Submitted to *Management Science*
manuscript DRAFT

Economic Analysis of Simulation Selection Problems

Stephen E. Chick
INSEAD; Technology & Operations Management Area; Blvd. de Constance; 77305 Fontainebleau France; stephen.chick@insead.edu

Noah Gans
OPIM Department – Wharton; Univ. of Pennsylvania; 3730 Walnut Street, Suite 500; Philadelphia, PA 19104 U.S.A.; gans@wharton.upenn.edu

Ranking and selection procedures are standard methods for selecting the best of a finite number of simulated design alternatives, based on a desired level of statistical evidence for correct selection. But the link between statistical significance and financial significance is indirect and poorly understood. This paper presents a new approach to the simulation selection problem, one that maximizes the expected net present value (NPV) of decisions made when using stochastic simulation. We provide a framework for answering these managerial questions: When does a proposed system design, whose performance is unknown, merit the time and money needed to develop a simulation to infer its performance? For how long should the simulation analysis continue before a design is approved or rejected? We frame the simulation selection problem as a "stoppable" version of a Bayesian bandit problem that treats the ability to simulate as a real option prior to project implementation. For a single proposed system, we solve a free boundary problem for a heat equation that approximates the solution to a dynamic program that finds optimal simulation project stopping times and that answers the managerial questions. For multiple proposed systems, we extend previous Bayesian selection procedures to account for discounting and simulation-tool development costs.

Key words: simulation, ranking and selection, economics of simulation, optimal stopping, free boundary problem
History: A previous version was accepted, then withdrawn from Management Science. This version: 20 Dec 2007

Managers must decide the operating characteristics of their companies' manufacturing, supply chain, or service delivery systems. Often the decision reflects the choice of one among a number of competing designs. To aid their decision-making, managers may use stochastic or discrete event simulation. For a fixed, finite set of alternative designs, one must decide how long to simulate each alternative and, given the simulation results, which design to implement.

A common approach for selecting the best of a finite set of simulated systems uses ranking and selection procedures, which seek to provide a desired level of statistical evidence that the system with the best performance is ultimately selected. A typical measure of statistical evidence is the probability of correct selection (PCS). Good ranking and selection procedures attempt to minimize the mean number of replications

that are needed to reach a desired level of statistical evidence for correct selection. This is a flexible approach that allows one to assess a wide variety of operational and other measures of system performance.

But statistical significance is not the same as financial significance, and when system performance and simulation results are themselves financial measures, the maximization of expected net present value (NPV) may be a more appropriate objective (Brealey and Myers 2001). That is, if a manager's goal is to maximize the expected NPV of high-level system design choices, then she is faced with two countervailing costs. On the one hand, uncertainty about the expected NPV of each alternative compels her to simulate more to reduce the opportunity cost associated with an incorrect selection. On the other, a simulation analysis itself may incur direct costs, and simulation-driven delays in project implementation may reduce the NPV of the system that is ultimately implemented, due to discounting. Discounting and NPV are relevant in both the private and public public sectors. In health technology assessments, for example, one typically discounts technology costs and health benefits through time (Gold et al. 1996).

Further, standard practice for sound simulation studies (e.g., Law and Kelton 2000, §1.7) does not provide formal guidance via economic principles about whether or not an alternative should be simulated at all.

In this paper, we formulate and solve a simulation selection problem in which the manager seeks to maximize the expected NPV of the system eventually selected, less discounting and analysis costs. Our formulation of the problem is Bayesian: we assume that the manager has prior beliefs concerning the distribution of the NPV of each of the alternatives and that she uses simulation output to update these beliefs. The system which the manager ultimately chooses to implement maximizes expected NPV with respect to the posterior distributions of her beliefs (rather than the actual, but unknown, NPV). Section 1 defines the problem and identifies our assumptions, and §2 compares the formulation with more traditional approaches found in the simulation literature.

Section 3 shows that, among procedures that sequentially select systems to simulate and then stop to implement a system, there exists a deterministic, stationary policy that is optimal. Section 4 then provides asymptotic approximations for the optimal expected discounted reward of the simulation selection problem when there is exactly one simulated alternative. The analysis indicates how long one must simulate before choosing to implement or reject the alternative, given simulation output that is normally distributed with a

known variance. The asymptotic regime is reasonable given typical discount rates and simulation run times. The approximation is determined by the solution of a free boundary problem for a heat equation that shares characteristics with financial and real options. That theory is applied to illustrative simulation selection scenarios in §5 to demonstrate the economic value of our approach, and to show how a manager can use our results to decide whether or not a design proposal warrants the time and money that is required to develop simulation tools.

Section 6 extends the scope of our analysis to problems with more than one simulated alternative. It begins by noting that well-known sufficient conditions for the existence of an optimal "allocation" index, which could simplify the characterization of the optimal simulation selection policy, do not appear to hold. The characterization of an optimal selection procedure policy for multiple systems therefore remains an open question. Nevertheless, §6 extends previous work for Bayesian selection procedures that account for the expected value of information, but not for discounting, to the present context with discounting. In numerical examples, the new policies are shown to be close to optimal.

In summary, this paper presents a new approach to the simulation selection problem, one that maximizes the expected net present value (NPV) of decisions made when using stochastic simulation. The framework is designed to help answer these managerial questions: When does a proposed system design, whose performance is unknown, merit the time and money needed to develop a simulation to infer its performance? For how long should the simulation analysis continue before a design is approved or rejected? The contributions include: the framing the simulation selection problem as a "stoppable" version of a Bayesian bandit problem, one that treats the ability to simulate as a real option prior to project implementation; the solution to a free boundary problem for a heat equation that approximates the solution to a dynamic program that finds optimal simulation project stopping times; and extending previous Bayesian selection procedures to account for discounting and simulation tool development costs.

The Appendices in the Online Companion provide mathematical proofs and specify the numerical methods used in the paper. They also describe how to handle simulation output from one-parameter members of the exponential family of distributions (e.g., Bernoulli, exponential), autocorrelated output from steady-state simulations that are amenable to analysis with batch means, simulation run times that differ from one system

to the next, and a trick to parallelize the algorithm for multiple CPUs. It also develops a framework for future work by linking the simulation selection problem to variations of the well-known bandit problem.

1. Simulation Selection Problem Description

A manager seeks to develop one of k projects, labeled $i = 1, \ldots, k$. The net present value (NPV) of each of the i projects is not known with certainty, however. The manager wishes to develop the project which maximizes her expected NPV, or to do nothing if the expected present value of all projects is negative. We represent the "do nothing" option as $i = 0$ with a sure NPV of zero.

1.1. Uncertain Project NPV's

Let X_i be the random variable representing the NPV of project i, where $X_0 \equiv 0$. If the manager is risk neutral and the distributions of all X_i's are known to her, then she will select the project with the largest expected NPV, $i^* = \arg\max_i \{\mathrm{E}[X_i]\}$.

We note that, although we model NPVs as simple random variables, the systems that generate them may be quite complex. For example, a particular project's sequence of cash flows may involve the composition of several interrelated random processes describing the evolution of investments, $\mathcal{I}(v)$, revenues, $\mathcal{R}(v)$, and operating costs, $\mathcal{O}(v)$, over time, v. Nevertheless, given a continuous-time discount rate $\delta > 0$, each realization of these processes, ω_i, yields a sample $X(\omega_i) = \int_{v=0}^{\infty} [\mathcal{R}(v, \omega_i) - \mathcal{O}(v, \omega_i) - \mathcal{I}(v, \omega_i)] e^{-\delta v} dv$. Here, v is the time elapsed from the moment a system is selected. (The letter t is used differently below.)

Fox and Glynn (1989) and Appendix C.2 suggest techniques for sampling the $X(\omega_i)$ if the time horizon is truly infinite. Projects that are to be used for only a finite time (e.g., a 5-year usable time horizon) can be implemented with terminating simulations, which effectively set $\mathcal{R}(v, \omega_i) - \mathcal{O}(v, \omega_i) - \mathcal{I}(v, \omega_i)$ to 0 during all but a finite interval. Fixed or random duration delays from the time a project is selected to the time of implementation (due to the need for project approval or startup delays) can similarly be implemented by setting $\mathcal{R}(v, \omega_i)$, $\mathcal{O}(v, \omega_i)$ or $\mathcal{I}(v, \omega_i)$ to 0 during an initial interval.

This approach to modeling delays is valid whenever they are statistically independent of the duration of the simulation analysis that led to the selection. This precludes fixed, pre-scheduled implementation dates, which can occur in practice. Nevertheless, the analysis below suggests that such pre-scheduled implementation

dates may themselves be suboptimal, given that they ignore a manager's option to implement earlier or to pursue additional analysis, depending on simulation results obtained up to that fixed date.

In this paper, we assume that the distributions of the X_i's are not known with certainty by the manager. Rather, she believes that a given X_i comes from one of a family of probability distributions, $P_{X_i|\theta_i}$, indexed by parameter θ_i from a parameter space Ω_{Θ_i}. We model her belief with a probability distribution on θ_i, which we call P_{Θ_i}. For example, the manager may believe that X_i is normally distributed with a known variance, σ_i^2, but unknown mean. Then P_{Θ_i} represents a probability distribution for the mean. To ease notation, we sometimes refer to the distribution as Θ_i. The expected NPV of project $i > 0$ is then $\mathrm{E}[X_i] = \mathrm{E}[X(\Theta_i)] \triangleq \iint X(\theta_i) dP_{X_i|\theta_i} dP_{\Theta_i}$. We denote the vector of distributions for the projects by $\mathbf{\Theta} = (\Theta_1, \ldots, \Theta_k)$.

1.2. Using Simulation to Select the Best Project

If the distributions of the X_i's are not known, then the manager may be able to use simulation as a tool to reduce distributional uncertainty, before having to decide which project to develop. She may decide to simulate the outcome of project i a number of times, and she views the result of each run as a sample of X_i.

We model the running of simulations as occurring at a sequence of discrete stages $t = 0, 1, 2, \ldots$. Let $\mathbf{X}_t$ be the set of all outputs seen through stage t. We represent Bayesian updating of prior beliefs and sample outcomes through time, $\{(\mathbf{\Theta}_t, \mathbf{X}_t) \,|\, t = 0, 1, \ldots\}$ as follows. If project $i > 0$ is simulated at stage t with sample outcome $x_{i,t}$, then $X_{i,t} = x_{i,t}$ and Bayes' rule determines the posterior distribution $\Theta_{i,t+1}$, which is a function of the parameter θ_i:

$$dP_{\Theta_{i,t+1}}(\theta_i \,|\, x_{i,t}, \Theta_{i,t}) \;=\; \frac{dP_{X_i \,|\, \theta_i}(x_{i,t} \,|\, \theta_i)\, dP_{\Theta_{i,t}}(\theta_i)}{\int_{\theta_i} dP_{X_i \,|\, \theta_i}(x_{i,t} \,|\, \theta_i)\, dP_{\Theta_{i,t}}(\theta_i)} \qquad \forall\, \theta_i \;\in\; \Omega_{\Theta_i}, \tag{1}$$

while $\Theta_{j,t+1} = \Theta_{j,t}$, and $X_{j,t}$ need not be defined for all $j \neq i$. Thus, the evolution of the manager's beliefs regarding the distribution of outcomes of each project, $\Theta_{i,t}$, is Markovian. We also assume that simulation results, hence the evolution of the manager's beliefs, are independent from one project to the next.

If, in theory, simulation runs could be performed at zero cost and in no time, then the manager might simulate each of the k systems infinitely, until all uncertainty regarding the θ_i's was resolved. At this point the problem would revert to the original case in which the distributions and means of the X_i are known.

But the simulation runs do take time and do cost money. We assume that the marginal cost of each run of system i is $\$c_i$ and takes η_i units of time to complete. Thus, given a continuous-time discount rate of $\delta > 0$, the decision to simulate system i once costs the manager c_i plus a reduction of $\Delta_i = \int_0^{\eta_i} e^{-\delta s} ds < 1$ times the expected NPV of the (unknown) project that is eventually chosen.

There may also be associated up-front costs associated with the development of the simulation tool, itself. It may require time and money to develop an underlying simulation platform, independent of which projects end up being evaluated. Additional costs may be required to be able to simulate particular projects.

This paper initially makes two assumptions regarding the costs of simulation that simplify the analysis. First, we assume that the up-front costs and delays to develop the simulation tools are sunk for all k projects. This is an implicit assumption of all other research on selection procedures. Second, we assume that $\eta_i \equiv \eta$ for all k projects. This allows us to define a common $\Delta \equiv \Delta_i$ for the projects as well. Section 4.4 relaxes the first assumption. Appendix C relaxes the second.

Even with these simplifications, the availability of a simulation tool to sample project outcomes makes the manager's problem much more complex. Rather than simply choosing the project that maximizes expected NPV, she must choose a sequence of simulation runs and ultimately select a project, so that the discounted stream of costs and terminal expected value, together, maximize expected NPV.

To track the manager's choices as they proceed, we define a number of indices. We let $T \in \{t = 0, 1, 2, \ldots\}$ be the stage at which the manager selects a system to implement. For $t < T$, we define $i(t) \in \{1, \ldots, k\}$ to be the index of the project simulated at time t. We set $I(T) \in \{0, \ldots, k\}$ to be the ultimate choice of project.

Then a *selection policy* is the choice of a sequence of simulation runs, a stopping time, and a final project. We define Π to be the set of all *non-anticipating* selection policies, whose choice at time $t = 0, 1, \ldots$ depends only on the history up to t: $\{\boldsymbol{\Theta}_0, \mathbf{X}_0, \ldots, \boldsymbol{\Theta}_{t-1}, \mathbf{X}_{t-1}, \boldsymbol{\Theta}_t\}$. Given prior distributions $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_k)$ and policy $\pi \in \Pi$, the expected discounted value of the future stream of rewards is

$$V^{\pi}(\boldsymbol{\Theta}) \;=\; \mathrm{E}_{\pi}\left[\sum_{t=0}^{T-1} -\Delta^t c_{i(t)} \;+\; \Delta^T X_{I(T),T} \;|\, \boldsymbol{\Theta}_0 = \boldsymbol{\Theta}\right]. \tag{2}$$

Formally, we define the manager's *simulation selection problem* to be the choice of a selection policy $\pi^* \in \Pi$ that maximizes this expected discounted value: $V^{\pi^*}(\boldsymbol{\Theta}) = \sup_{\pi \in \Pi} V^{\pi}(\boldsymbol{\Theta})$.

2. Literature Review

Two broad classes of research are related to this paper. One is the ranking and selection literature, the other is the bandit and optimal stopping literature. Both have substreams.

Branke et al. (2007) review several statistical approaches to ranking and selection. See also Nelson and Goldsman (2001) and Butler et al. (2001). To date, none of the approaches explicitly accounts for discounting costs due to delays in implementation as a result of simulation times, and only two papers explicitly account for the cost of sampling.

Chick and Inoue (2001) provide two-stage procedures whose second stage allocation trades off the cost of sampling with an approximation to the Bayesian expected value of information (EVI) of those samples. Sampling costs may differ for each system as may the unknown sample variances. Their work builds upon earlier results of Gupta and Miescke (1996), who examined the case of known sampling variances, and a fixed number of samples to be allocated. The EVI is measured with respect to one of two loss functions, the posterior probability of incorrect selection (PICS), or the posterior expected opportunity cost (EOC) of a potentially incorrect selection. The EOC is a first step for modeling financial value in selection procedures.

The indifference-zone (IZ) approach provides a frequentist guarantee of selection procedure effectiveness (Kim and Nelson 2006). Almost all IZ procedures focus on probability of correct selection (PCS) guarantees for each problem instance within a given class, and most IZ work ignores the sampling costs of replications. An exception is Hong and Nelson (2005), who account for the cost of switching from one system to another, and a common sampling cost for each system. In separate work that is related to this paper, Kim and Nelson (2006) use diffusion approximations for sequential IZ screening procedures to reduce the simulation time required to guarantee a desired PCS.

Branke et al. (2007) show that specific Bayesian procedures that allocate samples with an EOC criterion, and new adaptive stopping rules, perform very effectively for several classes of selection problems.

In another stream of literature, Gittins (1979) offers an early account of optimal dynamic allocation indices (later called Gittins indices) for infinite-horizon, discounted multi-armed bandit problems. Glazebrook (1979) provides sufficient conditions under which these index results apply to reward streams derived from stoppable arms. Gittins (1989) shows that Glazebrook's results hold under a slightly weaker set of assumptions.

In general, Gittins indices are difficult to compute exactly. Chang and Lai (1987) derive approximations for the Gittins index for the infinite-horizon discounted "Bayesian bandit" problem. Brezzi and Lai (2002) use a diffusion approximation for the Gittins index of a Bayesian bandit that is motivated by work of Chernoff (1961) on composite hypothesis testing (also see Breakwell and Chernoff 1964, Chernoff 1965).

This paper uses Chernoff-like diffusion approximations to solve the simulation selection problem (with $k = 1$ system) in an asymptotically optimal way. That asymptotically-optimal solution is shown to provide an improved approximation to the Gittins index of Brezzi and Lai's Bayesian bandit problem (see Appendix D).

We will show that the simulation selection problem, with $k \geq 1$ systems, is an example of what Glazebrook called a *stoppable family of alternative bandit processes*. We indicate that Glazebrook's sufficient conditions for a Gittins index to exist for "stoppable bandits" do not appear to be satisfied, so the existence of an Gittins index for the simulation selection problem is an open question. Still, we show that EOC-based sampling allocations like those in Chick and Inoue (2001), together with new stopping rules, are effective solutions for the simulation selection problem.

3. Preliminaries

This section shows that, given mild technical conditions, a simple class of stationary and deterministic policies, which we call "reasonable," is optimal for the simulation selection problem. It then further characterizes the simulation selection problem with $k = 1$ system, to prepare for our approximation results in §4.

We begin by noting that a policy is *stationary* if the action it prescribes, given state $\mathbf{\Theta}_t = (\Theta_{1,t}, \ldots, \Theta_{k,t})$, is independent of the time index, t. A policy is *deterministic* if the action it prescribes is never randomized. Blackwell (1965) has shown that, in infinite-horizon problems with discounted rewards, the following conditions ensure that there exists a deterministic, stationary policy that is optimal: 1) given any state and action, expected one-period rewards are finite; 2) the same, finite set of actions is available in all states.

The proof of Lemma 1 shows how the original problem formulation can meet these conditions. The potentially finite stopping time of the simulation selection problem is converted to an infinite horizon by converting the one-time reward $\mathrm{E}[X(\Theta_{I(T),T})]$ at the simulation problem stopping time T into a perpetuity $(1-\Delta)\mathrm{E}[X(\Theta_{I(T),T})]$ that is received at each period $t \geq T$. Without loss of generality, then, we can restrict our attention to the class of stationary, deterministic selection policies for the infinite-horizon problem.

LEMMA 1. *Suppose expected one-period rewards are uniformly bounded for the simulation selection problem in (2). Then there exists a deterministic, stationary policy* $\pi^* \in \Pi$ *that is optimal.*

See Appendix B in the Online Companion for proofs of all claims.

Now consider the infinite-horizon version of the simulation selection problem with a single project, i, and no outside alternative. For this problem we call the stopping time T_i, let $i(t) = i$ for $t < T_i$ and let $I(t) = i$ for $t \geq T_i$. Thus, π_i determines a stopping time, T_i, and an associated expected value,

$$V_i^{\pi_i}(\Theta_i) \;=\; \mathrm{E}_{\pi_i}\left[\sum_{t=0}^{\infty} \Delta^t \, R_t^{\pi_i} \,|\, \Theta_{i,0} = \Theta_i\right] \;=\; \mathrm{E}_{\pi_i}\left[\sum_{t=0}^{T_i-1} -\Delta^t c_i \;+\; \Delta^{T_i}\mathrm{E}[X(\Theta_{i,T_i})] \,|\, \Theta_{i,0} = \Theta_i\right]. \quad (3)$$

We denote the optimal stopping policy and stopping time for project i as π_i^* and T_i^*. If expected one-period rewards are uniformly bounded, then there exists a stationary, deterministic policy that is optimal. Further, the optimal value function satisfies the so-called Bellman equation (Bertsekas and Shreve 1996, Prop. 9.8):

$$\begin{aligned} V_i^{\pi_i^*}(\Theta_{i,t}) &= \max\left\{-c_i + \Delta\mathrm{E}[V_i^{\pi_i}(\Theta_{i,t+1}) \,|\, \Theta_{i,t}, t \neq T_i],\; (1-\Delta)\mathrm{E}[X(\Theta_{i,t})] + \Delta\mathrm{E}[V_i^{\pi_i^*}(\Theta_{i,t})]\right\} \\ &= \max\left\{-c_i + \Delta\mathrm{E}[V_i^{\pi_i}(\Theta_{i,t+1}) \,|\, \Theta_{i,t}, t \neq T_i],\; \mathrm{E}[X(\Theta_{i,t})]\right\}. \end{aligned} \quad (4)$$

We call $V_i^{\pi_i^*}(\Theta_{i,t})$ the optimal expected discounted reward (OEDR) for the option to simulate alternative i before deciding whether to implement it or not.

Since setting $T_i = \infty$ is a feasible (though not necessarily optimal) stationary policy, we know that $V_i^{\pi_i}(\Theta_{i,t}) \geq -c_i/(1-\Delta)$. In turn, from (4) it follows that an optimal policy will never choose the right maximand, and stop simulating, if $(1-\Delta)\mathrm{E}[X(\Theta_{i,t})] < -c_i$.

More generally, we call any stationary, deterministic stopping policy, π_i, *reasonable* if $T_i = t < \infty$ implies $(1-\Delta)\mathrm{E}[X(\Theta_{i,t})] \geq -c_i$. Thus a reasonable policy never stops when the one-period expected revenue from a project falls below the cost of sampling. Similarly, a reasonable policy for the entire simulation selection problem has $T = t$ and $I(t) = i$ only if $(1-\Delta)\mathrm{E}[X(\Theta_{I(t),t})] \geq -c_{I(t)}$.

LEMMA 2. *An optimal deterministic, stationary policy for the simulation selection problem is reasonable, almost surely.*

Thus, without loss of generality, we can restrict our attention to the analysis of reasonable policies.

4. A Value Function Approximation for One Alternative

A normative solution to the analysis of single simulated alternative requires the evaluation of the OEDR, $V_i^{\pi_i^*}(\Theta_{i,t})$. In this section we develop diffusion approximations that provide structural insight into the form of the OEDR and allow for its efficient computation. Our approach follows in the spirit of Chernoff (1961).

This section assumes $k = 1$, so to simplify notation we drop the system index, i. It also assumes that the simulation output X_j is i.i.d. $\texttt{Normal}(\theta, \sigma^2)$ for replication $j = 1, 2, \ldots$, with a known finite variance σ^2 and unknown mean θ. We suppose that θ has a $\texttt{Normal}(\mu_0, \sigma_0^2)$ prior distribution. While this assumption may not satisfy the uniform boundedness condition in Lemma 1, the analysis below results in a well-defined finite OEDR when σ_0^2 is finite.

The diffusion approximations are asymptotically appropriate when the discounting over the duration of a simulation replication is small, as is usually the case in simulation. Repeated sampling leads to realizations of a scaled Brownian motion with drift.

The calculation of the OEDR involves the solution of a so-called free boundary problem for a heat equation that is obtained from the diffusion approximation. The boundary is "free" since it is determined by equating the two maximands in the value function, rather than on a known, pre-specified boundary. A comparison of the maximands in the continuous-time analogue of (4) determines the free boundary between a continuation set, $\mathcal{C}$, in which it is optimal to continue simulating a project, and a stopping set, in which it is optimal to stop simulating and implement the project.

We motivate the diffusion approximation, present a standardized free boundary problem for that diffusion, and solve for the special case of $c = 0$. The solution when $c > 0$ is proven to be a function of the solution when $c = 0$. We then derive the solution to the optimal stopping problem when comparing $k = 1$ simulated alternative with an alternative that has a known deterministic NPV. This section concludes by showing whether or not a simulation tool for the $k = 1$ alternative should be implemented in the first place.

4.1. Diffusion Approximation for the Output of One System

Define $n_0 = \sigma^2/\sigma_0^2$, and redefine $t = n_0 + n$, where n is the number of simulation observations seen so far for the single system in question. Set $Y_t = n_0\mu_0 + \sum_{j=1}^{n} X_j$. This transformation conveniently makes the

posterior distribution of θ a `Normal` $(Y_t/t, \sigma^2/t)$ distribution, and will help to find an optimal stopping time for (3) when there is $k = 1$ system.

Proceeding informally at first, 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

$$B(y_t, t) = \max\{\lim_{h\to 0} -ch + e^{-\delta h} \times \mathrm{E}[B(Y_{t+h}, t+h) \mid y_t, t], y_t/t\}, \tag{5}$$

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 `Normal` $(\theta h, \sigma^2 h)$, and the posterior distribution of θ at time t is `Normal` $(y_t/t, \sigma^2/t)$. So the marginal distribution of U is `Normal` $(hy_t/t, \sigma^2(h + h^2/t))$, and

$$\begin{aligned} B(y_t, t) &= \max\{\lim_{h\to 0} -ch + e^{-\delta h} \times E_U[B + UB_y + hB_t + \frac{1}{2}U^2 B_{yy}] + o(h), y_t/t\} \\ &= \max\{\lim_{h\to 0} -ch + (1 - h\delta) \times (B + h\frac{y_t}{t}B_y + hB_t + h\frac{\sigma^2}{2}B_{yy}) + o(h), y_t/t\}, \end{aligned} \tag{6}$$

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 $\mathcal{C}$:

$$0 = -c - \delta B + \frac{y}{t}B_y + B_t + \frac{\sigma^2}{2}B_{yy}. \tag{7}$$

The boundary, $\partial\mathcal{C}$, of $\mathcal{C}$ will be determined by equating the two maximands in (5), as well as a smooth pasting condition (Chernoff 1961),

$$\begin{aligned} B(y, t) &= D(y, t), \text{ on } \partial\mathcal{C} \\ B_y(y, t) &= D_y(y, t), \text{ on } \partial\mathcal{C} \text{ (smooth pasting).} \end{aligned} \tag{8}$$

The basic problem is to solve for the value function, B, and the free boundary, $\partial\mathcal{C}$, determined by (7-8). When $c = 0$, (7-8) represent what might be called a perpetual American call option on regular (rather

than geometric) Brownian motion, with unknown drift. Equation (7) 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.

Insights and numerical solutions will be facilitated by rewriting (7) in reverse time, via the change of variables $w_s = y_t/\sigma t$, $s = 1/t$. (If $\sigma = 1$, then w is the posterior mean of θ and s is its posterior variance.) Set $t_0 = n_0$ and $s_0 = 1/t_0$. Then w_s is a Brownian motion in the $-s$ time scale, going backwards from s_0 to 0, with initial point (s_0, w_{s_0}) (Chernoff 1961), and (7) becomes

$$0 = -\frac{c + \delta B}{s^2} - B_s + \frac{1}{2}B_{ww}. \tag{9}$$

The boundary condition becomes $B = D$, with $D(w, s) = \max\{-c/\delta, \sigma w\}$ in (w, s) coordinates for $s \geq 0$, where only the second maximand can be chosen if $s > 0$. The first maximand represents simulating forever, and can only be selected, upon stopping, if $s = 0$. (This follows the idea of a reasonable policy.) The analysis below uses a similar (slightly different) normalization to approximate the expected reward, $B(y_t, t)$.

4.2. Standardized Free Boundary Problem for Optimal Stopping

The general free boundary problem that is determined by (7-8) depends on many parameters. In the spirit in which problems with normal distributions are analyzed using z-statistics, we can rescale specific instances of the diffusion process to obtain a standardized free boundary problem for optimal stopping. To do this we define a new time scale, $\tau = \gamma t$. Set $\tau_0 = \gamma t_0$. Let $Z_\tau = \alpha Y_t$ be a scaled motion with $z_{\tau_0} = z_0 = \alpha Y_{t_0}$.

This transformation 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 (6) and (7) that is asymptotically appropriate as $h = \gamma \to 0$ (e.g. Billingsley 1986, Section 37).

Let $\mu = \beta\theta$ be a rescaled drift parameter. So $\mathrm{E}[Z_\tau] = \mu\tau = \mathrm{E}[\alpha Y_t] = \alpha\theta t = \frac{\alpha\mu}{\beta\gamma}\tau$. If $\alpha/\beta\gamma = 1$ then the drift of Z_τ is μ. Also, $\mathrm{Var}[Z_\tau] = \alpha^2\mathrm{Var}[Y_t] = \alpha^2\sigma^2 t = \frac{\alpha^2\sigma^2}{\gamma}\tau$, so Z_τ has unit variance per time unit if $\alpha^2\sigma^2 = \gamma$. Those two moment relations constrain the set of suitable choices of γ, α, β. The third constraint, which is needed to identify the three parameters, is chosen after examining whether or not c equals 0.

4.2.1. Discounting Costs Only ($c = 0, \delta > 0$) Suppose the marginal cost of additional replications is essentially nil ($c = 0$), e.g. if analyst and computer time are considered to be sunk costs, but simulation delays discount a project's value ($\delta > 0$). In standardized coordinates, the reward function is

$$R(y_t, t) = e^{-\delta(t-t_0)} D(y_t, t) = e^{-(\tau-\tau_0)\delta/\gamma} \frac{\gamma}{\alpha} \frac{z_\tau}{\tau} = R(z, \tau).$$

When $T = \infty$ the reward is 0, the NPV of simulating forever without implementing.

The expectation in (3) is approximated asymptotically by $\mathrm{E}_{\tilde{\tau}^*}[R(Z_{\tilde{\tau}^*}, \tilde{\tau}^*)|\tau_0, z_0]$ for some suitable, measurable continuous-time policy $\tilde{\pi}$ with optimal stopping time $\tilde{\tau}^* \geq \tau_0$. That stopping time also maximizes $\frac{\alpha}{\gamma}\mathrm{E}_{\tilde{\tau}\geq\tau_0}[R(Z_{\tilde{\tau}}, \tilde{\tau})|\tau_0, z_0] = \mathrm{E}_{\tilde{\tau}\geq\tau_0}\left[e^{-(\tilde{\tau}-\tau_0)\delta/\gamma} Z_{\tilde{\tau}}/\tilde{\tau}|\tau_0, z_0\right]$.

We choose the parameters to standardize the loss function ($\delta/\gamma = 1$) and match the diffusion parameters ($\alpha/\beta\gamma = 1$ and $\alpha^2\sigma^2 = \gamma$). This parametrization requires

$$\alpha = \delta^{1/2}\sigma^{-1}, \beta = \delta^{-1/2}\sigma^{-1} \text{ and } \gamma = \delta, \tag{10}$$

and allows us to solve a standardized problem,

$$B(z_0, \tau_0) = \mathrm{E}_{\tilde{\tau}^*\geq\tau_0}\left[e^{-(\tilde{\tau}^*-\tau_0)} \frac{Z_{\tilde{\tau}^*}}{\tilde{\tau}^*}\middle|\tau_0, z_0\right] = \sup_{\tilde{\tau}\geq\tau_0} \mathrm{E}_{\tilde{\tau}}\left[e^{-(\tilde{\tau}-\tau_0)} \frac{Z_{\tilde{\tau}}}{\tilde{\tau}}\middle|\tau_0, z_0\right], \tag{11}$$

for stopping times $\tilde{\tau}$ of the Wiener process Z. Given $c = 0$, along with a standardized discount rate of 1, the diffusion equation for Problem (11) becomes $0 = -B + \frac{z}{\tau}B_z + B_\tau + \frac{1}{2}B_{zz}$, for $(z, \tau) \in \mathcal{C}$, with $D(z, \tau) = z/\tau$ and $B(z, \tau) = D(z, \tau)$ on $\partial\mathcal{C}$.

Finally, it is useful to rewrite these equations in the coordinates

$$s = 1/\tau \text{ and } w_s = z_\tau/\tau,$$

with $s_0 = 1/\tau_0$, and $W(s_0) = w_0 = z_{\tau_0}/\tau_0$. Then W is a Brownian motion in the $-s$ scale starting at (s_0, w_0). Each distribution for the unknown mean maps to a point in the (s, w) plane. Problem (11) becomes

$$B(w_0, s_0) = \sup_{0\leq S\leq s_0} \mathrm{E}_{\tilde{\tau}}\left[e^{-(1/S-1/s_0)} W_S \mid w_0, s_0\right]. \tag{12}$$

In summary, Problem (12) determines the OEDR $B_1(w, s)$ and free boundary $b_1(s)$ of a standardized simulation selection problem. The subscript "1" refers to this first case, $c = 0$. The free boundary is the

curve where the value function's maximands are equal, $B_1 = D$. The solution can be approached with the following free boundary problem (obtained from the PDE above with the chain rule).

$$\begin{aligned} 0 &= -\frac{B_1}{s^2} - B_{1,s} + \frac{1}{2}B_{1,ww} \qquad (13)\\ D(w,s) &= \max\{0, w\} \\ B_1 &= D \text{ and } B_{1,w} = D_w, \text{ on the free boundary } \partial\mathcal{C}. \end{aligned}$$

Let $\mathcal{V}_1$ denote the diffusion approximation for the OEDR in (y,t) coordinates. It equals the supremum of expected rewards over all stopping rules for the diffusion approximation of the simulation stopping problem.

THEOREM 1. *The free boundary $\partial\mathcal{C}$ of the continuation set for the standardized problem in (13) is a function $b_1(s) \geq 0$. The OEDR $B_1(w_0, s_0)$ can be converted to (y,t) coordinates to obtain the OEDR for the unscaled diffusion process, $\mathcal{V}_1(y_{t_0}, t_0) = \sigma\sqrt{\delta}B_1(w_0, s_0) \geq \max\{0, y_{t_0}/t_0\}$, for points in the continuation set $\mathcal{C} = \{(w,s) : w < b_1(s)\} = \{(y,t) : y/t < \sigma\sqrt{\delta}b_1(1/\delta t)\}$.*

Thus, if $y_t/t < \sigma\sqrt{\delta}b_1(1/\delta t)$ then it is optimal to simulate, and after a simulation replication the theorem can be used to update the OEDR for the posterior distribution, which becomes the prior distribution for the next stage. If $y_t/t \geq \sigma\sqrt{\delta}b_1(1/\delta t)$, then discounting costs outweigh the value of gathering additional information from more simulations, and there is a higher value $\mathcal{V}_1(y_t, t) = y_t/t$ to implementing immediately.

Theorem 2 characterizes the asymptotics of the stopping boundary – its proof shows that $b_1(s)$ is related to the optimal stopping boundary of a different problem that was considered by Brezzi and Lai (2002).

THEOREM 2. *$b_1(s) \doteq s/\sqrt{2}$ as $s \to 0$ and $b_1(s) \doteq s^{1/2}\left(2\log s - \log\log s - \log 16\pi\right)^{1/2}$ as $s \to \infty$.*

Appendix D shows how we computed B_1 and b_1 for the numerical examples below. The computations make use of the following lemma, which is also used below in the main paper. The lower bound is obtained by examining *one-stage policies* where ß replications are observed, and then the system is selected if the posterior mean exceeds $-c/\delta$ (=0 here), and is rejected in favor of infinite simulation replications otherwise.

LEMMA 3. *Let $\Psi[s] = \int_s^\infty (\xi - s)\phi(\xi)d\xi = \phi(s) - s(1 - \Phi(s))$ be the Newsvendor loss function, ϕ be the pdf, and Φ be the cdf of a standard normal distribution. Then*

$$B(y_0, t_0) \geq \underline{B}(y_0, t_0) \triangleq \sup_{\beta \geq 0} e^{-\delta\beta}\left(-\frac{c}{\delta} + \left(\frac{\sigma^2\beta}{t_0(t_0+\beta)}\right)^{1/2}\Psi\left[-\left(\frac{y_0}{t_0} + \frac{c}{\delta}\right) \Big/ \left(\frac{\sigma^2\beta}{t_0(t_0+\beta)}\right)^{1/2}\right]\right). \quad (14)$$

4.2.2. Both Sampling and Discounting Costs ($c, \delta > 0$) A similar analysis, with a conversion to standardized coordinates, can be applied when both sampling and discounting costs are relevant ($c > 0$, $\delta > 0$).

$$\begin{aligned} R(y_t, t) &= -\int_{t_0}^{t} ce^{-\delta(\xi - t_0)} d\xi + D(y_t, t) e^{-\delta(t-t_0)} \\ &= -\frac{c}{\delta}(1 - e^{-\frac{(\tau-\tau_0)\delta}{\gamma}}) + e^{-\frac{(\tau-\tau_0)\delta}{\gamma}} \frac{c}{\delta} \frac{\delta\gamma}{c\alpha} \frac{z_\tau}{\tau} = R(z_\tau, \tau). \end{aligned} \tag{15}$$

Set $\kappa = \delta\gamma/c\alpha$. Then (3) is approximated asymptotically by $\mathcal{V}(y_{t_0}, t_0) = \sup_{\tilde{\tau} \geq \tau_0} \mathrm{E}_{\tilde{\tau}}[R(Z_\tau, \tau) \mid z_0, \tau_0]$ for some suitable measurable, continuous-time selection policy $\tilde{\pi}$ with optimal stopping time $\tilde{\tau}^* \geq \tau_0$. 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

$$\frac{\delta}{c}\mathrm{E}_{\tilde{\tau} \geq \tau_0}[R(Z_{\tilde{\tau}}, \tilde{\tau}) | z_0, \tau_0] = \mathrm{E}_{\tilde{\tau} \geq \tau_0}\left[-1 + e^{-\frac{(\tilde{\tau}-\tau_0)\delta}{\gamma}}(1 + \kappa Z_{\tilde{\tau}}/\tilde{\tau}) \,\middle|\, z_0, \tau_0\right]. \tag{16}$$

The problem of finding the optimal $\tilde{\tau}^* \geq \tau_0$ to maximize (16) 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

$$\alpha = \delta^{1/2}\sigma^{-1}, \beta = \delta^{-1/2}\sigma^{-1}, \gamma = \delta \text{ and } \kappa = \delta^{3/2}\sigma c^{-1}. \tag{17}$$

Given (16), 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:

$$\frac{\delta}{c}\mathcal{V}(y_{t_0}, t_0) = \frac{\delta}{c}\mathrm{E}_{\tilde{\tau}^* \geq \tau_0}[R(Z^*_{\tilde{\tau}}, \tilde{\tau}^*) \mid z_0, \tau_0] = -1 + \sup_{\tilde{\tau} \geq \tau_0} \mathrm{E}_{\tilde{\tau} \geq \tau_0}\left[(1 + \kappa Z_{\tilde{\tau}}/\tilde{\tau}) e^{-(\tilde{\tau} - \tau_0)} \middle| z_0, \tau_0\right]. \tag{18}$$

Theorem 3 says that the OEDR when $c > 0$ is directly related to the OEDR in Theorem 1 (with $c = 0$), and that the continuation set for y/t is shifted by $-c/\delta$. This implies that only one free boundary problem must be solved to handle any values of $c \geq 0$ and $\delta > 0$.

THEOREM 3. *Let $b_1(s)$ be the free boundary and $B_1(w, s)$ be the OEDR from Theorem 1 for the case $c = 0, \delta > 0$. Set $W_s = Z_\tau/\tau$ and $s = 1/\tau$ as in §4.2.1. Then the optimal stopping time $\tilde{\tau}^* \geq \tau_0$ for the standardized problem in (18) with $c, \delta > 0$, derived from the parametrization of (17), is to stop when $W_s \geq b_1(s) - 1/\kappa$. Moreover the OEDR when $c, \delta > 0$ is*

$$\mathcal{V}_{2(\kappa)}(y_{t_0}, t_0) = \beta^{-1}\left(B_1(w_0 + 1/\kappa, s_0) - \frac{1}{\kappa}\right) = \sigma\sqrt{\delta} B_1\left(\frac{1}{\sigma\sqrt{\delta}}(\frac{y_{t_0}}{t_0} + \frac{c}{\delta}), \frac{1}{\delta t_0}\right) - \frac{c}{\delta}, \tag{19}$$

for points in the continuation set $\mathcal{C} = \{(w,s) : w < b_1(s) - 1/\kappa\} = \{(y,t) : y/t < \beta^{-1}b_1(1/\delta t) - c/\delta\}$.

Note that the formula $b(t) = \beta^{-1}b_1(1/\delta t) - c/\delta$ for the boundary of $\mathcal{C}$ is valid for both $c = 0$ and $c > 0$. We define $b^{-1}(m) = \sup\{t : b(t) \geq m\}$ for all $m > -c/\delta$. We also note that $b(t)$ is monotone decreasing and continuous for sufficiently large and sufficiently small t (Theorem 2), and that $b(t)$ is also monotone decreasing and continuous for all t (and thus invertible for $m > -c/\delta$) in numerical experiments (Appendix D).

4.3. Comparing a Single Simulated System to a Known Alternative

The analysis in §4.2 requires that one either simulate or implement a single system. In this case, given a simulated system whose E[NPV] is far below $-c/\delta$ with high probability, it is optimal to simulate forever, rather than to implement. Alternatively, one may wish either to simulate, to stop and implement the simulated system, or to stop and obtain a known deterministic NPV whose value is m. If the known deterministic alternative is to "do nothing"/maintain the status quo, then $m = 0$. An arbitrary $m \neq 0$ allows for comparisons with a known standard (Nelson and Goldsman 2001) or with the "retirement option" often used to characterize multi-armed bandit problems (Whittle 1980). We therefore address the following generalization of (4).

$$\begin{aligned} V^{\pi^*}(m,\Theta_t) &= \max\left\{m, -c + \Delta\mathrm{E}[V^{\pi}(m,\Theta_{t+1}) \,|\, \Theta_t, t \neq T],\ (1-\Delta)\mathrm{E}[X(\Theta_t)] + \Delta\mathrm{E}[V^{\pi^*}(\Theta_t)]\right\} \\ &= \max\left\{m, -c + \Delta\mathrm{E}[V^{\pi}(m,\Theta_{t+1}) \,|\, \Theta_t, t \neq T],\ \mathrm{E}[X(\Theta_t)]\right\}. \end{aligned} \tag{20}$$

Several results follow directly from the structure of (20) and the results of the previous subsections. First, since an optimal policy is reasonable, $V^{\pi^*}(m,\Theta_t) = V^{\pi^*}(\Theta_t)$ for all $m \leq -c/\delta$. We therefore focus on $m > -c/\delta$. Second, we can develop a diffusion approximation $B(m,y_0,t_0)$ to $V^{\pi^*}(m,\Theta_t)$ for the case of normally distributed outputs with a known variance. By examining policies that run ß replications, and then select a reward of $\max\{m, -c/\delta, y_{t_0+ß}/(t_0+ß)\}$, we obtain the following analog of Lemma 3:

$$B(m,y_0,t_0) \geq \sup_{ß \geq 0} e^{-\delta ß}\left(m + \left(\frac{\sigma^2 ß}{t_0(t_0+ß)}\right)^{1/2}\Psi\left[-\left(\frac{y_0}{t_0} - m\right)\left(\frac{\sigma^2 ß}{t_0(t_0+ß)}\right)^{-1/2}\right]\right). \tag{21}$$

Third, a better bound than (21) might be found by noting that the diffusion approximation for (20) is the same as in (7) in the continuation set $\mathcal{C}_m$. The boundary conditions change from (8) to

$$B(m,y,t) = D(m,y,t) \triangleq \max\{m, y/t, -c/\delta\}, \text{ and } B_y(m,y,t) = D_y(m,y,t), \text{ on } \partial\mathcal{C}_m. \tag{22}$$

The continuation set $\mathcal{C}_m$ is indexed by m as it may, in principal, differ from $\mathcal{C}$.

We define the *m-diffusion problem* to be the free boundary problem that is determined by the heat equation in (7) and the free boundary condition in (22). The lower bound in Theorem 4 is based upon the following stopping rule: do not continue sampling if one would stop if the mean were $\max\{y/t, m\}$. With this rule, the maximum number of replications that one should be willing to make is $\lceil b^{-1}(m)\rceil - t_0$ before one stops either to implement the simulated alternative or to take m.

THEOREM 4. *For a fixed m, let $B(m, y, t)$ be the solution to the m-diffusion problem given by (7) and (22). Let $B(y, t)$ be the solution to (7-8) in (19), with boundary $b(t) = \beta^{-1}b_1(1/\delta t) - c/\delta$ and continuation set $\mathcal{C} = \{(y, t) : y < b(t)\}$, where β, δ, b_1 are as above. Set $\tilde{t}(t) = tb^{-1}(m)/(b^{-1}(m) - t)$ and $\tilde{t}_0 = \tilde{t}(t_0)$.*

If $m \le -c/\delta$, then $B(m, y_0, t_0) = B(y_0, t_0)$. If $m > -c/\delta$, then $b^{-1}(m)$ is finite and

$$B(m, y_0, t_0) \ge \underline{B}(m, y_0, t_0) \stackrel{\Delta}{=} \begin{cases} \max\{y_0/t_0, m\} & \text{if } t_0 \ge b^{-1}(m) \\ m + \beta^{-1}B_1\left(\beta(\frac{y_0}{t_0} - m), 1/\delta\tilde{t}_0\right) & \text{if } t_0 < b^{-1}(m). \end{cases} \tag{23}$$

The second alternative of (23) depends upon c, as expected, because $\tilde{t}_0$ is a function of $b^{-1}(m)$, and $b(t) = \beta^{-1}b_1(1/\delta t) - c/\delta$ depends upon c. Note that $\underline{B}(m, y, t)$ is an easily computable function of $B_1(\cdot, \cdot)$.

An interesting question that we leave for future work is whether or not $B(m, y, t) = \underline{B}(m, y, t)$. This hypothesis was not rejected by our Monte Carlo tests in §5.

4.4. The Cost of Developing the Simulation Tool?

The analysis of the previous subsections assumes that the cost of the simulation tools is sunk and that the tools are immediately available for use. Now suppose that the simulation tools have not yet been developed, but that the manager has good estimates of the time and cost required to develop the simulation tools (scope, data collection, programming, validation, etc.), and an estimate of the run times of the simulations replications themselves (e.g., from prior experience with similar projects).

In particular, suppose that $u_0 \ge 0$ years and $\$g_0 \ge 0$ are required to develop the underlying simulation platform that enables the $k = 1$ alternative to be simulated. Then the NPV of having the option to simulate or implement the alternative is:

$$\bar{\mathcal{V}}(\Theta_0) = \max\left\{\mathrm{E}[X(\Theta_0)], -g_0 + e^{-\delta u_0}V^{\pi^*}(\Theta_0)\right\}. \tag{24}$$

The first maximand in (24) is the expected reward from implementing the alternative without building the associated simulation tool, and the second term combines the NPV of developing the simulation tool with the discounted value of the OEDR of the simulation selection problem in §3. The value of $V^{\pi^*}(\Theta_0)$, in turn, can be approximated by the diffusion results in the far right side of (19), which is valid for $c \geq 0$ (if there is no choice but to simulate or to implement the simulated alternative); or by Theorem 4 if there is an option to stop to implement either the simulated alternative or to select a known deterministic NPV of m.

If $\bar{\mathcal{V}}(\Theta_0) < 0$, then one would neither invest in developing the simulation tools, nor implement the alternative under consideration. If $\bar{\mathcal{V}}(\Theta_0) > 0$ and $\bar{\mathcal{V}}(\Theta_0)$ equals the second maximand of (24), then it is economically optimal to implement the simulation tool. If $\bar{\mathcal{V}}(\Theta_0) > 0$ and $\bar{\mathcal{V}}(\Theta_0)$ equals the first maximand of (24), then it is economically optimal to implement the alternative without developing a tool to simulate it.

5. Sample Simulation Selection Problems

This section applies our results to several illustrative examples with one alternative. Example 1 demonstrates that optimal stopping rule is more complex than existing stopping rules in ranking and selection. Example 2 shows that positive marginal sampling costs imply a finite amount of time that one should be willing to simulate. These two examples assume that the development cost of the simulation tool is sunk.

Two other examples illustrate the economic value of the approach. Example 3 analyzes whether or not it is optimal to invest in simulation tools in the first place. Example 4 demonstrates the economic value of having a flexible stopping time for simulation, as opposed to a rigid simulation analysis deadline.

Example 1 examines how large the simulation output mean must be before one stops to implement a system. Assume that a firm uses a discount rate of 10%/year, that the output of replications of a single simulated alternative has standard deviation $\sigma = \$10^7$ and requires $\eta = 20$ min to run at no marginal cost ($c = 0$), so that the results of §4.2.1 apply. The simulation time makes the discount rate per replication equal to $\delta = 20 \times 0.10/365/24/60$, so $1/\delta = 2.63 \times 10^5$ replications are required to get to scaled time $\tau = 1$.

Figure 1 indicates that simulation should stop after $t = 14$ replications if the sample mean is $y_t/t = \$10^7$ (corresponding to a z-score of $z = \frac{y_t/t}{\sigma/\sqrt{t}} = 3.7$). If the sample mean never crosses above the stopping boundary in Figure 1, when $c = 0$, then one would simulate forever in the absence of additional structure.

If the simulated system is implemented (with $y_t/t > 0$), then the posterior probability of incorrect selection, PICS, is the probability that the unknown mean NPV is less than the value of not implementing any system (NPV = 0). Recall that the posterior probability for the unknown mean is $\texttt{Normal}\,(y_t/t, \sigma^2/t)$, with density $p_t(\theta) = \frac{\sqrt{t}}{\sqrt{2\pi\sigma^2}} e^{-(\theta - y_t/t)^2 t/2\sigma^2}$. If $z = 3.7$ when $t = 14$, then PICS $= \int_{-\infty}^{0} p_t(\theta) d\theta = \Phi[-z] = 1 \times 10^{-4}$. If the simulated system is selected as best, but the mean turns out to be $\theta < 0$, then the opportunity cost is $0 - \theta$, and the posterior expected opportunity cost of potentially incorrect selection is EOC $= \int_{-\infty}^{0} (0 - \theta) p_t(\theta) d\theta = 69$.

One stops after 663 replications (9.2 days) if $y_t/t = \$10^6$ ($z = 2.57$; PICS $= 5.0 \times 10^{-3}$; EOC $= 615.6$), and after 1973 replications (274 days) if $y_t/t = \$10^5$ ($z = 1.4$; PICS $= 9.6 \times 10^{-2}$; EOC $= 2584$). In this example, then, a greater potential upside means that one is willing to "stop simulating and start building" sooner, but a more stringent level of evidence for correct selection is required (a higher z-score, meaning a lower PICS). We can compare this with highly effective Bayesian procedures that do not account for discounting Branke et al. (2007). Those earlier procedures specify a given fixed number of replications, or a PICS or EOC threshold that determines when stopping should occur. The optimal treatment of discounting indicates that those approaches are not optimal for E[NPV]. We also note that the optimal stopping boundary to maximize the E[NPV] of a selection differs from the shapes (e.g., triangular) of stopping regions for several frequentist IZ procedures.

Example 2 shows that the inclusion of marginal costs for sampling compels the analysis to end. Suppose that the variable cost per simulation run is \$3/hour (e.g. for computer time), and all other parameters are as in Example 1. The cost per replication $c = \$3 \times 20/60 = \$1$. We presume that the alternative to stop and "do nothing" is available, with $m = 0$, so that the results of §4.2.2 and §4.3 apply.

Figure 2 shows the original stopping boundary from Figure 1 as a line with long dashes, with the solid stopping boundary drawn $c/\delta \approx \$263$K below it, to account for the sampling costs (the y-axis is not in log-scale, to allow for negative values). The lower line means that one is willing to simulate for a shorter time (6.3 days instead of 9.2 days from Example 1 if $y_t/t = \$10^6$; 44.2 days instead of 274 days if $y_t/t = \$10^5$).

The horizontal dash-dot line corresponds to the "do nothing" option with a deterministic NPV of $m = 0$. It intersects the stopping line at 5120 replications (71.1 days), the longest amount of time that one would

Figure 1 One stops simulating to implement if the sample mean exceeds a stopping boundary ($\sigma = \$10^7$; $\delta = 5.71 \times 10^{-6}$, or 10% per year; $c = 0$).

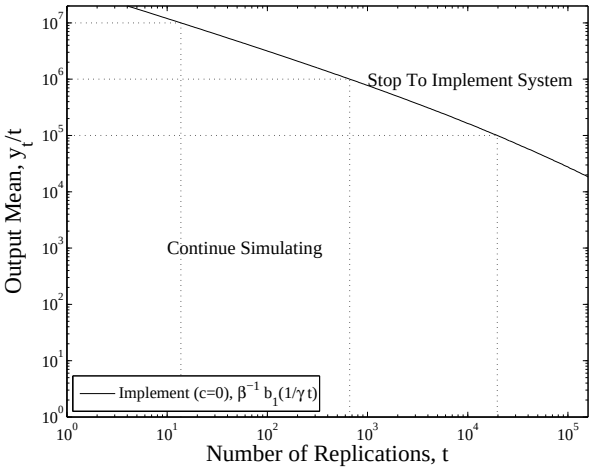
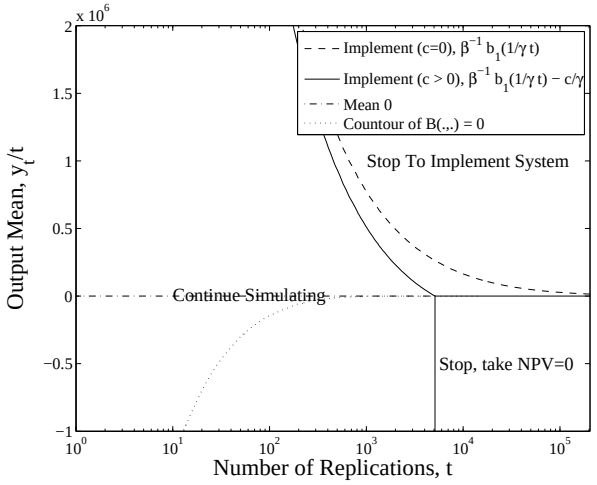


Figure 2 One stops sampling earlier in favor of implementing when the marginal cost of sampling is $1/hour ($\sigma = \$10^7$; $\delta = 5.71 \times 10^{-6}$, or 10% per year).



rationally simulate this system, if the goal were to maximize E[NPV]. Beyond that number of replications, one would take the zero option of the posterior mean $y_t/t < 0$, and one would implement if $y_t/t > 0$.

The dotted line in Figure 2 represents the contour $\mathcal{V}_{2(\kappa)}(y,t) = 0$. Below that line, the OEDR $\mathcal{V}_{2(\kappa)}(y,t)$ from (19) is negative (if the 0 option were not available, one would lose money by being forced to simulate a poor system). The OEDR $B(0,y,t)$ of (23), when the 0 option is available, is greater than 0 in that region (there is some potential for an up-side, up to the point where one would stop if the mean were 0).

A larger c means a willingness to run fewer replications. A larger σ pushes the stopping boundary proportionally higher above the base $-c/\delta$. Appendix C.2 further discusses how σ, c and δ interact to determine the continuation region in the context of stationary simulations.

Example 3 uses §4.4 to assess whether a simulation tool should be developed, assuming that it does not already exist. A manager is considering a system redesign ($k = 1$) as an alternative to continuing with an existing system (the "zero option", which brings no additional revenue beyond the status quo). A validated tool that could simulate the new alternative would require 3 months ($u_0 = 0.25$) of time and $g_0 = \$250$K to develop. The output of the tool would be the net improvement of the alternative over the mean NPV of continued operation of the current system. The marginal cost of simulation runs is assumed to be negligible ($c = 0$). The firm uses an annual discount rate of 10%. Based upon past simulation experience, a simulation

run is predicted to take $\eta = 20$ min, and experience with the existing system leads to an estimate $\sigma = \$10$Mil for the standard deviation of the simulated NPV of the alternative.

Should the manager invest time and money in developing the simulation platform? An application of (24) indicates that the answer depends upon the managers *a priori* assessment of how much better or worse the alternative might be. Suppose that the manager believes that the unknown E[NPV] has, *a priori*, a $\texttt{Normal}\left(\mu_0, \sigma_0^2\right)$ distribution. For instance, if the manager believes that the alternative has an equal chance of being better or worse, then $\mu_0 = 0$. If the amount of being better or worse is scaled like the random noise in the NPV of the existing system, then $\sigma^2 = \sigma_0^2$ and $t_0 = \sigma^2/\sigma_0^2 = 1$. A value of $t_0 = 4$ corresponds to specifying $\sigma_0 = \sigma/2 = \$5$Mil in this example.

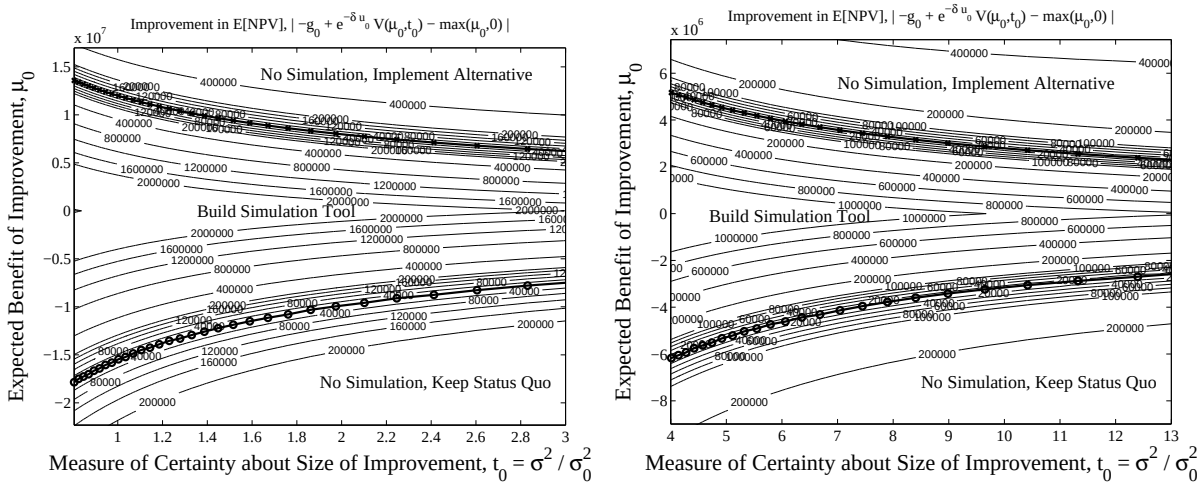
Figure 3 shows three main policy regions. The boundary of each region depends on the expected value, $-g_0 + e^{-\delta u_0} V^{\pi^*}(\Theta_0)$, of building the simulation tool followed by an application of the simulation selection procedure to learn more before deciding whether to implement. The contours represent the absolute value of the difference between $-g_0 + e^{-\delta u_0} V^{\pi^*}(\Theta_0)$ and the expected benefit of second policy, which never simulates. The policy to never simulate, and to immediately implement the alternative if and only if the alternative appears favorable, has an expected NPV of $\max(\mu_0, 0)$.

Below the lower bold line in Figure 3 with the "$\circ$" characters, which is defined by $-g_0 + e^{-\delta u_0} V^{\pi^*}(\Theta_0) = 0$, the zero option is more valuable than both maximands in the right hand side of (24). The manager should therefore not simulate and continue to operate the existing system, when below that line. The contours in that policy region show the expected loss of simulating, rather than immediately rejecting the alternative.

Above the upper bold line with the "*" characters, the first maximand of (24), which evaluates to μ_0, exceeds both the second maximand and 0. Therefore, implementing the alternative immediately is preferable to implementing the simulation tool when the expected performance is sufficiently high (if $t_0 = 1$, this happens when $\mu_0 > \$12$Mil; if $t_0 = 4$, this happens when $\mu_0 > \$5.1$Mil). The contours in that upper policy region represent the expected improvement in NPV by immediately implementing the alternative rather than investing in simulation.

In the middle band of Figure 3, where the alternative is believed to be neither a clear winner nor a serious loser, is worth the time and investment to develop the simulation tools for the analysis. Contours in that band

Figure 3 If the mean performance of the alternative is believed to be too low, one rejects the alternative; if the mean performance is believed to be high, one directly implements. In a middle range, one simulates.



represent the expected benefit of simulating rather than immediately implementing or rejecting the alternative, depending on the value of μ_0. For instance, if the manager represents uncertainty about the expected net improvement of the alternative with μ_0=\$4Mil and $t_0 = 4$ (a fair bit of uncertainty), then the improvement in E[NPV] by assessing the alternative optimally with simulation, relative to blindly implementing the alternative, is $-g_0 + e^{-\delta u_0} V^{\pi^*}(\mu_0, t_0) - \mu_0$ =\$250K. If μ_0=\$0 (may or may not be good) and $t_0 = 4$, the gain is \$1.7Mil. The more certain the manager is about the mean performance of the alternative (larger t_0), the narrower the policy region for building the simulation.

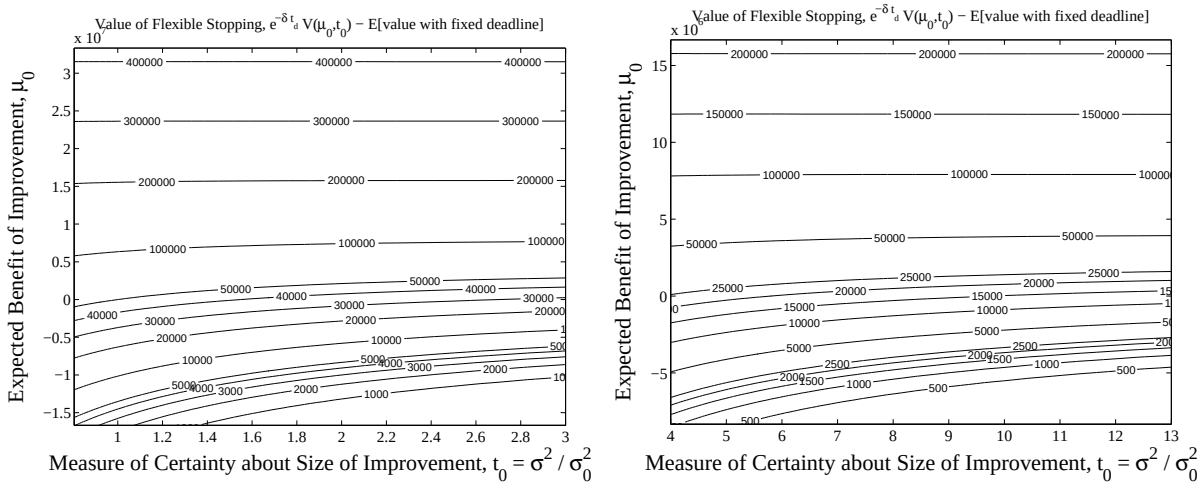
Example 4 presumes that the simulation tool from Example 3 has been fully developed. The manager faces a second question – how should the simulation analysis be performed? We consider two ways to perform that analysis. One way is to simulate nonstop until a deadline for a planned decision-making meeting occurs (after, say, a time t_d = 2 months = 0.166 years), followed by a decision to implement the alternative if the estimated E[NPV] is positive, and to reject the alternative otherwise. Since $c = 0$, the E[NPV] of that plan is

$$\text{E[NPV with fixed deadline]} = e^{-\gamma r}\left(\frac{\sigma^2 r}{t_0(t_0+r)}\right)^{1/2}\Psi\left[-\mu_0/\left(\frac{\sigma^2 r}{t_0(t_0+r)}\right)^{1/2}\right], \tag{25}$$

where $r = t_d \times 24 \times 60/\eta$ is the number of replications that can be run by the deadline (cf. Lemma 3).

Another way to analyze the alternative would be to use the simulation selection procedure that was developed above – simulate while in the continuation region, stop to implement the alternative if the stopping

Figure 4 The value of flexible stopping for simulation selection, rather than a rigid completion deadline.



boundary is reached; or reject the alternative if a deterministic fallback of value m were available – here we set $m = 0$. The alternative is assumed to be implementable two weeks ($t_p = 1/26$ years) after the moment the alternative is selected. This time delay represents coordination time after the flexible-length analysis, and incurs a discounting factor of $e^{-\delta t_p}$ times the value of the simulation selection option, $V(\mu_0, t_0)$.

Figure 4 shows the value of flexible stopping for simulation selection, rather than deciding after a rigid completion deadline. For example, if the manager's prior distribution for the unknown E[NPV] has $\mu_0 = \$0$ and $t_0 = 4$, then Figure 4 shows a value of \$25K for flexible stopping out of \$1.7Mil for the value of the simulation analysis option (as at the end of Example 3), for an expected net benefit of 1.5%. When $\mu_0 =$\$4Mil and $t_0 = 4$, that percentage increases to 55K/250K = 22% (cf. Example 3). The value of a simulation analysis with a flexible stopping time, relative to rigid deadlines, increases both with the belief that the alternative is better (larger μ_0), and with uncertainty about the mean NPV (larger $\sigma_0^2 = \sigma^2/t_0$).

6. Multiple Simulated Alternatives

Many simulation studies consider either a small, finite set of distinct systems, or a combinatorially large number of alternatives (e.g., that represent different parameter inputs into a system design structure). This section broadens the scope of our analysis to consider problems with $k > 1$ simulated alternatives.

We begin by recalling the link between simulation selection problems and the bandit literature and highlight a difficulty of establishing a so-called Gittins-index result that would greatly simplify the selection problem. We next present bounds for the optimal expected discounted reward of the simulation selection problem.

These bounds can be used to extend the analysis of Example 3 in 5, which determined whether or not simulation tools should be developed, to problems with $k > 1$ alternatives. We then extend the EOC analysis of Chick and Inoue (2001) to develop sequential sampling algorithms for the simulation selection problem and numerically show that these policies can deliver near-optimal expected rewards in a timely manner.

6.1. Simulation Selection and the Multi-Armed Bandit Problem

In the discounted multi-armed bandit problem, a decision-maker chooses repeatedly among a finite set of mutually-independent Markov chains that are indexed $i = 1, \ldots, k$. 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." Appendix A formalizes this background description.

The simulation selection problem defined in §1 is close to that of the multi-armed bandit. Both have discrete-time discounting, independent projects, and Markovian state transitions. At the same time, the simulation selection problem includes a stopping time, T, that is not part of the multi-armed bandit formalism. If, as in the simulation selection problem, a "zero" arm is included, then the bandit problem has $k+1$ actions available for all $t = 0, 1, \ldots$. In contrast, for $t \leq T$ the simulation selection problem has $2k+1$ actions available – decide $t < T$ and choose arm $i(t) \in \{1, \ldots, k\}$ to simulate, or decide $t = T$ and choose an arm $I(t) \in \{0, \ldots, k\}$ to implement – and for $t > T$ no actions are available.

The added stopping decision makes the simulation selection problem an example of what Glazebrook (1979) calls a *stoppable family of alternative bandit processes*. The fact that the simulation selection problem is such a "stoppable bandit" problem complicates the question of whether or not an index rule is optimal.

One intuitive solution to the problem with $k > 1$ systems follows a two-step hierarchical structure. First, use the results of §4 to identify an optimal stopping policy for each of the k projects. Next observe that, once

these optimal stopping policies are applied, each of k simulated systems behaves as a Markov chain. Then given these k Markov chains, obtained via the fixed application of the k optimal stopping rules, apply the Gittins-index result to sequentially select which system to simulate or implement at a given stage.

Glazebrook (1979, Theorem 3) identified a sufficient condition for such a hierarchical policy to be optimal for stoppable bandit problems. In the context of this paper, that sufficient condition requires that, when each of the k alternatives is in its stopping set when considered individually, the optimal policy for the simulation selection problem with k alternatives, considered together, would also stop. In Appendix A.3, we construct a simple counter-example to show that Glazebrook's sufficient condition does not hold. Thus, the question of whether or not there is an optimal allocation index for the simulation selection problem remains open. We therefore take a different tack.

6.2. Bounds for Deciding Whether To Develop Simulation Tools

While assessing the OEDR of the simulation selection problem is still an open question when $k > 1$, bounds on the OEDR might developed to assess whether or not to develop simulation tools in some settings. In particular, suppose that, after completing the development of a simulation platform that costs $\$g_0$ and takes u_0 time to build, all k systems could be simulated. This corresponds to the different system designs being specified by different inputs to the simulation platform and precludes problems for which simulation tools must be developed separately for each project.

The OEDR $V^{\pi^*}(\mathbf{\Theta})$ of the simulation selection problem is at least as large as the expected discounted reward of any given policy. That includes the so-called *one-stage allocation* policies. A one-stage allocation $\mathbf{r} = (r_1 \ldots, r_k)$ maps a given sampling budget of $ß \geq 0$ replications to the k systems, with a total of $r_i = r_i(ß) \geq 0$ replications to be run for alternative i, so that $\sum_{i=1}^{k} r_i = ß$. For example, the equal allocation sets $r_i = ß/k$ (relax the integer constraint if needed). After observing those samples, the one-stage allocation policy selects the alternative with the biggest (posterior) expected reward, if that reward exceeds

$$\mu_{00} \triangleq \max\{m, -c_i/\delta : i = 1, \ldots, k\}, \tag{26}$$

and otherwise selects the alternative that maximizes the right hand side of (26) (with $-c_i/\delta$ corresponding to simulating alternative i forever).

Suppose further that samples are normally distributed with known variance σ_i^2, but unknown mean whose distribution is $\texttt{Normal}\left(\mu_{0i}, \sigma_i^2/t_{0i}\right)$, with $\mu_{0i} = y_{0i}/t_{0i}$ following the notation in §4. Then the posterior mean that will be realized after the future sampling is done is a random variable (cf. (25)),

$$Ł_i \sim \texttt{Normal}\left(\mu_{0i}, \frac{\sigma_i^2 r_i}{t_{0i}(t_{0i}+r_i)}\right). \tag{27}$$

If we consider the allocation to be a function of ß and vary ß over all possible allocations, we obtain the following lower bound for the OEDR that generalizes Lemma 3 to $k > 1$ projects.

LEMMA 4. *Let* $V^{\pi^*}(\boldsymbol{\Theta})$ *maximize (2), and let* $\mathbf{r}$ *be a one-stage allocation. Then*

$$V^{\pi^*}(\boldsymbol{\Theta}) \geq \underline{OEDR}(\boldsymbol{\Theta}) \triangleq \sup_{ß \geq 0} \exp^{-\gamma ß} \mathrm{E}\left[\max\{\mu_{00}, Ł_1, Ł_2, \ldots, Ł_k\}\right] - \sum_{i=1}^{k} r_i c_i. \tag{28}$$

The expectation on the right hand size of (28), in turn, has some easy-to-compute bounds. The bound refers to the order statistics (i) for $i = 0, 1, 2, \ldots, k$ such that $\mu_{0(0)} \leq \mu_{0(1)} \leq \ldots \leq \mu_{0(k)}$.

LEMMA 5. *Let* $\mathbf{r}$ *be a one-stage allocation,* $\sigma^2_{Ł,0} = 0$*,* $\sigma^2_{Ł,i} = \frac{\sigma_i^2 r_i}{t_{0i}(t_{0i}+r_i)}$*, and* $\sigma^2_{Ł,i,(k)} = \sigma^2_{Ł,i} + \sigma^2_{Ł,(k)}$*. Then*

$$\mathrm{E}\left[\max\{\mu_{00}, Ł_1, Ł_2, \ldots, Ł_k\}\right] \geq \mu_{0(k)} + \max_{i:i\neq(k)} \sigma_{Ł,i,(k)} \Psi\left[(\mu_{0(k)} - \mu_{0i})/\sigma_{Ł,i,(k)}\right] \tag{29}$$

$$\mathrm{E}\left[\max\{\mu_{00}, Ł_1, Ł_2, \ldots, Ł_k\}\right] \leq \mu_{0(k)} + \sum_{i:i\neq(k)} \sigma_{Ł,i,(k)} \Psi\left[(\mu_{0(k)} - \mu_{0i})/\sigma_{Ł,i,(k)}\right]. \tag{30}$$

With perfect information and no discounting or sampling costs, the expected reward of $\mathbf{r}$ is

$$\overline{\mathrm{OEDR}}(\boldsymbol{\Theta}) \triangleq \mathrm{E}\left[\max\{\mu_{00}, Ł_1, Ł_2, \ldots, Ł_k\}\right]. \tag{31}$$

Observe that if $-g_0 + e^{-\gamma u_0}\underline{\mathrm{OEDR}}(\boldsymbol{\Theta}) > \mu_{0(k)}$, then it would be optimal to invest in the simulation tools that are required to simulate the k alternatives in question and to evaluate those alternatives, before selecting a project (including the 0 arm). That is because expected reward from developing the simulation tool, and using the allocation $r_i(ß)$ with the choice of ß that determined $\underline{\mathrm{OEDR}}(\boldsymbol{\Theta})$, would exceed 0.

If $-g_0 + e^{-\gamma u_0}\overline{\mathrm{OEDR}}(\boldsymbol{\Theta}) < \mu_{0(k)}$, however, it would be better to not implement the simulation tools, and to implement project (k) (which may include the 0 arm). In that case, even a simulator that incurs no costs and no discounting penalty could not deliver the required E[NPV] to justify the development of the tools.

6.3. Fully Sequential Algorithm with $k > 1$

Section 6.2 provides bounds that can help a manager to decide whether or not to build simulation tools. The bounds are based upon one-stage policies. The actual expected discounted reward, given that the simulation platform has been developed, is likely to improve with sequential algorithms. Section 6.1 indicates that a sequential Gittins-index policy may not be optimal. We therefore turn to other policies that are likely to be effective: those based upon maximizing the expected (undiscounted) reward over a finite horizon.

Gupta and Miescke (1996) showed that minimizing the expected opportunity cost is equivalent to maximizing the posterior mean that is realized once a finite total number of samples (with a known variance) is observed. Chick and Inoue (2001) presented a one-stage sampling allocation that asymptotically maximizes an upper bound on the expected opportunity cost (EOC) of incorrect selection when the samples are normally distributed with different and potentially unknown variances. Branke et al. (2007) showed how a sequential version of that one-stage EOC algorithm can be adapted into a fully sequential algorithm, Procedure $\mathcal{LL}$, which is highly efficient for a variety of selection problems. Thus, procedure $\mathcal{LL}$ seeks to maximize the expected undiscounted reward over a finite horizon.

Appendix E in the Online Companion to this paper shows how Procedure $\mathcal{LL}$ can be adapted and extended to the current context, where both discounting and sampling costs are included. The procedure assumes that samples are normally distributed with a known variance that may differ for each alternative.

The general idea of our sequential sampling procedure is simple. At each stage of sampling, the procedure first tests whether or not to continue sampling. It does this by using the $\mathcal{LL}$ allocation to test if there exists some ß ≥ 1 such that allocating ß replications leads to an expected discounted reward that exceeds the value of stopping immediately. If there is value to continuing, then one replication is run for the alternative that $\mathcal{LL}$ suggests would most warrant an additional replication. After that replication is run, the statistics for that system are updated, with the posterior distribution from the current stage becoming the prior distribution for the next stage. If there is no value to continuing for any ß ≥ 1, then the procedure stops.

The development of §6.2 immediately suggests a mechanism to assess whether there is value to additional sampling. One should continue to sample if $\underline{\text{OEDR}}(\boldsymbol{\Theta}) > \mu_{0(k)}$. This will happen if there is a one-stage allocation of size ß that leads to value for continuing to simulate. Unfortunately, the sequential recalculation

of $\underline{\text{OEDR}}(\boldsymbol{\Theta})$ that would be required by such a procedure is computationally burdensome. Fortunately, there is an easy to compute substitute. Substituting the right hand side of (29) for the expectation in the right hand side of (28) leads to an easily computable and analytically justifiable bound.

Stopping rule EOC$_1^\gamma$ (with implicit one-stage allocation $r_i = r_i(ß) \geq 0$ such that $\sum_{i=1}^k r_i = ß$): Continue sampling if and only if there is a budget $ß \geq 1$ such that

$$\exp^{-\gamma ß}\left(\mu_{0(k)} + \max_{i:i\neq(k)}\left\{\sigma_{Ł,i,(k)}\Psi\left[(\mu_{0(k)} - \mu_{0i})/\sigma_{Ł,i,(k)}\right]\right\}\right) - \sum_{i=1}^k r_i c_i > \mu_{0(k)}. \tag{32}$$

In numerical experiments, EOC$_1^\gamma$ may not be as effective as hoped. The problem is that the expected discounted reward function drops off more slowly if the budget is somewhat larger than optimal, as compared to the greater penalty for sampling somewhat less than is optimal, which may happen with EOC$_1^\gamma$. The following stopping rule, which may be somewhat less justifiable analytically, increases sampling slightly by plugging the right hand side of the upper bound in (30) into the expectation of (28).

Stopping rule EOC$_k^\gamma$ Continue sampling if and only if there is a budget $ß \geq 1$ such that

$$\exp^{-\gamma ß}\left(\mu_{0(k)} + \sum_{i:i\neq(k)}\sigma_{Ł,i,(k)}\Psi\left[(\mu_{0(k)} - \mu_{0i})/\sigma_{Ł,i,(k)}\right]\right) - \sum_{i=1}^k r_i c_i > \mu_{0(k)}. \tag{33}$$

Appendix E fully specifies how these stopping rules are used with the $\mathcal{LL}$ allocation to solve the simulation selection problem. Depending upon the stopping rule, we refer to Procedure $\mathcal{LL}(\text{EOC}_1^\gamma)$ or $\mathcal{LL}(\text{EOC}_k^\gamma)$. We note that the left hand sides of (32) and (33) are not monotonic in ß, so procedures that use these inequalities must test them for $ß \geq 1$, and not for $ß = 1$ alone.

6.4. Numerical Results

We now extend the numerical examples of §5 by allowing for $k > 1$ alternatives.

Example 5. We extend Examples 1 and 3 by assuming there are $k \geq 1$ projects, each with the same prior distribution for the unknown mean, independent $\texttt{Normal}\,(\mu_0, \sigma_i^2/t_0)$ for all i. We assume that the simulation output for each project is normally distributed with known variance $\sigma_i = 10^6$, a cpu time of $\eta = 20$ min/replication, an annual discount rate of 10%, and no marginal cost for simulations $c_i = 0$.

The top rows of Table 1 give the values of $\underline{\text{OEDR}}(\boldsymbol{\Theta})$ and $\overline{\text{OEDR}}(\boldsymbol{\Theta})$ as functions of the number of alternatives, when $\mu_0 = 0$ and $t_0 = 4$. These values of $\underline{\text{OEDR}}(\boldsymbol{\Theta})$ and $\overline{\text{OEDR}}(\boldsymbol{\Theta})$ can be compared with the

time and cost of developing a simulation platform, to decide if a platform warrants building or not, as in §6.2. The data show that the bounds are relatively close for this range of k.

Example 6. Suppose now that the simulation platform has been built, but that the problem is otherwise the same as in Example 5. Once the tool has been developed, there is no longer a constraint to use a one-stage algorithm with an equal allocation.

Table 1 also shows the expected NPV of using Procedure $\mathcal{LL}(\text{EOC}_1^\gamma)$ or Procedure $\mathcal{LL}(\text{EOC}_k^\gamma)$ to identify the best alternative. Those estimates are based on 6000 replications of each procedure to independently sampled problem instances, where a problem instance is a configuration of the unknown means that is sampled from the prior distribution for each unknown mean (except for $k = 1$, which is based upon 10^5 replications, and where the simulation results match the PDE solution with E[NPV]=$B(\mu_0, t_0) = 1.99 \times 10^6$). For these results, each procedure modified slightly to stop after a maximum of 75 days of observed replications, or if the stopping rule is satisfied, whichever comes first.

The top portion of Table 1 shows that $\mathcal{LL}(\text{EOC}_k^\gamma)$ and $\mathcal{LL}(\text{EOC}_1^\gamma)$ provide expected NPVs that are within the range from $\underline{\text{OEDR}}(\boldsymbol{\Theta})$ to $\overline{\text{OEDR}}(\boldsymbol{\Theta})$ (within the limits of stochastic noise in their estimates). There is a slight advantage for $\mathcal{LL}(\text{EOC}_k^\gamma)$ over $\mathcal{LL}(\text{EOC}_1^\gamma)$, as expected.

The middle portion of Table 1 shows that, on average, the sequential $\mathcal{LL}$ procedures, with either stopping rule, require much less time than is required by the optimal one-stage procedure that maximizes $\underline{\text{OEDR}}(\boldsymbol{\Theta})$. The procedure $\mathcal{LL}(\text{EOC}_k^\gamma)$ tends to sample more than $\mathcal{LL}(\text{EOC}_1^\gamma)$, as expected by the construction of the stopping rules. There is no corresponding time duration for $\overline{\text{OEDR}}(\boldsymbol{\Theta})$, since that figure assumes perfect information instantaneously at no cost.

The bottom portion of Table 1 shows the frequentist probability of correct selection for these procedures, estimated by the fraction of times the 'true' best alternative was selected by the procedure. With respect to this criterion, $\mathcal{LL}(\text{EOC}_k^\gamma)$ again beats $\mathcal{LL}(\text{EOC}_1^\gamma)$, which in turn beats the optimal one-stage allocation.

For the range of k tested, more systems means more opportunity to obtain a good system, which means better expected performance. We did not study combinatorially large k here.

E[NPV]$\times 10^6$	$k=3$	4	5	6	7	8	9	10
$\overline{\text{OEDR}}(\boldsymbol{\Theta})$	4.44	5.23	5.85	6.35	6.77	7.12	7.42	7.69
$\underline{\text{OEDR}}(\boldsymbol{\Theta})$	4.42	5.20	5.81	6.31	6.72	7.06	7.36	7.62
$\mathcal{LL}(\text{EOC}_k^\gamma)$	4.43	5.20	5.87	6.39	6.78	7.08	7.41	7.66
$\mathcal{LL}(\text{EOC}_1^\gamma)$	4.50	5.18	5.78	6.30	6.75	7.09	7.36	7.60
E[Days]								
$\underline{\text{OEDR}}(\boldsymbol{\Theta})$	17.4	20.0	22.4	24.5	26.4	28.3	30.0	31.6
$\mathcal{LL}(\text{EOC}_k^\gamma)$	10.1	8.3	6.6	6.2	6.4	6.3	6.2	6.1
$\mathcal{LL}(\text{EOC}_1^\gamma)$	10.2	8.2	6.4	6.1	5.9	5.2	5.4	5.4
PCS_{iz}								
$\underline{\text{OEDR}}(\boldsymbol{\Theta})$	0.945	0.935	0.925	0.917	0.909	0.902	0.895	0.889
$\mathcal{LL}(\text{EOC}_k^\gamma)$	0.967	0.955	0.945	0.938	0.930	0.921	0.916	0.914
$\mathcal{LL}(\text{EOC}_1^\gamma)$	0.965	0.950	0.943	0.934	0.923	0.905	0.904	0.889

Table 1 The expected discounted reward and average time until selecting a project as a function of the number of independent projects, k, allocation policy and stopping criterion.

7. Discussion and Conclusions

This paper responds to the question of how to link financial measures (a firm's discount rate, the marginal cost of simulations) to the optimal control of simulation experiments that are designed to inform operational decisions. It provides a theoretical foundation, analytical results, and numerical solutions to answer following questions: Should a manager invest time and money to develop simulation tools? For how long should competing systems be simulated before an alternative is selected, or all alternatives are rejected?

This work therefore provides a first link between a managerial perspective on simulation for project selection and the statistical simulation optimization literature. The approach was that of treating the ability to develop simulation tools, and the ability to simulate to gather more information about alternatives, as a real option to gather information before committing resources to a design alternative.

The diffusion model analysis for a simulation option with $k = 1$ alternative assumes normally distributed output with a known variance. The Online Companion indicates how the results can be extended to handle output from one-parameter members of the exponential family of distributions and autocorrelated output that allows for batch mean analysis. For $k \geq 1$ alternatives, we extended earlier ranking and selection work, that minimizes the expected opportunity cost of potentially incorrect selections, to adapt to the current context, that of maximizing the expected discounted NPV of a decision made with simulation. The Online Companion also allows for different runtime durations for each system and parallel simulation.

The paper raises several issues for future study. From a business perspective, we do not address the issue

of first-mover advantage or penalties for late implementation of projects due to project delays. From a simulation perspective, we did not account for common random numbers (CRN) across systems, a technique that can help sharpen contrasts between systems. Section 3 accounts for unknown variances, but not §4. We reserve CRN and unknown variances for future work.

Much current research focuses on probability of correct selection (PCS) guarantees, or asymptotic convergence results in simulation optimization. While these are useful properties, this paper suggests that an alternative approach may also be useful: maximizing the expected discounted NPV of decisions based on simulation analysis, even at the expense of potentially incorrect selections. Even with the limitations enumerated above, this new approach to simulation selection accounts for a much fuller accounting of the financial flows that are important to managers.

Acknowledgments

The research of Noah Gans was supported by the Fishman-Davidson Center for Service and Operations Management and the Wharton-INSEAD Alliance. We thank Shane Henderson for feedback on an early draft of this paper. We thank Ricki Ingalls for discussions about NPV in supply chain simulations.

References

Bertsekas, D. P., S. E. Shreve. 1996. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific.

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

Blackwell, D. 1965. Discounted dynamic programming. *Annals of Mathematical Statistics* **36** 226–235.

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.

Brealey, R. A., S. C. Myers. 2001. *Principles of Corporate Finance*. 6th ed. McGraw-Hill.

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

Butler, J., D. J. Morrice, P. W. Mullarkey. 2001. A multiple attribute utility theory approach to ranking and selection. *Management Science* **47**(6) 800–816.

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. 1965. Sequential tests for the mean of a normal distribution III (small t). *Ann. Math. Stats.* **36** 28–54.

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

Fox, B. L., P. W. Glynn. 1989. Simulating discounted costs. *Management Science* **35**(11) 1297–1315.

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.

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

Gold, M.R., J.E. Siegel, L.B. Russell, M. Weinstein. 1996. *Cost Effectiveness in Health and Medicine*. Oxford.

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.

Hong, L. J., B. L. Nelson. 2005. The tradeoff between sampling and switching: New sequential proceduress for indifference-zone selection. *IIE Transactions* **37** 623–634.

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

Law, A. M., W. D. Kelton. 2000. *Simulation Modeling & Analysis*. 3rd ed. McGraw-Hill, Inc, New York.

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

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

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, the relationship of the simulation selection problem to a stoppable version of the multi-armed bandit, and a numerical example that shows that the few existing results that characterize optimal policies for stoppable bandits does not apply to the simulation selection problem.

Appendix B provides mathematical proofs of the claims in the main paper.

Appendix C describes several technical extensions to the range of validity of the paper. It relaxes some assumptions about the distribution of the output, the duration of the replications for each alternative, and the sequential/parallel nature of sampling.

Appendix D summarizes how the optimal expected discounted reward (OEDR) and stopping boundaries for the simulation selection problem (with $k = 1$ alternative) were computed.

Appendix E 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.

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 $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_k)$ and one-period discount rate $0 \leq \Delta < 1$, 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 "$\boldsymbol{\Theta}_\mathbf{0} = \boldsymbol{\Theta}$" 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: $V^{\xi^*}(\boldsymbol{\Theta}) = \sup_{\xi \in \Xi}\left(V^{\xi}(\boldsymbol{\Theta})\right)$.

For the case in which expected one-period rewards are bounded, so that $R_i(\Theta_i) < \infty$ for almost all $\Theta_i \in \Omega_{\Theta_i}$, $i = 1, \ldots, k$, 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).

A.2. "Stoppable Bandit" Problems

The multi-armed bandit problem described in Appendix A.1 can be linked to the simulation selection problem, if the stopping policies of the later are fixed.

Observe that the application of any reasonable (or more generally, stationary) stopping policy, π_i, to project i induces it to behave as a simple Markov chain (with Bayes rule for state transitions and a reward of $-c_i$ when simulation data is observed, and no state change and a reward associated with the selected system if not). 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 is obtained (Glazebrook 1979, Corollary 1). That is, an allocation index exists, such that at each stage is it optimal to select the project with the largest index, and then to either simulate it or to implement it, with implementation if the stopping policy for the project in question indicates that stopping is optimal. That result is true for "stoppable bandits" in general.

Glazebrook (1979, Corollary 1) does not necessarily imply that a Gittins index policy is optimal for a stoppable bandit problem such as the simulation selection problem however. Because of the stopping problem embedded in the choice of project $i > 0$, the calculation of the Gittins index now involves two stopping times: given stationary stopping policy π_i, there exists a simulation-stopping time, T_i, whose distribution is determined *a priori*, via π_i; and given T_i there exists, in turn, an optimal Gittins-index stopping time, which we will call s_i. We let $G_i(\Theta_i, s \,|\, T_i)$ denote the analogue of (EC.4), so that $G_i(\Theta_i \,|\, T_i) = \sup_{s>0} G_i(\Theta_i, s \,|\, T_i)$ denotes the analogue of (EC.5), where s_i is the stopping time for which $G_i(\Theta_i \,|\, T_i)$ is achieved. This complication makes the identification of an optimal policy, whose existence is guaranteed by Lemma 1, more difficult.

A natural class of policies to consider for potential optimality is that of *hierarchical policies*. In a hierarchical policy, $\xi(\pi_1, \ldots, \pi_k)$, 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 $i = 1, \ldots, k$, the optimal policy for the resulting $k+1$ armed bandit, as in (EC.4) and (EC.5), uses the Gittins-index rule. We denote that optimal policy, for the given π_i, by $\xi^*(\pi_1, \ldots, \pi_k)$.

A special example of a hierarchical policy, $\xi^*(\pi_1^*, \ldots, \pi_k^*)$, 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).

Glazebrook (1979) provides sufficient conditions under which $\xi^*(\pi_1^*, \ldots, \pi_k^*)$ is optimal for stoppable bandit problems. We restate these results for the special case of the simulation selection problem.

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 $G_i(\Theta_{i,t} \,|\, T_i)$ 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, $G_i(\Theta_{i,t} \,|\, T_i^*) \geq G_i(\Theta_{i,t} \,|\, T_i)$ for all stationary π_i whenever $\Theta_{i,t}$ is such that $t \geq T_i^*$, then $\xi^*(\pi_1^*, \ldots, \pi_k^*)$ 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 $\xi^*(\pi_1^*, \ldots, \pi_k^*)$ 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 $k > 1$

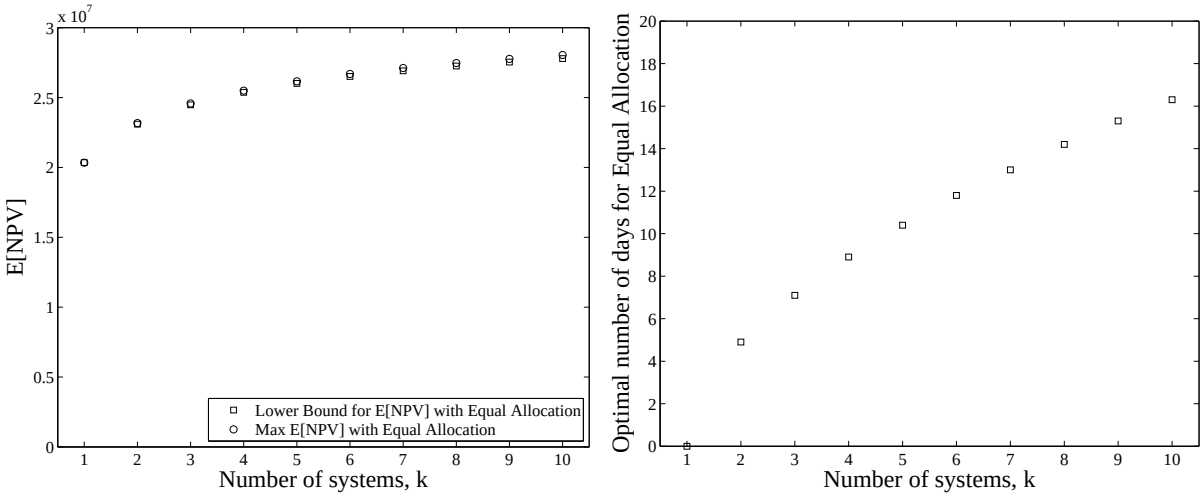
Appendix A.2 specifies sufficient conditions of Glazebrook (1979, Theorem 3) that would guarantee that the hierarchical policy, $\xi^*(\pi_1^*, \ldots, \pi_k^*)$, would be optimal for the simulation selection problem when $k > 1$. The policy $\xi^*(\pi_1^*, \ldots, \pi_k^*)$ has a natural appeal, since 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.

This section provides a simple simulation selection policy, for a specific simulation selection problem instance, that can outperform the hierarchical policy $\xi^*(\pi_1^*, \ldots, \pi_k^*)$. This implies that the optimal policy for the simulation selection problem, whose existence can be guaranteed by Lemma 1, is not $\xi^*(\pi_1^*, \ldots, \pi_k^*)$. 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, $b(t_{0i})$. The theoretical answer is no when $k = 1$. And for $k > 1$, one would expect the answer to be no *if* the sufficient conditions of Glazebrook (1979, Theorem 3) were true.

Figure EC.1 plots figures that are analogous to the first two rows of Table 1 except that $\mu_{0i} = 1.015 b(t_{0i})$, where $b(\cdot)$ is estimated as in Appendix D. We again chose $t_{0i} = 4$. The factor of 1.015 was chosen to avoid accidentally being inside the continuation set, as a result of numerical error. With that value of μ_{0i}, the lower bound from Lemma 3 was less than μ_{0i}. Further, the maximum estimated discounted reward when $k = 1$ is also achieved when no sampling is done (right panel) and the reward is $\mu_{0i} = 2.035 \times 10^7$, as expected.

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.



The most important feature of Figure EC.1 is the large benefit of additional sampling for several values of k, with $k > 1$. The figure indicates that, for the simulation selection problem, it is not optimal to use the optimal stopping set if Theorem 3 (which determines if a given project should be simulated or implemented) for each project individually, when considering simulation selection with $k > 1$. Stopping to implement the system should occur later and with a higher boundary when $k > 1$. The implication is that the sufficient conditions of Glazebrook (1979) – which would guarantee the optimality of the hierarchical selection procedure $\xi^*(\pi_1^*, \ldots, \pi_k^*)$ – do not seem to hold, since the optimal choice of stopping policy for each project appears to depend upon the state of uncertainty of the other projects.

Appendix B: 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 $j_r(t) \in \{1, \ldots, k\}$ denote decisions to simulate project $i = j_r(t)$ for one period and pay c_i. Actions $j_r(t) \in \{k+1, \ldots, 2k\}$ 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

$$R_t^{\pi_r} = \begin{cases} 0, & \text{if } j_r(t) = 0 \\ -c_j, & \text{if } j_r(t) \in \{1, \ldots, k\} \\ (1-\Delta)\,\mathrm{E}[X(\Theta_{j_r(t)-k,t})], & \text{if } j_r(t) \in \{k+1, \ldots, 2k\}. \end{cases} \tag{EC.6}$$

We then modify the definition of state transitions. For action $j_r(t) \in \{1,\ldots,k\}$ 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 $j_r(t) \in \{k+1,\ldots,2k\}$, we define the state of all k projects as unchanging; that is, $\Theta_{i,t+1} = \Theta_{i,t}$ for all projects $i = 1,\ldots,k$.

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

$$\begin{aligned} V^{\pi}(\boldsymbol{\Theta}) &= \mathrm{E}_{\pi}\left[\sum_{t=0}^{T-1} -\Delta^t c_{i(t)} \;+\; \Delta^T \mathrm{E}[X(\Theta_{I(T),T})] \,|\, \boldsymbol{\Theta}_0 = \boldsymbol{\Theta}\right] \\ &= \mathrm{E}_{\pi_r}\left[\sum_{t=0}^{\infty} \Delta^t \, R_t^{\pi_r} \,|\, \boldsymbol{\Theta}_0 = \boldsymbol{\Theta}\right] = V_r^{\pi_r}(\boldsymbol{\Theta}). \end{aligned} \tag{EC.7}$$

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 Θ_i, $i \in \{1,\ldots,k\}$. 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 $t_r = \inf\{t \,|\, j_r(t) \notin \{1,\ldots,k\}\}$ 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.7) 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.

To summarize, we can view the simulation selection problem as a stationary (infinite horizon) problem with $2k+1$ actions: simulate or implement each of k "stoppable" arms, or implement the zero arm. This differs from the original multi-armed bandit problem's $k+1$ possible actions available at each time $t \geq 0$, but matches Glazebrook's stoppable family of alternative bandit processes. In addition, given the use of such a stationary, deterministic policy the original meaning of the stopping time, T, continues to hold – it is the time at which the manager stops simulating and implements a project – and we will continue to refer to T in the context of the infinite-horizon problem as well.

To ease notation, we abandon the use of the "r" subscript in the paper. □

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 derivation of the free boundary problem in (13) from the standardized problem in Problem (11) follows by construction and the parametrization of (10), using standard derivations of Brownian motion and stopping times for Brownian motion (Billingsley 1986, Section 37).

We note two points that prove 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 $w = \beta y/t$ and $\beta^{-1} = \sigma\sqrt{\delta} = \delta/\alpha$. The fact that $\mathcal{V}_1 = \beta^{-1} B_1(w_0, s_0)$ is the correct scaling follows from recalling that the original problem was multiplied by α/δ to

obtain the standardized problem. The fact that $\mathcal{V}_1 \geq \max\{0, y_{t_0}/t_0\}$ follows because, at worst, one can simulate forever to get 0, or can stop immediately to get y_{t_0}/t_0, in expectation. □

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 (13) 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 $M = \mathcal{V}(y_{t_0}, t_0) = \sup_{T \geq t_0} \mathrm{E}[R(Y_T, T) \mid y_{t_0}, t_0]$ be the OEDR for the original problem in (y, t) coordinates. The simulation selection problem technically lets Y be observed at discrete times. Here, we abuse notation and consider the stopping time T to be in continuous time (for a Wiener process, 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 $y_0 = y_{t_0}$. Then

$$\begin{aligned} M &= \sup_{T \geq t_0} \mathrm{E}\left[D(Y_T, T) e^{-\delta(T - t_0)} \right] \\ &= \sup_{T \geq t_0} \mathrm{E}\left[\int_T^\infty D(Y_T, T) \delta e^{-\delta(\xi - t_0)} d\xi \right], \text{ so} \\ 0 &= \sup_{T \geq t_0} \mathrm{E}\left[-\int_{t_0}^T \delta M e^{-\delta(\xi - t_0)} d\xi + \int_T^\infty \delta (D(Y_T, T) - M) e^{-\delta(\xi - t_0)} d\xi \right]. \end{aligned} \tag{EC.8}$$

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. §4.1). Recall that $\gamma = \delta$. (EC.8) implies that

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

$$= \sup_{0\le S\le s_0} \mathrm{E}\left[\frac{W(S)}{\beta} e^{-(\frac{1}{S}-\frac{1}{s_0})} - M\right]. \tag{EC.9}$$

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

$$M = \sigma\sqrt{\delta} \sup_{0\le S\le s_0} \mathrm{E}\left[W(S)e^{-(\frac{1}{S}-\frac{1}{s_0})} \mid w(s_0)=w_0\right]. \tag{EC.10}$$

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 with $\sigma=1$, is

$$w_0' = \sup_{0\le S'\le s_0'} \mathrm{E}\left[W'(S')e^{-(\frac{1}{S'}-\frac{1}{s_0'})} \mid w'(s_0')=w_0'\right], \tag{EC.11}$$

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

$$M'-u_0 = \sigma\sqrt{\delta} \sup_{0\le S'\le s_0'} \mathrm{E}\left[W'(S')e^{-(\frac{1}{S'}-\frac{1}{s_0'})} \mid w'(s_0')=(M'-u_0)/\sigma\sqrt{\delta}\right], \tag{EC.12}$$

for general σ (cf. 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 $b_{BL}(s')\ge 0$.

The random processes in the expectations in (EC.10) and (EC.12) 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. Recall that $c=0$. Define $T_ß$ to be the one-stage stopping rule that says to continue sampling for exactly $ß\ge 0$ replications, then implement if $y_{t_0+ß}/(t_0+ß) \ge -c/\delta = 0$ (for an expected reward of $y_{t_0+ß}/(t_0+ß)$); and never stop otherwise (e.g. simulate forever if $y_{t_0+ß}/(t_0+ß) < -c/\delta = 0$, for reward $-c/\delta = 0$).

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

$$E_{T_ß}[e^{-\delta ß}\max\{0, Y_{t_0+ß}/(t_0+ß)\}|y_{t_0}, t_0] = e^{-\delta ß}\left(\frac{\sigma^2 ß}{t_0(t_0+ß)}\right)^{1/2} \Psi\left[-\frac{y_{t_0}}{y_0}\left(\frac{\sigma^2 ß}{t_0(t_0+ß)}\right)^{-1/2}\right].$$

Inequality (14) is justified because B is defined as a supremum over all stopping rules, including $T_ß$ for all $ß\ge 0$. □

Proof of Theorem 3. Define $\mathcal{C}_1$ to be the stopping boundary for the problem when $c=0$, and let $\mathcal{C}_{2(\kappa)}$ be the stopping boundary when $c>0$, with all other parameters the same.

Set $\mathcal{V}(y_{t_0},t_0)=\sup_{T\geq t_0}\mathrm{E}[R(Y_T,T)\mid y_{t_0},t_0]$. Recall $1/\beta\kappa=c/\delta$ and (18).

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

The OEDR $\mathcal{V}_{2(\kappa)}$ for the simulation selection problem with $c,\delta>0$ is therefore

$$\mathcal{V}_{2(\kappa)}=\frac{c\kappa}{\delta}B_1(w_0+1/\kappa,s_0)-\frac{c}{\delta}=\sigma\sqrt{\delta}B_1\left(\frac{1}{\sigma\sqrt{\delta}}(\frac{y_0}{t_0}+\frac{c}{\delta}),\frac{1}{\delta t_0}\right)-\frac{c}{\delta}.$$

Suppose that (y_t,t) is on $\partial\mathcal{C}_1$, the boundary of the continuation set when $c=0$. For that fixed t, y_t is the smallest y such that $\beta^{-1}B_1(\beta y/t,1/\gamma t)=y/t$ (Theorem 1). Define $\hat{y}$ so that $\hat{y}/t=y/t-c/\delta$, so that

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

has the form $\mathcal{V}_{2(\kappa)}=\hat{y}/t$. So $(y,t)\in\mathcal{C}_1$ if and only if $(\hat{y},t)\in\mathcal{C}_{2(\kappa)}$. The continuation set is therefore shifted down by c/δ as claimed. □

Proof of Theorem 4. The proof, for $m\leq -c/\delta$, comes from noting that m never need be chosen by an optimal policy, since one can always do at least as well as 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 (22) simplifies to $D(m,y,t)=\max\{m,y/t\}$. We define $b^{-1}(m)=\sup\{t:b(t)\geq m\}$ for all $m>-c/\delta$. 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 D 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 $-\infty$). So while we define $b^{-1}(m)=\sup\{t:b(t)\geq m\}$, we hypothesize that $b(\cdot)$ is in fact invertible for $m>-c/\delta$.

We first examine the first alternative in (23), which assumes that $t_0 \geq b^{-1}(m)$ and $m > -c/\delta$. This means that the retirement reward exceeds the stopping boundary of the original problem without the retirement reward, $m \geq b(t_0)$. One can therefore achieve an NPV of $\max\{y_0/t_0, m\}$ by stopping immediately and selecting the better of those two options. This justifies the lower bound in the first alternative of (23). 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 $y_0/t_0 = m$ 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 $y_{b^{-1}(m)}/b^{-1}(m) \geq m$. 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 (23)).

Figure EC.2 gives a conceptual schematic 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 $\max\{m, y_{b^{-1}(m)}/b^{-1}(m)\}$.

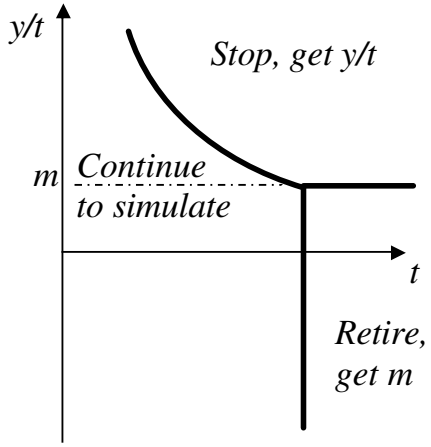
In particular, the m-diffusion satisfies the same diffusion equation as before (as in (7)), but now has a terminal condition, at time $t = b^{-1}(m)$, with expected NPV $\max\{m, y_{b^{-1}(m)}/b^{-1}(m)\}$. This differs from the 'terminal condition' from the original case, which informally is to pick 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 $w_0 = \beta y_0/t_0$. 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 a Browning 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 that 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 $1/\delta b^{-1}(m)$.

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).



In doing so, we note that m will take the role of $-c/\delta$ in (8), and the role of t in that equation will be replaced by the value of t that corresponds to $\tilde{s}_0$. To find that 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 for the form of the original process with $c \neq 0$, with βm taking the role of $-\kappa$, and 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 (7-8) and the solution in (19), 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 (23). The value of $\underline{B}(m, y_0, t_0)$ 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 Ł$_i$ to be observed, given that r_i replications will be observed, is as in (27) follows directly from (de Groot 1970, Sec. 11.9). The expectation $\mathrm{E}[\max\{\mu_{00}, Ł_1, Ł_2, \ldots, Ł_k\}]$ 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 ß $= \sum_{i=1}^{k} r_i$ samples. (Note that if $\mu_{00} = \max\{-c_i/\delta\} < 0$, this choice corresponds to simulating alternative $\arg\max\{-c_i/\delta\}$ forever). The factor of $\exp^{-\delta ß}$ discounts that reward appropriately.

That specific one-stage sampling policy has a value that is not greater than the policy that is optimal over all non-anticipative sampling policies. Also, the discounted cost of sampling is not more expensive than the undiscounted cost of sampling, $\sum_{i=1}^{k} r_i c_i$. Because of these two facts, the right hand side of (28) is a lower bound for $V^{\pi^*}(\mathbf{\Theta})$. □

Proof of Lemma 5. The fact that the expectation $\mathrm{E}\left[\max\{\mu_{00}, \text{Ł}_1, \text{Ł}_2, \ldots, \text{Ł}_k\}\right]$ 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 (29) and the upper bound in (30). 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 (not proven) in (Chick and Inoue 1998). □

Appendix C: Extensions

Section 4 assumed jointly independent Gaussian output with known variances, simulations runs for each alternative that are of the same duration, and sequential simulation sampling, as might be experienced with a single CPU. This section shows that some of the results appear to hold more generally. It points to references that provide sufficient conditions for the results to hold. A full analysis is beyond the scope of this paper.

C.1. One-Parameter Members of the Exponential Family of Distributions

Chang and Lai (1987) show that their Gittins-index results for the Bayesian bandit asymptotically apply to independent samples from one-parameter members of the exponential family of distributions, with pdf $f(x \mid \theta) = \exp\{\theta x - \varphi(\theta)\}$. They require several technical conditions, including a conjugate prior distribution, an information number $\varphi''(\theta)$ that is bounded away from 0 and ∞, and φ'' uniformly continuous on $(a_1 - r, a_2 + r)$ for some $r > 0$ and some $a_1 < a_2$. Although the reward in their problem differs from ours, the asymptotic convergence issues appear to be the same.

Let $\mathcal{E}_t$ represent the state of information at time t. Denote the posterior mean by $\mu_{\theta,t} = \mathrm{E}[\theta \mid \mathcal{E}_t]$, the posterior variance by $\varsigma_t^2 = \mathrm{Var}[\theta \mid \mathcal{E}_t]$, and the (conditional) variance of a sample by $\sigma^2_{\mu_{\theta,t}} = \mathrm{Var}[X \mid \mu_{\theta,t}]$. Under mild regularity conditions, the posterior distribution $p_t(\theta) = p(\theta \mid \mathcal{E}_t)$ of θ at time t is asymptotically $\texttt{Normal}\left(\mu_{\theta,t}, \varsigma_t^2\right)$ as $t \to \infty$. If the results of Chang and Lai (1987) apply here, the OEDR of the simulation selection problem can be asymptotically approximated using $\mu_{\theta,t}, \varsigma_t^2, \sigma^2_{\mu_{\theta,t}}$ for large t and small δ.

For Bernoulli sampling with probability θ and a Beta(α, β) prior distribution for θ, for some $\alpha, \beta > 1$, $\mu_{\theta,t} = \alpha/(\alpha+\beta)$; $\varsigma_t^2 = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$; and $\sigma^2_{\mu_{\theta,t}} = \mu_{\theta,t}(1-\mu_{\theta,t})$. With that setup, $t = \sigma^2_{\mu_{\theta,t}}/\varsigma_t^2 = \alpha+\beta+1$ is the effective number of observations, and $\sqrt{\delta}\sigma_{\mu_{\theta,t}} B_1(\mu_{\theta,t}/\sqrt{\delta}\sigma_{\mu_{\theta,t}}, 1/\delta t)$ is an asymptotically appropriate OEDR when $c = 0, \delta > 0$, with stopping boundary $\sigma_{\mu_{\theta,t}}\sqrt{\delta} b_1(1/\delta t)$.

C.2. 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 it 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 treating them as perpetuities.

C.3. 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 C.2 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.

C.4. Suboptimal Solution for Parallel Simulation

The framework employed here for the simulation selection problem requires sequential sampling, which implicitly prohibits parallel simulation. When there are $k > 1$ alternatives, the use of the $\mathcal{LL}$ algorithm in Appendix E can be used to allocate more than one replication per sampling stage. In particular, one can allocate one replication per stage per parallel CPU that can be used to perform the sampling. At each step, each parallel CPU can be used to run one of the allocated samples.

Appendix D: 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 (13) 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 (13) 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 (13) 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.

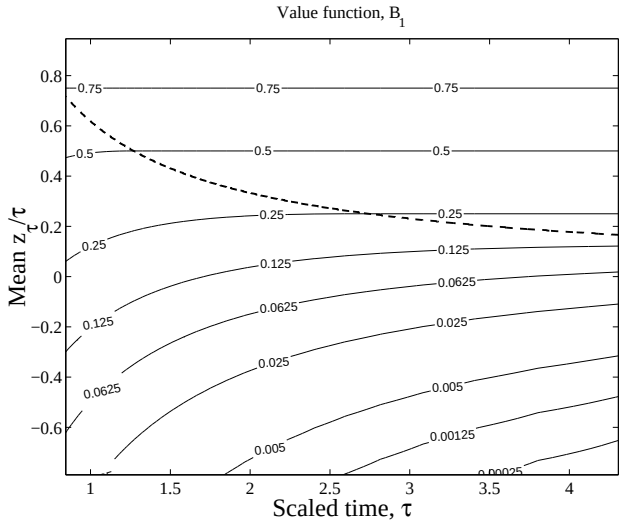


Figure EC.3 Contours of the OEDR $B_1(w,s) = B_1(z_\tau/\tau, 1/\tau)$, with dashed free boundary, $b_1(s) = b_1(1/\tau)$.

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

$$b_{BL}(s)/\sqrt{s} \approx \begin{cases} \sqrt{s/2} & \text{if } s \le 0.2 \\ 0.49 - 0.11 s^{-1/2} & \text{if } 0.2 < s \le 1 \\ 0.63 - 0.26 s^{-1/2} & \text{if } 1 < s \le 5 \\ 0.77 - 0.58 s^{-1/2} & \text{if } 5 < s \le 15 \\ [2\log s - \log\log s - \log 16\pi]^{1/2} & \text{if } 15 < s. \end{cases} \tag{EC.13}$$

For small $\tau = 1/s$, that approximation matches our computation well. That approximation 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.

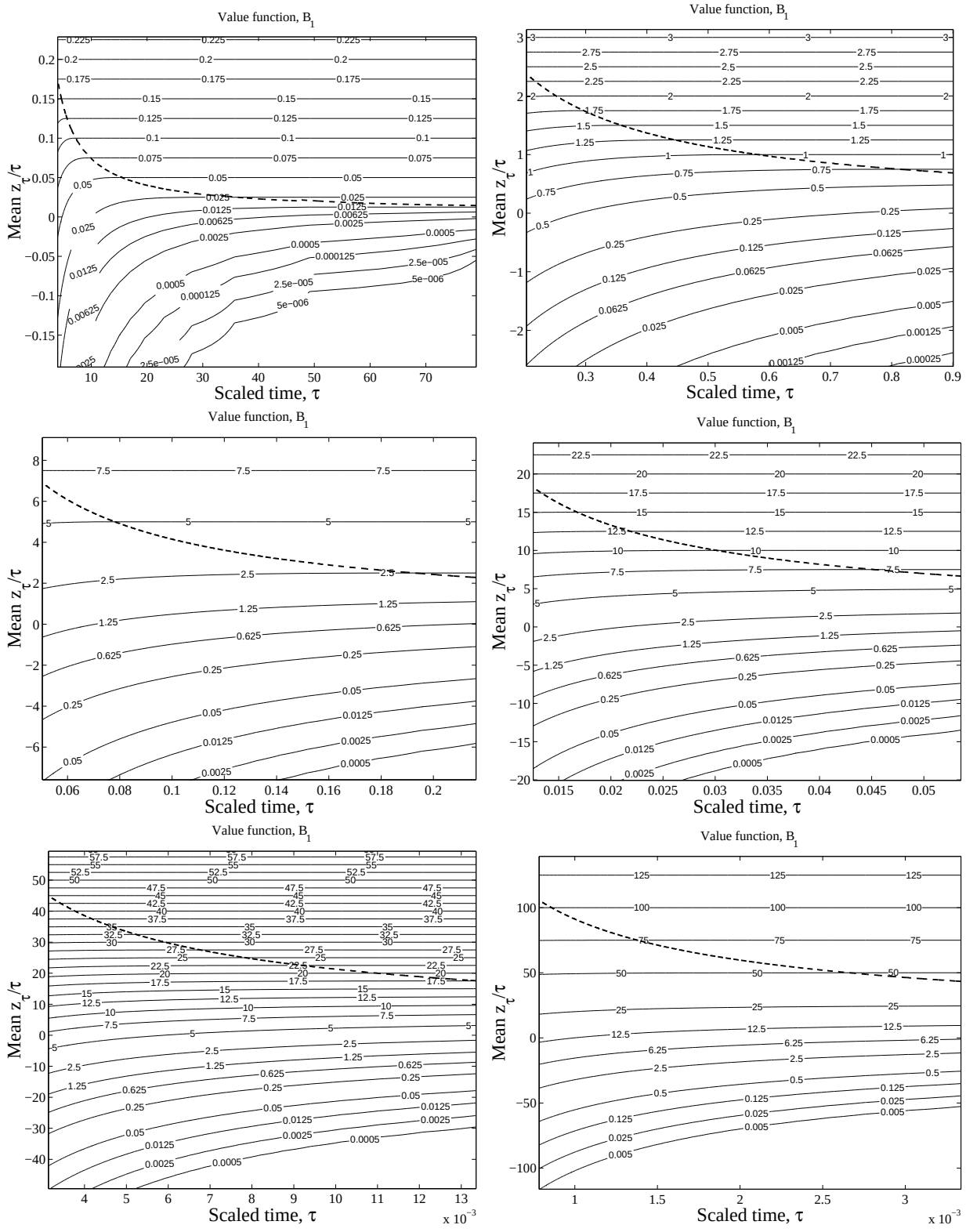


Figure EC.4 Contours for standardized OEDR, $B_1(w,s) = B_1(z_\tau/\tau, 1/\tau)$, with dashed free boundary $b_1(s) = b_1(1/\tau)$.

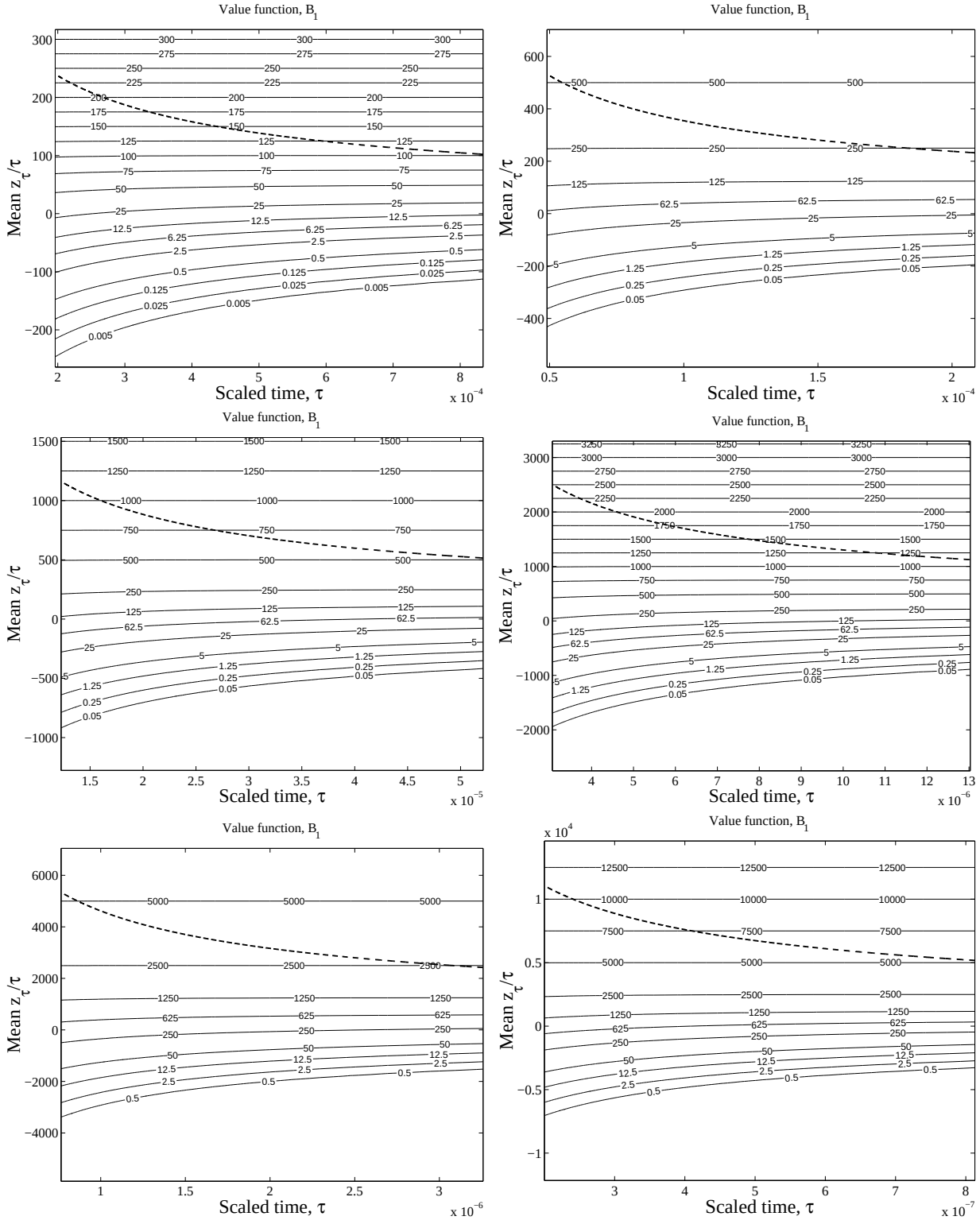


Figure EC.5 Contours for standardized OEDR, $B_1(w,s) = B_1(z_\tau/\tau, 1/\tau)$ with dashed free boundary $b_1(s) = b_1(1/\tau)$.

We propose and recommend an easy-to-compute alternative to (EC.13) that reduces the difference between the approximation in (EC.13) and the more accurate free boundary solution for $b_1(1/\tau)$. To develop the approximation, we used 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:

$$\tilde{b}_1(s) \approx \begin{cases} s/\sqrt{2} & \text{if } s \leq 1/7 \\ \exp\left[-0.02645(\log s)^2 + 0.89106 \log s - 0.4873\right] & \text{if } 1/7 < s \leq 100 \\ \sqrt{s}\left[2\log s - \log\log s - \log 16\pi\right]^{1/2} & \text{if } 100 < s. \end{cases} \tag{EC.14}$$

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, $y_t/t + \beta^{-1}\tilde{b}_1(1/\gamma t)$.

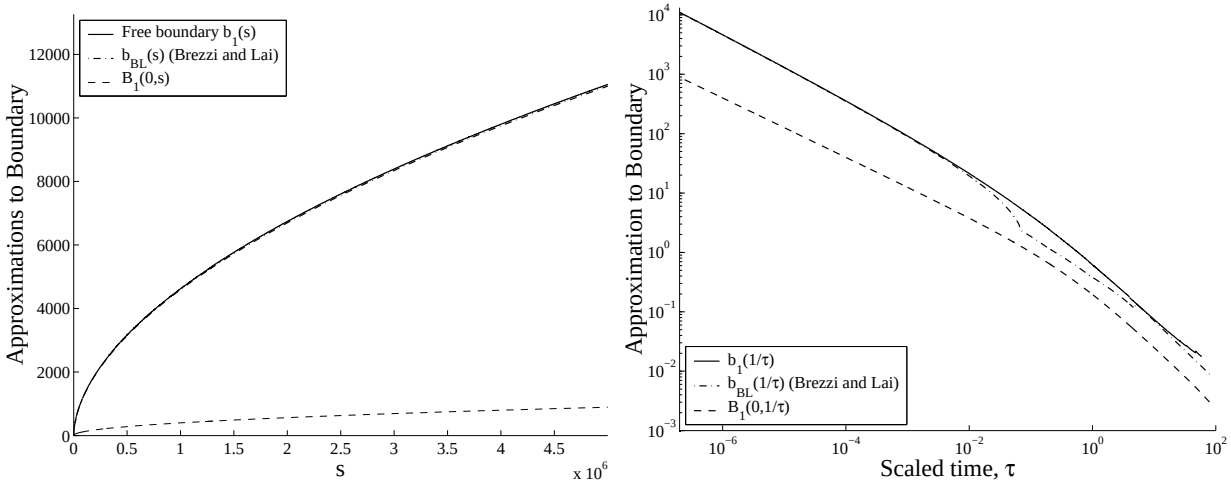


Figure EC.6 Free boundary $b_1(s) = b_1(1/\tau)$.

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 used the lower bound of Theorem 3. We did not have a special bias correction for B_1 as we did 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 E: Simulation Selection Procedures with $k > 1$ Project Alternatives

This section describes how to adapt one-stage $\mathcal{LL}$ allocations to the present context. One-stage $\mathcal{LL}$ allocations 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 are 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. That result 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. First, the specification of prior distributions obviates the need to take a first stage of sampling. Second, for a small to 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. Third, the allocation can be made sequential by updating statistics and repeatedly allocating replications until a stopping rule is satisfied. Fourth, 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.

These adaptations, in the notation of the current paper, 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 $\Theta_i \sim \texttt{Normal}\left(\mu_{0i}, \sigma_i^2/t_{0i}\right)$, 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 $\mu_{0(0)} \leq \mu_{0(1)} \leq \ldots \leq \mu_{0(k)}$.
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 $\lambda_{ik}^{-1} \leftarrow \hat{\sigma}^2_{(i)}/t_{0,(i)} + \hat{\sigma}^2_{(k)}/t_{0,(k)}$. If $(k) \notin \mathcal{S}$ then set $\lambda_{ik} \leftarrow t_{0,(i)}/\hat{\sigma}^2_{(i)}$.

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

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

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 $r_{(i)} \leq 0$, 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 $i = 1, \ldots, k$. 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 $\mu_{0(0)} \leq \mu_{0(1)} \leq \ldots \leq \mu_{0(k)}$.

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.

The stopping rules EOC_1^γ and EOC_k^γ formally test whether or not the sampling budget ß that maximizes an approximation to the expected discounted value of sampling, assuming that, for any given ß, the allocation for future samples was calculated to by evaluating Steps 3a-3e (with $r \leftarrow$ ß). The determination of the optimal value of ß incurs a computational cost that is associated, for example, with a line-search optimization algorithm for ß. A computational speed-up can be obtained by simply checking if there exists a ß ≥ 1 such that the expected discounted value of sampling is positive. If that is the case, then the optimal ß certainly has a positive expected discounted value of sampling. In our implementation, we initially solve for the optimal ß. If that value exceeds 1, we continue sampling. In the next iteration, we check if a sampling budget of $\max\{1, ß - 1\}$ 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 ß ≥ 1 with 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 ß. For example, when comparing $k = 1$ simulated alternative with a known deterministic NPV of 0, and when the simulated mean is just below the stopping boundary, the expected reward of a one-step algorithm with ß $= 1$ replication might not justify additional sampling, but some values of ß > 1 may justify additional sampling.

It is therefore not optimal to perform a one-step lookahead allocation by only testing if $ß = 1$ additional replication is sufficient to justify continuing.

In the numerical experiments of §6.4, we implemented 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.

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., 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.

Inoue, K., S. E. Chick. 1998. Comparison of Bayesian and frequentist assessments of uncertainty for selecting the best system. D. J. Medeiros, E. J. Watson, M. Manivannan, J. Carson, eds., *Proc. 1998 Winter Simulation Conference*. IEEE, Inc., Piscataway, NJ, 727–734.

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.