (n_t\mu_t+X_{i,t+1})/(n_t+1), n_t+1)\mid\mu_t,n_t],\mu_t\big\}. \quad (7)$$

We could solve the discrete-time optimal stopping problem (6) directly using Bellman's recursion (7), beginning from the implicit horizon given by Proposition 3, but this recursion depends on c, σ, and m, so the whole solution would need to be recomputed numerically each time these values were changed. Instead, we solve a single standardized problem whose solution can be easily transformed to give a nearly optimal stopping boundary for any c, σ, and m. With this approach, the practitioner does not need to maintain a working implementation of the DP with which he can recompute the optimal stopping boundary for his given values of c, σ, and m. He can just store the optimal stopping boundary or use a convenient analytic approximation to it that we provide below.

The transformation to a standardized problem is accomplished in §2.1, which approximates (6) using a diffusion that rescales both time and the values of X. Such a diffusion approximation is common in the sequential sampling and simulation literatures, and the resulting continuous-time problem is of interest in its own right. Then, §2.2 uses a standard technique, that of finite differences, to approximate the solution of the resulting continuous state-space optimization problem with the solution to a related discretized problem on a lattice. An easy-to-use analytic approximation to the resulting optimal stopping boundary is also provided.

At first glance, it may seem odd to convert a discrete-time optimal stopping problem to a continuous-time problem and to solve the resulting continuous-time problem with numerical methods that require a discrete-time grid. However, conversion to continuous time is required for the rescaling that supports standardization, and the resulting problem is then most easily solved numerically with a lattice. Although the lattice does rediscretize the problem, the discretization scheme of the lattice and of the original problem differ substantially.

2.1. Diffusion Approximation for Sampling Selection When $k=1$

We first give a continuous-time diffusion approximation to the discrete-time problem (6) that depends upon the parameters c, σ, and m. We then transform this continuous-time problem ((8), below) to a second equivalent continuous-time problem ((11), below) that does not depend on these parameters.

The process $\{(\mu_0 n_0+\sum_{j=1}^t X_j, n_0+t)\colon t=0,1,\ldots\}$, given U, is a random walk with independent Gaussian increments. This discrete-time process has the same distribution as the continuous-time process $\{(Y_t, n_0+t)\colon t\geq 0\}$ restricted to integral times t, where, given U, $\{Y_t\colon t\geq 0\}$ is a Brownian motion with $Y_0 = y_0 \triangleq \mu_0 n_0$, drift U, and volatility σ. We may further couple these discrete-time and continuous-time processes by letting $Y_t = y_0+\sum_{j=1}^t X_j$, so that the value of the discrete-time process after $t=0,1,\ldots$ observations is equal to value of the continuous-time process at time t. From this construction, we have

$$Y_t \mid U \sim \texttt{Normal}(\mu_0 n_0 + Ut, \sigma^2 t)$$

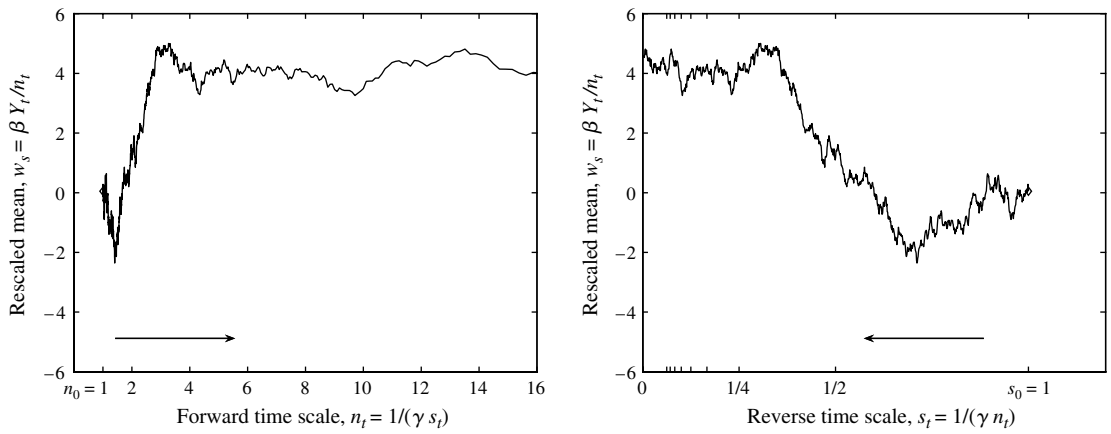
for real-valued $t\geq 0$ and $U\sim\texttt{Normal}(\mu_0,\sigma^2/n_0)$, so that the posterior distribution of U given continuous observations of Y_s is

$$U\mid\{(Y_s,s)\quad\text{for } s\in[0,t]\}\sim\texttt{Normal}(Y_t/n_t,\sigma^2/n_t).$$

Thus, the posterior distribution of U, conditional on continuous observations of the continuous-time process for integral t, matches that of the discrete-time process in §1 because $\mu_t = y_t/n_t$ and $n_t = n_0+t$.

We now approximate the discrete-time sampling selection problem (6) with a continuous-time problem. Let $\tilde{T}$ be a stopping time for the diffusion $\{Y_t\colon t\geq 0\}$. If $\tilde{T}$ were restricted to integer times, then $\tilde{T}$ would be equal to some stopping time T for the discrete-time process. We would also have equality

Figure 1 **A Sample Path of the Rescaled Posterior Mean W_s in Two Time Scales: Forward Time in the Effective Number of Samples n_t (Increasing from n_0 in the Left-Hand Plot) and Reverse Time $s = 1/(\gamma n_t)$ (Decreasing from $s_0 = 1/(\gamma n_0)$ to 0 in the Right-Hand Plot)**



Note. Here, $\beta = \gamma = 1$.

between $Y_{\tilde{T}}/n_{\tilde{T}}$ and $\mathrm{E}[X_{T+1} \mid \Theta_T]$ when stopping. By allowing $\tilde{T}$ to be real valued rather than an integer, we obtain the first continuous-time diffusion approximation to (6),

$$\begin{aligned} &V^*(m, \mu_0, n_0) \\ &= \sup_{\pi} \mathrm{E}_\pi[-cT \ + \max\{m, \mathrm{E}[X_{T+1} \mid \Theta_T]\} \mid \Theta_0 = (\mu_0, n_0)] \\ &\approx \sup_{\tilde{T} \geq 0} \mathrm{E}[-c\tilde{T} + \max\{m, Y_{\tilde{T}}/n_{\tilde{T}}\} \mid \Theta_0 = (\mu_0, n_0)]. \end{aligned} \tag{8}$$

The solution to (8) depends upon the values of m, c, and σ^2. We now rescale the diffusion to obtain a single standardized stopping problem, from which we can obtain the solution to (8) with a simple transformation of variables. Let β and γ be constants. They are arbitrary now, but we will fix their values below to achieve the desired rescaling. Then, let $s = 1/(\gamma n_t)$ index the progression of the inference. This s coordinate is inversely proportional to the posterior variance for the unknown mean and is therefore proportional to the information about the unknown mean. The behavior of the posterior mean is simplified if we index by s rather than by t. We let $W_s = \beta Y_t/n_t$ be the rescaled posterior mean, let $\tilde{m} = \beta m$ be the rescaled value of the standard, and label the initial conditions $s_0 = 1/(\gamma n_0)$ and $w_{s_0} = \beta\mu_0 = \beta y_0/n_0$.

Consider the process $\{(W_s, s)\colon s_0 \geq s \geq 0\}$ in the $-s$ scale (beginning at $s = s_0$ and decreasing to $s = 0$). It is a Gaussian process, and calculation of its mean and covariance at arbitrary values of s shows that it is a Brownian motion with some volatility that depends on β and γ. It has no drift because W_s is proportional to the posterior mean of U, and the posterior mean is a martingale in Bayesian inference. We now set β and γ to achieve a volatility of 1, i.e., to achieve $\mathrm{Var}[W_s \mid (w_{s_0}, s_0)] = s_0 - s = 1/(\gamma n_0) - 1/(\gamma n_t) = (1/\gamma)(t/(n_0(n_0 + t)))$. By standard results for the predictive distribution of the posterior mean (de Groot 1970), $\mathrm{Var}[\beta Y_t/n_t \mid y_0, n_0] = \beta^2\sigma^2 t/(n_0(n_0 + t))$. Thus, setting $\beta^2\sigma^2\gamma = 1$ assures that the volatility is 1 and $\{(W_s, s)\colon s_0 \geq s \geq 0\}$ is a standard Brownian motion in the $-s$ scale.

Figure 1 illustrates the behavior of the two equivalent processes $\{(W_s, s)\colon s_0 \geq s \geq 0\}$ and $\{(\beta Y_t/n_t, n_t)\colon t \geq 0\}$. Recall that $W_{s_0} = \beta Y_0/n_0$ is the prior mean at time $t = 0$. As we collect more samples, t and n_t increase, the corresponding $s = 1/(\gamma n_t)$ shrinks toward 0, and the posterior mean $W_s = \beta Y_t/n_t$ moves toward the true sampling mean. A similar rescaling has been used for other optimal sequential sampling problems (Chernoff 1961, Brezzi and Lai 2002, Chick and Gans 2009).

To complete this transformation, we let $S = 1/(\gamma n_{\tilde{T}})$, which is a stopping time in the $-s$ scale. This standardizes (8) as follows:

$$\begin{aligned} &B(m, \mu_0, n_0) \\ &\quad \triangleq \sup_{\tilde{T} \geq 0} \mathrm{E}[-c\tilde{T} + \max\{m, Y_{\tilde{T}}/n_{\tilde{T}}\} \mid Y_0 = \mu_0 n_0] \\ &\quad = m + \sup_{S \in [0, s_0]} \mathrm{E}\Bigg[-\frac{c}{\gamma}\left(\frac{1}{S} - \frac{1}{s_0}\right) \\ &\qquad\qquad + \max\{0, \beta^{-1} W_S - m\} \Big| W_{s_0} = w_{s_0}\Bigg] \\ &\quad = m + \beta^{-1} \sup_{S \in [0, s_0]} \mathrm{E}\Bigg[-\frac{c\beta}{\gamma}\left(\frac{1}{S} - \frac{1}{s_0}\right) \\ &\qquad\qquad + \max\{0, W_S\} \Big| W_{s_0} = w_{s_0} - \tilde{m}\Bigg]. \end{aligned} \tag{9}$$

If in addition to setting $\beta^2\sigma^2\gamma = 1$ to set the underlying volatility to one we also set $c\beta = \gamma$, then (9) simplifies even further. To accomplish this, we set

$$\beta = c^{-1/3}\sigma^{-2/3} \quad \text{and} \quad \gamma = c^{2/3}\sigma^{-2/3}. \tag{10}$$

This value of β is the cube root of the sampling efficiency (which is itself inversely proportional to the product of the sampling variance and the cost per sample; Hammersley and Hanscomb 1964, p. 22).

With these parameter values, the optimal stopping problem in (9) is

$$\tilde{B}(w_{s_0}-\tilde{m}, s_0) \triangleq \sup_{S\in[0,s_0]} \mathrm{E}\left[-\left(\frac{1}{S}-\frac{1}{s_0}\right)+\max\{0, W_S\}\,\middle|\, W_{s_0}=w_{s_0}-\tilde{m}\right]. \tag{11}$$

This development proves the following key result for our analysis, summarized below in Proposition 4: only the standardized problem ($c=1$, $\sigma=1$, $m=0$) in (11), which provides $\tilde{B}(\cdot,\cdot)$, need be solved to obtain $B(\cdot,\cdot,\cdot)$ for any $c>0$, $\sigma>0$, and m. This is important because the function B is a diffusion approximation to the solution V^* to the original discrete-time problem (6).

Proposition 4. *$B(m,\mu_0,n_0)=m+\beta^{-1}\tilde{B}(\beta(\mu_0-m),n_0)$, where β is as in* (10).

The key to the optimal solution when $k=1$, then, is to solve the optimal stopping problem (11). Standard techniques (Chernoff 1961, Bather 1970) show that the function $\tilde{B}$ in (11) is the solution to a free boundary problem. A free boundary problem is a partial differential equation (PDE) whose boundary is implicitly determined by an indifference between stopping to get a reward and continuing to sample. This characterization of the solution can be derived from a more general dynamic programming principle for continuous-time stochastic control (see, e.g., Pham 2009, §5.2.1).

The appendix shows that $\tilde{B}$ satisfies the following free boundary problem:

$$0=-\frac{1}{s^2}-\tilde{B}_s(w,s)+\frac{1}{2}\tilde{B}_{ww}(w,s), \quad \text{for all } (w,s)\in\mathscr{C}, \tag{12}$$

$$\tilde{B}(w,s)=D(w,s) \text{ on the boundary } \partial\mathscr{C} \text{ of } \mathscr{C}, \tag{13}$$

$$\tilde{B}_w(w,s)=D_w(w,s) \text{ at regular points of } \partial\mathscr{C}, \tag{14}$$

where $D(w,s)=\max\{0,w\}$ is the (normalized) reward for taking the better of 0 and w when stopping, and one or more variables in the subscript of $\tilde{B}$ or D indicate differentiation with respect to those variables.

By solving for $\tilde{B}$, we find the optimal stopping boundary and the continuation region that it defines, $\mathscr{C}$. Inside $\mathscr{C}$ it is optimal to continue sampling. Outside $\mathscr{C}$ it is optimal to stop sampling and implement the better of the known standard, with reward m, or the alternative whose mean reward is unknown.

The following pair of structural results further characterize $\mathscr{C}$ in terms of a simple function of one variable. First, in Proposition 5, the symmetry of the normal distribution implies an interesting symmetry structure for $\tilde{B}(w,s)-\max\{w,0\}$, the expected benefit of the optimal sampling plan relative to selecting the better of w and 0. Then, Proposition 6 uses this symmetry structure to show that the continuation region is symmetric about $w=0$ and can be described by a function of one variable.

Proposition 5. $\tilde{B}(w,s)-\max\{w,0\}=\tilde{B}(-w,s)-\max\{-w,0\}$.

Proposition 6. *$\mathscr{C}=\{(w,s)\colon |w|<b(s)\}$ for some function $b(s)\geq 0$ for $s\geq 0$.*

The description of $\mathscr{C}$ in Proposition 6 can be directly mapped using Proposition 4 to the first unstandardized continuous-time problem (8) and, via the diffusion approximation, to the original sampling selection problem (6). As a result, the upper and lower optimal stopping boundaries of the original problem in (μ_t, n_t) coordinates can be approximated by

$$m\pm\beta^{-1}b(1/(\gamma n_t)). \tag{15}$$

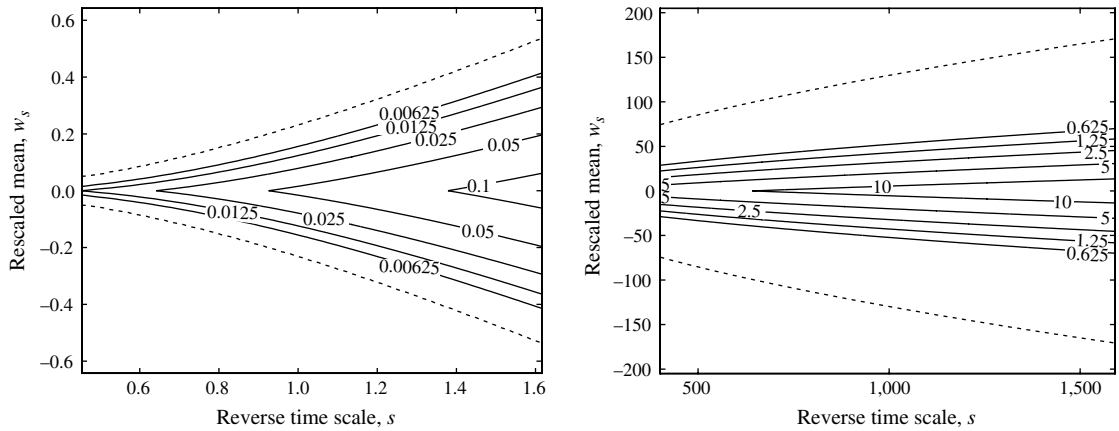
Thus, sampling continues as long as the posterior mean μ_t is in the range $m\pm c^{1/3}\sigma^{2/3}b(\sigma^{2/3}/(c^{2/3}n_t))$.

Interestingly, this optimal continuation set has both an upper and a lower stopping boundary, which differs from the optimal continuation set for the selection problem when there is a positive discount rate. If the discount rate is positive, then there is no lower stopping boundary, and one samples forever with positive probability if the reward m of the known standard is not more than the (nonpositive) discounted reward of sampling forever (Chick and Gans 2009). In contrast, without discounting, a positive sampling cost implies that the penalty for sampling forever is infinite, and if an alternative is sufficiently worse than m, one would prefer to stop sampling and accept a known expected reward of m.

2.2. Empirical Results for Diffusion Approximation when $k=1$

A finite differences (FD) scheme with a trinomial tree was implemented to solve the standardized free boundary problem in (12) to (14) that is central to this paper. There can be a small bias when estimating $\tilde{B}(w,s)$ with FD. That bias can be corrected, to first

Figure 2 **The Continuation Set of the Standardized Free Boundary Problem in (12)–(14) Is Between the Dashed Lines**



Note. The solid lines give level sets of the expected reward of continuing when it is optimal to do so, $\tilde{B}(w,s)-\max(0,w)$.

order, by running the FD with several levels of discretization (values of Δs). We observed a linear bias as Δs was varied over numerous values of Δs and corrected the bias in estimates of $\tilde{B}(w,s)$ reported below, assuming a linear bias.

Figure 2 displays the optimal stopping boundary with dashed lines for two ranges of s. The figure also displays the contours of the expected benefit for continuing to sample when it is optimal to do so, $\tilde{B}(w_s,s)-\max(w_s,0)$. The continuation set has upper and lower stopping boundaries given by Proposition 6. For small values of s (left panel of Figure 2) the upper boundary of the continuation set appears to be convex in s and is narrow. Small values of s indicate that a large number of samples have been observed, so the scaled value βU of the true sampling mean is likely to be close to the scaled posterior mean, w. Only a narrow range of values of w merit additional sampling.

For larger values of s (right panel of Figure 2), the upper boundary of the continuation set appears to be concave in s and includes a wider range of values of w in the continuation region. This is because there is still a great deal of uncertainty about U when s is large, and sampling has more value.

These figures and others like it (not shown) show that the upper boundary $b(s)$ of the optimal continuation set of the standardized diffusion is approximately $\tilde{b}(s)$, where

$$\tilde{b}(s) = \begin{cases} 0.233s^2 & \text{if } s\le 1,\\ 0.00537s^4-0.06906s^3 & \\ \quad +0.3167s^2-0.02326s & \text{if } 1<s\le 3,\\ 0.705s^{1/2}\ln(s) & \text{if } 3<s\le 40,\\ 0.642[s(2\ln(s))^{1.4}-\ln(32\pi)]^{1/2} & \text{if } 40<s. \end{cases} \tag{16}$$

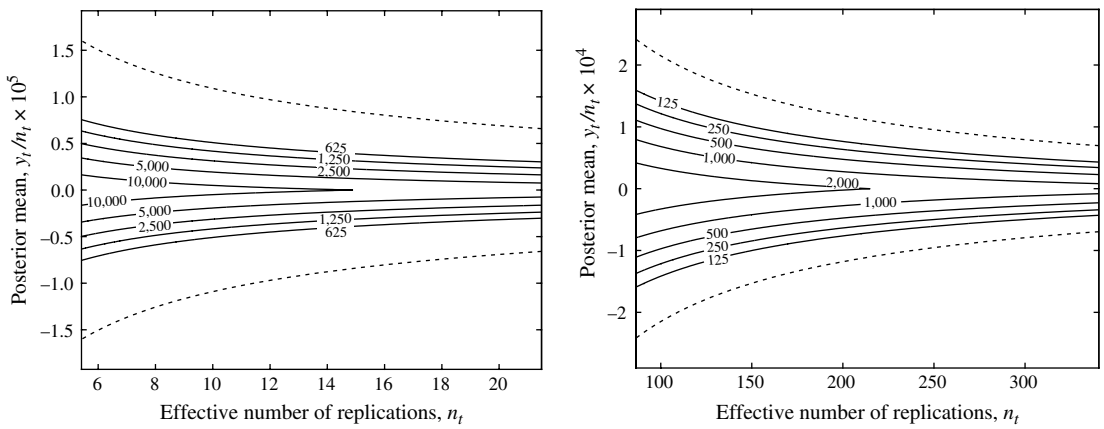
This function is easily computed and is a good approximation to the optimal stopping boundary over the range of values that we tested ($0.08<s<6{,}300$). The form of $\tilde{b}(s)$ is based on numerical approximations and not analytical results. We hypothesize that it is also satisfactorily accurate for $s\in(0,0.08)$.

Figure 3 shows the optimal stopping boundary with dashed lines in coordinates that are more natural for a decision maker (the posterior mean $\mu_t=y_t/n_t$ and effective number of samples n_t, rather than w_s,s). The figure presumes that $c=1$, $\sigma=10^5$, and $m=0$. The contours in Figure 3 describe the expected net benefit of following the optimal policy, rather than stopping. That expected net benefit is $B(m,\mu_t,n_t)-\max\{m,\mu_t\}$. Above the upper dashed line and below the lower dashed line it is optimal to stop immediately. In those regions, the expected net benefit of following the optimal policy is 0, because it is optimal to stop there. Within the continuation set, the benefit of using the optimal stopping boundary increases with the amount of uncertainty about the unknown mean (small n_t) and as the mean of the distribution for the unknown mean approaches that of the known alternative (μ_t approaches m). From the figure, the benefit of sampling optimally when $m=0$, $n_0=6$, and $\mu_0=0$, rather than selecting an alternative without sampling, exceeds 10,000 in this example. That benefit exceeds 30,000 when $n_0=1$ (data not shown).

The approximation in (16) for the diffusion's optimal stopping boundary suggests a continuation set of width $O(1/n_t^2)$ for sufficiently large n_t. At the same time, Proposition 3 says that there is a fixed finite number of samples beyond which it is never optimal to sample. This apparent contradiction is resolved by noting that the bound in Proposition 3 requires that integral numbers of samples be taken, and that the bound in (16) allows for fractional samples.

The optimal stopping boundary in this section is useful for two reasons. First, the PDE need not

Figure 3 **Expected Reward from Continuing to Sample When It Is Optimal to Do So as a Function of the Effective Number of Samples n_t and the Posterior Mean μ_t ($c=1$, $\sigma=10^5$, $m=0$)**



be implemented numerically to solve the problem when $k=1$. Instead, the optimal stopping boundary for the original problem is readily approximated via $\tilde{b}(s)$. Second, the optimal stopping boundary for $k=1$ alternative will be useful for the case of multiple alternatives.

3. Sequential Sampling to Select the Best of Multiple Alternatives

We now present a new sequential sampling procedure for the sampling selection problem of §1 when there are $k>1$ alternatives. The key elements of such a procedure, as exemplified in Figure 4, are (i) a stopping rule that determines when to stop sampling, (ii) an allocation rule that identifies which alternative, $i(t)$, to sample when continuing at times $t=0,1,\ldots$, and (iii) a selection rule that chooses an alternative to implement when sampling has stopped.

The optimal selection rule when $k\geq 1$ is known. Bellman's equation in (5) indicates that, conditional on selecting an alternative for implementation, the selection rule that picks the alternative with the largest posterior mean reward is optimal. We use this selection rule uniformly.

The identification of a stopping rule and allocation rule remains. Although §1 shows that any optimal policy satisfies Bellman's recursion, the curse of dimensionality prevents us from solving it numerically when k is even moderately large. We therefore propose a solution that takes advantage of the structural results and approximations for the case of $k=1$ alternative from §2. This solution, which we call ESP, has two components: a stopping rule and an allocation rule. These rules can either be based on only the stopping boundary $b(\cdot)$, for which we write ESP_b, or on the full solution $\tilde{B}(\cdot)$, for which we write ESP_B.

Figure 4 **A Generic Sequential Sampling Procedure**

1. Specify the parameters and prior distributions for each alternative and set $t=0$.
2. While the stopping rule is not satisfied
   (a) Use the allocation rule to identify which alternative $i(t)\in\{1,2,\ldots,k\}$ to sample.
   (b) Take one sample for alternative $i(t)$, update its statistics, and increment t.
3. Select the alternative $I(T)\in\{0,1,\ldots,k\}$ with the largest posterior mean.

The *ESP_b stopping rule* we now propose assesses whether there is a dynamic policy that allocates all samples to a single alternative before making a selection and that achieves a positive net value of sampling. In other words, the ESP_b stopping rule continues to sample if and only if there is an alternative $i(t)\in\{1,2,\ldots,k\}$ such that comparing that one alternative with an unknown mean with a known value m_t' would result in continuation for the case of $k=1$ alternative. The relevant value of m_t' is the maximum of the posterior expected rewards of the other alternatives, given the information to time t. Based on (15), this is equivalent to continuing to sample if and only if there is an alternative $i>0$ such that

$$c_i^{1/3}\sigma_i^{2/3}b(\sigma_i^{2/3}/(c_i^{2/3}n_{i,t}))>\Delta_{i,t},$$

where $\Delta_{i,t}=|\mu_{i,t}-\max_{j\neq i}\mu_{j,t}|$ is the difference in expected value between alternative i and the best of the other alternatives (including 0) conditional on information up to time t.

The development in §2 for $k=1$ also suggests two allocation rules for the case of $k>1$. The first such allocation is based on the expected net benefit of being able to optimally sample from one alternative before stopping and select an alternative for implementation. Propositions 4 and 5 describe the expected net benefit of the best nonanticipative sampling policy that allocates only to alternative i, in a comparison of alternative i relative to the best of the means of the other

alternatives, relative to the expected reward of stopping immediately to select an alternative for implementation. That expectation is $\tilde{B}(-\beta\Delta_{i,t}, 1/(\gamma n_{i,t}))/\beta$, which gives rise to our first new allocation rule, the ESP_B allocation.

The *ESP_B allocation* samples the alternative $i>0$ that maximizes $\tilde{B}(-\beta\Delta_{i,t}, 1/(\gamma n_{i,t}))/\beta$.

Selecting the largest such figure of merit requires the full solution $\tilde{B}()$ of the standardized PDE. The development in §2 suggests the following allocation, called the ESP_b allocation, as an alternative. It does not require $\tilde{B}()$, is much faster to compute, and is much easier to implement.

The *ESP_b allocation* samples the alternative that is furthest inside the continuation set as measured in standardized coordinates, $\operatorname{argmax}_{i>0} b(\sigma_i^{2/3}/(c_i^{2/3} n_{i,t})) - \Delta_{i,t}/(c_i^{1/3}\sigma_i^{2/3})$.

Thus we have two new procedures, identified by their allocation and stopping rules, respectively: the $\text{ESP}_B, \text{ESP}_b$ procedure and the $\text{ESP}_b, \text{ESP}_b$ procedure. These procedures are equivalent and optimal, up to a diffusion approximation, for $k=1$. When $k>1$, the ESP_b stopping rule considers a strict subset of possible future sampling policies when deciding whether to stop, and therefore may stop prematurely. Furthermore, these two allocations might not sample the optimal choice determined by Bellman's equation when $k>1$ and sampling continues. Both procedures, however, perform extremely well in numerical results presented below. The $\text{ESP}_b, \text{ESP}_b$ procedure is very easy to implement, especially when using the approximation $\tilde{b}$, and so we recommend its use in practice over the more cumbersome $\text{ESP}_B, \text{ESP}_b$ procedure.

The ESP_b and ESP_B allocations do not always require the figures of merit for all alternatives to be recomputed after each sample is observed. For example, if alternative j does not have one of the two largest $\mu_{i,t}$ and is sampled, and the new sample does not make it one of the two best, then only the figure of merit for alternative j needs to be recomputed. This feature can save considerable computation for large values of k.

Step 1 in Figure 4, the specification of parameters and the prior distribution, remains to be discussed. The sampling variance and average cost per sample can be estimated from test simulations during code development. The prior distributions for the unknown means may be elicited using expert judgment (O'Hagan et al. 2006). One may instead assume a noninformative prior distribution for the unknown parameters and observe n_0 samples for each alternative in an initialization phase so that the resulting posterior distribution can be used as the prior distribution for the sequential sampling phase, which begins with resetting $t=0$. For the case of normal samples with unknown mean and known variance, the prior distribution would have a mean that is the sample mean from the initialization phase and a variance of σ^2/n_0.

4. Tools to Assess the Performance of New ESPs

This section first presents several bounds, allocation rules, and stopping rules that can be used as comparators for the newly proposed procedures in §3. We summarize the main intuition for the rules here and refer the reader to the original papers from which they are derived and Chick and Frazier (2009) for further details. They will be used to assess the performance of the new procedure of §3 in numerical experiments below.

4.1. Bounds for the Value Function

We first present bounds on the value function $V^*(\vec{\Theta}_t)$ that provide an upper bound on the expected optimality gap for the procedures tested, as well as an assessment of the improvement that can be expected relative to the naive approach of allocating an equal number of samples per alternative in a round-robin fashion. Additionally, because $V^*(\vec{\Theta}_t)$ is the expected value of developing a simulator (and using an optimal selection algorithm), bounds on $V^*(\vec{\Theta})$ can be used to decide whether to develop a simulator in the first place.

One upper bound on the value function $V^*(\vec{\Theta}_t)$ is the expected reward of having perfect information about each alternative at zero cost. The bound follows directly from Proposition 1.

Proposition 7. $V_{\max}(\vec{\Theta}_t) \triangleq \mathrm{E}[\max_{i=0,1,\ldots,k} U_i \mid \vec{\Theta}_t] \geq V^*(\vec{\Theta}_t)$.

For a lower bound, we note that the value function $V^*(\vec{\Theta}_t)$ of the sampling selection problem is at least as large as the expected reward of any *one-stage allocation* policy. Formally, a one-stage allocation maps each sampling budget $ß \geq \min_i c_i$ to an allocation of those samples across the k alternatives, with a total of $\tau_i = \tau_i(ß) \geq 0$ samples for alternative i, so that $\sum_{i=1}^k c_i \tau_i \leq ß$. We require that each $\tau_i(ß)$ be nondecreasing in ß. We set $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_k)$ and allow ß and τ_i to be real valued for mathematical convenience. In implementations, the values of τ_i will be rounded to integer values.

To describe the reward obtained from a one-stage allocation, we need to describe the distribution of the posterior means that arise from one stage of sampling. Standard results (de Groot 1970) show that the posterior mean $Z_i = \mathrm{E}[X_{i,t+t'+1} \mid \vec{\Theta}_{t+t'}]$ that will result given that τ_i samples will be observed for alternative $i>0$

(of the t' samples total) is a random variable that is distributed as

$$Z_i \sim \texttt{Normal}\left(\mu_{i,t}\frac{\sigma_i^2\tau_i}{n_{i,t}(n_{i,t}+\tau_i)}\right). \tag{17}$$

Note that $Z_0 = m$ almost surely because the mean for alternative 0 is assumed known. The variance of Z_i increases in τ_i for $i > 0$ because more sampling leads to potentially greater changes in the posterior mean.

The following lower bound on the value function arises from checking the so-called equal allocation.

Proposition 8. *If $\tau_i(ß) = \lfloor ß/k \rfloor$ for each i, then*

$$V^*(\vec{\Theta}_t) \geq V_{\min}(\vec{\Theta}_t) \triangleq \sup_{ß \geq \min_i c_i} \mathrm{E}\left[\max_{i=0,1,\ldots,k} Z_i\right] - ß.$$

4.2. Allocation Rules for Comparison

We now summarize allocation rules that have been proposed in the literature so that the new procedure from §3 can be compared with other procedures. Each allocation rule identifies which alternative to sample from next, given a decision to sample at least one more time before stopping to select an alternative.

Except for the equal allocation, these allocation rules each assess the EVI of certain one-stage policies, and some account for the cost of the samples themselves. The allocations differ in that they consider different classes of one-stage policies and approximations to the EVI of sampling when closed forms are not known. Although the class of one-stage policies considered may allow multiple samples when assessing the EVI of sampling, each of these allocations allocates only one sample at a time.

The *equal allocation* samples the k alternatives one at a time in a round-robin fashion, so that each alternative has approximately the same number of samples at any time. This is used as a comparator in many studies.

The *KG_1/LL_1 allocation* greedily allocates one sample at a time ($\tau_i = 1$ with no samples for the others) to a single alternative to maximize the EVI from that sample, under the hypothetical assumption that only one more sample can be collected before an alternative is selected for implementation. This policy is variously known as the knowledge gradient (KG) allocation (Frazier et al. 2008 derived this for known variances) and the LL_1 allocation (Chick et al. 2010 derived this for unknown variances). The name LL_1 connotes linear loss (a synonym for opportunity cost) when one sample is allowed.

The *$KG_ß$ allocation* allocates one sample at a time to the alternative that maximizes the average EVI per sample, assuming that a budget ß of samples will all be allocated to that alternative. This alternative is

$$\operatorname*{argmax}_{i=1,2,\ldots,k} \frac{1}{ß}\tilde{\sigma}_i(\beta)\Psi\left(\frac{\Delta_{i,t}}{\tilde{\sigma}_i(\beta)}\right), \tag{18}$$

where $(\tilde{\sigma}_i(\beta))^2 = \sigma_i^2\beta/(n_{i,t}(n_{i,t}+\beta))$, $\Psi(s) = \phi(s) - s\Phi(-s)$, ϕ is the probability density function (pdf) of the standard normal distribution, and Φ is corresponding cumulative distribution function (cdf). Although a measurement is valued according to the average value obtained over a budget ß of samples, only one sample is allocated at a time. This is a straightforward extension of the KG_1/LL_1 allocation, and may outperform KG_1/LL_1 (Frazier and Powell 2010).

The *KG_* allocation* is similar to $KG_ß$, but recomputes ß separately for each alternative at each time period to maximize the EVI per sample, $\max_{ß\geq 1}(1/ß)\tilde{\sigma}_i(\beta)\Psi(\Delta_{i,t}/(\tilde{\sigma}_i(\beta)))$. Thus, KG_* allocates one sample at a time to the alternative whose EVI per sample is largest, over all deterministic single-alternative budgets. Computing the solution $ß_i^*$ to this optimization problem over ß for each alternative $i > 0$ can be cumbersome, so instead we employ the following asymptotic approximation $\tilde{ß}_i^*$ to $ß_i^*$ as in Frazier and Powell (2010):

$$\tilde{ß}_i^* = c_i \max\left\{1, \frac{n_{i,t}}{4}\left(\frac{\Delta_{i,t}^2 n_{i,t}}{\sigma_i^2} - 1 + \left(\left(\frac{\Delta_{i,t}^2 n_{i,t}}{\sigma_i^2}\right)^2 + 6\frac{\Delta_{i,t}^2 n_{i,t}}{\sigma_i^2} + 1\right)^{1/2}\right)\right\}. \tag{19}$$

The *sequential LL allocation* allocates one sample at a time with a variation of the LL allocation. For a given budget ß, the one-stage policy that minimizes an easily computed upper bound on the EVI of those samples is computed (Chick and Inoue 2001, Corollary 1). The optimal value $ß^*$ of ß that maximizes the resulting EVI of such an allocation, less the sampling cost, is then determined. At each stage $t = 0, 1, \ldots$, the alternative i that is allocated the greatest computational effort by that allocation is selected, $i(t) = \operatorname{argmax}\{c_i\tau_i(ß^*)\}$. Although this sounds like a large number of approximations, we note that a sequential LL allocation, in combination with the $\mathrm{EOC}_{c,k}$ stopping rule in §4.3 below, tied for the "best" performance among a variety of procedures in what is believed to be the largest numerical study of selection procedures to date (Branke et al. 2007). It tied with the OCBA_{LL} allocation (He et al. 2007) with the $\mathrm{EOC}_{c,k}$ stopping rule, which is derived from different principles but can be shown to have certain asymptotic similarities.

In summary, all of the procedures proposed here, with the exception of the equal allocation, have been shown to perform well in numerical experiments elsewhere, compared to several other procedures in the

literature, both frequentist and Bayesian. In some cases we have adapted or extended them slightly to improve their performance in the context of this paper (e.g., to handle differing sampling costs).

4.3. Stopping Rules for Comparison

We now summarize stopping rules that have been proposed in the literature so that the new procedure from §3 can be compared with other procedures. Each stopping rule can be thought of as checking whether a different, tractable class of one-stage allocation policies brings enough expected reward to justify continuing to sample as opposed to stopping and selecting an alternative for implementation.

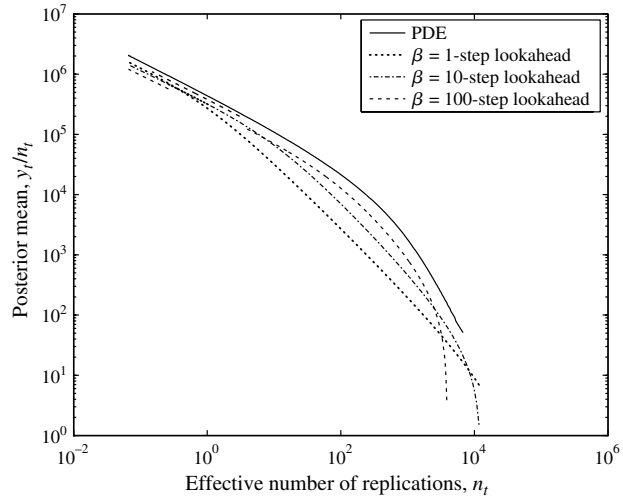
The $EOC_{c,k}$ *stopping rule* considers allocations across more than one alternative when deciding whether to stop sampling. It is analogous to a similar rule in Chick and Gans (2009) for discounted rewards. It checks whether a certain upper bound on the EVI from continuing to sample using the one-stage LL allocation above exceeds the cost of those samples, for at least one sampling budget ß. That bound is based on the EVI of each of k pairwise comparisons of the current best alternative with each other alternative.

The $KG_{ß}$ *stopping rule* we consider slightly extends the knowledge gradient stopping rule of Frazier and Powell (2008) by allowing for different sampling costs and multiple samples. With this rule, sampling continues if and only if there is an alternative i such that allocating the whole sampling budget to that alternative in a single stage of sampling results in a net expected benefit as compared to not sampling at all; that is, it assesses the EVI from observing $\tau_i = ß$ samples from alternative i and no samples from the others, and continues if and only if the EVI exceeds ß.

We also consider the KG_* *stopping rule* of Frazier and Powell (2010). This rule continues to sample if there exists *any* sampling budget $ß \geq \max_i c_i$ for which the $KG_{ß}$ stopping rule would continue. It therefore allows sampling to continue in more situations than does the $KG_{ß}$ stopping rule (because $KG_{ß}$ uses a fixed ß.)

Figure 5 shows the relationship of the continuation sets for these stopping rules for the special case of $k=1$. The figure presumes that $c=1$, $\sigma^2=10^5$, and $m=0$. The optimal stopping boundary somewhat exceeds the KG_* stopping boundary, and the relative error percentage increases somewhat as the effective number of samples n_t increases. The KG_* continuation set coincides, theoretically, with the $EOC_{c,k}$ continuation set for the special case of $k=1$. The KG_* continuation set in turn contains the continuation sets of the $KG_{ß}$ stopping rules. The nesting of the continuation sets is explained by the nesting of the sets of policies by which each stopping rule is defined. Those continuation sets are not nested for ß = 1, 10, and 100.

Figure 5 **Upper Portion of Stopping Boundaries with $k=1$ Alternative and $c=1$, $\sigma=10^5$, and $m=0$**



5. Numerical Examples: Sequential Sampling for Selection with Known Variances

Simulation results are presented first for $k=1$ and then for $k>1$. With $k=1$ we compare the known optimal result for the diffusion with the approximations in the above stopping rules to assess their degree of suboptimality. With $k>1$, we assess the performances of the various procedures by comparing them with bounds on their performance. In summary, the new ESP approach allows for improvements beyond the performance of any other procedure we have found in the literature, at least for these experiments.

The expected reward of a selection procedure was estimated with Monte Carlo (MC) experiments. In each experiment, a large number of random problem instances were generated from the prior distribution $\vec{\Theta}_0$ for the unknown parameters of each alternative. Selection procedures were applied to each problem instance. Results from each procedure are compared with each other and with the theoretical upper bound $V_{\max}(\vec{\Theta}_0)$ from Proposition 7 and the benchmark value $V_{\min}(\vec{\Theta}_0)$ of the one-stage equal allocation from Proposition 8.

We implemented the ESP_b stopping rule, the ESP_B and ESP_b allocations, and all comparators in Matlab and Java with one exception. Because the ESP_B allocation is more cumbersome to implement and slower to compute than other allocations, we did not implement it in Java. Results for the ESP_b stopping rule and ESP_b allocation as well as for the comparator procedures in §4 were consistent with these two implementations, confirming correct implementation. The results in this section examine the ESP_b stopping rule, the ESP_b allocation, and the comparators. We assess the performance of the ESP_B allocation in §6.4 below.

Table 1 **Performance of Stopping Rules for $k=1$, $c=1$, $\mu_0=0$, and $\sigma=10^5$ Calculated Using Monte Carlo Simulation with 10^6 Samples**

Stopping rule	E[cT]	E[OC]	E[$cT+OC$]	E[Reward]	Suboptimality (%)
			Panel A		
ESP_b	64.33±0.13	32.8±0.4	97.1±0.4	39,797.1±0.4	—
$\text{EOC}_{c,k}$	29.90±0.06	127.0±1.2	156.9±1.2	39,737.3±1.2	0.15
KG_* ($\tilde{ß}^*$)	29.39±0.06	128.6±1.2	158.0±1.2	39,736.2±1.2	0.15
KG_1	6.88±0.01	1,103.1±5.4	1,110.0±5.4	38,784.3±5.4	2.54
			Panel B		
ESP_b	321.85±0.25	263.0±1.0	584.9±1.0	3,404.6±1.0	—
$\text{EOC}_{c,k}$	142.53±0.11	612.1±1.8	754.6±1.8	3,234.8±1.8	4.99
KG_* ($\tilde{ß}^*$)	136.51±0.11	633.9±1.9	770.4±1.9	3,219.0±1.9	5.45
KG_1	10.53±0.01	2,504.5±4.6	2,515.0±4.6	1,474.4±4.6	56.69

Notes. For panel A, the PDE estimate is $V^*(0,0,n_0)\approx B(0,0,n_0)=3{,}407$, and the bounds are $V_{\max}=3{,}989$ and $V_{\min}=3{,}170$. In panel B, the PDE estimate is $V^*(0,0,n_0)\approx B(0,0,n_0)=39{,}800$, and the bounds are $V_{\max}=39{,}894$ and $V_{\min}=39{,}613$. In panel A, $n_0=1$, and in panel B, $n_0=100$.

5.1. Numerical Results: One Sampled Alternative

Table 1 shows MC simulation results for these stopping rules when there is $k=1$ simulated alternative. We estimated (plus or minus one standard error) the expected sampling cost, E[cT], the expected opportunity cost, $\text{E}[OC]=\text{E}[\max_{i=0,1,\ldots,k} U_i - U_{I(T)} \mid \Theta_0]$, and the total expected penalty for not knowing the means, E[$cT+OC$]. The expected reward, E[Reward] $=\text{E}[-cT+U_{I(T)} \mid \Theta_0]$, was estimated indirectly by $V_{\max}(\vec{\Theta}_0)-\text{E}[cT+OC]$.

We set $c=1$, $\mu_0=0$, $\sigma=10^5$, and $m=0$ for both $n_0=1$ and $n_0=100$. The relative suboptimality of the stopping rules is calculated in a comparison with the MC value of the reward with the optimal stopping boundary. In each test, the optimal stopping rule (as computed via the approximate stopping boundary $\tilde{b}(\cdot)$ in (16)) performs best, with $\text{EOC}_{c,k}$ second best, followed by KG_* and KG_1 in this order. This ordering in reward, opportunity cost, and number of samples is explained by the nesting of the continuation regions as in §4.3. We can make several observations from these results.

First, the diffusion approximation $B(0,0,n_0)$ of the value of the optimal policy $V^*(0,0,n_0)$ is very close to the simulated value of the optimal stopping rule, ESP_b, as approximated with $\tilde{b}(\cdot)$. When $n_0=1$, the relative difference is $2.9/39{,}800=0.007\%$, and when $n_0=100$, the relative difference is $2.4/3{,}407=0.07\%$. Thus, the approximation $\tilde{b}(\cdot)$ for the stopping boundary appears satisfactory. Furthermore, $B(0,0,n_0)$ is slightly higher than the MC estimates based on discrete-time sampling, as expected, because the diffusion allows stopping at real-valued times but the MC simulation enforces integer-valued stopping times.

Second, the KG_* and $\text{EOC}_{c,k}$ stopping rules, both of which test a range of one-stage allocations to assess whether to continue or not, perform very well, especially when n_0 is very small. They provide a significant benefit relative to the KG_1 stopping rule, which examines only one one-stage allocation. The benefit of stopping by considering a range of potential stopping rules is therefore evident. (The KG_* and $\text{EOC}_{c,k}$ stopping rules are equivalent when $k=1$, but the numerical results differ slightly because the implementation of KG_* uses the approximation $\tilde{ß}^*$ in (19) for the optimal ß, whereas $\text{EOC}_{c,k}$ optimizes the budget numerically.)

Third, the poor performance of KG_1 when $n_0=100$ can be explained by recalling Figure 5. When t is in the range from 80 to 3,000, the KG_1 stopping boundary is rather lower than the optimal stopping boundary. That is the range of t that is most likely to lead to stopping when $n_0=100$, given the value of E[cT].

5.2. Numerical Results: Multiple Sampled Alternatives

Table 2 displays the performance of various allocation and stopping rules for $k>1$. Here and elsewhere, procedures are written with the allocation rule first and the stopping rule second, e.g., KG_*,ESP_b is the KG_* allocation rule with the ESP_b stopping rule. The table does not report the expected percentage of suboptimality, because the optimal expected reward is not known precisely when $k>1$. Small values of E[$cT+OC$] correspond to better sequential sampling procedures. For reference, the table also reports the penalty for the optimal one-stage equal allocation, $V_{\max}-V_{\min}$, and the expected reward from perfect and costless information, $V_{\max}$. We now summarize general conclusions supported by the data.

Most strikingly, the ESP_b stopping rule performs best of all stopping rules for any given allocation rule. It outperforms the KG stopping rules because it considers a broader class of stopping policies. It outperforms the EOC-based stopping rules, we hypothesize, because the EOC-based stopping rules consider allocations that spread samples across the k alternatives, which in turn incur sampling costs that are detrimental to a comparison between a given alternative and

Table 2 **Expected Total Penalty for Not Knowing the Mean Rewards,** $\mathrm{E}[cT+OC]$**, for Several Allocation and Stopping Rules Calculated Using Monte Carlo Simulation with** 10^6 **Samples (for** $k=2,\ldots,20$**) or** 10^5 **Samples (for** $k=50,100$**)**

Allocation stop rule	$k=2$	$k=3$	$k=5$	$k=10$	$k=20$	$k=50$	$k=100$
KG_1, KG_1	2,393±9	3,508±12	5,136±15	7,140±18	8,767±19	10,862±67	12,500±72
KG_*, KG_* ($\tilde{\beta}^*$)	426±3	674±4	1,040±5	1,445±6	1,761±6	2,245±23	2,666±25
LL, $\mathrm{EOC}_{c,k}$	320±2	429±2	577±3	821±4	1,095±4	1,577±17	2,168±22
Equal, $\mathrm{EOC}_{c,k}$	321±2	433±2	629±2	1,040±3	1,815±3	4,220±16	8,425±29
KG_*, $\mathrm{EOC}_{c,k}$	319±2	424±2	575±3	799±3	1,057±4	1,489±16	2,027±19
KG_1, $\mathrm{EOC}_{c,k}$	318±2	419±2	546±3	728±3	916±3	1,223±11	1,577±11
KG_1, ESP_b	231±1	348±2	506±2	694±3	875±3	1,158±10	1,516±10
KG_*, ESP_b	228±1	327±2	458±2	600±2	722±3	905±8	1,111±9
ESP_b, ESP_b	233±1	344±2	505±2	700±3	856±3	1,075±11	1,308±12
$V_{\max}-V_{\min}$	522	736	1,116	1,947	3,463	7,799	14,891
$V_{\max}$	68,104	88,815	116,970	153,887	186,748	224,907	250,759

Notes. The $V_{\max}-V_{\min}$ (the expected total penalty of the optimal one-stage equal allocation) and $V_{\max}$ (the expected reward of perfect costless information) are also reported. Here, $m=0$, $c_i=1$, $\sigma_i=10^5$, and $n_{i,0}=1$ for $i=1,2,\ldots,k$, and estimates are plus or minus one standard error.

the best of the rest. For each value of k, the best three allocation–stopping rule pairs use the ESP_b stopping rule, even though these allocations perform poorly with some other stopping rules. This highlights the importance of a good stopping rule.

The ESP_b stopping rule does very well with the ESP_b allocation and even better with the KG_* allocation. These combinations are better than the most effective procedures (LL or OCBA_{LL} or $\mathrm{KG}_1/\mathrm{LL}_1$ allocation with the $\mathrm{EOC}_{c,k}$ stopping rule) from a large and recent empirical study (Branke et al. 2007).

The equal allocation samples more than the other procedures, for any given stopping rule, because it does not prioritize alternatives for sampling. This is demonstrated for the $\mathrm{EOC}_{c,k}$ stopping rule in the table.

The KG_* allocation with the KG_* stopping rule performs only somewhat better than the optimal one-stage equal allocation (whose penalty for not knowing the mean rewards is in the row labeled $V_{\max}-V_{\min}$). The KG_1 stopping rule with the KG_1 allocation performs the poorest of the sampling procedures. For $k\leq 50$, it is even worse than the optimal one-stage equal allocation.

The differences in performance between the policies grows with k. As k grows, the approximations made by the KG_* and KG_1 stopping rules, which consider allocations to only a single alternative, become more restrictive and less justified. They cause the difference in values between the best single-alternative allocation and the best allocation overall to become larger. The $\mathrm{EOC}_{c,k}$ stopping rule considers allocations that sample multiple alternatives. This explains why KG_1 and KG_* stop earlier than $\mathrm{EOC}_{c,k}$ as k increases.

Although the KG_1 and KG_* *stopping* rules did not perform well, the corresponding KG_1 and KG_* *allocation* rules perform very well when paired with other stopping rules. For example, KG_*, ESP_b was the best of all the procedures in this test, and the KG_1 and KG_* allocations perform better than all other allocation rules when paired with the $\mathrm{EOC}_{c,k}$ stopping rule. This dichotomy is a consequence of the KG method's strength in estimating the *differences* in value between different sampling choices and its weakness in estimating the *overall* value of sampling, that is, it is important for an allocation rule to correctly estimate the ordering of the expected values of sampling more than the expected values themselves.

5.3. Numerical Results: Special Configurations

The performance reported in Table 2 is the average performance over a large number of different configurations drawn from the prior distribution. On any given configuration, however, the relative ordering of the policies may differ. To assess this variability, we also examined performance on variants of the slippage configuration (with $u_i=u_1-\delta$ for $i=2,3,\ldots,k$ and some $\delta>0$) and the monotone decreasing means configuration (with $u_i=u_1-\delta(i-1)$ for $i=2,3,\ldots,k$), which are commonly studied in the literature. To summarize our results, we found that on the slippage configuration, the LL, $\mathrm{EOC}_{c,k}$ procedure performs well, perhaps because the $\mathrm{EOC}_{c,k}$ stopping rule slightly overestimates the Bayesian value of continuing, which can be advantageous on difficult configurations. On the monotone decreasing means configuration, procedures with the ESP_b stopping rule, and in particular the KG_*, ESP_b procedure, perform best. Although the relative ordering of the policies varies with configuration, we recommend ESP_b, ESP_b and KG_*, ESP_b for use because of their excellent performance in the average case.

6. Sequential Sampling for Selection When Variances Are Unknown

The general setup in §1 assumed normally distributed samples with known variances. Much of the analysis

in that section can be extended to account for samples with other distributions. Propositions 1 and 2 and cited results from Bertsekas and Shreve (1978, Chap. 9) hold more generally under the following conditions: (i) a finite set of sufficient statistics for each X_i exists so that we can define a fixed-length parameter vector $\mathbf{\Theta}_{i,t}$ for the posterior on the unknown parameters R_i of alternative i's sampling distribution at time t; (ii) the samples $X_{i,t}$ are conditionally independent given R_i; (iii) the unknown parameters $\mathbf{R}=(R_1, R_2, \ldots, R_k)$ are a priori independent across alternatives, $P(\mathbf{R})=\prod_{i=1}^{k} P(R_i)$; (iv) the integral $\mathrm{E}[X_{i,t}]=\int\int X_{i,t}\, dP_{X_{i,t}|\Theta_{i,0}}$ exists and is finite.

This section uses this extension to account for normally distributed samples with unknown variance, which is much more common in practice than having known variances. In doing so, we will refer to several properties of the Student t distribution. Let $\phi_\nu(x)$ denote the pdf of a standard Student t distribution with ν degrees of freedom, let $\Phi_\nu(x)$ denote the corresponding cdf, and let $\Psi_\nu[s]=\int_s^\infty (x-s)\phi_\nu(x)\,dx$ denote the standard Student t linear loss function. One can show that $\Psi_\nu[s]=((\nu+s^2)/(\nu-1))\phi_\nu(s)-s\Phi_\nu(-s)$ for $\nu>1$. If T_ν is a standard Student t random variable with ν degrees of freedom, then we say that $\mu+T_\nu/\sqrt{\kappa}$ is a three parameter Student t random variable, denoted $\mathtt{St}(\mu,\kappa,\nu)$, with precision κ. When $\nu>2$, the variance is $\kappa^{-1}\nu/(\nu-2)$.

6.1. Special Case: Normally Distributed Samples with Unknown Variances

We now suppose that the samples $X_{i,t}$ are normally distributed and conditionally independent, given the unknown means and unknown variances. Let ς_i be the random variable that represents the unknown variance and whose realization is σ_i^2. Then, conditioned on the unknown means U_i and unknown variances ς_i,

$$\{X_{i,t}\colon t=1,2,\ldots\}\,|\,U_i,\varsigma_i \overset{\text{iid}}{\sim} \mathtt{Normal}(U_i,\varsigma_i) \qquad \text{for } i=1,2,\ldots,k.$$

We further presume that the prior distribution for each unknown mean and variance is in the family of conjugate priors for normally distributed samples with unknown means and variances (de Groot 1970),

$$\begin{aligned} \varsigma_i &\sim \mathtt{InvGamma}(\xi_{i,0},\chi_{i,0}),\\ U_i\,|\,\varsigma_i &\sim \mathtt{Normal}(\mu_{i,0},\varsigma_i/\eta_{i,0}), \end{aligned} \tag{20}$$

where $\xi_{i,0}>1$ and $\chi_{i,0}>0$ are shape and scale parameters, respectively, of an inverted gamma distribution with a finite mean $\mathrm{E}[\varsigma_i]=\chi_{i,0}/(\xi_{i,0}-1)$ and variance $\mathrm{Var}[1/\varsigma_i]=\xi_{i,0}/\chi_{i,0}^2$, and where $\mu_{i,0}$ and $\eta_{i,0}$ describe the a priori mean and variance, respectively, of the unknown sampling mean. It follows that U_i is an $\mathtt{St}(\mu_{i,0},\xi_i\eta_{i,0}/\chi_{i,0},2\xi_{i,0})$ random variable. We further presume that the prior distributions are independent from one i to the next.

In general, the prior distribution for the unknown means and variances can be specified by selecting parameters $\xi_{i,0},\chi_{i,0},\eta_{i,0}>0$ and $\mu_{i,0}$ for each i. If each $\xi_{i,0}$ exceeds 1, then the a priori variance of each U_i exists. An alternative method for obtaining a prior distribution is to use a reference prior distribution (Bernardo and Smith 1994) together with an initial stage of data. That corresponds to observing τ_0 samples for each alternative and then setting $\mu_{i,0}=\bar{x}_{i,\tau_0}$, $\xi_{i,0}=(\tau_0-1)/2$, $\chi_{i,0}=\hat{\sigma}^2{}_i(\tau_0-1)/2$, where $\hat{\sigma}^2{}_i=\sum_{j=1}^{\tau_0}(x_{i,j}-\bar{x}_{i,\tau_0})^2/(\tau_0-1)$, and $\eta_{i,0}=\tau_0$. One would then reset $t=0$ and start the sampling procedure.

With the conjugate prior distribution in (20), the posterior distribution has the same form. If $\mathbf{x}_t$ is the vector of samples after a total of t samples have been observed, and $l_{i,t}$ of those samples have been from alternative $i>0$, the posterior distribution of the unknown parameters is

$$\begin{aligned} \varsigma_i\,|\,\mathbf{x}_t &\sim \mathtt{InvGamma}(\xi_{i,t},\chi_{i,t}),\\ U_i\,|\,\varsigma_i,\mathbf{x}_t &\sim \mathtt{Normal}(\mu_{i,t},\varsigma_i/\eta_{i,t}), \end{aligned} \tag{21}$$

where $\bar{x}_{i,t}$ is the average of the $l_{i,t}$ samples for alternative i, $\xi_{i,t}=\xi_{i,0}+l_{i,t}/2$,

$$\chi_{i,t}=\chi_{i,0}+\frac{1}{2}\left(\frac{\eta_{i,0}l_{i,t}}{\eta_{i,0}+l_{i,t}}(\mu_{i,0}-\bar{x}_{i,t})^2+\sum_{j=1}^{l_{i,t}}(x_{i,j}-\bar{x}_{i,t})^2\right),$$

$$\eta_{i,t}=\eta_{i,0}+l_{i,t}, \quad \text{and} \quad \mu_{i,t}=(\eta_{i,0}\mu_{i,0}+l_{i,t}\bar{x}_{i,t})/\eta_{i,t}$$

(de Groot 1970).

The bounds $V_{\min}(\vec{\Theta})\le V^*(\vec{\Theta})\le V_{\max}(\vec{\Theta})$ from Propositions 7 and 8 carry over directly under the assumption that the U_i are distributed as in (21). The proofs of those bounds do not rely on the sampling distributions. The random variables Z_i required for $V_{\min}$ can be described in closed form. Suppose that t' additional samples are to be observed, with τ_i samples to be observed for alternative i so that $t'=\sum_i\tau_i$. Bernardo and Smith (1994, p. 440) show that the posterior mean $Z_i=\mathrm{E}[X_{i,t+\tau_i+1}\,|\,\vec{\Theta}_t]$, given that τ_i samples will be observed for alternative $i>0$, has a distribution analogous to (17), namely,

$$Z_i\sim\mathtt{St}\left(\mu_{i,t},\frac{\xi_{i,t}\eta_{i,t}(\eta_{i,t}+\tau_i)}{\chi_{i,t}\tau_i},2\xi_{i,t}\right). \tag{22}$$

For alternative $i=0$, which has a known mean m, we have $\theta_0=\mu_{0,t}=m$, $U_0=m$, and $Z_0=m$ as before.

6.2. One Alternative with an Unknown Variance

This subsection adapts our analysis from §2 with $k=1$ to handle unknown sampling variances. We drop the

subscript i in this subsection. The posterior distribution for the unknown parameters of the alternative at time t is therefore $\Theta_t = (\xi_t, \chi_t, \mu_t, \eta_t)$ for $t = 0, 1, \ldots$, where $\eta_t = \eta_0 + t$ is the effective number of samples.

The diffusion approximation in §2 provides a convenient policy with plug-in estimators. We estimate the unknown sampling variance after t samples are observed with its posterior expectation, $\mathrm{Var}[X_{t+1} \mid \Theta_t] = \mathrm{E}[\varsigma_i \mid \Theta_t] = \chi_t/(\xi_t - 1)$. The optimal stopping boundary used by ESP_b is approximated by recalling (15) and substituting $\chi_t/(\xi_t - 1)$ for σ^2 and η_t for n_t, so that sampling continues as long as the posterior mean μ_t remains in the interval defined by

$$m \pm c^{1/3}\left(\frac{\chi_t}{\xi_t - 1}\right)^{1/3} b\left(\left(\frac{\chi_t}{\xi_t - 1}\right)^{1/3} \Big/ (c^{2/3}\eta_t)\right). \tag{23}$$

These substitutions require that $\eta_t > 0$ and $\xi_t > 1$ for $t = 0, 1, \ldots$. If $\eta_0 > 0$ or $\xi_0 > 1$ are not satisfied (e.g., because of the selection of an improper prior distribution), then a finite stopping boundary is not initially defined, and sampling should continue at least until $\xi_t > 1$ and $\eta_t > 0$.

6.3. Multiple Alternatives with Unknown Variances

The ESPs from §3 for the case of $k > 1$ alternatives with known variances are easily adapted to the case of unknown variances using the plug-in estimators $\chi_t/(\xi_t - 1)$ for σ^2 and η_t for n_t, as in (23). The bounds for comparison and the allocation and stopping rules in §4 can also be readily adapted to the case of unknown variances. To adapt the KG-type rules, we use these plug-in estimators. In particular, for KG_*, we substitute these plug-in estimators into the expression (19) for $\tilde{ß}_i^*$, and then again substituting $\tilde{ß}_i^*$ for ß in (18). For LL and $\mathrm{EOC}_{c,k}$, the papers that derived these rules treated unknown variances directly.

6.4. Numerical Analysis with Unknown Variances

6.4.1. One Alternative. The approximation in §6.2 for the optimal stopping boundary when the sampling variance is unknown and when $k = 1$ appears to be effective. We report here results when the prior distribution for the unknown variance is $\mathtt{InvGamma}(\xi_{\eta_0} = 10, \chi_{\eta_0} = 9 \times 10^{10})$, so the expectation of the variance is 10^{10}. The prior distribution for the unknown mean, given a variance ς_i, is $\mathtt{Normal}(\mu_{\eta_0} = 0, \varsigma_i/\eta_0)$, with $\eta_0 = 5$. Samples cost $c = 1$. The reward of the known standard is $m = U_0 = 0$. This determines $\vec{\Theta}_0$.

With these parameters, the maximum one could hope to obtain, in expectation, is $V_{\max}(\vec{\Theta}_0) = 17{,}598$. When the estimated stopping boundary in §6.2 is used, the expected sampling cost is $\mathrm{E}[cT] = 119.2 \pm 0.56$, and the expected opportunity cost is $\mathrm{E}[OC] = \mathrm{E}[\max_j U_j - U_{I(T)} \mid \Theta_T] = 73.2 \pm 34$, for an expected net reward of $V_{\max}(\vec{\Theta}_0) - \mathrm{E}[cT + OC] = 1.7403 \times 10^4 \pm 34$, as estimated by Monte Carlo with 10^5 samples. This uses fewer samples and incurs less opportunity cost, in expectation, than the optimal one-stage equal allocation, which observes 206 samples and has an expected opportunity cost of 209.7 and an expected net reward of $V_{\min}(\vec{\Theta}_0) = 17{,}180$.

6.4.2. Several Alternatives. Table 3 summarizes the performance of the allocations and stopping rules when the variance is unknown and when there are $k > 1$ alternatives. The bounds for the value function ($V_{\max}$ and $V_{\min}$) are estimated with Monte Carlo methods (3×10^6 samples for a relative standard error of the means of 0.06%). The performance of $\mathrm{ESP}_b, \mathrm{ESP}_b$ is assessed with $50{,}000/k$ samples. The performance of the other procedures is estimated with at least $15{,}000/k$ samples. The standard error for the penalty for not knowing the mean ($\mathrm{E}[cT + OC]$) of $\mathrm{ESP}_b, \mathrm{ESP}_b$ for $k = 100$ is 281, and the relative error decreases with k.

Most of the conclusions made in the known variance case remain true in the unknown variance case. The ESP_b stopping rule is better than all of the other stopping rules for any given allocation rule. The gap between the reward possible with perfect information and the best one-stage equal allocation, $V_{\max} - V_{\min}$,

Table 3 **Expected Total Penalty for Not Knowing the Mean Rewards,** $\mathrm{E}[cT + OC]$**, When Variances Are Unknown, for Several Allocation and Stopping Rules Calculated Using Monte Carlo simulation**

Allocation stop rule	$k=2$	$k=3$	$k=5$	$k=10$	$k=20$	$k=50$	$k=100$
$\mathrm{KG}_1/\mathrm{LL}_1, \mathrm{KG}_1$	4,041	5,950	9,155	13,059	17,654	26,085	37,113
$\mathrm{KG}_*, \mathrm{KG}_*(\tilde{ß}^*)$	701	994	1,447	2,317	2,389	3,238	3,315
$\mathrm{LL}, \mathrm{EOC}_{c,k}$	549	791	1,149	1,432	2,151	2,483	2,963
$\mathrm{KG}_*, \mathrm{EOC}_{c,k}$	552	873	1,081	1,595	2,095	2,435	3,679
$\mathrm{KG}_1, \mathrm{ESP}_b$	399	558	754	1,227	1,897	3,660	5,740
$\mathrm{KG}_*, \mathrm{ESP}_b$	366	555	774	965	1,420	1,984	2,747
$\mathrm{ESP}_b, \mathrm{ESP}_b$	376	545	791	1,088	1,607	1,915	2,335
$\mathrm{ESP}_B, \mathrm{ESP}_b$	393	543	759	1,010	1,423	2,066	3,198
$V_{\max} - V_{\min}$	781	1,066	1,609	2,600	4,051	7,062	10,340
$V_{\max}$	30,110	39,350	52,120	69,180	84,900	104,320	117,930

Notes. Here, $m = 0$, $c_i = 1$, $\mu_{i,0} = 0$, $\eta_{i,0} = 5$, $\xi_{i,0} = 10$, and $\chi_{i,0} = 9 \times 10^{10}$ so that $\mathrm{E}[\varsigma_i] = (10^5)^2$ for $i = 1, 2, \ldots, k$. The $V_{\max} - V_{\min}$ and $V_{\max}$ are also reported.

also increases with k. If that gap is sufficiently small in the eyes of the analyst, as it might be for small k, and predictability is required in the number of samples collected, then the optimal one-stage equal allocation can be used. Otherwise, the ESP_b stopping rule is recommended. The size of the penalty for not knowing the means, $\mathrm{E}[cT+OC]$, is better for that stopping rule, relative to other stopping rules, for larger k.

One substantive difference between the numerical results for the two cases of known and unknown sampling variances is the performance of the two best allocation rules with the ESP_b stopping rule. For the example with the known sampling variance in §5.2, the KG_* allocation is best, with the ESP_b allocation second best (with other sampling policies being less effective). Here, the KG_* and ESP_b allocations are equally good (within statistical error), with a relative improvement in ESP_b for larger k. The reason appears to be that $\tilde{ß}^*$, as computed using plug-in estimators in (19), is relatively further from the true optimal $ß^*$.

We also assess the ESP_B allocation, which has longer run times than the other procedures because of numerical extrapolation of the value of $\tilde{B}(w,s)$ for various w,s. Conceptually, the ESP_B allocation should be at least as good as the ESP_b allocation. In Table 3, the $\mathrm{ESP}_B,\mathrm{ESP}_b$ procedure is tied for best with the $\mathrm{KG}_*,\mathrm{ESP}_b$ and $\mathrm{ESP}_b,\mathrm{ESP}_b$ procedures, within the sampling noise of the experiments (less than 1.5 standard errors). The performance of ESP_B might be improved by refining the FD grid from which a numerical extrapolation of $\tilde{B}(w,s)$ is made or with future research on theoretical approximations to its value. We do not recommend ESP_B until a more computationally efficient implementation becomes available. Instead, we recommend the $\mathrm{ESP}_b,\mathrm{ESP}_b$ procedure because it is easy to implement using (16) and approximately as effective in these experiments. The $\mathrm{KG}_*,\mathrm{ESP}_b$ procedure is also very effective.

7. Conclusion

The vast preponderance of selection procedures have been concerned with statistical thresholds for sampling and stopping and have ignored the cost of sampling. In many applications, however, the cost of sampling is important. This paper uses the cost of sampling and the expected economic benefit of the alternative that is ultimately selected to sequentially determine which alternatives to sample and when to stop to implement an alternative. This is accomplished by examining a broad class of nonanticipative sampling policies, not just one-stage policies as is common in the literature.

When comparing $k=1$ alternative with normally distributed samples and a known sampling variance to a known mean, the ESP_b stopping rule is optimal, up to an asymptotic approximation. For $k>1$ alternatives, the optimal policy is shown to exist but is unknown. We provide somewhat suboptimal policies for $k>1$, with or without known variances, that are motivated by the $k=1$ optimality analysis and are more effective in our numerical examples than any other procedure of which we are aware. With their readily computed approximations, the $\mathrm{ESP}_b,\mathrm{ESP}_b$ and $\mathrm{KG}_*,\mathrm{ESP}_b$ procedures are also simple to implement. Because of their excellent performance and ease of implementation, we wholeheartedly recommend their use in practice.

From a practical perspective, there is only one main parameter to assess: the cost per sample as measured in the same units as the output of the simulation. Such sampling costs are a more direct economic means to control sequential sampling and selection than are techniques that ignore sampling costs.

Acknowledgments

The authors thank Assaf Zeevi and the anonymous reviewing team for their constructive feedback.

Appendix. Mathematical Proofs

This appendix contains proofs of the mathematical results that were omitted in the main text. The proofs of some results are written to handle a more general case than that of normal distributions with unknown means. In particular, we use the notation for more general distributions that was introduced at the start of §6.

Before presenting the proofs, we note some formalism about the definition of a policy in §1. A policy π is a sequence of universally measurable kernels, where the kernel at time t gives the distribution of the random variables $i(t)$ (and therefore of T) as a function of the history up to time t, as summarized by $\vec{\Theta}_t$. The values of Θ_i are presumed to be in some space Ω_{Θ_i}. We define Π to be the set of all such policies. We allow universally measurable kernels (which include those that are Borel measurable) because uncountably large Ω_{Θ_i} create the possibility that the value function will be universally measurable but not Borel measurable (Bertsekas and Shreve 1978, Chap. 7).

Proof of Proposition 1. Let $\mathcal{F}=(\mathcal{F}_t)_{t\geq 0}$ be the filtration generated by $(\vec{\Theta}_t)_{t\geq 0}$. By construction, T is a stopping time of this filtration. Beginning with the definition of the sampling selection problem in (3) and conditioning on $\mathcal{F}_T$ and $U_1,U_2,\ldots,U_k$, the tower property of conditional expectation provides

$$V^\pi(\vec{\Theta}_0)=\mathrm{E}_\pi\left[\sum_{t=0}^{T-1}-c_{i(t)}\ +\ U_{I(T)}\,\middle|\,\vec{\Theta}_0\right].$$

We then add and subtract $\mathrm{E}[\max_i U_i]$, which is finite by the integrability of each U_i and does not depend on π. Thus

$$V^\pi(\vec{\Theta}_0)\ =\ \mathrm{E}\left[\max_i U_i+|\,\vec{\Theta}_0\right]$$
$$+\mathrm{E}_\pi\left[\sum_{t=0}^{T-1}-c_{i(t)}+U_{I(T)}-\max_i U_i\,\middle|\,\vec{\Theta}_0\right].$$

Observing that $U_{I(T)}-\max_j U_i=L_{I(T)}$ completes the proof.

Lemmas and Definitions for Proof of Proposition 3

Here and throughout the proof of Proposition 3, $\vec{\Theta}$ represents a generic prior or posterior distribution, to which we may assign any $\vec{\Theta}_0$ or $\vec{\Theta}_t$. Before proving Proposition 3, we state a few definitions and lemmas. Let $\Pi_i = \{\pi \in \Pi\colon \mathbb{P}^\pi\{i(0)=i, T>0\}=1\}$ be the set of measurement policies that sample alternative i at time 0. Then, the so-called Q-factor for first sampling alternative i and behaving optimally afterward may be written as follows:

$$Q(\vec{\Theta}, i) = \sup_{\pi\in\Pi_i} \mathrm{E}_\pi\left[-\sum_{t=0}^{T-1} c_{i(t)} + U_{I(T)} \,\middle|\, \vec{\Theta}_0 = \vec{\Theta}\right].$$

For any given $i \in \{1, 2, \ldots, k\}$, let Π_i' be the set of policies that measure alternative i at time 0 and whose decisions $i(t)$ at each subsequent time $t = 1, 2, \ldots$ depend on both U_i and $\vec{\Theta}_t$. Thus, Π_i' is similar to the set of policies Π_i except that policies in Π_i' may depend on the additional piece of information U_i. Then define

$$A(\vec{\Theta}, i) = \sup_{\pi\in\Pi_i'} \mathrm{E}_\pi\left[-\sum_{t=0}^{T-1} c_{i(t)} + U_{I(T)} \,\middle|\, \vec{\Theta}_0 = \vec{\Theta}\right].$$

This definition is identical to that of $Q(\vec{\Theta}, i)$ except the supremum is taken over policies in Π_i', rather than Π_i. We may interpret $A(\vec{\Theta}, i)$ as the best value possible in a problem in which one is forced to first measure alternative i once, revealing the *true* value U_i of this alternative, and subsequent measurements are made with noise as in the original problem to discover the values of the other alternatives. Notice that i is fixed in $A(\vec{\Theta}, i)$, and thus the true value of only this single alternative is revealed.

Now define $J^\pi(\vec{\Theta}, i)$ to be the value of policy π in a problem where alternative i has a *fixed* value $\mu_{i,0}$ so that

$$J^\pi(\vec{\Theta}, i) = \mathrm{E}_\pi\left[-\sum_{t=0}^{T-1} c_{i(t)} + \max\Big(\mu_{i,0}, \max_{j\neq i}\mu_{j,T}\Big) \,\middle|\, \vec{\Theta}_0 = \vec{\Theta}\right].$$

Lemma 1. $\sup_{\pi\in\Pi} J^\pi(\vec{\Theta}, i) - c_i + \mathrm{E}[(U_i - \mu_{i,0})^+ \mid \vec{\Theta}_0 = \vec{\Theta}] \geq A(\vec{\Theta}, i)$ *for any given* $i \in \{1, 2, \ldots, k\}$.

Proof of Lemma 1. Observe that the objective function

$$-\sum_{t=0}^{T-1} c_{i(t)} + \max\Big(\mu_{i,0}, \max_{j\neq i}\mu_{j,T}\Big), \tag{24}$$

of which $J^\pi(\vec{\Theta}, i)$ is an expectation, is independent of U_i given $\mu_{i,0}$ because beliefs are independent across alternatives.

Because the objective function (24) does not depend on U_i, the supremum of its expected value over all policies depending on U_i, $\sup_{\pi\in\Pi_i'} J^\pi(\vec{\Theta})$, is unchanged if we restrict to the policies in Π_i (the policies in Π_i' without explicit dependence on U_i); that is, $\sup_{\pi\in\Pi_i'} J^\pi(\vec{\Theta}, i) = \sup_{\pi\in\Pi_i} J^\pi(\vec{\Theta}, i)$.

Furthermore, this lack of dependence on U_i implies that $\sup_{\pi\in\Pi_i} J^\pi(\vec{\Theta}, i) = \sup_{\pi\in\Pi} J^\pi(\vec{\Theta}, i) - c_i$ because the initial measurement of alternative i required by membership in Π_i carries only the cost c_i, and the observation of alternative i can be ignored. Combining this with $\sup_{\pi\in\Pi_i'} J^\pi(\vec{\Theta}, i) = \sup_{\pi\in\Pi_i} J^\pi(\vec{\Theta}, i)$ provides

$$\sup_{\pi\in\Pi_i'} J^\pi(\vec{\Theta}, i) = \sup_{\pi\in\Pi} J^\pi(\vec{\Theta}, i) - c_i. \tag{25}$$

Now, for any fixed policy $\pi \in \Pi_i'$, we have

$$\begin{aligned}
&J^\pi(\vec{\Theta}, i) + \mathrm{E}\big[(U_i - \mu_{i,0})^+ \mid \vec{\Theta}_0 = \vec{\Theta}\big] \\
&= \mathrm{E}_\pi\left[-\sum_{t=0}^{T-1} c_{i(t)} + \max\Big(\mu_{i,0}, \max_{j\neq i}\mu_{j,T}\Big) + (U_i - \mu_{i,0})^+ \,\middle|\, \vec{\Theta}_0 = \vec{\Theta}\right] \\
&\geq \mathrm{E}_\pi\left[-\sum_{t=0}^{T-1} c_{i(t)} + \max\Big(\mu_{i,0} + (U_i - \mu_{i,0})^+, \max_{j\neq i}\mu_{j,T}\Big) \,\middle|\, \vec{\Theta}_0 = \vec{\Theta}\right] \\
&\geq \mathrm{E}_\pi\left[-\sum_{t=0}^{T-1} c_{i(t)} + \max\Big(U_i, \max_{j\neq i}\mu_{j,T}\Big) \,\middle|\, \vec{\Theta}_0 = \vec{\Theta}\right].
\end{aligned}$$

Take the supremum of both sides of the inequality over all policies $\pi \in \Pi_i'$. The left-hand side becomes $\sup_{\pi\in\Pi_i'} J^\pi(\vec{\Theta}, i) + \mathrm{E}[(U_i - \mu_{i,0})^+ \mid \vec{\Theta}_0 = \vec{\Theta}]$, whereas the right-hand side becomes $A(\vec{\Theta}, i)$. Replacing $\sup_{\pi\in\Pi_i'} J^\pi(\vec{\Theta}, i)$ using (25) shows the claimed inequality. Q.E.D.

Lemma 2. $\sup_{\pi\in\Pi} J^\pi(\vec{\Theta}, i) \leq V^*(\vec{\Theta})$.

Proof of Lemma 2. Fix i, and let $\Pi'' = \{\pi \in \Pi\colon i(t) \neq i \ \forall t \ P_\pi\text{-a.s.}\}$ be the set of policies that never measure alternative i. Under the objective function $-\sum_{t=0}^{T-1} c_{i(t)} + \max(\mu_{i,0}, \max_{j\neq i}\mu_{j,T})$, of which $J^\pi(\vec{\Theta}, i)$ is an expectation, measurements of alternative i incur a cost c_i but have no benefit because $\max(\mu_{i,0}, \max_{j\neq i}\mu_{j,T})$ is unimproved by them. Thus, any policy $\pi \in \Pi$ has a corresponding policy $\pi'' \in \Pi''$ (obtained by skipping decisions by π to measure alternative i, but otherwise behaving identically) for which $J^{\pi''}(\vec{\Theta}, i) \geq J^\pi(\vec{\Theta}, i)$. Thus,

$$\sup_{\pi\in\Pi} J^\pi(\vec{\Theta}, i) = \sup_{\pi\in\Pi''} J^\pi(\vec{\Theta}, i).$$

For any policy $\pi \in \Pi''$, the posterior distribution on the value of alternative i is the same as the prior distribution, so $\mu_{i,T} = \mu_{i,0}$. Thus,

$$\begin{aligned}
J^\pi(\vec{\Theta}, i) &= \mathrm{E}_\pi\left[-\sum_{t=0}^{T-1} c_{i(t)} + \max\Big(\mu_{i,T}, \max_{j\neq i}\mu_{j,T}\Big) \,\middle|\, \vec{\Theta}_0 = \vec{\Theta}\right] \\
&= V^\pi(\vec{\Theta}).
\end{aligned}$$

The supremum over a larger set results in a value at least as large as the supremum over a smaller set and $\Pi'' \subset \Pi$, so

$$\begin{aligned}
\sup_{\pi\in\Pi} J^\pi(\vec{\Theta}, i) = \sup_{\pi\in\Pi''} J^\pi(\vec{\Theta}, i) &= \sup_{\pi\in\Pi''} V^\pi(\vec{\Theta}) \\
&\leq \sup_{\pi\in\Pi} V^\pi(\vec{\Theta}) = V^*(\vec{\Theta}). \quad \text{Q.E.D}
\end{aligned}$$

Proof of Proposition 3. Let π^* be any optimal policy. We will show that $T \leq k + \sum_{i=1}^k [\sigma_i^2/(2\pi c_i^2) - n_{i,0}]$ almost surely under $\mathbb{P}^\pi$. Observe that for any fixed $i = 1, 2, \ldots, k$,

$$\begin{aligned}
Q(\vec{\Theta}, i) &\leq A(\vec{\Theta}, i) \leq \sup_{\pi\in\Pi} J^\pi(\vec{\Theta}, i) + \mathrm{E}[(U_i - \mu_{i,0})^+ \mid \vec{\Theta}_0 = \vec{\Theta}] - c_i \\
&\leq V^*(\vec{\Theta}) + \mathrm{E}[(U_i - \mu_{i,0})^+ \mid \vec{\Theta}_0 = \vec{\Theta}] - c_i.
\end{aligned}$$

The first inequality follows by noting that $Q(\vec{\Theta}, i)$ and $A(\vec{\Theta}, i)$ are both supremums of the same quantity but over different sets, with $\Pi_i \subset \Pi_i'$. The second and third inequalities follow from Lemmas 1 and 2, respectively.

Consider any $\vec{\Theta}$ for which $\mathrm{E}[(U_i - \mu_{i,0})^+ \mid \vec{\Theta}_0 = \vec{\Theta}] < c_i$. The inequality (26) above implies $Q(\vec{\Theta}, i) < V^*(\vec{\Theta})$. Because Bellman's optimality equation tells us that $V^*(\vec{\Theta})$ is the maximum of $Q(\vec{\Theta}, j)$ over all $j = 0, 1, \ldots, k$, we must have $Q(\vec{\Theta}, j) > Q(\vec{\Theta}, i)$ for some $j \neq i$. This implies that the policy π^* chooses not to measure alternative i when the current posterior distribution is $\vec{\Theta}$. Instead, it either stops measuring, or it measures some alternative other than i.

Consider further the inequality $\mathrm{E}[(U_i - \mu_{i,t})^+ \mid \vec{\Theta}_t] < c_i$. By applying the preceding argument to $\vec{\Theta}_t$ instead of $\vec{\Theta}$, we see that π^* will not measure alternative i when this inequality is satisfied. We have,

$$\mathrm{E}[(U_i - \mu_{i,t})^+ \mid \vec{\Theta}_t] = \sigma_{i,t} \mathrm{E}\left[\left(\frac{U_i - \mu_{i,t}}{\sigma_{i,t}}\right)^+ \middle| \vec{\Theta}_t\right] = \sigma_{i,t}/\sqrt{2\pi},$$

because $(U_i - \mu_{i,t})/\sigma_{i,t}$ is a standard normal random variable under P_i, and the expectation of the positive part of a standard normal random variable is $1/\sqrt{2\pi}$. Thus, the inequality $\mathrm{E}[(U_i - \mu_{i,t})^+ \mid \vec{\Theta}_t] < c_i$ holds iff $\sigma_{i,t}/\sqrt{2\pi} < c_i$, which holds iff $n_{i,t} > (\sigma_i^2/2\pi c_i^2)$, where we have noted the definition $\sigma_{i,t}^2 = \sigma_i^2/n_{i,t}$ and rearranged terms. This implies that we never measure alternative i when $n_{i,t} > (\sigma_i^2/2\pi c_i^2)$. Because $n_{i,t}$ increases by 1, and only when alternative i is measured, we have that $n_{i,t} \leq 1 + (\sigma_i^2/2\pi c_i^2)$ for all t.

The total number of measurements at a particular time is the sum of the number of times each individual alternative has been measured; that is, $t = \sum_{i=1}^k n_{i,t} - n_{i,0}$ for $t \leq T$. Applying this uniform bound on $n_{i,t}$ to the time $t = T$ provides $T = \sum_{i=1}^k n_{i,T} \leq \sum_{i=1}^k (1 + (\sigma_i^2/2\pi c_i^2) - n_{i,0})$, the claimed deterministic upper bound on T. Q.E.D.

Derivation of Free Boundary Problem in (12) to (14)

Recall that the choice $\beta^2\sigma^2\gamma = 1$ causes (W_s, s) to be a Brownian motion without drift and with unit volatility in the $-s$ scale, going from (w_{s_0}, s_0) back to time $s = 0$.

We now turn to the supremum $\sup_{S \in [0, s_0]} \beta^{-1} \mathrm{E}[-((1/S) - (1/s_0)) + \max\{0, W_S\} \mid W_{s_0} = w_{s_0} - \tilde{m}]$ in (11). The supremum exists because the expectation over which the supremum is taken is finite for at least one stopping time, $S = s_0$ almost surely, and is uniformly bounded above by $\mathrm{E}[\max\{0, W_0\} \mid W_{s_0} = w_{s_0} - \tilde{m}]$. This uniform bound is derived by dropping the sampling cost $-((1/S) - (1/s_0))$ and noting that $\max\{0, W_s\}$ is a submartingale in the $-s$ time scale, implying that $\sup_{S \in [0, s_0]} \mathrm{E}[\max\{0, W_S\} \mid W_{s_0} = w_{s_0} - \tilde{m}]$ is attained when $S = 0$ almost surely.

We now set $w = w_{s_0} - \tilde{m}$ and $s = s_0$ and check whether (w, s) is in the continuation set or not. If (w, s) is in the continuation set, then by definition $S < s$ almost surely (one stops at a time less than s, going in reverse time). Because (w, s) is in the continuation set, we can examine $\tilde{B}(w, s)$ by examining the evolution of the process from time s to time $s - h$ for some small $h > 0$. While sampling for a time h, a cost of $1/(s-h) - 1/s = h/(s(s-h))$ is incurred, and the process goes to $(w + \Delta w, s - h)$ if it is not stopped, for some Δw whose distribution is $\mathtt{Normal}(0, h)$ by properties of a reverse time Brownian motion. Therefore,

$$\begin{aligned}
\tilde{B}(w, s) &= -\frac{h}{s(s-h)} + \mathrm{E}[\tilde{B}(w + \Delta w, s - h) \mid w_s = w, s] + o(h) \\
&= -\frac{h}{s(s-h)} + \mathrm{E}\Big[\tilde{B}(w, s) - h\tilde{B}_s(w, s) + \Delta w \tilde{B}_w(w, s) \\
&\qquad + \frac{(\Delta w)^2}{2} \tilde{B}_{ww}(w, s) + o(h) + o((\Delta w)^2)\Big] \\
&= -\frac{h}{s(s-h)} + \tilde{B}(w, s) - h\tilde{B}_s(w, s) + \frac{h}{2}\tilde{B}_{ww}(w, s) + o(h),
\end{aligned}$$

where the $o(h)$ in the first equality is due to the potential of stopping before h time elapses, the second equality comes from the usual Taylor series expansion for functions of a Brownian motion, and the third equality follows by examining the first two moments of the random variable Δw. We subtract $\tilde{B}(w, s)$ from both sides, divide through by h, and let $h \to 0$ to get the claimed heat equation in (12), $0 = -(1/s^2) - \tilde{B}_s(w, s) + (1/2)\tilde{B}_{ww}(w, s)$.

To get the boundary conditions, note first that the reward in the expectation of (12) simplifies to $\max\{0, w\}$ for points (w, s) outside of the continuation set, because the term $((1/S) - (1/s))$ equals zero when stopping immediately. On the boundary one is indifferent between stopping and continuing. Thus, if $D(w, s) = \max(0, w)$, then $\tilde{B}(w, s) = D(w, s)$ on the boundary. Chernoff (1961, §§4 and 6) provides a heuristic argument for the "smooth pasting" condition, $\tilde{B}_w(w, s) = D_w(w, s)$, when the boundary is regular (when there are not multiple values of w on the boundary for a given s). Bather (1970) formalized Chernoff's (1961) argument for a broad class of optimal stopping problems. Q.E.D.

Proof of Proposition 5. We show the result by first showing that the expected reward for stopping in a given continuation set is closely related to the expected reward for stopping in the mirror image of that continuation set. For a given stopping rule S that stops sampling only outside of a given continuation set, consider the stopping rule S' that stops at $(-w, s)$ if and only if S stops at (w, s). Then for $w_{s_0} \geq 0$ we have

$$\begin{aligned}
&\tilde{B}(w_{s_0}, s_0) \\
&\quad = \sup_{S \in [0, s_0]} \mathrm{E}_S\left[-\left(\frac{1}{S} - \frac{1}{s_0}\right) + \max\{0, W_S\} \middle| W_{s_0} = w_{s_0}\right] \\
&\quad = \sup_{S \in [0, s_0]} \mathrm{E}_S[W_S \mid W_{s_0} = w_{s_0}] \\
&\qquad + \mathrm{E}_S\left[-\left(\frac{1}{S} - \frac{1}{s_0}\right) - \min\{0, W_S\} \middle| W_{s_0} = w_{s_0}\right] \\
&\quad = w_{s_0} + \sup_{S \in [0, s_0]} \mathrm{E}_S\left[-\left(\frac{1}{S} - \frac{1}{s_0}\right) - \min\{0, W_S\} \middle| W_{s_0} = w_{s_0}\right] \\
&\quad = w_{s_0} + \sup_{S \in [0, s_0]} \mathrm{E}_{S'}\left[-\left(\frac{1}{S'} - \frac{1}{s_0}\right) + \max\{0, W_{S'}\} \middle| W_{s_0} = -w_{s_0}\right] \\
&\quad = w_{s_0} + \tilde{B}(-w_{s_0}, s_0),
\end{aligned}$$

where the first equality is by definition and the second equality follows by the linearity of expectations and the

observation that $x = \max\{0, x\} + \min\{0, x\}$ for all x. The third equality follows because $\mathrm{E}_S[W_S \mid W_{s_0} = w_{s_0}] = w_{s_0}$ does not depend on S, because the process is a martingale on the compact set $[0, s_0]$. The fourth equality follows because $-\min\{0, x\} = \max\{0, -x\}$ and $W_S = -W_{S'}$ by construction and the fact that the statistics of $W_S - w_{s_0}$ are the same as for $-(W_S - w_{s_0})$ by properties of the normal distribution. The last equality follows because a supremum over $S \in [0, s_0]$ is the same as the supremum over $S' \in [0, s_0]$.

Thus, the claimed property is true for $w_{s_0} \geq 0$. The case of $w_{s_0} < 0$ follows similarly. Q.E.D.

Proof of Proposition 6. By the arguments in Proposition 5, the optimal stopping boundary is symmetric about $w = 0$, so that $(w, s) \in \mathscr{C}$ if and only if $(-w, s) \in \mathscr{C}$ for $s \geq 0$ (because the choice of $s_0 > 0$ is arbitrary). Therefore, we can prove the claim by showing that, for any $w, a \geq 0$, $(w, s + a) \in \mathscr{C}$ implies that $(w, s) \in \mathscr{C}$.

Fix some $w_{s_0}, a \geq 0$ and suppose $(w_{s_0} + a, s_0) \in \mathscr{C}$. This implies the existence of a stopping rule S of W in the $-s$ scale such that $S < s_0$ almost surely and $\mathrm{E}[-((1/S) - (1/s_0)) - \min\{0, W_S\} \mid W_{s_0} = w_{s_0} + a] \geq w_{s_0} + a$. We write

$$
\begin{aligned}
0 &\leq \mathrm{E}\left[-\left(\frac{1}{S} - \frac{1}{s_0}\right) + \max\{0, W_S\} \,\middle|\, W_{s_0} = w_{s_0} + a\right] - (w_{s_0} + a) \\
&= \mathrm{E}_S\left[-\left(\frac{1}{S} - \frac{1}{s_0}\right) - \min\{0, W_S + a\} \,\middle|\, W_{s_0} = w_{s_0}\right] - (w_{s_0} + a) \\
&\leq \mathrm{E}_S\left[-\left(\frac{1}{S} - \frac{1}{s_0}\right) - \min\{0, W_S\} \,\middle|\, W_{s_0} = w_{s_0}\right] - w_{s_0}.
\end{aligned}
$$

The second line is due to the fact that the stochastic process $\{W_s\colon s \in [0, s_0]\}$ with $W_{s_0} = w_{s_0} + a$ is equal in distribution to $\{W_s + a\colon s \in [0, s_0]\}$ with $W_{s_0} = w_{s_0}$. The third line is due to $\min\{0, W_S + a\} - a \leq \min\{0, W_S\}$. Because $S < s_0$ almost surely, the final line implies that $(w_{s_0}, s_0) \in \mathscr{C}$. Thus, $(|w| + a, s) \in \mathscr{C}$ implies that $(|w|, s) \in \mathscr{C}$ for all $a \geq 0$. Thus we can define $b(s) = \inf\{w : w > 0 \text{ and } (w, s) \notin \mathscr{C}\}$ as required. By construction, $b(s) \geq 0$. Q.E.D.

Proof of Proposition 8. The proof follows directly from the original problem definition in (3), the definition of Z_i in (17), and by noting that, upon stopping, $Z_{I(T)} = \mathrm{E}[X_{I(T),T+1} \mid \vec{\Theta}_t]$ is the posterior mean of the alternative that has the highest posterior mean after sampling stops at time $T = t + t'$. Q.E.D.

References

Bather, J. 1970. Optimal stopping problems for Brownian motion. *Adv. Appl. Probab.* **2** 259–286.

Bechhofer, R. E., T. J. Santner, D. M. Goldsman. 1995. *Design and Analysis for Statistical Selection, Screening, and Multiple Comparisons.* John Wiley & Sons, New York.

Bernardo, J. M., A. F. M. Smith. 1994. *Bayesian Theory.* Wiley, Chichester, UK.

Bertsekas, D. P., S. E. Shreve. 1978. *Stochastic Optimal Control: The Discrete Time Case.* Academic Press, Belmont, MA.

Branke, J., S. E. Chick, C. Schmidt. 2007. Selecting a selection procedure. *Management Sci.* **53**(12) 1916–1932.

Brezzi, M., T. L. Lai. 2002. Optimal learning and experimentation in bandit problems. *J. Economic Dynam. Control* **27**(1) 87–108.

Chan, H. P., T. L. Lai. 2006. Sequential generalized likelihood ratios and adaptive treatment allocation for optimal sequential selection. *Sequential Analysis* **25**(2) 179–201.

Chernoff, H. 1961. Sequential tests for the mean of a normal distribution. *Proc. Fourth Berkeley Sympos. Math. Statist. Probab.*, Vol. 1, University of California Press, Berkeley, CA, 79–91.

Chernoff, H., S. N. Ray. 1965. A Bayes sequential sampling inspection plan. *Ann. Math. Statist.* **36**(5) 1387–1407.

Chick, S. E., P. Frazier. 2009. The conjunction of the knowledge-gradient and the economic approach to simulation selection. M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, R. G. Ingalls, eds. *Proc. Winter Simulation Conf.*, IEEE, Piscataway, NJ, 528–539.

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

Chick, S. E., K. Inoue. 2001. New two-stage and sequential procedures for selecting the best simulated system. *Oper. Res.* **49**(5) 732–743.

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

de Groot, M. H. 1970. *Optimal Statistical Decisions*. McGraw-Hill, New York.

Frazier, P. I., W. B. Powell. 2008. The knowledge-gradient stopping rule for ranking and selection. S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler, eds. *Proc. Winter Simulation Conf.*, IEEE, Piscataway, NJ, 305–312.

Frazier, P. I., W. B. Powell. 2010. Paradoxes in learning and the marginal value of information. *Decision Anal.* **7**(4) 378–403.

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

Gupta, S. S., K. J. Miescke. 1996. Bayesian look ahead one-stage sampling allocations for selecting the best population. *J. Statist. Planning Inference* **54** 229–244.

Gupta, S. S., K. Nagel, S. Panchapakesan. 1979. *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations.* Wiley, New York.

Hammersley, J. M., D. C. Hanscomb. 1964. *Monte Carlo Methods.* Methuen, London.

He, D., S. E. Chick, C.-H. Chen. 2007. The opportunity cost and OCBA selection procedures in ordinal optimization for a fixed number of alternative systems. *IEEE Trans. Systems, Machines, Cybernetics C: Appl. Reviews* **37**(5) 951–961.

Kim, S.-H., B. L. Nelson. 2006. Selecting the best system. S. G. Henderson, B. L. Nelson, eds. *Handbooks in Operations Research and Management Science: Simulation.* Elsevier, Amsterdam, 501–534.

Lai, T. -L. 1987. Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15**(3) 1091–1114.

Law, A. M. 2007. *Simulation Modeling and Analysis,* 4th ed. McGraw-Hill, New York.

O'Hagan, A., C. Buck, A. Daneshkhah, J. Eiser, P. Garthwaite, D. Jenkinson, J. Oakley, T. Rakow. 2006. *Uncertain Judgements: Eliciting Experts' Probabilities.* John Wiley & Sons, Chichester, UK.

Pham, H. 2009. *Continuous-Time Stochastic Control and Optimization With Financial Applications*. Springer, Berlin.

Siegmund, D. 1985. *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag, New York.