Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Moench, O. Rose, eds.

THE KNOWLEDGE-GRADIENT STOPPING RULE FOR RANKING AND SELECTION

Peter Frazier
Warren B. Powell

Department of Operations Research & Financial Engineering
Princeton University
Engineering Quadrangle
Olden St. Princeton, N.J. 08544, U.S.A.

ABSTRACT

We consider the ranking and selection of normal means in a fully sequential Bayesian context. By considering the sampling and stopping problems jointly rather than separately, we derive a new composite stopping/sampling rule. The sampling component of the derived composite rule is the same as the previously introduced LL1 sampling rule, but the stopping rule is new. This new stopping rule significantly improves the performance of LL1 as compared to its performance under the best other generally known adaptive stopping rule, EOC Bonf, outperforming it in every case tested.

1 INTRODUCTION

The problem of ranking and selection is ubiquitous in simulation applications, arising whenever we must select which of several options is the best. The essential content of the problem is how to allocate simulation effort among the options to most efficiently and accurately make this selection. Intelligent ranking and selection techniques often perform markedly better than naive allocation rules like equal allocation.

The problem of ranking and selection has a long history, both within and apart from simulation, marked by the seminal work Bechhofer, Kiefer, and Sobel (1968), the comprehensive monograph Bechhofer, Santner, and Goldsman (1995), and a great deal of more recent work. See Swisher, Jacobson, and Yücesan (2003) for a review of its application within simulation.

Within this body of work, a number of staged and fully sequential Bayesian ranking and selection techniques have been recently proposed including Chen, Dai, and Chen (1996), Chen, Lin, Yücesan, and Chick (2000), Chick and Inoue (2001b), Chick and Inoue (2001a), Chen, He, and Fu (2006), He, Chick, and Chen (2007). These techniques optimize average-case instead of worst-case performance, and so are generally less conservative than techniques using the indifference zone formulation.

Recently, Chick, Branke, and Schmidt (2007a) and Chick, Branke, and Schmidt (2007b) introduced a new myopic Bayesian rule, LL1, for ranking and selection with independent normal rewards of unknown mean and variance. This myopic rule looks a single measurement into the future and chooses the measurement that would be best if this next measurement were the last. The rule adopts the linear loss function, which penalizes according to the difference in value between the chosen option and the best, contrasting it with another common choice, $0-1$ loss, which penalizes a constant 1 for failing to find the best alternative. Other algorithms designed for the linear loss objective function under independent normal samples include LL(S) (Chick and Inoue 2001b) and OCBA for linear loss (He, Chick, and Chen 2007).

The rules LL1 and LL(S) are derived similarly, except that LL(S) considers the effect of a block of measurements while LL1 considers the effect of only one single measurement. Blocks of measurements are more difficult to analyze and necessitate the introduction of the Bonferonni inequality to approximate their effect, while the effect of a single measurement may be computed analytically. Essentially, LL1 allocates single measurements exactly while LL(S) allocates multiple measurements approximately.

Extensive numerical comparisons were made in Chick, Branke, and Schmidt (2007b) with these two procedures under two different stopping rules: the naive or fixed stopping rule, which simply stops after a fixed number of samples have been taken; and the EOC Bonf stopping rule, introduced in Branke, Chick, and Schmidt (2005), which more intelligently decides when to stop based on an approximation to the expected loss that would occur due to remaining uncertainty. As shown in Branke, Chick, and Schmidt (2005) and again in Branke, Chick, and Schmidt (2007), using the EOC Bonf stopping rule instead of the fixed one generally improves the performance of ranking and selection procedures.

These numerical comparisons in Chick, Branke, and Schmidt (2007b) between LL(S) and LL1 revealed that LL(S) outperforms LL1 in a broad class of problem configurations when both operate under the EOC Bonf stopping rule. In other situations, for example with the fixed stopping rule taking a small number of samples, LL(S) performed better but the great benefit provided by using an adaptive stopping rule like EOC Bonf led the authors to conclude that LL1 may not be broadly applicable. Further, they supposed that its myopic assumption was the culprit behind its poor performance.

In this article, we show that the performance of the LL1 procedure may be markedly improved by an adaptive stopping rule other than EOC Bonf. This new stopping rule is derived by including the cost of each sample into the overall objective and then solving this new objective using the same myopic assumption used by LL1. We call this technique of making a myopic assumption in an information acquisition problem the knowledge-gradient method (Frazier, Powell, and Dayanik 2007, Frazier, Powell, and Dayanik 2008), since it evaluates the expected change in the value of knowledge that would result from a single measurement, and chooses its measurements to maximize this change or gradient. We call the stopping rule that results the knowledge-gradient (KG) stopping rule.

The performance of LL1 when stopped using this new stopping rule is commensurate with LL(S) when stopped using the EOC Bonf stopping rule, and from this we infer that the culprit behind LL1's poor performance under the EOC Bonf stopping rule may not have been myopia, but instead a negative interaction between sampling and stopping decisions which is alleviated by the KG stopping rule.

2 BAYESIAN FORMULATION

We briefly review the Bayesian formulation of the ranking and selection problem with normal means, as well as the decisions made by the LL(S) and LL1 rules.

Suppose we have M alternatives, and samples from alternative x are iid normal with mean μ_x and precision β_x. We will denote the vector of means $(\mu_1,\ldots,\mu_M)$ by μ and the corresponding vector of precisions by β.

We know neither the means nor variances, and so we adopt a normal-gamma prior (see, e.g., DeGroot (1970)) in which β_x is gamma distributed with precision a_x^0 and scale b_x^0, and μ_x is normally distributed with mean μ_x^0 and precision $\beta_x\rho_x^0$ when conditioned on β_x. Under this prior, we assume that (μ_x,β_x) is independent of $(\mu_{x'},\beta_{x'})$ for $x \neq x'$. The vectors a^0, b^0, ρ^0, and μ^0 composed of a_x^0, b_x^0, ρ_x^0, and μ_x^0 with x ranging from 1 to M then completely characterize the prior.

Commonly we assume that the prior is noninformative, so $a_x^0 = -1/2$, $b_x^0 = 0$, $\rho_x^0 = 0$ for each x. With these parameters, μ_x^0 does not affect the posterior and so its value is irrelevant.

We will then make a sequence of sampling decisions $x^0, x^1, \ldots$ and from each observe a corresponding sample $W^1, W^2, \ldots$, where $W^{n+1} \sim \text{Normal}(\mu_{x^n}, \beta_{x^n})$ and is conditionally independent of the previous $(W^k)_{k\leq n}$ given x^n, μ_{x^n}, and β_{x^n}.

As our prior is conjugate to our sampling distribution, our samples result in a sequence of posterior distributions on μ,β which are again normal-gamma distributed. We will denote the parameter vectors of the posterior at time n by a^n, b^n, ρ^n, μ^n. More precisely, we have

$$\begin{aligned}
\beta_x \mid x^0,\ldots,x^{n-1}, W^1,\ldots,W^n \quad &\sim \text{Gamma}(a_x^n, b_x^n) \\
\mu_x \mid x^0,\ldots,x^{n-1}, W^1,\ldots,W^n, \beta_x &\sim \text{Normal}(\mu_x^n, 1/\rho_x^n\beta_x^n).
\end{aligned}$$

These posterior parameter vectors may be computed recursively by the following update as in DeGroot (1970). For all $x \neq x^n$ we leave the parameters unchanged, and for $x = x^n$ we compute the new parameters via

$$\begin{aligned}
a_x^{n+1} &= a_x^n + 1/2, \\
b_x^{n+1} &= b_x^n + (W^{n+1} - \mu_x^n)^2/2(\rho_x^n + 1), \\
\rho_x^{n+1} &= \rho_x^n + 1, \\
\mu_x^{n+1} &= \left(\rho_x^n\mu_x^n + W^{n+1}\right)/2(\rho_x^n + 1).
\end{aligned}$$

If the noninformative prior is taken, then these parameters may be interpreted further. In this case, $\rho_x^n = 2a_x^n + 1$ is the number of times we have sampled alternative x by time n, and μ_x^n and $2b_x^n$ are respectively the sample mean and sum of square deviations of these samples. In addition, the maximum likelihood estimator of the sampling variance $1/\beta_x$ given by the sum of square deviations divided by the number of samples minus 1 is equal to b_x^n/a_x^n.

Together with this information collection process we define a filtration $(\mathscr{F}^n)_{n=0}^\infty$, where $\mathscr{F}^n$ is the sigma-algebra generated by $x^0, W^1, \ldots, x^{n-1}, W^n$, so that the posterior at time n is the prior conditioned on $\mathscr{F}^n$.

We will suppose that we take samples until some stopping time τ, and then choose the alternative that appears to be the best based on the accumulated evidence. The chosen alternative is any from the set $\arg\max_x \mu_x^\tau$. We then receive a reward equal to the true value of the selected alternative. Conditioned on $\mathscr{F}^\tau$, which is the information acquired by time τ, this reward has expected value $\max_x \mu_x^\tau$.

For now we will suppose that we have no control over this stopping rule τ, and that it is simply a given stopping time of the filtration. For example, it could be a constant, or it could be the EOC Bonf stopping rule. With τ given, we would like to choose a sequence of sampling decisions $\pi = (x^0, x^1, \ldots)$ so as to maximize the expected value of our reward, with our only requirement being that x^n must be adapted to $\mathscr{F}^n$ for each n. Then the optimal Bayesian sampling rule would be given by the solution to

$$\sup_{\pi} \mathbb{E}^{\pi}\left[\max_x \mu_x^{\tau}\right], \tag{1}$$

where again the supremum is over all policies adapted to the filtration. Note that maximizing this reward is equivalent to minimizing the expected opportunity cost, where opportunity cost is defined to be $\mu_{x^*} - \mu_{i^*}$, with $x^* \in \arg\max_x \mu_x$ and $i^* \in \arg\max_x \mu_x^{\tau}$. This opportunity cost is the difference in value between the best alternative and the one that we have chosen. This corresponds to the linear loss function discussed above.

We have assumed in (1) that τ is given. Indeed, the derivations of most existing sampling rules do not explicitly consider the role that the choice of stopping rule plays in overall performance. A common assumption for the purposes of analysis (see, for example, (Frazier, Powell, and Dayanik 2007)) is that τ is some fixed constant. Our goal in this article, however, is to show that there is value in deriving the sampling and stopping rules together, and so later, in Section 3, we will consider the optimization problem in which we control both the sampling rule π and the stopping rule τ.

Although (1) can in principle be solved through dynamic programming, the computational challenges are prohibitive. Instead, a number of heuristic approaches have been presented. We briefly review two of these approaches: LL1, and LL(S).

LL1, introduced in Chick, Branke, and Schmidt (2007a), allocates its measurements one-at-a-time by supposing at time n that τ will equal $n+1$ and allocating x^n in a way that would be optimal were this assumption true. There, the optimization problem is solved explicitly by noting first that $\mu_{x'}^{n+1} = \mu_{x'}^n$ for all $x' \neq x^n$, and the marginal distribution of $\mu_{x^n}^{n+1}$ is student-t. From this, we can define a quantity v_x^n as the marginal value of measuring x and calculate it as,

$$\begin{aligned} v_x^n &:= \mathbb{E}\left[\max_{x'} \mu_{x'}^{n+1} \mid \mathscr{F}^n, x^n = x\right] - \max_{x'} \mu_{x'}^n \\ &= \lambda_{\{x\cdot\}}^{-1/2} \Psi_{\rho_x^n}\left(\lambda_{\{x\cdot\}}^{1/2} |\mu_x^n - \max_{x' \neq x} \mu_{x'}^n|\right). \end{aligned} \tag{2}$$

Here $\lambda_{\{x\cdot\}}$ and Ψ_d are defined by

$$\lambda_{\{x\cdot\}} := \rho_x^n(\rho_x^n + 1)a_x^n / b_x^n,$$

$$\Psi_d(s) := \int_{u=s}^{\infty} \phi_d(u)\ du = \frac{d+s^2}{d-1}\phi_d(s) - s\Phi_d(-s),$$

where Φ_d and ϕ_d are respectively the cdf and pdf of the student-t distribution with d degrees of freedom.

The LL1 policy is then given by

$$x^n \in \arg\max_x v_x^n \tag{3}$$

The LL(S) rule, introduced in Chick and Inoue (2001b), considers the effect of blocks of measurements. It is parameterized by the block size, which is commonly denoted by τ, but which we will refer to as B to avoid confusing it with our stopping time τ. At the beginning of each stage, the LL(S) allocation considers the marginal benefit of the next B measurements,

$$\mathbb{E}\left[\max_{x'} \mu_{x'}^{n+B} \mid \mathscr{F}^n, x^n, \ldots, x^{n+B-1}\right] - \max_{x'} \mu_{x'}^n, \tag{4}$$

as a function of the alternatives sampled, $x^n, \ldots, x^{n+B-1}$.

Ideally, the LL(S) algorithm would like to optimize (4) over $x^n, \ldots, x^{n+B-1}$, but since computing (4) is computationally intensive and optimizing over it is even more so, LL(S) uses the Bonferonni inequality to approximate the optimal allocation and allocates according to that approximation. A full description of the LL(S) algorithm may be found in Chick and Inoue (2001b).

Note that in this formulation the decision of which alternatives to measure between times n and $n+B-1$ may depend only with the information available at time n, while under LL1 each measurement decision is made with the full information available. Also note that when $B = 1$ the objective function (4) from which LL(S)'s allocation is derived is identical to (1), but LL1 optimizes this expression exactly while LL(S)'s use of the Bonferonni inequality results in an approximation to the optimal.

3 OPTIMAL STOPPING RULE

We have formulated the objective function that would result from having an externally imposed stopping rule τ. In most applications, however, we may control our measurement budget in order to trade measurement cost against the value of obtaining more information. To model this trade-off, we will suppose that the total cost of measurement is some convex non-decreasing function $C : \mathbb{N} \to \mathbb{R}$ of the number of measurements taken.

We will call our sampling rule π and our stopping rule τ. The sampling rule must be adapted to the filtration, as before, and the stopping rule τ will again be required to be a stopping time of the filtration generated by π, by which we mean that the event $\{\tau \leq n\}$ is $\mathscr{F}^n$ measurable for each n. This is a non-anticipativity requirement and simply prevents basing the decision to stop on information that would be obtained in the future. Further details on stopping times may be found, for example, in Kallenberg (1997).

Our objective function is then,

$$\sup_{\pi,\tau} \mathbb{E}^{\pi}\left[\max_x \mu_x^{\tau} - C(\tau)\right]. \tag{5}$$

The form of the cost function assumed is a generalization of that used in the sequential probability ratio test in Wald and Wolfowitz (1948), which assumes that the cost function C is linear in the amount of reward obtained. Since we assume that C is convex, but not necessarily strictly so, this allows linear costs. Requiring only that C is convex and non-decreasing also allows the choice $C(n) = \infty\, \mathbf{1}_{\{n\geq N\}}$, by which we mean that $C(n)$ is 0 for $n < N$ and infinite for $n \geq N$. If we take this choice we recover the fixed-budget objective function, which allows free measurements up to time N, and no subsequent measurements.

Note the contrast between this formulation and that in Chick and Gans (2008), which assumed the cost of measurement was implicit in a discounting of the final reward obtained, giving a net final reward of $e^{-r\tau}\max_x \mu_x^{\tau}$. In that formulation, the cost of measurement depends on the final reward obtained, and in the formulation proposed here it does not. Each objective function is appropriate for its own applications.

4 KNOWLEDGE-GRADIENT STOPPING RULE

Just as solving the Bayesian ranking and selection problem with a given stopping rule τ is computationally intractable, so is solving the more difficult problem (5) in which we also optimize over τ. This justifies the introduction of a heuristic, which we derive using a method which we refer to as the knowledge-gradient (KG) method.

To apply the KG method, we fix a time n and suppose that we have not yet stopped by this time. We further suppose that *if we continue*, then we will still be required to stop at the next time $n+1$. This is the same assumption used to derive the LL1 policy, and what we call the knowledge-gradient method is referred to as the myopic or greedy assumption in Chick, Branke, and Schmidt (2007a). Our choice of name "knowledge-gradient" refers to the fact that the single-sample assumption induces a direct measure of the value of the knowledge we have before and after a measurement, and the difference in these values can be regarded as something like the gradient in knowledge achieved by a measurement.

Given the KG assumption, we can then compute what the optimal decision would be if this assumption were true. The optimal decision is the best among either stopping now at n, or measuring any alternative x, incurring the measurement cost, and stopping at time $n+1$. Stopping now by taking $\tau = n$ has value

$$\max_{x'} \mu_{x'}^{n} - C(n),$$

while measuring alternative x and then stopping has value

$$\mathbb{E}\left[\max_{x'} \mu_{x'}^{n+1} - C(n+1) \mid \mathscr{F}^n, x^n = x\right] = \left(\max_{x'} \mu_{x'}^{n}\right) + \nu_x^n - C(n+1), \tag{6}$$

as can be seen directly from (2). The x that maximizes (6) is exactly the x^n maximizing (2), and is thus the same as the decision of the LL1 sampling rule. Thus, the decision we face in our sampling and stopping problem is between sampling the alternative suggested by the LL1 sampling rule, or stopping. Furthermore, the difference in value between sampling this best x and stopping now is equal to

$$-\left(C(n+1) - C(n)\right) + \max_x \nu_x^n \tag{7}$$

and so we should sample if this difference is positive, and stop if it is negative. This gives us the composite KG sampling/stopping rule as

1. If $C(n+1) - C(n) \geq \max_x \nu_x^n$, then stop sampling.
2. Otherwise, sample $x^n \in \arg\max_x \nu_x^n$.

This derivation shares much with that of LL1, with the crucial difference being that it applies the KG method to the sampling and stopping problem together, rather than simply applying it to the sampling problem and then imposing another stopping rule.

We show in the following proposition that the τ chosen by the KG stopping rule bounds from below the τ chosen by the best stopping rule for the LL1 sampling rule. Note that the value of the KG stopping rule is also (trivially) a lower bound on the value of the best policy, but that this proposition bounds the *decision* made by the optimal policy.

Proposition 1. *Let τ^{KG} be the stopping rule defined by the KG stopping rule and let τ^* be a stopping rule that is optimal for the problem*

$$\sup_{\tau} \mathbb{E}^{\pi=\mathrm{LL1}}\left[\max_x \mu_x^{\tau} - C(\tau)\right]. \tag{8}$$

Then $\tau^* \geq \tau^{KG}$ almost surely.

Another way to understand this proposition is as telling us that, if the KG stopping rule suggests continuing at a given time n, then so would the optimal stopping rule. The only mistake that the KG rule makes is in sometimes stopping too soon. We provide a formal proof of the proposition in the appendix, but the essential intuition is that the KG stopping rule uses the exact value of stopping but underestimates the value of continuing. Thus, when comparing these values, it errs on the side of stopping too soon.

Although the KG stopping rule can stop too soon, there is numerical evidence to suggest that the cost of this early stopping is low. We present this evidence in Section 5. In addition, we make a theoretical argument here that is based on the idea that the net value of continued measurement decreases on average.

In particular, consider the case when the expected net marginal value of continuing in this time period given by (7) is negative. This is the situation in which the KG stopping rule stops, and is the only case in which it can err. The only reason an optimal stopping rule τ^* would continue in this situation would be an expectation that net marginal values of continuing will be positive in the future, compensating for the net loss incurred in taking the current measurement. Thus continuing in this situation incurs an immediate loss with the *possibility* of future profit. But if we expect future net marginal values of continuing to be even more negative than they are now, there is little possibility of future profit and the KG method's single-sample assumption is reasonable.

In Figure 1, the marginal value $(\max_x \mu_x^{n+1}) - (\max_x \mu_x^n)$ of the information obtained from the sample taken at time n is plotted against n for one particular simulation from the slippage configuration described in Section 5.1. Samples are taken according to the LL1 sampling rule. In this simulation the marginal value indeed tended to decrease, and this tendency is displayed in general for other sampling rules.

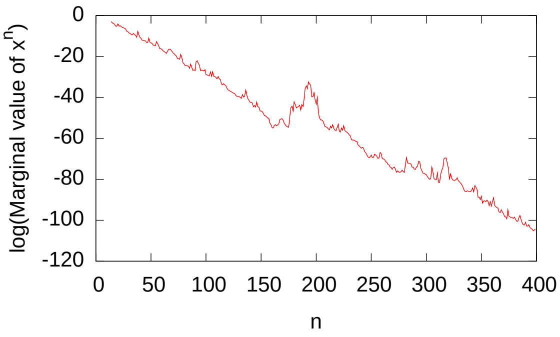

Figure 1: The logarithm of the marginal value of measurement x^n, $\log((\max_x \mu_x^{n+1}) - (\max_x \mu_x^n))$, plotted vs. n. Sampling decisions are made by LL1, and the underlying configuration is the slippage configuration described in Section 5.1

Finally, we note the KG stopping rule will better approximate the optimal stopping rule τ^* of Proposition 1 as the function C becomes more strictly convex. This is because the possibility of continuing after $n+1$ becomes increasingly remote as the marginal cost of continued measurement increases, better justifying the heuristic's single-sample assumption. The KG stopping rule is perfect, for example, in the case when $C(n) = \infty \ \mathbf{1}_{\{n \geq N\}}$ because it correctly continues for $n < N$, and stops at time N.

5 NUMERICAL RESULTS

We now explore the relative quality of KG and other stopping rules through numerical simulation on several test cases. The selection of test cases owes a great deal to the work in Chick, Branke, and Schmidt (2007a) and Chick, Branke, and Schmidt (2007b). We concentrated our effort on the test cases used in that work, since that is where the LL1 policy was first presented, and the test cases presented there show that LL1 with the EOC Bonf stopping rule underperforms. In particular, we provide results here for LL1 and LL(S) sampling rules under KG, EOC Bonf, and fixed stopping rules. This makes for a total of six choices of sampling and stopping rules. We explore these choices on three configurations: the slippage configuration (SC); the monotone decreasing means configuration (MDM); and random problem instances (RPI). These configurations are described in more detail below.

In each case we simulated the sampling and stopping rules on the configuration 10^5 times in order to obtain the results pictured. We then varied the parameter of the stopping rule used in order to obtain different trade-offs between accuracy and sample size. At each such trade-off we estimated the expected sample size $E[\tau]$ and the expected opportunity cost denoted E[OC], where again by opportunity cost we mean the difference in value between the best alternative and the one that appears best based on sampling. We used the noninformative prior for all sampling and stopping rules.

When computing the KG stopping rule, we fixed the function C to be a linear function so that $C(n) = cn$ for some constant c. We then varied the constant c in order to obtain different trade-offs between sampling and opportunity cost. This suggests one drawback of the KG stopping rule. In many applications, we may have difficulty quantifying our cost of measurement function C and instead we would like our stopping rule to satisfy an upper bound α on the expected opportunity cost upon stopping. Unlike the EOC Bonf stopping rule, for which we could set its stopping parameter to α and at least obtain an expected opportunity cost upon stopping that is reasonably close to α, when using the KG stopping rule it is not clear what value of c we should choose. Further research is needed to relate values of c to target expected opportunity costs upon stopping.

When computing the decisions of the LL(S) sampling rule we set its block-size parameter B to 1. This decision was based on a series of tuning experiments in which we tested LL(S) at values of B between 1 and 10 on the slippage configuration described below. These tuning experiments revealed a small but significant difference between the performance of the policies, with performance improving as B decreased to 1. Note that except for the Bonferonni approximation, LL(S) with $B = 1$ is equivalent to LL1, and that whenever LL(S) with this value of B outperforms LL1 it can only be because the approximation is *helping* the policy. Hence the fact that the performance of LL(S) improves as B decreases, at least in the slippage configuration, is interesting evidence that the single-sample assumption made by LL1 is not a liability.

The results for LL1 and LL(S) on EOC Bonf and fixed stopping rules replicate results originally presented in Chick, Branke, and Schmidt (2007b), while the results for the KG stopping rule are new. We will see from these experiments that the KG stopping rule improves the performance of LL1 to the point where it is comparable to LL(S) with the EOC Bonf stopping rule.

5.1 Slippage Configuration

Under the slippage configuration (SC), the best alternative is given a sampling mean $\delta > 0$, and the remaining alternatives all have sample mean 0. We chose $\delta = 0.5$. Some flexibility is generally given to the sampling variances as well, but we set them all equal to each other at a value of 1. The configuration had $M = 5$ alternatives.

The slippage configuration draws its name from the indifference zone formulation of the ranking and selection problem, where it is the configuration that marks the transition from the preference to the indifference zone. In this sense, it is the most difficult configuration that we should be able to identify. Since we are dealing with a Bayesian formulation of the problem in which linear loss in the objective function, the slippage configuration loses some of this meaning, but nevertheless it is an important test case.

We picture the relative performance of LL(S) and LL1 under KG, EOC Bonf, and fixed stopping rules in Figure 2. We see in these results that LL1 performs better under the KG sampling rule than it does under EOC Bonf, and that both adaptive stopping rules perform better than their fixed counterparts under both sampling rules. We also see that LL1 under KG stopping performs better than does LL(S) with EOC Bonf stopping in this problem setting.

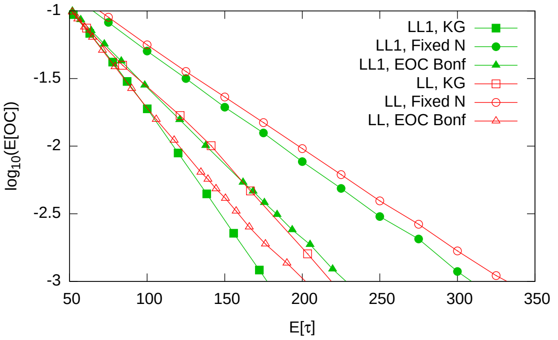

Figure 2: Slippage configuration with $\delta = 0.5$, 5 alternatives, and sampling variance 1.

5.2 Monotone Decreasing Means

As the name implies, under the monotone decreasing means configuration (MDM) the alternatives are arranged in monotonically decreasing order. In particular, the sampling mean of alternative i is equal to δi. We chose $\delta = 0.5$. The sampling variances were all 1 and the number of alternatives was $M = 10$.

We picture the relative performance of our sampling and stopping rules for the MDM configuration in Figure 3. We see again in these results that LL1 performs better under the KG sampling rule than it does under EOC Bonf, and that both adaptive stopping rules outperform better their fixed counterparts. Unlike in the SC configuration, however, we see in this configuration that LL1 under KG stopping performs is outperformed by LL(S) with EOC Bonf stopping.

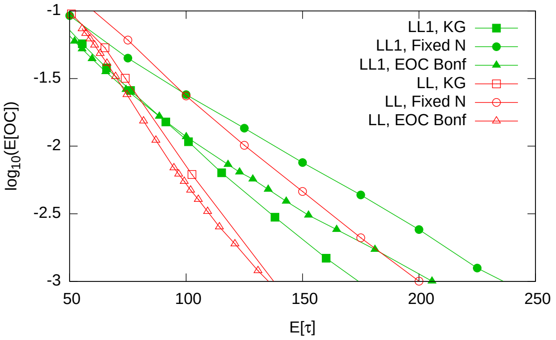

Figure 3: Monotone decreasing means configuration with $\delta = 0.5$, 10 alternatives, and sampling variance 1.

5.3 Random Problem Instances

Since SC and MDM configurations represent idealized special cases and are not necessarily typical of problems that might be met in application, we attempted to replicate more naturalistic configurations by randomly generating them from a normal-gamma prior. Specifically, we generated the sampling precision β_x independently for each alternative from a gamma prior with shape parameter 99 and scale parameter 100. We then generated the sampling mean μ_x independently for each alternative from a normal distribution with mean 0 and variance $1/(\beta_x \eta)$. We chose $\eta = 1/2$. Configurations had $M = 5$ alternatives.

We randomly generated 20 problem configurations according to this prior, paying special attention to the relative performance of LL1 with KG stopping, LL1 with EOC Bonf stopping, and LL(S) with EOC Bonf stopping. We found that KG stopping outperformed EOC Bonf stopping under LL1 sampling in every situation. LL1 with KG stopping performed comparably to LL(S) with EOC Bonf stopping, sometimes outperforming it and sometimes being outperformed, but always by a small margin.

In Figure 4 we see results from a typical randomly generated problem configuration. Again, these results are typical in the particular sense that LL1 performed better with KG stopping than with EOC Bonf stopping, and that LL1 with KG stopping performed similarly to LL(S) with EOC Bonf stopping. In this particular case LL1 with KG stopping performed better, but this advantage was reversed in other configurations not pictured.

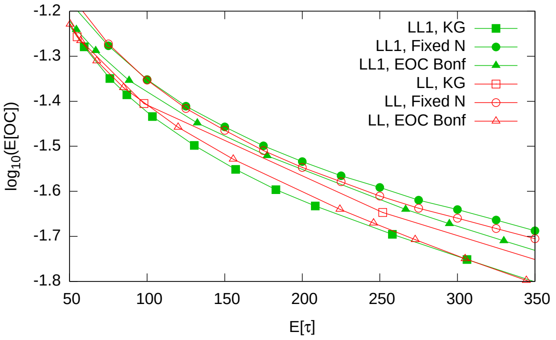

Figure 4: Random problem instance generated according to a normal-gamma prior with 5 alternatives. The sampling distribution had means $\mu = [0.16, 0.21, -1.40, -1.20, -0.16]$ and precisions $\beta = [1.07, 1.04, 0.89, 1.05, 0.97]$.

We draw two conclusions from these numerical experiments. First, LL1 performs much better with the KG stopping rule than it does with EOC Bonf. Second, LL1 under KG stopping performs commensurately with LL(S) under EOC Bonf stopping. On any given test case one might outperform the other by a small margin, but the advantage switches from test case to test case, and neither choice has a clear overall advantage. We see from this that a careful choice of stopping rule is critical to LL1's performance.

6 CONCLUSION

We have shown that the LL1 procedure, introduced as a sampling rule that can be derived exactly under the knowledge-gradient assumption, can be understood in a broader context as the sampling portion of a composite sampling and stopping rule that can again be derived exactly under the same knowledge-gradient assumption. Furthermore, we can obtain significantly better efficiency by sampling and stopping according to this composite rule as compared to sampling with LL1 but using another adaptive stopping rule. The resulting performance is commensurate with other well-regarded sampling/stopping rules like LL(S) with EOC Bonf stopping.

This is valuable first because it provides a new sampling/stopping rule that works well overall and is likely to work even better in small sample situations. Secondly, and we believe more importantly, it provides a general framework under which composite sampling/stopping rules may be derived. The knowledge-gradient assumption, which is that the current time-period is our last opportunity to sample, depended in no way on the normality of the sample distributions, the particular form of the prior, or on the independence of the samples through time. As long as we can evaluate the one-dimensional integral needed to compute the marginal value of a single measurement, we can create a knowledge-gradient based heuristic for the problem at hand. The quality of the resulting heuristic must of course be evaluated in each new situation to which it is applied, but the results described here add to the evidence accumulated in other problem settings (see Frazier, Powell, and Dayanik (2007), Frazier, Powell, and Dayanik (2008)) suggesting that the heuristic performs well in many important problems.

A APPENDIX

Proof of Proposition 1. We will show the result by considering the problem (8) as a dynamic program. Our state space will consist of the current time n and the posterior distribution at that time, which is parameterized by the vectors a^n, b^n, ρ^n, μ^n. For compactness we will use S^n to denote this tuple of random vectors $(a^n, b^n, \rho^n, \mu^n)$ and s to denote one possible value that this tuple might take.

Let V denote the value function for this problem, which is a function from our state space to the real numbers defined by

$$V(s,n) = \sup_{\tau \geq n} \mathbb{E}\left[\max_x \mu_x^\tau - C(\tau) \mid S^n = s\right].$$

Now, fix n and let s be a state for which τ^{KG} would continue sampling if S^n were equal to s. To show the proposition, it is enough to show that τ^* would also continue if it found S^n equal to this s.

Since τ^{KG} would continue sampling on the event $\{S^n = s\}$, we have by the definition of τ^{KG} that on this event

$$\max_x \mu_x^n - C(n) < \mathbb{E}\left[\max_x \mu^{n+1} - C(n+1) \mid S^n = s\right]. \quad (9)$$

The value of stopping at n is given by $\max_x \mu_x^n - C(n)$ while the value of continuing at n and subsequently following an optimal policy is given by $\mathbb{E}\left[V(S^{n+1}, n+1) \mid S^n\right]$. Any optimal policy will always continue at n if S^n is such that the value of continuing is strictly better than the value of stopping. Hence, it is enough to show for our particular value of s that, on the event $\{S^n = s\}$,

$$\max_x \mu_x^n - C(n) < \mathbb{E}\left[V(S^{n+1}, n+1) \mid S^n = s\right]. \quad (10)$$

To see that this is indeed the case, we note that

$$
\begin{aligned}
&\mathbb{E}\left[V(S^{n+1},n+1) \mid S^n = s\right] \\
&\quad = \mathbb{E}\left[\sup_{\tau \geq n+1} \mathbb{E}\left[\max_x \mu_x^\tau - C(\tau) \mid S^{n+1}\right] \mid S^n = n\right] \\
&\quad \geq \mathbb{E}\left[\mathbb{E}\left[\max_x \mu_x^{n+1} - C(n+1) \mid S^{n+1}\right] \mid S^n = n\right] \\
&\qquad\qquad = \mathbb{E}\left[\max_x \mu_x^{n+1} - C(n+1) \mid S^n = s\right],
\end{aligned}
$$

where the inequality in the penultimate line is a consequence of the fact that the deterministic time $n+1$ is a stopping time contained in the set over which the supremum is taken, and the final line is due to the tower property of conditional expectation. Finally, using this inequality with (9) shows (10), which was the required result. □

REFERENCES

Bechhofer, R., J. Kiefer, and M. Sobel. 1968. *Sequential identification and ranking procedures*. Chicago: University of Chicago Press.

Bechhofer, R., T. Santner, and D. Goldsman. 1995. *Design and analysis of experiments for statistical selection, screening and multiple comparisons*. New York: J.Wiley & Sons.

Branke, J., S. Chick, and C. Schmidt. 2005. New developments in ranking and selection: an empirical comparison of the three main approaches. In *Proc. 2005 Winter Simulation Conference*, ed. M. Kuhl, N. Steiger, F. Argstrong, and J. Joines, 708–717. Piscataway, NJ: Winter Simulation Conference: IEEE, Inc.

Branke, J., S. Chick, and C. Schmidt. 2007. Selecting a selection procedure. Submitted.

Chen, C., L. Dai, and H. Chen. 1996. A gradient approach for smartly allocating computing budget for discrete event simulation. In *Proc. 1996 Winter Simulation Conference*, 398–405. Winter Simulation Conference.

Chen, C., D. He, and M. Fu. 2006. Efficient Dynamic Simulation Allocation in Ordinal Optimization. *IEEE Transactions Automatic Control* 51:2005–2009.

Chen, C., J. Lin, E. Yücesan, and S. Chick. 2000. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems* 10 (3): 251–270.

Chick, S., J. Branke, and C. Schmidt. 2007a. New greedy myopic and existing asymptotic sequential selection procedures: preliminary empirical results. In *Proc. 2007 Winter Simulation Conference*, 289–296. Piscataway, NJ: Winter Simulation Conference: IEEE, Inc.

Chick, S., J. Branke, and C. Schmidt. 2007b. New myopic sequential sampling procedures. Submitted.

Chick, S., and N. Gans. 2008. Economic analysis of simulation selection problems. Submitted.

Chick, S., and K. Inoue. 2001a. New procedures to select the best simulated system using common random numbers. *Management Science* 47 (8): 1133–1149.

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

DeGroot, M. H. 1970. *Optimal Statistical Decisions*. John Wiley and Sons.

Frazier, P., W. B. Powell, and S. Dayanik. 2007. A knowledge gradient policy for sequential information collection. Submitted.

Frazier, P., W. B. Powell, and S. Dayanik. 2008. The knowledge gradient policy for correlated normal rewards. Submitted.

He, D., S. Chick, and C. Chen. 2007, SEP. Opportunity cost and OCBA selection procedures in ordinal optimization for a fixed number of alternative systems. *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews* 37 (5): 951–961.

Kallenberg, O. 1997. *Foundations of modern probability*. New York: Springer.

Swisher, J., S. Jacobson, and E. Yücesan. 2003. Discrete-event simulation optimization using ranking, selection, and multiple comparison procedures: A survey. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 13 (2): 134–154.

Wald, A., and J. Wolfowitz. 1948. Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics* 19 (3): 326–339.

AUTHOR BIOGRAPHIES

PETER FRAZIER received a B.S. degree from the California Institute of Technology in Physics and Engineering & Applied Science, an M.A. from Princeton University in Operations Research & Financial Engineering (ORFE), and is currently a Ph.D. candidate in the ORFE Department at Princeton. He was a finalist in the 2007 INFORMS Decision Analysis Society Student Paper competition. His research interest is in the optimal acquisition of information, with applications in simulation, medicine, and energy. His web address is <www.princeton.edu/˜pfrazier> and his email address is <pfrazier@princeton.edu>.

WARREN B. POWELL is a professor in the department of Operations Research and Financial Engineering at Princeton University, and director of CASTLE Laboratory (<www.castlelab.princeton.edu>). An Informs Fellow, he has coauthored over 100 refereed publications in stochastic optimization, stochastic resource allocation and related applications. He is the author of *Approximate Dynamic Programming*, and is currently involved in applications in energy, transportation, finance and homeland security. His email address is <powell@princeton.edu>.