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

Online Companion For:
Economic Analysis of Simulation Selection Problems

Appendix A provides additional background that describes the multi-armed bandit problem and the relationship of the simulation selection problem to a stoppable version of the multi-armed bandit. It also provides a numerical example that shows that the few existing results that characterize optimal policies for stoppable bandits do not apply to the simulation selection problem.

Appendix B motivates the free boundary equation whose solution approximates the optimal expected discounted reward when $k = 1$. Appendix C provides mathematical proofs of the claims in the main paper.

Appendix D describes several technical extensions that expand the range of validity of the paper. It relaxes some assumptions about the independence of the output from a single system, as well as the duration of the replications for each alternative.

Appendix E summarizes how the optimal expected discounted reward (OEDR) and stopping boundaries for the simulation selection problem with $k = 1$ alternative were computed. Appendix F specifies the simulation selection procedures that are used in §6.3.

Appendix A: Supplement: Multi-Armed Bandits and the Simulation Selection Problem

The simulation selection problem is closely related to a class of sequential decision problem known as the multi-armed bandit problem. In this section, we review relevant theory, and we apply the theory to demonstrate that simulation selection problems can be reduced to a variation of multi-armed bandits that is called a stoppable bandit problem. We then present a numerical example that indicates that well-known sufficient conditions, used to justify the optimality of indexed-based rules in stoppable-bandit problems, do not hold in our case.

A.1. The Multi-Armed Bandit Problem

This section supplements the discussion in §3 by providing formal definitions of the multi-armed bandit problem and of optimal allocation index rules.

In the discounted multi-armed bandit problem, a decision-maker chooses repeatedly among a finite set of mutually-independent Markov chains that are indexed $i = 1, 2, \ldots, k$. A choice of chain i at stage t yields an expected reward that is specific to the state of chain i, and it initiates a state transition for chain i. The $k-1$ chains not chosen at stage t remain in their current states and earn no rewards. The objective is to maximize the expected sum of discounted rewards over an infinite horizon (Gittins 1989).

For the case in which expected one-period rewards are bounded for each chain, Gittins and co-workers proved that an index can be computed for each arm, independently of all other arms, such that it is optimal to select the arm whose index is greatest among all arms. This allocation index has come be known as a "Gittins index."

Formally, we define the multi-armed bandit's parameters as follows. Markov chain i has state space Ω_{Θ_i}, with states $\Theta_i \in \Omega_{\Theta_i}$. The state space has σ-algebra, $\mathcal{F}_i$, of subsets of Ω_{Θ_i}, which includes all elements $\Theta_i \in \Omega_{\Theta_i}$. We define the product space of joint outcomes across all k Markov chains as $(\Omega, \mathcal{F})$. If chain i is chosen at time t, so that $i(t) = i$, then chain i advances according to an $\mathcal{F}_i$– and $\mathcal{F}$–measurable 1-step transition law

$$P_i(\Theta_{i,t+1} \,|\, \Theta_{i,t}) \tag{EC.1}$$

and earns chain i's transition-dependent expected reward, defined by the similarly measurable function

$$R_t \;=\; R_i(\Theta_{i,t}). \tag{EC.2}$$

Alternatively, if chain i is not chosen at time t, then $\Theta_{i,t+1} \equiv \Theta_{i,t}$ and chain i provides no reward.

An *allocation policy* is decision rule for making the infinite sequence of choices $\{i(1), i(2), \ldots\}$ and we let Ξ be the set of all $\mathcal{F}$–measurable non-anticipative allocation policies. Given initial states $\boldsymbol{\Theta} = (\Theta_1, \Theta_2, \ldots, \Theta_k)$ and one-period discount rate $0 \le \Delta < 1$, the choice of allocation policy $\xi \in \Xi$ yields

$$V^{\xi}(\boldsymbol{\Theta}) \;=\; \mathrm{E}_{\xi}\left[\sum_{t=0}^{\infty} \Delta^t R_t \,|\, \boldsymbol{\Theta}_0 = \boldsymbol{\Theta}\right], \tag{EC.3}$$

when the expectation exists. The "$\boldsymbol{\Theta_0} = \boldsymbol{\Theta}$" in (EC.3) highlights the expected discounted value's dependence on the initial set of prior states. An optimal allocation policy, $\xi^* \in \Xi$, maximizes the expected discounted value: $V^{\xi^*}(\boldsymbol{\Theta}) = \sup_{\xi \in \Xi}(V^{\xi}(\boldsymbol{\Theta}))$.

For the case in which expected one-period rewards are bounded, so that $R_i(\Theta_i) < \infty$ for almost all $\Theta_i \in \Omega_{\Theta_i}$, $i = 1, 2, \ldots, k$, Gittins and co-workers proved two important sets of results which are relevant for our problem. First, Gittins and Jones (1974) demonstrated that there exists a state-dependent index for each arm, $G_i(\Theta_i)$, which is independent of all other arms, such that it is optimal to choose at each stage, t, the arm whose index is the greatest among all arms. Second, Gittins and Glazebrook (1977) and Gittins (1979) demonstrated that this so-called Gittins index has an appealing form. Let

$$G_i(\Theta_i, s) \;=\; \left(\frac{\mathrm{E}\left[\sum_{t=0}^{s-1} \Delta^t R_{i(t)}(\Theta_{i(t),t}) \,|\, \Theta_{i,0} = \Theta_i\right]}{\mathrm{E}\left[\sum_{t=0}^{s-1} \Delta^t \,|\, \Theta_{i,0} = \Theta_i\right]} \right), \tag{EC.4}$$

for some random stopping time $s > 0$. Then the Gittins index for an arm in state Θ_i, $G_i(\Theta_i)$, is the supremum of (EC.4) among all such stopping times:

$$G_i(\Theta_i) \;=\; \sup_{s>0} G_i(\Theta_i, s). \tag{EC.5}$$

In words, the Gittins index is the supremum of the expected discounted value per unit of discounted time over all stopping times $s > 0$. Gittins (1979) demonstrates that there exists an optimal stopping time such that the supremum in (EC.5) is achieved and that, by playing the arm with the highest index at each time t, the decision maker maximizes the expected discounted value defined in (EC.3).

In Appendix A.2 we demonstrate how the simulation selection problem reduces to a multi-armed bandit when the selection problem's stopping policies are fixed. We then discuss the optimality of a hierarchical family of controls for this class of problems.

A.2. Simulation Selection as a "Stoppable Bandit" Problem

In the simulation selection problem, the application of any reasonable (or more generally, stationary) stopping policy, π_i, to project i induces it to behave as a simple Markov reward process: in each period in which a system is simulated, Bayes' rule governs that system's state transition, and a reward of $-c_i$ is earned; and in each period in which a system is implemented, the state transition returns to the current state with probability one (equivalently, there is no state transition), and the one-period reward is that associated with actual payout from the selected system. Therefore, if each of the k projects' stopping problems is *a priori* defined to be operated according to a specific reasonable policy, then the stoppable bandit effectively behaves as a traditional multi-armed bandit problem, and a Gittins index result holds (Glazebrook 1979, Corollary 1). That is, given the application of fixed set of stopping policies, $\{\pi_1, \pi_2, \ldots, \pi_k\}$, an allocation index exists, such that at each stage is it optimal to select the project, i, with the largest index and then simulate or implement according to π_i. This fact is true for "stoppable bandits" in general.

Thus, a natural class of policies to consider for the simulation selection problem is that of *hierarchical policies*. In a hierarchical policy, $\xi(\pi_1, \pi_2, \ldots, \pi_k)$, project $i = 1, 2, \ldots, k$ is operated according to a reasonable policy, π_i. Given a set of fixed π_i, the system is operated as a $k+1$ armed bandit with policy ξ (the extra arm corresponding to the "do nothing" option). Given the use of specific reasonable (or more generally, stationary) policies π_i for projects $i = 1, 2, \ldots, k$, the optimal policy for the resulting $k + 1$ armed bandit uses the Gittins-index rule, as in (EC.5). We denote that optimal policy for the given π_i as $\xi^*(\pi_1, \pi_2, \ldots, \pi_k)$.

A special example of a hierarchical policy, $\xi^*(\pi_1^*, \pi_2^*, \ldots, \pi_k^*)$, uses the stationary stopping policies, π_i^*, that are optimal for each of the k individual projects defined in (3) and then uses the Gittins-index rule for the $k + 1$ armed bandit problem that results from the use of the π_i^*. After the π_i^* are determined, this is implemented at each time t by 1) calculating each stopping problem's Gittins index, assuming that policy π_i^* is applied to problem i starting in state $\Theta_{i,t}$; and then 2) selecting the arm i, with the highest Gittins index, and operating it according to π_i^* for one period (i.e., simulate or implement system i).

While attractive, the policy $\xi^*(\pi_1^*, \pi_2^*, \ldots, \pi_k^*)$ need not be optimal for stoppable bandit problems. It may be the case that, while hierarchical policies are optimal, the optimal stopping policies do not produce the right Gittins indices; that is, there exists some $\{\pi_1, \pi_2, \ldots, \pi_k\}$ such that $\xi^*(\pi_1, \pi_2, \ldots, \pi_k)$ outperforms $\xi^*(\pi_1^*, \pi_2^*, \ldots, \pi_k^*)$. Even worse, it may be the case that hierarchical policies themselves are not optimal; that is, there is no set of stopping policies $\{\pi_1, \pi_2, \ldots, \pi_k\}$ such that $\xi^*(\pi_1, \pi_2, \ldots, \pi_k)$ is optimal. Nevertheless, Glazebrook (1979) provides sufficient conditions under which $\xi^*(\pi_1^*, \pi_2^*, \ldots, \pi_k^*)$ is optimal for stoppable bandit problems.

We restate Glazebrook (1979)'s results for the special case of the simulation selection problem. In the lemma below, we let T_i denote the random simulation stopping time associated with stationary simulation stopping policy π_i. In turn, given some *a priori* application of π_i and resulting T_i, we can extend the definitions of the Gittins index, so that $G_i(\Theta_i, s \,|\, T_i)$ denotes the analogue of (EC.4) and $G_i(\Theta_i \,|\, T_i) = \sup_{s>0} G_i(\Theta_i, s \,|\, T_i)$ denotes the analogue of (EC.5).

LEMMA EC.1. *(Glazebrook 1979, Theorem 3, adapted to simulation selection context) Suppose, for all initial Θ_i and stationary stopping policies π_i, $R_t^{\pi_i}$ is uniformly bounded above for all t. Let T_i be the stopping time induced by π_i, and let $G_i(\Theta_{i,t} \mid T_i)$ be the associated Gittins index when project i is in state $\Theta_{i,t}$. Let T_i^* be the stopping time associated with an optimal stationary, deterministic stopping policy, π_i^*, for project i. If, for each i, $G_i(\Theta_{i,t} \mid T_i^*) \geq G_i(\Theta_{i,t} \mid T_i)$ for all stationary π_i whenever $\Theta_{i,t}$ is such that $t \geq T_i^*$, then $\xi^*(\pi_1^*, \pi_2^*, \ldots, \pi_k^*)$ is an optimal simulation selection policy.*

In words, if, in states in which the optimal stopping rule has stopped, the Gittins index for each project cannot be improved upon through the use of a sub-optimal stopping rule, then $\xi^*(\pi_1^*, \pi_2^*, \ldots, \pi_k^*)$ is optimal.

These "stoppable bandits" are, in turn, special cases of what Whittle (1980) calls bandit superprocesses. Whittle (1980) provided related optimality results for those superprocesses, that were later shown (Glazebrook 1982) to be equivalent in some sense to the above stoppable bandit result.

A.3. Counterexample for Glazebrook's Optimality Condition when $k > 1$

Appendix A.2 specifies sufficient conditions of Glazebrook (1979, Theorem 3) that guarantee that the hierarchical policy, $\xi^*(\pi_1^*, \pi_2^*, \ldots, \pi_k^*)$, would be optimal for the simulation selection problem when $k > 1$. The policy $\xi^*(\pi_1^*, \pi_2^*, \ldots, \pi_k^*)$ has a natural appeal, since it selects the optimal stopping policy π_i^* for each arm i, which converts each arm into a Markov chain, and then applies the optimal allocation (Gittins) index to the resulting Markov chains.

In this section we provide a counterexample that implies that Glazebrook (1979, Theorem 3)'s sufficient condition does not hold for the simple simulation selection problem. Specifically, we construct a problem instance in which a simple policy outperforms the hierarchical policy $\xi^*(\pi_1^*, \pi_2^*, \ldots, \pi_k^*)$. Thus, that the optimal policy for the simulation selection problem, whose existence can be guaranteed by Lemma 1, is not $\xi^*(\pi_1^*, \pi_2^*, \ldots, \pi_k^*)$. This supports the claim, in §6.1, that the existence of a 'Gittins index' is still an open question for the simulation selection problem in (2).

Example 7. This example extends Example 5 of §6.4 in the main paper. Using the setup of Example 5, it assesses if there is value to continuing to simulate if the expectation of the unknown mean, μ_{0i}, of all systems is on or above the stopping boundary, $b(t_{0i})$. When $k = 1$, the theoretical answer is no, and for $k > 1$, the answer should also be no *if* the sufficient conditions of Glazebrook (1979, Theorem 3) were true. In this example, we set $\mu_{0i} = 1.015 b(t_{0i})$, where $b(\cdot)$ is estimated as in Appendix E. We again choose $t_{0i} = 4$, for $i = 1, 2, \ldots, k$. This makes $\mu_{0i} = \mu_0 = 2.035 \times 10^7$ for each i. As a check for potential numerical round off issues in the computation of $b(\cdot)$, we verified that, for this value of μ_0, the lower bound from Lemma 3 is indeed less than μ_0, as expected. The factor 1.015 clearly places the μ_0 into the stopping set for each system.

Figure EC.1's left panel plots data that are analogous to the first two rows of Table 1, and its right panel plots the optimum amount of time one should simulate when using the equal allocation scheme in these cases. The figure demonstrates that there can be a large benefit to additional sampling when $k > 1$. In particular, the figure's left panel shows that, for the range of $k > 1$ that is in the figure, the lower bound $\underline{\text{OEDR}}(\boldsymbol{\Theta})$ for E[NPV], from (20), is greater than μ_0, the value of implementing a system without first sampling. Moreover, the right panel shows that one should sample for a nonzero time in order to obtain the maximal value of $\underline{\text{OEDR}}(\boldsymbol{\Theta})$ with a one-stage sampling allocation, for $k = 2, 3, \ldots, 10$. When $k = 1$ the maximum estimated discounted reward is achieved when no sampling is done and the reward is μ_0, as expected.

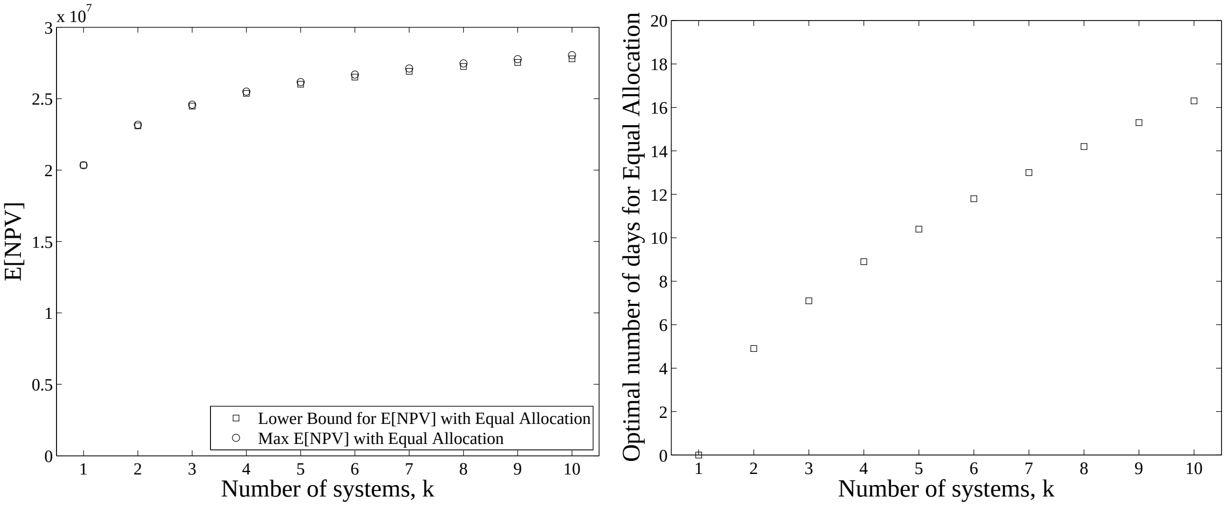
Thus, the figure indicates that, when a considering simulation selection problem with $k > 1$, it is not optimal to use the optimal stopping policies of the individual projects, as determined in Theorem 3. Stopping to implement the system should occur later and with a higher boundary. The implication is that the sufficient conditions of Glazebrook (1979) – which would guarantee the optimality of the hierarchical selection procedure $\xi^*(\pi_1^*, \pi_2^*, \ldots, \pi_k^*)$ – do not seem to hold. The optimal choice of stopping policy for each project may, in fact, depend upon the states of the other projects.

Appendix B: Discussion of Diffusion Approximation

This section provides additional discussion of some of the developments in §4 that establish a diffusion approximation. It is intended to be read with the general overview at the beginning of §4 in mind, and it also makes use of the notation that is established in the beginning of §4.1.

We first show how the diffusion equation in (6) can be derived from the continuous-time analog of (4). Note that (4) solves for $V^{\pi^*}(\Theta_i)$, the supremum of $V^{\pi_i}(\Theta_i)$ over all non-anticipative policies π_i in (3), and that (5) is the continuous-time analog of the supremum over non-anticipative polices for (3). We therefore have two types of representations of the simulation stopping problem in either discrete or continuous time – one involves a supremum over non-anticipative stopping rules. The other involves the solution of a free boundary problem for a heat equation (or diffusion). The representation as a supremum over stopping rules makes the reward structure through time more explicit. The free boundary problem representation will be more useful for computing a numerical solution to the simulation selection problem with $k = 1$. This section describes how these representations are related.

Figure EC.1 Lower bounds for E[NPV] when samples are allocated equally to each alternative (left panel), when replications take place for the optimal one-stage duration of time (right panel), when the mean is slightly above the stopping boundary.



Proceeding informally, suppose that observations are obtained continuously rather than at discrete intervals, so that Y_t is a Brownian motion with unknown drift θ and variance σ^2 per unit time. The analog of (4) with an infinitesimal number (h) of replications observed is then

$$B(y_t,t) = \max\{\lim_{h\to 0} -ch + e^{-\delta h} \times \mathrm{E}[B(Y_{t+h},t+h) \mid y_t,t], y_t/t\}, \tag{EC.6}$$

where B is the continuous-time analog of the value function, V.

Set $D(y_t,t) = y_t/t$ and $U = Y_{t+h} - y_t$. In the continuation set, $\mathcal{C} = \{(y_t,t) : B(y_t,t) > D(y_t,t)\}$, the first maximand is selected and simulation sampling continues. Note that $B(Y_{t+h}, t+h) = B(y_t,t) + UB_y(y_t,t) + hB_t(y_t,t) + U^2 B_{yy}(y_t,t)/2 + o(h)$, where the subscripts on B indicate partial derivatives, and $e^{-\delta h} = 1-\delta h + o(h)$. The distribution of U, given θ, is $\texttt{Normal}\,(\theta h, \sigma^2 h)$, and the posterior distribution of θ at time t is $\texttt{Normal}\,(y_t/t, \sigma^2/t)$. So the marginal distribution of U is $\texttt{Normal}\,(hy_t/t, \sigma^2(h+h^2/t))$, and

$$\begin{aligned} B(y_t,t) &= \max\{\lim_{h\to 0} -ch + e^{-\delta h} \times E_U[B + UB_y + hB_t + \frac{1}{2}U^2 B_{yy}] + o(h), y_t/t\} \\ &= \max\{\lim_{h\to 0} -ch + (1-h\delta) \times (B + h\frac{y_t}{t}B_y + hB_t + h\frac{\sigma^2}{2}B_{yy}) + o(h), y_t/t\}, \end{aligned} \tag{EC.7}$$

where B, B_y, B_t and B_{yy} in the first maximand are all evaluated at (y_t,t).

Therefore the following PDE describes the evolution of the value function in the continuation set $\mathcal{C}$,

$$0 = -c - \delta B + \frac{y}{t}B_y + B_t + \frac{\sigma^2}{2}B_{yy}. \tag{EC.8}$$

which justifies (6) of the main paper. The boundary, $\partial\mathcal{C}$, of $\mathcal{C}$ will be determined by equating the two maximands in (EC.6), as well as by maintaining a smooth pasting condition (Chernoff 1961),

$$\begin{aligned} B(y,t) &= D(y,t), \text{ on } \partial\mathcal{C} \\ B_y(y,t) &= D_y(y,t), \text{ on } \partial\mathcal{C} \text{ (smooth pasting).} \end{aligned} \tag{EC.9}$$

When $c = 0$, (EC.8)-(EC.9) represent what might be called a perpetual American call option on regular (rather than geometric) Brownian motion with unknown drift.

Equation (EC.8) is related to the PDE considered by Breakwell and Chernoff (1964, p. 164, $0 = 1 + \frac{y}{t}B_y + B_t + \frac{1}{2}B_{yy}$). It differs from Breakwell and Chernoff's PDE in a few respects, including that paper's lack of discounting, a different terminal value function, D, and minimization of losses rather than maximization of gains.

A natural question is to when the diffusion approximation might be expected to be reasonable. We examine the transformation in §4.2 to assess this. The transformation $\tau = \gamma t$ means that, as replications are observed, the scaled times $\tau \in \{\gamma t_0, \gamma(t_0+1), \gamma(t_0+2), \ldots\}$ become dense on $[0,\infty)$ as $\gamma \to 0$. The transformation leads to a diffusion limit as in (EC.7) and (EC.8) that is asymptotically appropriate as $h = \gamma \to 0$ (e.g. Billingsley 1986, Section 37). As long as the effect of discounting over the duration of one simulation replication is very close to 0, as is typically the case in real simulation applications, one would therefore expect the diffusion approximation to be reasonable.

Appendix C: Mathematical Proofs

Proof of Lemma 1. We can modify the original problem formulation of the simulation selection so that it meets Blackwell's conditions. We distinguish the revised problem from the original simulation selection problem through use of the subscript r.

We let the same $2k+1$ actions be available in every state, and we modify the one-period reward. Action $j_r(t) = 0$ at time t represents the decision to "do nothing" and receive an NPV of 0. Actions $j_r(t) \in \{1,2,\ldots,k\}$ denote decisions to simulate project $i = j_r(t)$ for one period and pay c_i. Actions $j_r(t) \in \{k+1, k+2, \ldots, 2k\}$ represent decisions to take expected one-period reward, $(1-\Delta)\mathrm{E}[X(\Theta_{i,t})]$, from project $i = j_r - k$ for the current period. Observe that a perpetuity based on this one-period reward has expected discounted value $\mathrm{E}[X(\Theta_{i,t})]$. Thus, for policy $\pi_r \in \Pi$, expected one-period rewards become

$$R_t^{\pi_r} = \begin{cases} 0, & \text{if } j_r(t) = 0 \\ -c_j, & \text{if } j_r(t) \in \{1,2,\ldots,k\} \\ (1-\Delta)\,\mathrm{E}[X(\Theta_{j_r(t)-k,t})], & \text{if } j_r(t) \in \{k+1, k+2, \ldots, 2k\}. \end{cases} \tag{EC.10}$$

We then modify the definition of state transitions. For action $j_r(t) \in \{1,2,\ldots,k\}$ transitions remain as before: the state of project $i = j_r(t)$ changes according to Bayes' rule, (1), and the states of other projects remain unchanged. For $j_r(t) = 0$ and $j_r(t) \in \{k+1, k+2, \ldots, 2k\}$, we define the state of all k projects as unchanging; that is, $\Theta_{i,t+1} = \Theta_{i,t}$ for all projects $i = 1, 2, \ldots, k$.

We note that any policy, π, in the original problem has a feasible analogue, π_r, in the revised problem with the same expected discounted value. For $t < T$ in the original problem, let $j_r(t) = i(t)$ in the revised problem. For $t \geq T$ in the original problem, let $j_r(t) = I(T) + k$ in the revised version when $I(T) > 0$, and $j_r(t) = I(T) = 0$ otherwise. Then the application of policy $\pi \in \Pi$ in the original problem yields

$$\begin{aligned} V^\pi(\mathbf{\Theta}) &= \mathrm{E}_\pi\left[\sum_{t=0}^{T-1} -\Delta^t c_{i(t)} \;+\; \Delta^T X(\Theta_{I(T),T}) \,|\, \mathbf{\Theta}_0 = \mathbf{\Theta}\right] \\ &= \mathrm{E}_{\pi_r}\left[\sum_{t=0}^{\infty} \Delta^t\, R_t^{\pi_r} \,|\, \mathbf{\Theta}_0 = \mathbf{\Theta}\right] = V_r^{\pi_r}(\mathbf{\Theta}). \end{aligned} \tag{EC.11}$$

When there exists an $\Upsilon < \infty$ such that $|R_t^{\pi_r}| < \Upsilon$ for all $t \geq 0$ and every $\pi_r \in \Pi$, expected one-period rewards are uniformly bounded, and Blackwell's conditions are met. In the context of the simulation-selection problem this is equivalent to the condition that there exists an $\Upsilon < \infty$ such that $\max\{|c_i|, |\mathrm{E}[X(\Theta_i)]|\} < \Upsilon$ for all $\Theta_i, i \in \{1,2,\ldots,k\}$. Given these conditions, there exists a stationary, deterministic policy that is optimal for the infinite-horizon version of the problem.

Just as each policy $\pi \in \Pi$ has an analog, $\pi_r \in \Pi$, with the same expected discounted value in the infinite-horizon problem, every stationary, deterministic policy in the revised problem has a feasible analog in the original problem that has the same expected discounted value. To see this, suppose that $t_r = \inf\{t \,|\, j_r(t) \notin \{1,2,\ldots,k\}\}$ in the revised problem. If there is no such time, let $t_r = \infty$. By definition, from t_r to $t_r + 1$ the system state does not change, so for any stationary, deterministic policy it must be that $j_r(t) = j_r(t_r)$ all $t > t_r$. Thus, in the original problem we can set $T = t_r$, $i(t) = j_r(t)$ for all $t < t_r$, and $I(T) = \max\{0, j_r(t_r) - k\}$. (EC.11) again shows that the two expected discounted values are the same.

Using this correspondence, it is not difficult to show that we can map optimal solutions from the infinite-horizon formulation to the original simulation selection problem. In particular, suppose π_r^* is a stationary deterministic policy that is optimal for the infinite-horizon problem. Then its analog, π^*, is feasible for the original problem statement and has the same expected discounted value. Now, by contradiction, suppose that π^* is not optimal in the original problem. Then there must be another policy, π', which has a higher expected discounted value. But π', itself, has a feasible analogue in the infinite-horizon problem, π_r', with the same expected discounted value. Therefore, π_r^* must not have been optimal for the revised problem statement, a contradiction. See also Glazebrook (1979, Lemma 1). This concludes the proof of Lemma 1. □

Proof of Lemma 2. Suppose that a deterministic, stationary policy for the simulation selection problem were not reasonable (almost surely). Then there would be sample paths with $T < \infty$, $I(T) = i$, and $(1-\Delta)\mathrm{E}[X(\Theta_{i,T})] < -c_i$, with probability greater than 0. On these sample paths, performance can be strictly improved upon by never stopping and setting $i(t) = i$ for all $t \geq T$. A deterministic, stationary policy that is not reasonable, almost surely, is therefore not optimal. □

Proof of Theorem 1. The proof first provides some added details for the development in §4.2.1, then shows that the solution of the standardized is proportional to the solution of the original problem, as claimed. Properties of the boundary of continuation set are then established.

With $c = 0$ and the change of variables to (z_τ, τ) coordinates, the expectation in (5) is

$$B(y_{t_0}, t_0) = \mathrm{E}_{\tilde{\tau}^* \geq \tau_0}\left[e^{-(\tilde{\tau}^*-\tau_0)\delta/\gamma}\frac{\gamma}{\alpha}\frac{Z^*_{\tilde{\tau}}}{\tilde{\tau}^*}|\tau_0, z_0\right] = \sup_{\tilde{\tau}\geq\tau_0}\mathrm{E}_{\tilde{\tau}}\left[e^{-(\tilde{\tau}-\tau_0)}\frac{\gamma}{\alpha}\frac{Z_{\tilde{\tau}}}{\tilde{\tau}}\middle|\tau_0, z_0\right],$$

for some measurable continuous-time policy $\tilde{\pi}$ with optimal stopping time $\tilde{\tau}^* \geq \tau_0$. The choice of $\delta/\gamma = 1$ standardizes the discount factor to be 1. With the other parameter choices in (8), we can write

$$\frac{\alpha}{\gamma}B(y_{t_0}, t_0) = B_1(z_0, \tau_0) \triangleq \mathrm{E}_{\tilde{\tau}^* \geq \tau_0}\left[e^{-(\tilde{\tau}^*-\tau_0)}\frac{Z_{\tilde{\tau}^*}}{\tilde{\tau}^*}\middle|\tau_0, z_0\right] \qquad \text{(EC.12)}$$

Upon converting to (w_0, s_0) coordinates, Problem (EC.12) becomes

$$B_1(w_0, s_0) = \sup_{0\leq S\leq s_0}\mathrm{E}_S\left[e^{-(1/S-1/s_0)}W_S \mid w_0, s_0\right]. \qquad \text{(EC.13)}$$

Problem (EC.13) determines the OEDR $B_1(w, s)$ of a standardized simulation selection problem which has $c = 0$, $\delta = 1$ and $\sigma = 1$. The solution to that problem can be approached with the following free boundary problem in (9). The fact that solving (9) solves (EC.13) can be found either by deriving (9) from (6) via the chain rule and the given change of variables, or by using techniques like those in Appendix B. The fact that $B(y_{t_0}, t_0) = \sigma\sqrt{\delta}B_1(w_0, s_0)$, as claimed, follows from (EC.12), the equivalence of (EC.12) to (EC.13), and the choice $\beta^{-1} = \delta/\alpha = \sigma\sqrt{\delta}$. That $B(y_{t_0}, t_0) \geq \max\{0, y_{t_0}/t_0\}$ follows because, at worst, one can simulate forever to get 0, or can stop immediately to get y_{t_0}/t_0, in expectation.

Two points imply that $\partial\mathcal{C}$ is defined by a single function $b_1(s) \geq 0$. First, note that $(w_0, s_0) \in \mathcal{C}$ for all negative w_0, since implementing has NPV $w_0 < 0$ and simulating forever has a higher NPV, 0. Two, suppose that $a > 0$ and $(w_0, s_0) \in \mathcal{C}$, which means that $B_1(w_0, s_0) > D(w_0, s_0)$. Then

$$\begin{aligned} B_1(w_0 - a, s_0) &= \sup_{S\in[0,s_0]}\mathrm{E}\left[W_S e^{-(1/S-1/s_0)}|w(s_0) = w_0 - a\right] \\ &= \sup_{S\in[0,s_0]}\mathrm{E}\left[(W_S - a)e^{-(1/S-1/s_0)}|w(s_0) = w_0\right] \\ &\geq -a + \mathrm{E}\left[W_S e^{-(1/S-1/s_0)}|w(s_0) = w_0\right] \\ &= -a + B_1(w_0, s_0) \\ &> -a + w_0 = D(w_0 - a, s_0), \end{aligned}$$

where the last inequality follows because $(w_0, s_0) \in \mathcal{C}$ by assumption. Therefore $(w_0 - a, s_0) \in \mathcal{C}$, so there is a single nonnegative $b_1(s)$ that defines the boundary of the continuation set, $\mathcal{C} = \{(w, s) : w < b_1(s)\}$. The scaling of $\mathcal{C}$ in (y, t) coordinates follows from the fact that $\beta^{-1}w = y/t$ and $\beta^{-1} = \sigma\sqrt{\delta}$. □

Before proving Theorem 2, we compare the optimal stopping boundary and OEDR of the above problem to those of Bayesian bandit problems. In a Bayesian bandit problem, the unknown distribution Θ_t evolves according to Bayes' rule as samples X_t are observed, as with the simulation selection problem. But the reward structures, and therefore the OEDRs, of the two problems differ: The Bayesian bandit generates a reward X_t at each time t, while the simulation selection problem with $c = 0$ provides no reward until simulation stops, and a project is implemented. Nonetheless, the optimal stopping boundaries for the two problems are closely related:

THEOREM EC.1. *When $c = 0$ and $\delta > 0$, the optimal stopping boundary for the continuous-time, standardized free boundary problem in (9) satisfies $b_1(s) = b_{BL}(s)$, where $-b_{BL}(s)$ is the optimal stopping time of the asymptotic approximation of Brezzi and Lai (2002) for the infinite-horizon discounted Bayesian bandit problem with independent, normally distributed samples, unknown mean, and known variance.*

Proof of Theorem EC.1. Let $M = B(y_{t_0}, t_0) = \sup_{T \geq t_0} \mathrm{E}[R(Y_T, T) \mid y_{t_0}, t_0]$ be the OEDR for the original problem in (y, t) coordinates. While, technically, the simulation selection problem defines Y as being observed at discrete times, here we abuse notation and consider the stopping time T to be in continuous time (for a Wiener process, which is asymptotically valid as $\gamma \to 0$). This is done to show the relationship of the original problem with the standardized Brownian motion approximation in (w, s) coordinates.

Let all expectations in the proof be conditional on $Y_{t_0} = y_{t_0}$. Then

$$\begin{aligned} M &= \sup_{T \geq t_0} \mathrm{E}\Big[D(Y_T, T) e^{-\delta(T - t_0)}\Big] \\ &= \sup_{T \geq t_0} \mathrm{E}\Big[\int_T^\infty D(Y_T, T)\delta e^{-\delta(\xi - t_0)} d\xi\Big], \text{ so} \\ 0 &= \sup_{T \geq t_0} \mathrm{E}\Big[-\int_{t_0}^T \delta M e^{-\delta(\xi - t_0)} d\xi + \int_T^\infty \delta(D(Y_T, T) - M) e^{-\delta(\xi - t_0)} d\xi\Big] \end{aligned} \tag{EC.14}$$

because $M = \int_{t_0}^\infty M\delta e^{-\delta(\xi - t_0)} d\xi$. Apply the change of coordinates $W(s) = \beta Y_t / t$ and $s = 1/\gamma t$, as for the standardized problem, so that W is a Brownian motion in the $-s$ scale going from $s_0 = 1/\gamma t_0$ to 0 (cf. §4.2.1). Recall that $\gamma = \delta$. Then (EC.14) implies that

$$\begin{aligned} 0 &= \sup_{T \geq t_0} \mathrm{E}\Big[-\int_{t_0}^T \delta M e^{-\delta(\xi - t_0)} d\xi + \int_T^\infty \delta(\frac{W(S)}{\beta} - M) e^{-\delta(\xi - t_0)} d\xi\Big] \\ &= \sup_{T \geq t_0} \mathrm{E}\Big[\int_{t_0}^T M de^{-\delta(\xi - t_0)} - \int_T^\infty (\frac{W(S)}{\beta} - M) de^{-\delta(\xi - t_0)}\Big] \\ &= \sup_{0 \leq S \leq s_0} \mathrm{E}\Big[M e^{-(\frac{1}{S} - \frac{1}{s_0})} - M - (\frac{W(S)}{\beta} - M) e^{-\delta(\xi - t_0)}\Big|_{\xi = \infty} + (\frac{W(S)}{\beta} - M) e^{-(\frac{1}{S} - \frac{1}{s_0})}\Big] \\ &= \sup_{0 \leq S \leq s_0} \mathrm{E}\Big[\frac{W(S)}{\beta} e^{-(\frac{1}{S} - \frac{1}{s_0})} - M\Big]. \end{aligned} \tag{EC.15}$$

Formally, we need to worry about the payoff when $T = \infty$ (or $S = 0$). But the reward when $S = 0$ is 0 for any finite W, due to infinite discounting, and can therefore be safely ignored. Recall that $\beta^{-1} = \sigma\sqrt{\delta}$, and make explicit the implicit condition above, to obtain

$$M = \sigma\sqrt{\delta} \sup_{0 \leq S \leq s_0} \mathrm{E}\Big[W(S) e^{-(\frac{1}{S} - \frac{1}{s_0})} \mid w(s_0) = w_0\Big]. \tag{EC.16}$$

By Theorem 1, the stopping boundary is $w_0 = b_1(s_0)$, or when $y_{t_0}/t_0 = \sigma\sqrt{\delta} b_1(s_0)$.

Chang and Lai (1987, Eq. (2.6)) show that a standardized problem for the infinite-horizon discounted Bayesian bandit problem, with normally distributed output and $\sigma = 1$, is

$$w'_0 = \sup_{0 \leq S' \leq s'_0} \mathrm{E}\Big[W'(S') e^{-(\frac{1}{S'} - \frac{1}{s'_0})} \mid w'(s'_0) = w'_0\Big], \tag{EC.17}$$

where (W', S') is also a Brownian motion in the $-s$ scale; $W'(s') = (Y_\tau/\tau - u_0)/\sqrt{\delta}$; $w'_0 = (M' - u_0)/\sqrt{\delta}$; and $u_0 = y_{\tau_0}/\tau_0$ is the mean of the prior distribution for the expected reward from a given bandit arm. Then

$$M' - u_0 = \sigma\sqrt{\delta} \sup_{0 \leq S' \leq s'_0} \mathrm{E}\Big[W'(S') e^{-(\frac{1}{S'} - \frac{1}{s'_0})} \mid w'(s'_0) = (M' - u_0)/\sigma\sqrt{\delta}\Big], \tag{EC.18}$$

for general σ. (See Brezzi and Lai 2002, Eq. 6 and 8, which find an inf over stopping rules with $w'_0 = (u_0 - M')/\sqrt{\delta}$; see their Eq. 16 to incorporate σ.) Lai and coauthors show that $M' - u_0 = \sigma\sqrt{\delta} b_{BL}(s'_0)$, where $-b_{BL}(s')$ is the optimal stopping boundary for the standardized Bayesian bandit problem: one is indifferent between the 0 option and stopping when $u_0 = -\sigma\sqrt{\delta} b_{BL}(s'_0)$, or $w'_0 = b_{BL}(s'_0)$), and $b_{BL}(s') \geq 0$.

The random processes in the expectations in (EC.16) and (EC.18) are both Brownian motions in a reverse time scale with the same support (if $s_0 = s'_0$). Only the conditioning statements differ. We can therefore equate w_0 and w'_0 in the conditioning statements where one is indifferent between stopping and continuing at time $s_0 = s'_0$. That is, $w_0 = b_1(s_0) = b_{BL}(s_0) = w'_0$, as claimed. □

Proof of Theorem 2. The stated asymptotic approximations are a result from Chang and Lai (1987) and Brezzi and Lai (2002) for $b_{BL}(s)$. By Theorem EC.1 above, the result therefore holds for $b_1(s) = b_{BL}(s)$. □

Proof of Lemma 3. Define T_β to be the one-stage stopping rule that: samples for exactly $\beta \geq 0$ replications and either implements, if $y_{t_0+\beta}/(t_0+\beta) \geq -c/\delta$ (for an expected reward of $y_{t_0+\beta}/(t_0+\beta)$), or otherwise never stops (e.g. simulate forever if $y_{t_0+\beta}/(t_0+\beta) < -c/\delta$, for reward $-c/\delta$).

The predictive distribution of $Y_{t_0+\beta}/(t_0+\beta)$, given (y_{t_0}, t_0), is normal with mean y_{t_0}/t_0 and variance $\sigma^2\beta/t_0(t_0+\beta)$ (de Groot 1970, Sec. 11.9). The expected discounted reward of stopping rule T_β is therefore

$$\begin{aligned} E_{T_\beta}[e^{-\delta\beta}\max\{-c/\delta, Y_{t_0+\beta}/(t_0+\beta)\}|y_{t_0}, t_0] &= E_{T_\beta}\left[e^{-\delta\beta}\left(-c/\delta + \max\{0, Y_{t_0+\beta}/(t_0+\beta) + c/\delta\}\right)|y_{t_0}, t_0\right] \\ &= e^{-\delta\beta}\left(\frac{-c}{\delta} + \left(\frac{\sigma^2\beta}{t_0(t_0+\beta)}\right)^{1/2}\Psi\left[-\left(\frac{y_{t_0}}{t_0} + \frac{c}{\delta}\right)/\left(\frac{\sigma^2\beta}{t_0(t_0+\beta)}\right)^{1/2}\right]\right). \end{aligned}$$

Inequality (10) is justified because B is defined as a supremum over all stopping rules, including T_β for all $\beta \geq 0$. □

Proof of Theorem 3. We proceed as in §4.2.1, for the development that leads to Theorem 1, except that $c > 0$. We start by rewriting the total reward function:

$$\begin{aligned} R(y_t, t) &= -\int_{t_0}^{t} c e^{-\delta(\xi - t_0)} d\xi + D(y_t, t)e^{-\delta(t-t_0)} \\ &= -\frac{c}{\delta}(1 - e^{-\frac{(\tau-\tau_0)\delta}{\gamma}}) + e^{-\frac{(\tau-\tau_0)\delta}{\gamma}} \frac{c}{\delta}\frac{\delta\gamma}{c\alpha}\frac{z_\tau}{\tau} = R(z_\tau, \tau). \end{aligned} \tag{EC.19}$$

Then (5) is $B(y_{t_0}, t_0) = \sup_{\tilde{\tau} \geq \tau_0} \mathrm{E}_{\tilde{\tau}}[R(Z_{\tilde{\tau}}, \tilde{\tau}) \mid z_0, \tau_0]$ for some suitable measurable, continuous-time selection policy, $\tilde{\pi}$, with optimal stopping time $\tilde{\tau}^* \geq \tau_0$.

Set $\kappa = \delta\gamma/c\alpha$. Since $\mathrm{E}[R(Y_T, T)] = \mathrm{E}[R(Z_{\tilde{\tau}}, \tilde{\tau})]$ when $\tilde{\tau} = \gamma T$, the stopping time $\tilde{\tau}^*$ also maximizes

$$\frac{\delta}{c}\mathrm{E}_{\tilde{\tau} \geq \tau_0}[R(Z_{\tilde{\tau}}, \tilde{\tau})|z_0, \tau_0] = \mathrm{E}_{\tilde{\tau} \geq \tau_0}\left[-1 + e^{-\frac{(\tilde{\tau}-\tau_0)\delta}{\gamma}}\left(1 + \kappa Z_{\tilde{\tau}}/\tilde{\tau}\right)\Big|z_0, \tau_0\right]. \tag{EC.20}$$

The problem of finding the optimal $\tilde{\tau}^* \geq \tau_0$ to maximize (EC.20) over stopping times $\tilde{\tau} \geq \tau_0$ of the Wiener process Z_τ can be reduced to a family of standardized problems indexed by κ if δ/γ is chosen to equal 1 to simplify the exponent, and if the diffusion's two moment constraints are satisfied ($\alpha/\beta\gamma = 1$ and $\alpha^2\sigma^2 = \gamma$). We adopt that parametrization here, namely

$$\alpha = \delta^{1/2}\sigma^{-1}, \beta = \delta^{-1/2}\sigma^{-1}, \gamma = \delta \text{ and } \kappa = \delta^{3/2}\sigma c^{-1}. \tag{EC.21}$$

Given (EC.20), the general solution is reduced to finding $\tilde{\tau}^*$ for a standardized problem whose sampling costs, discount factor and variance are all equal to 1:

$$\frac{\delta}{c}B(y_{t_0}, t_0) = \frac{\delta}{c}\mathrm{E}_{\tilde{\tau}^* \geq \tau_0}[R(Z^*_{\tilde{\tau}}, \tilde{\tau}^*) \mid z_0, \tau_0] = -1 + \sup_{\tilde{\tau} \geq \tau_0} \mathrm{E}_{\tilde{\tau} \geq \tau_0}\left[(1 + \kappa Z_{\tilde{\tau}}/\tilde{\tau})e^{-(\tilde{\tau}-\tau_0)}\Big|z_0, \tau_0\right]. \tag{EC.22}$$

With a change to (w, s) coordinates, (EC.22) implies

$$\begin{aligned} \frac{\delta}{c}B(y_{t_0}, t_0) + 1 &= \sup_{0 \leq S \leq s_0} \mathrm{E}\left[(\kappa W_s + 1)e^{-(1/S - 1/s_0)} \mid w_0, s_0\right] \\ &= \kappa \sup_{0 \leq S \leq s_0} \mathrm{E}\left[(W_s + 1/\kappa)e^{-(1/S - 1/s_0)} \mid w_0, s_0\right] \\ &= \kappa \sup_{0 \leq S \leq s_0} \mathrm{E}\left[e^{-(1/S - 1/s_0)}W_s \mid w_0 + 1/\kappa, s_0\right] \\ &= \kappa B_1(w_0 + 1/\kappa, s_0). \end{aligned}$$

Since $1/\beta\kappa = c/\delta$, the OEDR for the simulation selection problem with $c, \delta > 0$ is therefore as claimed,

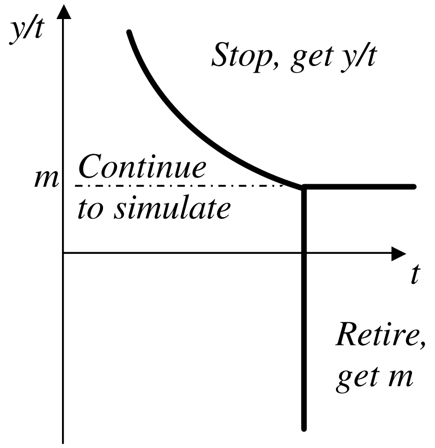
$$B(y_{t_0}, t_0) = \frac{c\kappa}{\delta}B_1(w_0 + 1/\kappa, s_0) - \frac{c}{\delta} = \sigma\sqrt{\delta}B_1\left(\frac{1}{\sigma\sqrt{\delta}}(\frac{y_{t_0}}{t_0} + \frac{c}{\delta}), \frac{1}{\delta t_0}\right) - \frac{c}{\delta}.$$

Let $\mathcal{C}$ be the stopping boundary for the original problem with $c > 0$. Define $\mathcal{C}_{c=0}$ to be the stopping boundary for the problem with all parameter the same, except with $c = 0$. Suppose that $(y_t, t) \notin \mathcal{C}_{c=0}$ is not in the continuation set when $c = 0$, so that $\beta^{-1}B_1(\beta y_t/t, 1/\gamma t) = y_t/t$ (Theorem 1). Let $\hat{y}$ satisfy $\hat{y}/t = y_t/t - c/\delta$. Then

$$B(\hat{y}, t) = \beta^{-1}B_1(\beta(\hat{y}/t + c/\delta), 1/\gamma t) - c/\delta = \beta^{-1}B_1(\beta y_t/t, 1/\gamma t) - c/\delta = y_t/t - c/\delta = \hat{y}/t$$

has the form $B(\hat{y}, t) = \hat{y}/t$. So $(\hat{y}, t) \in \mathcal{C}$ if $(y_t, t) \in \mathcal{C}_{c=0}$. A similar argument shows that $(\hat{y}, t) \notin \mathcal{C}$ if $(y_t, t) \notin \mathcal{C}_{c=0}$. The continuation set when $c > 0$ is therefore shifted down by c/δ, as claimed. □

Figure EC.2 There are two stopping regions when a retirement of $m > -c/\delta$ is allowed, one for retirement (terminal reward m) and the other for implementing a system (terminal expected reward y/t).



Proof of Theorem 4. The proof, for $m \le -c/\delta$, comes from noting that m never need be chosen by an optimal policy, since one can always do at least as well as m by simulating forever, which has expected NPV of $-c/\delta$. The optimal reward is therefore the same, whether or not such a retirement option is available at all.

For the balance of the proof, suppose that $m > -c/\delta$, so that the expected value of stopping in (14) simplifies to $D(m, y, t) = \max\{m, y/t\}$. We define $b^{-1}(m) = \sup\{t : b(t) \ge m\}$ for all $m > -c/\delta$.

We note that, while we define $b^{-1}(m) = \sup\{t : b(t) \ge m\}$, we hypothesize that $b(\cdot)$ is in fact invertible for $m > -c/\delta$. By Theorem 2, $b(t)$ is continuous and monotone decreasing for sufficiently small and sufficiently large t, with $\lim_{t\to 0} b(t) = \infty$ and $\lim_{t\to\infty} b(t) = -c/\delta$. The fact that $\lim_{t\to\infty} b(t) = -c/\delta$ means that $b^{-1}(m)$ is finite for $m > -c/\delta$. Furthermore, we note that $b(t)$ is continuous and monotone decreasing (and hence invertible) in our numerical experiments of Appendix E, below, and in the numerical experiments of Brezzi and Lai (2002) (cf. our Theorem 2). Chernoff (1961, p. 89) notes that for a related (undiscounted) free boundary problem the boundary is decreasing, continuous and differentiable (except for a set of t of measure 0 where the slope may be $-\infty$).

We first examine the first alternative in (15), which assumes that $t_0 \ge b^{-1}(m)$ and $m > -c/\delta$. This means that the retirement reward exceeds the stopping boundary of the original problem without the retirement reward, $m \ge b(t_0)$. One can therefore achieve an NPV of $\max\{y_{t_0}/t_0, m\}$ by stopping immediately and selecting the better of those two options. This justifies the lower bound in the first alternative of (15). For the lower bound not to be tight, in this case, there would need to be some value to simulating at least once, with the m option, even though one would stop if $y_{t_0}/t_0 = m$ and the retirement option of value m were not available.

In order to justify the remaining alternative, suppose now that $t_0 < b^{-1}(m)$ and $m > -c/\delta$. Consider the following terminal condition: If one has not stopped before time $b^{-1}(m)$, then when the diffusion hits time $b^{-1}(m)$, retire with a sure NPV of m if $y_{b^{-1}(m)}/b^{-1}(m) < m$, and implement the simulated project for an expected NPV of $y_{b^{-1}(m)}/b^{-1}(m)$ if $y_{b^{-1}(m)}/b^{-1}(m) \ge m$. One might stop at some time $t < b^{-1}(m)$ if the m-diffusion suggests that the mean is sufficiently high as to warrant early stopping. We consider the optimal policy for that subclass, call it Ξ', of all possible non-anticipative stopping times. (Since we consider a subclass of possible stopping times, we will obtain a lower bound for the second alternative in (15).)

Figure EC.2 gives a conceptual diagram of the continuation region for this stopping policy. In (y, t) coordinates, one proceeds up to a maximum time of $t = b^{-1}(m)$, and one is forced to take a terminal reward of $\max\{m, y_{b^{-1}(m)}/b^{-1}(m)\}$.

In particular, the m-diffusion satisfies the same diffusion equation in (6), but now has a terminal condition, at time $t = b^{-1}(m)$, with expected NPV $\max\{m, y_{b^{-1}(m)}/b^{-1}(m)\}$. This differs from the "terminal condition" from the original case, which informally is to pick the best alternative of $\max\{-c/\delta, \lim_{t\to\infty} y_t/t\}$, as $t \to \infty$. More formally, that terminal condition was to retire with terminal value $\max\{w_0, 0\}$ when analyzed in the (w, s) reverse-time coordinates.

Stopping at time $t = b^{-1}(m)$ corresponds to stopping at time $s_f = 1/\delta b^{-1}(m)$ in (w, s) coordinates and gives a reward $\max\{\beta m, w\}$. The process in (w, s) coordinates is a standard Brownian motion in the $-s$ scale that starts at time $s_0 = 1/\tau_0 = 1/\delta t_0$ and with position $w_0 = \beta y_{t_0}/t_0$. The statistics for that process are equivalent to the statistics of a process that starts at time $s_0 - s_f$ and that has the same terminal reward at time 0. (Shifting time by $1/\delta b^{-1}(m)$ does not change the statistics of the increments of a Brownian motion.)

The solution to the optimal policy in the subclass Ξ', therefore, is directly equated to the optimal policy in the original class, but for a modified problem. That modified problem has a time shift of $1/\delta b^{-1}(m)$ in the $-s$ scale.

We will use this fact to express the value function of the m-process in terms of the original process by shifting the m-diffusion process in the $-s$ time scale to run from time $\tilde{s}_0$ to time 0, rather than from time s_0 to time $1/\delta b^{-1}(m)$. In doing so, we note that m will take the role of $-c/\delta$ in (7). The role of t in that equation will be replaced by the value of t that corresponds to $\tilde{s}_0$. To find the value of t, we set $\tilde{s}(t) = 1/\delta t - 1/\delta b^{-1}(m) = (b^{-1}(m) - t)/(\delta t b^{-1}(m))$ and note that this corresponds to $\tilde{t}(t) = 1/\delta\tilde{s}(t) = tb^{-1}(m)/(b^{-1}(m) - t)$, as in the statement of the theorem.

The analogous terminal condition for the m-process is to terminate at time $1/\delta b^{-1}(m)$ with terminal reward $\tilde{D}(w, 1/\delta b^{-1}(m)) = \max\{\beta m, w\}$. Note that this form for the m-process is the same as that for the original process with $c \neq 0$, with βm taking the role of $-\kappa$. The diffusion runs in the $-s$ scale from time $s_0 = 1/\delta t_0$ to $1/\delta b^{-1}(m)$, for a total time duration in the $-s$ scale of $\tilde{s}_0 \triangleq 1/\delta t_0 - 1/\delta b^{-1}(m)$.

By recalling the problem in (6)-(7) and the solution in (11), then, and noting that m takes the role of $-c/\delta$ and that $\tilde{t}(t)$ takes the role of t, we arrive at a justification of (15). The value of $\underline{B}(m, y_{t_0}, t_0)$ solves the free boundary problem for the class of non-anticipative stopping policies Ξ' that require stopping by time $b^{-1}(m)$. □

Proof of Lemma 4. The fact that the distribution of the posterior mean $\mathcal{Z}_i$ is distributed as in (19), given that r_i replications are observed, follows directly from de Groot (1970, Sec. 11.9). The expectation $\mathrm{E}\left[\max\{\mu_{00}, \mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_k\}\right]$ is therefore the expected reward from selecting the alternative with the largest posterior mean, or the known NPV of μ_{00}, after having observed a total of $\beta = \sum_{i=1}^{k} r_i$ samples. The factor of $e^{-\delta\beta}$ discounts that reward appropriately.

In turn, two facts imply that the right hand side of (20) is a lower bound for $V^{\pi^*}(\boldsymbol{\Theta})$. First, the expected value of the specific one-stage sampling policy cannot be greater than that of the policy that is optimal over all non-anticipative sampling policies. Second, the discounted cost of sampling is not more expensive than the undiscounted cost of sampling, $\sum_{i=1}^{k} r_i c_i$. □

Proof of Lemma 5. The fact that the expectation $\mathrm{E}\left[\max\{\mu_{00}, \mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_k\}\right]$ can be decoupled into a sum of $\mu_{0(k)}$ and an expected opportunity cost for a potentially incorrect selection was shown by Gupta and Miescke (1996, Equation (11)).

Chick et al. (2001, Theorem 1) proved that the expected opportunity cost for a potentially incorrect selection has lower and upper bounds that justify the lower bound in (21) and the upper bound in (22). Those bounds come from assessing the expected loss in a pairwise comparison of the current best with any other single alternative (to get the lower bound), and from the sum of the expected losses when summing over all pairwise comparisons of the current best with each alternative (for the upper bound). The result was stated but not proven in Chick and Inoue (1998). □

Appendix D: Extensions

Section 4 assumed jointly independent Gaussian output with known variances, and simulations runs for each alternative that are of the same duration. This section shows that some of §4's results appear to hold more generally, although some loss of optimality may result. An assessment of how much loss of optimality may accrue, and a more detailed analysis for necessary and sufficient conditions for the theoretical results to be relevant, are left for future work.

D.1. Autocorrelated Output

The infinite-horizon expected NPV of a simulated system can sometimes be estimated by simulating the mean of a stationary process and applying a discount-factor correction. For example, if the initial state is appropriately modeled by sampling from the stationary distribution, and the stationary mean is A, then the mean infinite-horizon NPV is A/δ. Such processes are typically autocorrelated, however.

Autocorrelated output can often be analyzed using "batches", so that time averages from consecutive, finite time periods are treated as if they were statistically uncorrelated. Kim and Nelson (2006) justify this asymptotically in a diffusion-approximation framework when certain technical conditions, such as those for a *functional central limit theorem*, are valid. We presume that such technical conditions hold in this subsection.

One would hope that the boundary (as a function of the time spent simulating) specified by our approach would be invariant to the batch length if batching were used. Invariance occurs if β were invariant and γ were doubled whenever the length of a batch is doubled (so the number of batches is halved). That would keep $\beta^{-1}b_\ell(1/\tau)$ constant, as $\tau = \gamma t = (2\gamma)(t/2)$. Doubling the length of the runs would change parameters to $\delta' = 2\delta$ and $\sigma' = \sigma/\sqrt{2}$, so that $\beta' = 1/(\sqrt{\delta'}\sigma') = \sqrt{2}/(\sqrt{2}\sqrt{\delta}\sigma) = \beta$, and $\gamma' = \delta' = 2\delta = 2\gamma$, as required. The OEDR $\mathcal{V}_1 = \beta^{-1}B_1(\beta y_t/t, 1/\gamma t)$ is also invariant by the same argument. Shifts in the continuation set are also invariant: $-c'/\delta' = -2c/2\delta = -c/\delta$. Factors other than 2 are handled similarly.

Our approach is therefore compatible with a batch mean analysis, when the asymptotic variance is known. Note that some non-stationary investments, such as up-front construction costs for an implemented project, can be converted to the required stationary-process format by being treated as perpetuities.

D.2. Different Durations of Simulation Runs for Each System

The stoppable bandit results that justify the simulation selection analysis in §3-4 are based on discrete-time sampling with a common discount factor. While this assumption is violated if the time duration of replications for different systems differs, there exist simple methods for finding a common time scale.

If the simulations are steady-state simulations, then the rescaling technique described in Appendix D.1 can be used to change the duration of each system to a common value, as required, with the side-effect of changing the variance of the output of each system. If the replications are independent, rather than from steady-state simulations, batches of different numbers of replications from each system can be averaged to make the duration of running a batch for each system about the same. This again changes the output variance, but removes the original restrictive assumption of a common simulation duration.

The batching of simulation runs necessarily introduces an element of suboptimality into the sampling algorithm (as entire batches of runs must be taken, rather than allowing for stopping before the entire batch is observed). Nevertheless, to the extent that simulation run-times and costs are small, relative to time scales and costs being simulated, the resulting degradation in overall performance should be minimal.

Appendix E: Computational Issues

The assessment of a project's OEDR requires the computation of $B_1(w,s)$ and the determination of the stopping boundary $b_1(s)$. An analytical solution for B_1 and b_1 is challenging to derive.

A numerical solution of (9) requires initial conditions $B_1(w,s_n)$ for some fixed s_n and all w so that recursive calculations for $s > s_n$ can be made. We would like to have initial conditions at $s_n = 0$, but (9) poses numerical stability problems as $s \to 0$. We therefore need initial conditions for some time $s_n > 0$ to approximate B_1 and b_1. In the absence of a readily-computable analytic form for the exact initial conditions, we can use a lower bound for $B_1(w,s_n)$ as an approximation. Lemma 3 from the main paper provides that lower bound.

We next use the ideas of Chernoff and Petkau (1986) to numerically compute (9) in the $-s$ scale from some time $s_n > 0$ through a series of times $s_n < s_{n-1} < s_{n-2} < \cdots s_1 < s_0$. Chernoff and Petkau (1986) and Brezzi and Lai (2002) approximate similar diffusions with a binomial grid, working from time s_{i+1} to s_i with a small increment Δ_s. We started with $s_n = 5 \times 10^{-3}$.

The differences between our implementation and those of Chernoff and Petkau (1986) and Brezzi and Lai (2002) are that we: a) use an explicit finite difference method with trinomial trees (rather than binomial trees), with initial time step $\Delta_s = 24 \times 10^{-6}$ and an equal probability of going to 0 or up or down by $\Delta_w = \sqrt{3\Delta_s/2}$; b) employ an undiscounted terminal value function ($D(w,s) = \max\{0,w\}$) that is then discounted backwards in time, rather than a discounted terminal value ($\max\{0,w\}e^{-(1/s-1/s_0)}$) that is not discounted backwards in time, so that plotted values of $B(w,s)$ are valued in currency at time s, rather than being discounted back to time s_0; and c) initialize values of $B_1(w,s_n)$ using the lower bound in Lemma 3.

The discrete-time, discrete-space binomial grid does not directly allow for estimates of smooth boundaries, so we have implemented an effective correction term proposed by Chernoff and Petkau (1986). (See also Brezzi and Lai 2002.) We pass through many orders of magnitude to arrive at $s_0 = 5 \times 10^6$, and in order to obtain a fully usable range, after iterating from s_{i+1} to s_i we quadruple the size of the time step Δ_s. We also double the space increment Δ_w to preserve the unit variance of the random walk per time unit. This procedure causes a slight ripple in estimates of the boundary, so we restart the diffusion at a value slightly smaller than s_i before iterating to s_{i-1}.

Figure EC.3 shows the OEDR $B_1(w,s) = B_1(z_\tau/\tau, 1/\tau)$, and the free boundary, $b_1(s) = b_1(1/\tau)$, for one range of τ. Figure EC.4 and Figure EC.5 cover other ranges of τ. The graphs were generated in 4 min of CPU time in Matlab on a 1.6Mhz PC with 384Mb RAM.

Figure EC.6 plots the free boundary $b_1(1/\tau)$ over a wide range of values of τ. Theorem EC.1 indicated that $b_1 = b_{BL}$, where b_{BL} is related to the Gittins index of a Bayesian bandit problem. Brezzi and Lai (2002) approximated $b_{BL}(s)$ by

$$b_{BL}(s)/\sqrt{s} \approx \begin{cases} \sqrt{s/2} & \text{if } s \le 0.2 \\ 0.49 - 0.11s^{-1/2} & \text{if } 0.2 < s \le 1 \\ 0.63 - 0.26s^{-1/2} & \text{if } 1 < s \le 5 \\ 0.77 - 0.58s^{-1/2} & \text{if } 5 < s \le 15 \\ [2\log s - \log\log s - \log 16\pi]^{1/2} & \text{if } 15 < s. \end{cases} \tag{EC.23}$$

For small $\tau = 1/s$, that approximation matches our computation well. It is less accurate for intermediate values, and it improves upon our numerical calculations for $\tau > 5$. The most relevant range for τ in the illustrative examples of §5 makes use of smaller values of τ.

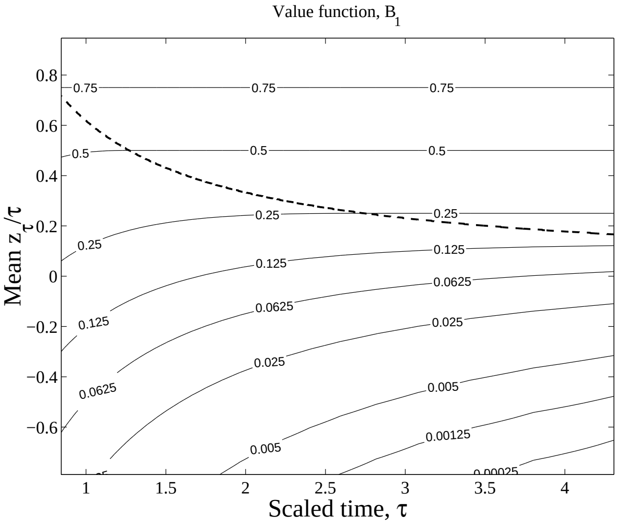


Figure EC.3 Contours of the OEDR $B_1(w,s)=B_1(z_\tau/\tau,1/\tau)$**, with dashed free boundary,** $b_1(s)=b_1(1/\tau)$**.**

We propose and recommend an easy-to-compute alternative to (EC.23) that reduces the difference between the approximation in (EC.23) and the more accurate free boundary solution for $b_1(1/\tau)$. To develop the approximation, we use Matlab to fit a low-order polynomial to the $\tau, b_1(1/\tau)$ in the log-log scale over the range $\tau \in [.01, 7]$. That range contains the range of $1/s$ values in question (from 1/15 to 1/.2). The alternative approximation that conforms quite closely to $b_1(1/\tau)$ in Figure EC.6 is:

$$\tilde{b}_1(s) \approx \begin{cases} s/\sqrt{2} & \text{if } s \le 1/7 \\ \exp\left[-0.02645(\log s)^2 + 0.89106 \log s - 0.4873\right] & \text{if } 1/7 < s \le 100 \\ \sqrt{s}\left[2\log s - \log\log s - \log 16\pi\right]^{1/2} & \text{if } 100 < s. \end{cases} \tag{EC.24}$$

This can be used to provide a quickly computed asymptotic approximation of the Gittins index of the Bayesian bandit problem of Brezzi and Lai (2002), with normal samples (unknown mean, known variance), namely, $y_t/t + \beta^{-1}\tilde{b}_1(1/\gamma t)$.

Good asymptotic approximations for $B_1(w,s)$ as $s \to 0$, $s \to \infty$, and $w \to -\infty$ would be helpful for obtaining a rapidly-computable OEDR over a broader range of values than is presently covered in Figure EC.4 and Figure EC.5. For extreme values of $\tau = 1/s$, outside of the range for which we computed the plots, we use the lower bound of Theorem 3. We did not have a special bias correction for B_1 as we do for b_1. Such bias corrections and approximations for more extreme values of s are left for future work.

Figure EC.6 also plots $B_1(0, 1/\tau)$, the OEDR when the sample mean is 0. The Gittins index of the standardized Bayesian bandit problem is $b_{BL}(1/\tau)$ when the sample mean is 0, and the figure confirms that the Gittins index of the Bandit problem and the OEDR of the simulation selection problem differ, although the boundaries are related.

Appendix F: Simulation Selection Procedures with $k > 1$ Project Alternatives

This section describes how to adapt one-stage $\mathcal{LL}$ allocations to the present context. These policies allocate a finite number of samples to k alternatives in a way that maximizes the expected (undiscounted) reward at the end of sampling. Because the optimal solution is only known for some special cases (e.g., $k = 2$), some allocations have been derived that maximize bounds on the expected opportunity cost of a potentially incorrect selection, when an asymptotically large number of samples is to be allocated.

Chick et al. (2001, Corollary 2) derives such an one-stage $\mathcal{LL}$ allocation. It assumes normally distributed outputs with unknown means and known sampling variances that may differ for each system. This policy is analogous to the one-stage $\mathcal{LL}$ allocation in Chick and Inoue (2001) that handles the case of unknown means and variances that may differ for each system. Branke et al. (2007) specified how the one-stage $\mathcal{LL}$ allocation in Chick and Inoue (2001) can be converted to a fully sequential algorithm. We use a similar conversion here to adapt the one-stage allocation of Chick et al. (2001) to a fully sequential algorithm.

With four adaptations, the one-stage allocation of Chick et al. (2001, Corollary 2) can be used to solve the simulation selection algorithm. One, we note that specifying the prior distributions for the performance of each system obviates the need for the "usual" first stage of sampling that is found in many ranking and selection procedures. Two, for a small to

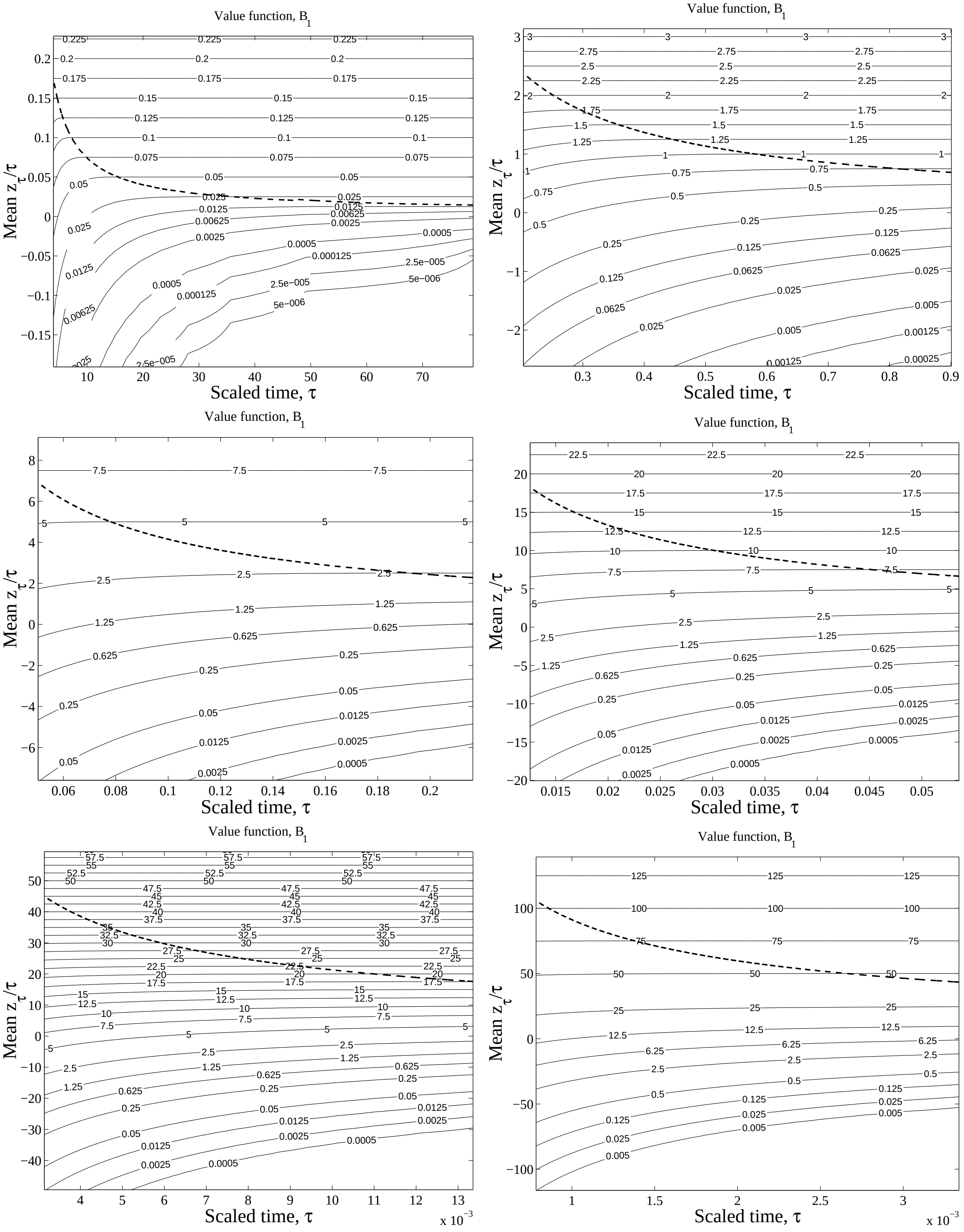


Figure EC.4 Contours for standardized OEDR, $B_1(w,s) = B_1(z_\tau/\tau, 1/\tau)$, with dashed free boundary $b_1(s) = b_1(1/\tau)$.

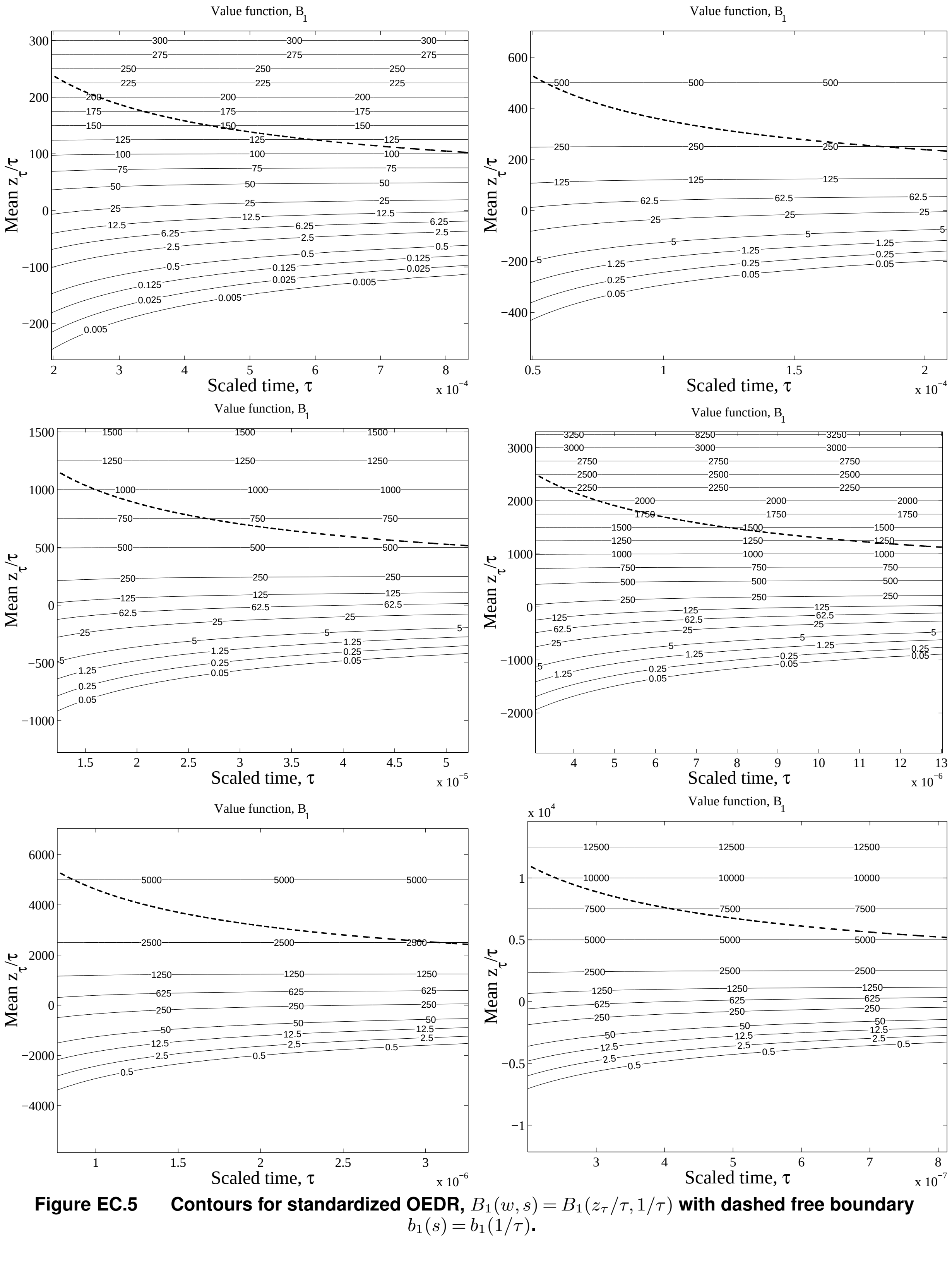


Figure EC.5 **Contours for standardized OEDR, $B_1(w,s) = B_1(z_\tau/\tau, 1/\tau)$ with dashed free boundary $b_1(s) = b_1(1/\tau)$.**

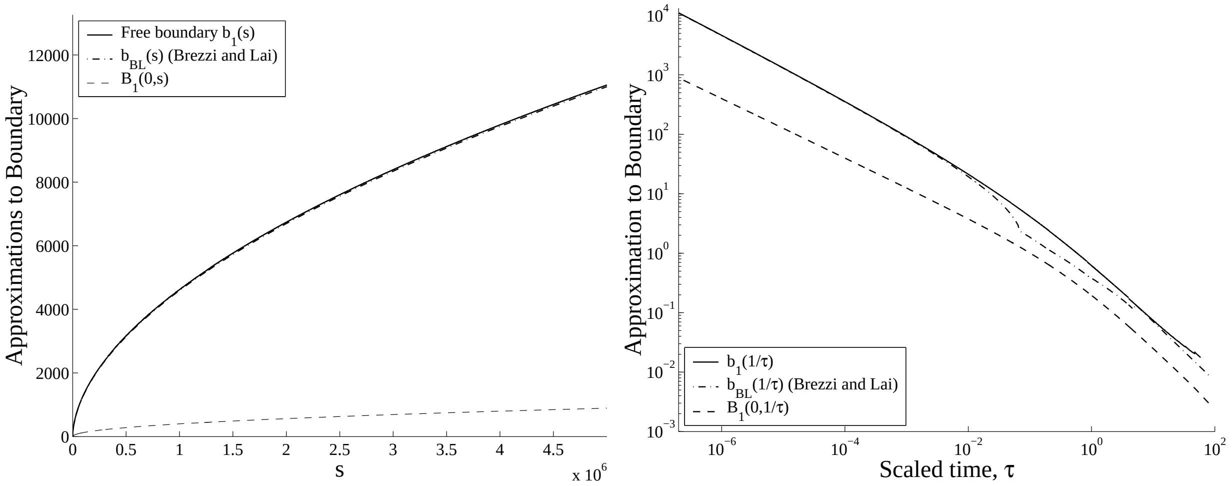


Figure EC.6 Free boundary $b_1(s) = b_1(1/\tau)$.

medium number of samples, some of the allocations can be negative. Techniques such as those used in the $\mathcal{LL}$ of Chick and Inoue (2001, for unknown variances) can be used to remedy any violations of a non-negativity constraint. Three, the allocation can be made sequential by updating statistics and repeatedly allocating replications until a stopping rule is satisfied. Four, the allocation can be extended to account for discounting by incorporating new stopping rules, such as EOC_1^γ and EOC_k^γ in §6.3, that discount the value of information from additional sampling.

In the notation of the current paper, these adaptations culminate in the following algorithm. The specification of a prior distribution replaces the first-stage of sampling that appears in a majority of other ranking and selection procedures.

Procedure $\mathcal{LL}$ (known variances).

1. Specify prior distributions for the unknown means Θ_i, with $\Theta_i \sim \texttt{Normal}\,(\mu_{0i}, \sigma_i^2/t_{0i})$, for each alternative. Set $y_{0i} = \mu_{0i}t_{0i}$ for each i, as in §4.1. Include $\mu_{00} = 0$ as an option so that the "do nothing" option is available (set σ_0^2 to be very small, e.g. 10^{-6} and t_{00} to be very large, e.g., 100 years worth of replications, for numerical reasons).
2. Determine the order statistics, so that $\mu_{0(0)} \le \mu_{0(1)} \le \ldots \le \mu_{0(k)}$.
3. WHILE stopping rule not satisfied DO another stage:

(a) Initialize the set of systems considered for additional replications, $\mathcal{S} \leftarrow \{0, 1, \ldots, k\}$.

(b) For each (i) in $\mathcal{S}\backslash\{(k)\}$: If $(k) \in \mathcal{S}$ then set $\lambda_{ik}^{-1} \leftarrow \hat{\sigma}_{(i)}^2/t_{0,(i)} + \hat{\sigma}_{(k)}^2/t_{0,(k)}$. If $(k) \notin \mathcal{S}$ then set $\lambda_{ik} \leftarrow t_{0,(i)}/\hat{\sigma}_{(i)}^2$.

(c) Tentatively allocate a total of r replications to systems $(i) \in \mathcal{S}$ (set $r_{(j)} \leftarrow 0$ for $(j) \notin \mathcal{S}$):

$$r_{(i)} \leftarrow \frac{(r + \sum_{j\in\mathcal{S}} t_j)(\sigma_{(i)}^2\gamma_{(i)})^{\frac{1}{2}}}{\sum_{j\in\mathcal{S}}(\sigma_j^2\gamma_j)^{\frac{1}{2}}} - t_{(i)}, \text{ where } \gamma_{(i)} \leftarrow \begin{cases} \lambda_{ik}^{1/2}\phi(d_{ik}^*) & \text{for } (i) \ne (k) \\ \sum_{(j)\in\mathcal{S}\backslash\{(k)\}}\gamma_{(j)} & \text{for } (i) = (k) \end{cases}$$

and $d_{ik}^* = \lambda_{ik}^{1/2}(\mu_{(k)} - \mu_{(i)})$.

(d) If any $r_i < 0$ then fix the nonnegativity constraint violation: remove (i) from $\mathcal{S}$ for each (i) such that $r_{(i)} \le 0$, and go to Step 3b. Otherwise, round the r_i so that $\sum_{i=1}^k r_i = r$ and go to Step 3e.

(e) Run r_i additional replications for system i, for $i = 1, 2, \ldots, k$. Update the sample statistics, $t_{0,i} \leftarrow t_{0,i} + r_i$; $y_{0i} \leftarrow y_{0i}+$ sum of r_i outputs for system i; $\mu_{0i} \leftarrow y_{0i}/t_{0i}$; and the order statistics, so that $\mu_{0(0)} \le \mu_{0(1)} \le \ldots \le \mu_{0(k)}$.

4. Select the system with the best estimated mean, $\mathfrak{D} = (k)$.

The value of r in Step 3c is taken to be $r = 1$ replication per stage for a fully sequential algorithm. The value of r can be increased if more replications per iteration are desired, e.g., if several replications per stage are run, or if several replications can be run in parallel during each stage. A computational speed-up can be obtained for the allocation, when $r = 1$, by ignoring the potential requirement to iterate through Steps 3a-3e, and by directly allocating one replication to the alternative that maximizes $r_{(i)}$ in the first pass through Step 3c.

Each stopping rule, EOC_1^γ and EOC_k^γ, formally identifies the sampling budget$\beta \ge 1$ that maximizes an approximation to the expected discounted value of continuing to run an additionalβ replications before selecting a system to implement. The approximations to the expected discounted value require that the β samples be allocated to the k systems with a one-stage allocation. We do that here by assigning $r \leftarrow \beta$ and allocating the samples with Steps 3a-3e. The determination of the optimal value of β incurs a computational cost that is associated, for example, with a line-search optimization algorithm for β. A computational speed-up can be obtained by simply checking if there exists a $\beta \geq 1$ such that the expected discounted value of sampling is positive. If that is the case, then the optimal β certainly has a positive expected discounted value of sampling. In our implementation, we initially solve for the optimal β. If that value exceeds 1, we continue sampling. In the next iteration, we check if a sampling budget of $\max\{1, \beta - 1\}$ leads to a positive expected discounted value of sampling. If this is so, we continue to sample. If not, we recheck the optimal value of $\beta \geq 1$ with a line search again.

Importantly, we note that the left hand sides of the inequalities that determine the stopping rules EOC_1^γ and EOC_k^γ are *not* monotonic in β. For example, when comparing $k = 1$ simulated alternative with a known deterministic NPV of 0, if the simulated mean is just below the stopping boundary, the expected reward of a one-step algorithm with $\beta = 1$ replication might not justify additional sampling. Nevertheless, some values of $\beta > 1$ may justify additional sampling. It is therefore not optimal to perform a one-step lookahead allocation by only testing if $\beta = 1$ additional replication is sufficient to justify continuing.

In the numerical experiments of §6.4, we implement the above algorithm with $r = 1$ replication allocated per stage, and with the preceding computational speedups.

References

Billingsley, P. 1986. *Probability and Measure*. 2nd ed. John Wiley & Sons, Inc., New York.

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

Breakwell, J., H. Chernoff. 1964. Sequential tests for the mean of a normal distribution II (large t). *Ann. Math. Stats.* **35** 162–163.

Brezzi, M., T. L. Lai. 2002. Optimal learning and experimenation in bandit problems. *J. Economic Dynamics & Control* **27** 87–108.

Chang, F., T. L. Lai. 1987. Optimal stopping and dynamic allocation. *Adv. Appl. Prob.* **19** 829–853.

Chernoff, H. 1961. Sequential tests for the mean of a normal distribution. *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Univ. California Press, 79–91.

Chernoff, H., A. J. Petkau. 1986. Numerical solutions for Bayes sequential decision problems. *SIAM J. Sci. Stat. Comput.* **7**(1) 46–59.

Chick, S. E., M. Hashimoto, K. Inoue. 2001. Bayesian sampling allocations for selecting the best population with different sampling costs and known variances. M. Xie, T. Z. Irony, Y. Hayakawa, eds., *System and Bayesian Reliability*. World Scientific, 333–349.

Chick, S. E., K. Inoue. 1998. Sequential allocation procedures that reduce risk for multiple comparisons. D. J. Medeiros, E. J. Watson, M. Manivannan, J. Carson, eds., *Proc. 1998 Winter Simulation Conference*. IEEE, Inc., Piscataway, NJ, 669–676.

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

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

Gittins, J. C. 1979. Bandit problems and dynamic allocation indices. *J. Royal Stat. Soc. B* **41** 148–177.

Gittins, J. C. 1989. *Multi-Armed Bandit Allocation Indices*. Wiley, New York.

Gittins, J. C., K. D. Glazebrook. 1977. On Bayesian models in stochastic scheduling. *J. Appl. Prob.* **14** 556–565.

Gittins, J. C., D. M. Jones. 1974. A dynamic allocation index for the sequential design of experiments. J. Gani, K. Sarkadi, J. Vincze, eds., *Progress in Statistics*. North-Holland, 241—-266.

Glazebrook, K. D. 1979. Stoppable families of alternative bandit processes. *J. Appl. Prob.* **16** 843–854.

Glazebrook, K. D. 1982. On a sufficient condition for superprocesses due to Whittle. *J. Appl. Prob.* **19**(1) 99–110.

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

Kim, S.-H., B. L. Nelson. 2006. On the asymptotic validity of fully sequential selection procedures for steady-state simulation. *Operations Research* **54**(3) 475–488.

Whittle, P. 1980. Multi-armed bandits and the Gittins index. *J. Royal Stat. Soc. B* **42** 143–149.