SIAM J. CONTROL OPTIM.
Vol. 47, No. 5, pp. 2410–2439
© 2008 Society for Industrial and Applied Mathematics

A KNOWLEDGE-GRADIENT POLICY FOR SEQUENTIAL INFORMATION COLLECTION*

PETER I. FRAZIER†, WARREN B. POWELL†, AND SAVAS DAYANIK†

Abstract. In a sequential Bayesian ranking and selection problem with independent normal populations and common known variance, we study a previously introduced measurement policy which we refer to as the knowledge-gradient policy. This policy myopically maximizes the expected increment in the value of information in each time period, where the value is measured according to the terminal utility function. We show that the knowledge-gradient policy is optimal both when the horizon is a single time period and in the limit as the horizon extends to infinity. We show furthermore that, in some special cases, the knowledge-gradient policy is optimal regardless of the length of any given fixed total sampling horizon. We bound the knowledge-gradient policy's suboptimality in the remaining cases, and show through simulations that it performs competitively with or significantly better than other policies.

Key words. ranking and selection, Bayesian statistics, sequential decision analysis

AMS subject classifications. 62F07, 62F15, 62L05

DOI. 10.1137/070693424

1. Introduction.

We consider a ranking and selection problem in which we are faced with $M \geq 2$ alternatives, each of which can be measured sequentially to estimate its constant but unknown underlying average performance. The measurements are noisy, and as we obtain more measurements, our estimates become more accurate. We assume normally distributed measurement noise and independent normal Bayesian priors for each alternative's underlying average performance. We have a budget of N measurements to spread over the M alternatives before deciding which is best. The goal is to choose the alternative with the best underlying average performance.

Information collection problems of this type arise in a number of applications:

(i) Choosing the chemical compound from a library of existing test compounds that has the greatest effectiveness against a particular disease. A compound's effectiveness may be measured by exposing cultured cells infected with the disease to the compound and observing the result. The compound found most effective will be developed into a drug for treating the disease.

(ii) Choosing the most efficient of several alternative assembly line configurations. We may spend a certain short amount of time testing different configurations, but once we put one particular configuration into production, that choice will remain in production for a period of several years.

(iii) Selecting the best of several policies applied to a stochastic Markov decision process. The policies may be evaluated only through Monte Carlo simulation, so a method of ranking and selection is needed to determine which policy is best. This selection may be as part of a larger algorithm for finding the optimal policy as in evolutionary policy iteration [3].

*Received by the editors May 31, 2007; accepted for publication (in revised form) April 29, 2008; published electronically September 8, 2008.

http://www.siam.org/journals/sicon/47-5/69342.html

†Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ 08544 (pfrazier@princeton.edu, powell@princeton.edu, sdayanik@princeton.edu). The second author's research was partially supported by AFOSR contract FA9550-08-1-0195. The third author's research was partially supported by the Center for Dynamic Data Analysis for Homeland Security, ONR Award N00014-07-1-0150.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

In this article we study a measurement policy introduced in [16] under the name of the $(R_1, \ldots, R_1)$ policy, and referred to herein as the knowledge-gradient (KG) policy. We briefly describe this policy and leave further description for section 4.1. Let μ_x^n and $(\sigma_x^n)^2$ denote the mean and variance of the posterior predictive distribution for the unknown value of alternative x after the first n measurements. Then the KG policy is the policy that chooses its $(n+1)$st measurement $X^{KG}((\mu_1^n, \sigma_1^n), \ldots, (\mu_M^n, \sigma_M^n))$ from within $\{1, \ldots, M\}$ to maximize the single-period expected increase in value, $\mathbb{E}_n\left[(\max_{x'} \mu_{x'}^{n+1}) - (\max_{x'} \mu_{x'}^n)\right]$, where $\mathbb{E}_n$ indicates the conditional expectation with respect to what is known after the first n measurements. That is,

$$X^{KG}((\mu_1^n, \sigma_1^n), \ldots, (\mu_M^n, \sigma_M^n)) \in \underset{x^n \in \{1,\ldots,M\}}{\arg\max} \mathbb{E}_n\left[(\max_{x'} \mu_{x'}^{n+1}) - (\max_{x'} \mu_{x'}^n)\right].$$

In this expression the expectation is implicitly a function of x^n, the measurement decision at time n. If the maximum is attained by more than one alternative, then we choose the one with the smallest index. As the terminal reward is given by $\max_{x=1,\ldots,M} \mu_x^N$, this policy is like a gradient ascent algorithm on a utility surface with domain parameterized by the state of knowledge $((\mu_1, \sigma_1), \ldots, (\mu_M, \sigma_M))$. It may also be viewed as a single-step Bayesian look-ahead policy.

In this work we continue the analysis of [16]. We demonstrate that the KG policy, introduced there as the most rudimentary of a collection of potential policies and studied for its simplicity but neglected thereafter, is actually a powerful and efficient tool for ranking and selection that should be considered for application alongside current state-of-the-art policies. As discussed in detail in section 2, a number of other sequential Bayesian look-ahead policies have been derived in recent years by solving a sequence of single-stage optimization problems just as the KG policy does, and, among these, the optimal computing budget allocation for linear loss of [18] and the LL(S) policy of [12] assume situations most similar to the one assumed here. The KG policy differs, however, from these other policies in that it solves its single-stage problem exactly, while the other policies must use approximations. We believe that solving the look-ahead problem exactly offers an advantage.

After formulating the problem in section 3 and defining the policy in section 4, we show in section 5 that the KG policy is optimal in the limit as $N \to \infty$ in the sense that the policy incurs no opportunity cost in the limit as infinitely many measurements are allowed. Also, by its construction and as noted in [16], KG is optimal when there is only one measurement remaining. This provides optimality guarantees at two extremes: N large and N small. While many policies are asymptotically optimal without performing particularly well in the finite sample case, a policy with both kinds of optimality satisfies a more stringent performance check. For example, the equal-allocation policy is asymptotically optimal, but it is not optimal when $N = 1$, except in certain special cases, and performs poorly overall. In the other extreme, myopic policies for generic Markov decision processes often perform poorly because they ignore long-term rewards. By being optimal for both $N = 1$ and $N = \infty$, KG avoids the problem that most afflicts other myopic policies, while retaining single-sample optimality.

In accordance with our belief that optimality at two extremes suggests good performance in the region between, we provide a bound on the policy's suboptimality for finite N in section 6. In section 7 we introduce the KG persistence property and use it to show both optimality for the case when $M = 2$ and for a further special case in which the means and variances are ordered. Our proof that KG is optimal when

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$M = 2$ confirms a claim made by Gupta and Miescke [15], who showed its optimality among deterministic policies for $M = 2$, but did not offer a formal proof for optimality among sequential policies. Finally, in section 8, we demonstrate in numerical experiments that KG performs competitively against the other policies discussed here. In particular, the KG policy is best according to the measure of average performance across a number of randomly generated problems, and the margin by which it outperforms the best competing policies on the most favorable problems is significantly larger than the margin by which it is outperformed on the most unfavorable problems.

2. Literature review. The KG policy was introduced in [16] as the simplest of a collection of look-ahead policies and was studied because its simplicity provided tractability, but this simple policy has seldom been studied or applied in the years since. Instead, a number of more complex Bayesian look-ahead policies have been introduced. A series of researchers beginning with [4] and continuing with [5], [9], [7], [8], [6] proposed and then refined a family of policies known as the optimal computing budget allocation (OCBA). These policies are derived by formulating a static optimization problem in which one chooses the measurements to maximize the probability of later correctly selecting the best alternative. OCBA policies solve this optimization problem by approximating the objective function with various bounds and relaxations, and by assuming that the predictive mean will remain unchanged by measurement. They then solve the approximate problem using gradient ascent or greedy heuristics, or with an asymptotic solution that is exact in the limit as the number of measurements in the second stage is large. All OCBA policies assume normal samples with known sampling variance, but in practice one may estimate this variance through sampling.

Any OCBA policy can be extended to multistage or fully sequential problems by performing the second stage of the two-stage policy repeatedly, at each stage calling all previous measurements the first stage and the set of measurements to be taken next the second stage. It is in this extension that one sees the similarity to the one-step Bayesian look-ahead approach of KG, which extends the one-stage policy which is optimal with one measurement remaining to a sequential policy by supposing at each point in time that the current measurement will be the last.

The OCBA policies mentioned above are designed to maximize the probability of correctly selecting the best alternative, while KG is designed to maximize the expected value of the chosen alternative. These different objective functions are also termed $0-1$ loss and linear loss, respectively. They are similar but not identical, $0-1$ loss perhaps being more appropriate when knowledge of the identity of the best is intrinsically valuable (and where accidentally choosing the second best is nearly as harmful as choosing the worst), and linear loss being more appropriate when value is obtained directly by implementing the chosen alternative.

Recently [18] introduced an OCBA policy designed to minimize expected linear loss. Although more similar to KG than other OCBA policies, it differs in that it uses the Bonferroni inequality to approximate the linear loss objective function for a single stage, and then solves the approximate problem using a second approximation which is accurate in the limit as the second stage is large. This is in contrast to KG, which solves the single-stage problem exactly. The OCBA policy in [18] does not assume, as the other OCBA approaches do, that the posterior predictive mean is equal to the prior predictive mean, and in this regard it is more similar to the approach of [12] discussed below.

A set of Bayesian look-ahead ranking and selection policies distinct from OCBA

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

were introduced in [12]. They differ by not assuming the predictive means equal through time and by allowing the sampling variance to be unknown. This causes the posterior predictive mean to be student-t distributed, inducing an optimization problem governing the second-stage allocation with an objective function that is somewhat different from that in OCBA formulations. This objective function, corresponding to expected loss, is bounded below, and this lower bound is then approximately minimized. The resulting solution minimizes the lower bound exactly in the limit as sampling costs are small, or as the number of second-stage measurements is large.

Six policies are derived in total by considering both $0-1$ and linear loss under three different settings: two-stage measurements with a budget constraint; two-stage without a budget constraint; and sequential. Among these policies, the one most similar to KG is LL(S), which uses linear loss in a sequential setting, allocating τ measurements at a time.

In [10] an unknown-variance version of the KG policy was developed under the name LL_1. The authors compared LL_1 to LL(S) using Monte Carlo simulations and found that LL_1 performed well for a small sampling budget, but degraded in performance as the sampling budget increased. We briefly discuss how these results relate to our own in section 8.

In addition to the Bayesian approaches to sequentially ranking and selecting normal populations described thus far, a substantial amount of progress has been made using a frequentist approach. We do not review this literature in detail, but state only that an overview may be found in [1] and that a more recent policy which performs quite well in the multistage setting with normal rewards is given in [23], [22]. Other sequential and staged policies for independent normal rewards with frequentist guarantees include those in [25], [27], [17], [26], and [24].

Sequential tests also exist which choose measurements based upon confidence bounds for the value Y_x. Such tests include interval estimation [19], which was developed for on-line bandit-style learning in a reinforcement learning setting, and upper confidence bound estimation [3], which was developed for estimating value functions for Markov decision processes. Both tests form frequentist confidence intervals for each Y_x and then select the alternative with the largest upper bound on its confidence interval for measurement. Such policies have general applicability beyond the independent normal setting discussed here.

3. Problem formulation. We state a formal model for our problem, including transition and objective functions. We then formulate the problem as a dynamic program.

3.1. A formal model. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\{1, \ldots, M\}$ be the set of alternatives. For each $x \in \{1, \ldots, M\}$ define a random variable Y_x to be the true underlying value of alternative x. We assume a Bayesian setting for the problem in which we have a multivariate normal prior predictive distribution for the random vector Y, and we further assume that the components of Y are independent under the prior and that $\max_{x=1,\ldots,M} |Y_x|$ is integrable. We will be allotted exactly N measurements, and time will be indexed using n with the first measurement decision made at time 0. At each time $0 \le n < N$, we choose an alternative x^n to measure. Let ε^{n+1} be the measurement error, which we assume is normally distributed with mean 0 and a finite known variance $(\sigma^\varepsilon)^2$ that is the same across all alternatives. We also assume that errors are independent of each other and of the random vector Y. Then define $\hat{y}^{n+1} = Y_x + \varepsilon^{n+1}$ to be the measurement value observed. At time N, we choose an implementation decision x^N based on the measurements recorded, and we

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

receive an implementation reward $\hat{y}^{N+1}$. We assume that the reward is unbiased, so that $\hat{y}^{N+1}$ satisfies $\mathbb{E}\left[\hat{y}^{N+1}|Y, x^N\right] = Y_{x^N}$. Define the filtration $(\mathcal{F}^n)_{n=0}^N$ by letting $\mathcal{F}^n$ be the sigma-algebra generated by $x^0, \hat{y}^1, x^1, \ldots, x^{n-1}, \hat{y}^n$. We will use the notation $\mathbb{E}_n[\cdot]$ to indicate $\mathbb{E}[\,\cdot \mid \mathcal{F}^n]$, the conditional expectation taken with respect to $\mathcal{F}^n$. Measurement and implementation decisions x^n are restricted to be $\mathcal{F}^n$-measurable so that decisions may depend only on measurements observed and decisions made in the past.

Let $\mu^0 := \mathbb{E}[Y]$ and $\Sigma^0 := \mathrm{Cov}[Y]$ be the mean and covariance of the predictive distribution for Y so that Y has prior predictive distribution $\mathcal{N}(\mu^0, \Sigma^0)$ and Σ^0 is a diagonal covariance matrix. Note that our assumed integrability of $\max_x |Y_x|$ is equivalent to assuming integrability of every Y_x because $|Y_{x'}| \leq \max_x |Y_x|$ and $\max_x |Y_x| \leq |Y_1| + \cdots + |Y_M|$, which is equivalent to assuming Σ^0_{xx} finite for every x.

We will use the Bayes rule to form a sequence of posterior predictive distributions for Y from this prior and the successive measurements. Let $\mu^n := \mathbb{E}_n[Y]$ be the mean vector and $\Sigma^n := \mathrm{Cov}[Y \mid \mathcal{F}^n]$ the covariance matrix of the predictive distribution after n measurements have been made. Because the error term ε^{n+1} is independent and normally distributed, the predictive distribution for Y will remain normal with independent components, and Σ^n will be diagonal almost surely. We write $(\sigma^n_x)^2$ to refer to the diagonal component Σ^n_{xx} of the covariance matrix. Then $Y_x \sim \mathcal{N}(\mu^n_x, (\sigma^n_x)^2)$ conditionally on $\mathcal{F}^n$. We will also write $\beta^n_x := (\sigma^n_x)^{-2}$ to refer to the precision of the predictive distribution for Y_x, $\beta^n := (\beta^n_1, \ldots, \beta^n_M)$ to refer to the vector of precisions, and $\beta^\epsilon := (\sigma^\varepsilon)^{-2}$ to refer to the measurement precision. Note that $\sigma^\epsilon < \infty$ implies $\beta^\epsilon > 0$.

Our goal will be to choose the measurement policy $(x^0, \ldots, x^{N-1})$ and implementation decision x^N that maximizes $\mathbb{E}[Y_{x^N}]$. The implementation decision x^N that maximizes $\mathbb{E}_N[Y_{x^N}] = \mu^N_x$ is any element of $\arg\max_x \mu^N_x$, and the value achieved is $\max_x \mu^N_x$. Thus, letting Π be the set of measurement strategies $\pi = (x^0, \ldots, x^{N-1})$ adapted to the filtration, we may write our problem's objective function as

$$\sup_{\pi \in \Pi} \mathbb{E}^\pi \left[\max_x \mu^N_x\right]. \tag{1}$$

3.2. State space and transition function. Our state space is the space of all possible predictive distributions for Y. It can be shown by induction that these are all multivariate normal with independent components. We formally define the state space $\mathbb{S}$ by $\mathbb{S} := \mathbb{R}^M \times (0, \infty]^M$, and it consists of points $s = (\mu, \beta)$ where, for each $x \in \{1, \ldots, M\}$, μ_x and β_x are, respectively, the mean and precision of a normal distribution. We will write $S^n := (\mu^n, \beta^n)$ to refer to the state at time n. The notation S^n will refer to a random variable, while s will refer to a fixed point in the state space.

Fix a time n. We use the Bayes rule to update the predictive distribution of Y_x conditioned on $\mathcal{F}^n$ to reflect the observation $\hat{y}^{n+1} = Y_x + \varepsilon^{n+1}$, obtaining a posterior predictive distribution conditioned on $\mathcal{F}^{n+1}$. Since ε^{n+1} is an independent normal random variable and the family of normal distributions is closed under sampling, the posterior predictive distribution is also normal. Thus our posterior predictive distribution for Y_x is $\mathcal{N}(\mu^{n+1}_x, 1/\beta^{n+1}_x)$, and writing it as a function of the prior and the observation reduces to writing μ^{n+1} and β^{n+1} as functions of μ^n, β^n, and $\hat{y}^{n+1}$. The Bayes rule tells us that these functions are

$$\mu^{n+1}_x = \begin{cases} \left[\beta^n_x \mu^n_x + \beta^\epsilon \hat{y}^{n+1}\right] / \beta^{n+1}_x & \text{if } x^n = x, \\ \mu^n_x & \text{otherwise,} \end{cases} \tag{2}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\beta_x^{n+1} = \begin{cases} \beta_x^n + \beta^\epsilon & \text{if } x^n = x, \\ \beta_x^n & \text{otherwise.} \end{cases} \tag{3}$$

Conditionally on $\mathcal{F}^n$, the random variable μ^{n+1} has a multivariate normal distribution whose mean and variance we can compute. First, we use the tower property of conditional expectation and the definitions of μ^n and μ^{n+1} as the predictive means of Y given $\mathcal{F}^n$ and $\mathcal{F}^{n+1}$, respectively, to write $\mathbb{E}_n\left[\mu^{n+1}\right] = \mathbb{E}_n\left[\mathbb{E}_{n+1}\left[Y\right]\right] = \mathbb{E}_n\left[Y\right] = \mu^n$. Then we compute the variance of μ^{n+1} componentwise. For those alternatives $x \neq x^n$ that we do not measure, our posterior is equal to our prior and $\mu^{n+1} = \mu^n$. This shows that $\mathrm{Var}\left[\mu_x^{n+1} \mid \mathcal{F}^n\right] = 0$ if $x \neq x^n$. For $x = x^n$ this variance is generally positive. Let us define

$$\tilde{\sigma}_x^n := \sqrt{\mathrm{Var}\left[\mu_x^{n+1} \mid \mathcal{F}^n, x^n = x\right]}, \tag{4}$$

so that $(\tilde{\sigma}_x^n)^2$ is equal to $\mathrm{Var}\left[\mu_x^{n+1} \mid \mathcal{F}^n, x^n = x\right]$. This variance may be interpreted as the variance of the *change* in the predictive mean $\mu_x^{n+1} - \mu_x^n$ caused by a measurement as $\mathrm{Var}\left[\mu_x^{n+1} \mid \mathcal{F}^n, x^n = x\right] = \mathrm{Var}\left[\mu_x^{n+1} - \mu_x^n \mid \mathcal{F}^n, x^n = x\right]$. As shown in the following proposition, it is also equal to the reduction in predictive variance, i.e., the reduction in "uncertainty," caused by a measurement.

PROPOSITION 3.1. *For every $x = 1, \ldots, M$, we have $(\tilde{\sigma}_x^n)^2 = (\sigma_x^n)^2 - (\sigma_x^{n+1})^2$.*

Proof. We begin with the relation

$$(\mu_x^{n+1} - Y_x) = (\mu_x^{n+1} - \mu_x^n) + (\mu_x^n - Y_x).$$

Squaring both sides, taking the expectation with respect to $\mathcal{F}^{n+1}$, and noting that $(\sigma_x^{n+1})^2 = \mathbb{E}_{n+1}\left[(Y_x - \mu_x^{n+1})^2\right]$ gives

$$\begin{aligned} (\sigma_x^{n+1})^2 &= \mathbb{E}_{n+1}\left[(\mu_x^n - Y_x)^2\right] \\ &\quad + 2\mathbb{E}_{n+1}\left[(\mu_x^n - Y_x)(\mu_x^{n+1} - \mu_x^n)\right] + \mathbb{E}_{n+1}\left[(\mu_x^{n+1} - \mu_x^n)^2\right] \\ &= \mathbb{E}_{n+1}\left[(\mu_x^n - Y_x)^2\right] + 2(\mu_x^n - \mu_x^{n+1})(\mu_x^{n+1} - \mu_x^n) + (\mu_x^{n+1} - \mu_x^n)^2 \\ &= \mathbb{E}_{n+1}\left[(\mu_x^n - Y_x)^2\right] - (\mu_x^{n+1} - \mu_x^n)^2. \end{aligned}$$

Since $\sigma_x^{n+1} \in \mathcal{F}^n$, we may take the expectation with respect to $\mathcal{F}^n$ to get

$$\begin{aligned} (\sigma_x^{n+1})^2 &= \mathbb{E}_n\left[\mathbb{E}_{n+1}\left[(\mu_x^n - Y_x)^2\right]\right] - \mathbb{E}_n\left[(\mu_x^{n+1} - \mu_x^n)^2\right] \\ &= \mathbb{E}_n\left[(\mu_x^n - Y_x)^2\right] - \mathbb{E}_n\left[(\mu_x^{n+1} - \mu_x^n)^2\right] \\ &= (\sigma_x^n)^2 - (\tilde{\sigma}_x^n)^2. \qquad \square \end{aligned}$$

To more easily compute $\tilde{\sigma}_x^n$, define a function $\tilde{\sigma} : (0, \infty] \mapsto [0, \infty)$ by

$$\tilde{\sigma}(\beta_x) = \sqrt{(\beta_x)^{-1} - (\beta_x + \beta^\epsilon)^{-1}}. \tag{5}$$

Then we have that $\tilde{\sigma}_x^n = \tilde{\sigma}(\beta_x^n)$ by Proposition 3.1 applied to the identities $(\sigma_x^{n+1})^2 = (\beta_x^{n+1})^{-1} = (\beta_x^n + \beta^\epsilon)^{-1}$ and $(\sigma_x^n)^2 = (\beta_x^n)^{-1}$.

Remark 3.1. For $\beta_x \in (0, \infty)$, we have that $(\tilde{\sigma}(\beta_x))^2 = \beta^\epsilon / [(\beta_x + \beta^\epsilon)\beta_x]$ is strictly decreasing in β_x, and thus so is $\tilde{\sigma}(\beta_x)$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Since $\mu_{x^n}^{n+1}$ is a normal random variable with conditional mean $\mu_{x^n}^n$ and conditional variance $(\tilde{\sigma}(\beta_{x^n}^n))^2$ under $\mathcal{F}^n$, we can write in terms of an $\mathcal{F}^n$ adapted sequence $Z^1, \ldots, Z^N$ of standard normal random variables,

$$\mu^{n+1} = \mu^n + \tilde{\sigma}(\beta_{x^n}^n) Z^{n+1} e_{x^n}, \tag{6}$$

$$\beta^{n+1} = \beta^n + \beta^\epsilon e_{x^n}, \tag{7}$$

where e_x is a vector in $\mathbb{R}^M$ with all components zero except for component x, which is equal to 1. We also define a function $T : \mathbb{S} \times \{1, \ldots, M\} \times \mathbb{R} \mapsto \mathbb{S}$ by

$$T((\mu, \beta), x, z) := (\mu + \tilde{\sigma}(\beta_x) z e_x, \beta + \beta^\epsilon e_x), \tag{8}$$

so that $S^{n+1} = T(S^n, x^n, Z^{n+1})$. This is our transition function.

We briefly recall and summarize the random variables which play a role in the measurement process. The underlying and unknown value of alternative x is denoted Y_x and is randomly fixed at the beginning of the measurement process. At time n, μ_x^n is our best estimate of Y_x, and β_x^n is the precision with which we make this estimate. The result of our time n measurement causes us to update this estimate to μ_x^{n+1}, which we now know with precision β_x^{n+1}. This change from μ_x^n to μ_x^{n+1} is random, and furthermore is normally distributed with mean 0 and standard deviation $\tilde{\sigma}(\beta_x^n)$ when we measure alternative x.

One may think of Y_x as fixed and of μ_x^n as converging toward Y_x while β_x^n converges to infinity under some appropriately exploratory sampling strategy. It is also appropriate, however, to fix μ_x^n and β_x^n (this is the essential content of conditioning on $\mathcal{F}^n$) and think of Y_x as an unknown quantity. From this viewpoint, Y_x is random and, furthermore, is normally distributed with predictive mean μ_x^n and precision β_x^n. This randomness does not imply that Y_x must be chosen again according to the predictive normal distribution, but instead the predictive normal distribution only quantifies our uncertain knowledge of the value Y_x adopted when it was first chosen.

3.3. Dynamic program. We apply a dynamic programming approach to our problem. In this approach, the value function is defined as the value of the optimal policy given a particular state S^n at a particular time n, and may also be determined recursively through Bellman's equation. If the value function can be computed efficiently, the optimal policy may then also be computed from it. Although in this problem the "curse of dimensionality" makes direct computation of the value function difficult even for M as small as 3, the dynamic programming principle still provides a valuable method for studying the problem.

The terminal value function $V^N : \mathbb{S} \mapsto \mathbb{R}$ is given by (1) as

$$V^N(s) := \max_{x \in \{1,\ldots,M\}} \mu_x \quad \text{for every } s = (\mu, \beta) \in \mathbb{S}. \tag{9}$$

The dynamic programming principle tells us that the value function at any other time $0 \le n < N$ is given recursively by

$$V^n(s) = \max_{x \in \{1,\ldots,M\}} \mathbb{E}\left[V^{n+1}(T(s, x, Z^{n+1}))\right], \quad s \in \mathbb{S}. \tag{10}$$

We define the Q-factors, $Q^n : \mathbb{S} \times \{1, \ldots, M\} \mapsto \mathbb{R}$, as

$$Q^n(s, x) := \mathbb{E}\left[V^{n+1}(T(s, x, Z^{n+1}))\right], \quad s \in \mathbb{S}, \tag{11}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

and the dynamic programming principle tells us that any policy whose measurement decisions satisfy

$$X^{*n}(s) \in \arg\max_{x\in\{1,\dots,M\}} Q^n(s,x), \quad s \in \mathbb{S}, \tag{12}$$

is optimal. Finally, we define the value of a measurement policy $\pi \in \Pi$ as

$$V^{n,\pi}(s) := \mathbb{E}^{\pi}\left[V^N(S^N) \mid S^n = s\right], \quad s \in \mathbb{S}. \tag{13}$$

This same object might also be thought of as the reward-to-go from state s at time n under policy π.

Later we will need several preliminary results concerning the benefit of measurement. First, the following proposition states that, under the optimal policy, it is always better to make a measurement than to measure nothing at all. Here, the value of measuring alternative x when $S^n = s$ at time n is $Q^n(s,x)$, and the value of making no measurement is $V^{n+1}(s)$. The proof is left until Appendix A.

PROPOSITION 3.2. *$Q^n(s,x) \geq V^{n+1}(s)$ for every $0 \leq n < N$, $s \in \mathbb{S}$, and $x \in \{1,\dots,M\}$.*

We see as a corollary to this proposition that the optimal policy will never measure an alternative with zero variance (i.e., with infinite precision) unless all the other alternatives also have zero variance. In other words, there is no value to measuring something that we know perfectly. This is stated precisely in the following corollary.

COROLLARY 3.1. *Let $i,j \in \{1,\dots,M\}$, $n < N$, and $s = (\mu,\beta) \in \mathbb{S}$. If $\beta_j = \infty$, then $Q^n(s,i) \geq Q^n(s,j)$.*

Proof. Since $\tilde{\sigma}(\beta_j) = \tilde{\sigma}(\infty) = 0$ and $\beta_j + \beta^{\epsilon} = \beta_j$,

$$T(s,j,Z^{n+1}) = (\mu + \tilde{\sigma}(\beta_j)Z^{n+1}e_j, \beta + \beta^{\epsilon}e_j) = (\mu,\beta) = s.$$

Then, by Proposition 3.2,

$$Q^n(s,j) = \mathbb{E}\left[V^{n+1}(T(s,j,Z^{n+1}))\right] = V^{n+1}(s) \leq Q^n(s,i). \qquad \square$$

We also have a second corollary to the proposition. Proposition 3.2 allowed arbitrarily specifying the alternative to which the extra measurement would be applied, while this corollary points out that the extra measurement may be made according to the optimal policy, in which case $Q^n(s,x)$ is equal to $V^n(s)$. We will use this corollary in section 6 to bound the suboptimality of KG.

COROLLARY 3.2. *$V^{n+1}(s) \leq V^n(s)$ for all states $s \in \mathbb{S}$.*

Proof. In Proposition 3.2, take the extra measurement x to be the measurement made by the optimal policy in state s. $\square$

Let us say that a policy π is *stationary* if $X^{\pi,n}(s) = X^{\pi,0}(s)$ for all $s \in \mathbb{S}$ and all $n = 1,\dots,N-1$. In this case we denote $X^{\pi,n}$ simply by X^{π}. Corollary 3.2 showed that the value of the optimal policy increases as more measurements are allowed, and we will see in Theorem 3.1 below that this monotonicity also holds for stationary policies.

THEOREM 3.1. *$V^{\pi,n}(s) \geq V^{\pi,n+1}(s)$ for every stationary policy π and every state $s \in \mathbb{S}$.*

The proof is left until Appendix A. We will need this theorem when showing both asymptotic optimality and bounded suboptimality of KG.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

4. The knowledge-gradient policy. In our problem, the entire reward is received after the final measurement. We may formulate an equivalent problem in which the reward is given in pieces over time, but the total reward given is identical. We define the KG policy as that policy which maximizes the single period reward under this alternate formulation. We will see later that this KG policy is optimal in several cases and has bounded suboptimality in all others. This policy was first introduced in [16] under the name of the $(R_1, \dots, R_1)$ policy.

4.1. Definition. The problem given by (1) has a terminal reward $V^N(S^N) := \max_x \mu_x^N$, but no rewards at any other times. We restructure these rewards by writing $V^N(S^N)$ as a telescoping sequence,

$$\max_x \mu_x^N = \left[V^N(S^N) - V^N(S^{N-1})\right] + \cdots + \left[V^N(S^{n+1}) - V^N(S^n)\right] + V^N(S^n).$$

Thus, the problem that provides single period reward $V^N(S^n)$ at time n and $V^N(S^k) - V^N(S^{k-1})$ at times $k = n+1, \dots, N$ is equivalent to problem (1) because the total reward provided is the same in each case. The KG policy π^{KG} is defined as the policy that chooses its measurements to maximize the expectation of the single period reward provided under this alternate formulation, $\mathbb{E}_n\left[V^N(T(S^n, x, Z^{n+1})) - V^N(S^n)\right]$. Since the $(Z^n)_{n=1}^N$ are independent and identically distributed normal random variables, we may take Z to be a generic standard normal random variable and write the decision function of the KG policy $X^{KG} : \mathbb{S} \mapsto \{1, \dots, M\}$ as

$$X^{KG}(s) \in \operatorname*{arg\,max}_{x \in \{1,\dots,M\}} \mathbb{E}\left[V^N(T(s, x, Z)) - V^N(s)\right] \quad \text{for every } s \in \mathbb{S}, \tag{14}$$

where ties in the arg max are broken by choosing the alternative with the smaller index. Note that KG is stationary in time so we drop the time index n when we write X^{KG}. Since $V^N(s)$ does not depend on x, the KG policy may be rewritten as

$$X^{KG}(s) \in \operatorname*{arg\,max}_{x \in \{1,\dots,M\}} \mathbb{E}\left[V^N(T(s, x, Z))\right] = \operatorname*{arg\,max}_{x \in \{1,\dots,M\}} Q^{N-1}(s, x). \tag{15}$$

Remark 4.1. As noted in [16], KG is optimal by construction when $N = 1$. This is because $V^{N-1} = V^{KG,N-1}$ by (12) and (15), where $V^{KG,n}$ denotes the value of the KG policy at time n and is defined according to (13) with the policy π fixed to KG.

If we think of $V^N(\cdot)$ as a utility function, or as a measure of the amount of "knowledge" contained in a state, we see from (14) that the KG policy chooses its decisions in the direction of steepest expected ascent of this measure. This is the reason behind the name *knowledge-gradient policy*. One may also view it as a single-step look-ahead policy.

4.2. Computation. It was already known in [16] that an exact and computationally tractable expression exists for X^{KG}. We present it here.

For each $x \in \{1, \dots, M\}$ define a function $\zeta_x : \mathbb{S} \mapsto [0, \infty)$ by

$$\zeta_x(s) := -\left|\frac{\mu_x - \max_{x' \neq x} \mu_{x'}}{\tilde{\sigma}(\beta_x)}\right|. \tag{16}$$

Except for the sign, $\zeta_x(S^n)$ is the minimum distance, in terms of the number of standard deviations $\tilde{\sigma}(\beta_x^n)$, that a measurement of alternative x must alter μ_x^{n+1} from its premeasurement value of μ_x^n to make $\arg\max_{x'} \mu_{x'}^{n+1} \neq \arg\max_{x'} \mu_{x'}^n$—that is, to

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

change the identity of the alternative with the largest conditional expected value. In addition, define the function $f : \mathbb{R} \mapsto \mathbb{R}$ as

$$f(z) := z\Phi(z) + \varphi(z), \tag{17}$$

where $\Phi(z)$ is the normal cumulative distribution function and $\varphi(z)$ is the normal probability density function. Then the following theorem provides an efficient way to compute KG's decisions. The proof may be found in Appendix A.

THEOREM 4.1. *For every $s = (\mu, \beta) \in \mathbb{S}$, we have*

$$Q^{N-1}(s, x) = \max_{x'} \mu_{x'} + \tilde{\sigma}(\beta_x) f(\zeta_x(s)), \tag{18}$$

$$X^{KG}(s) \in \arg\max_{x \in \{1,\ldots,M\}} \tilde{\sigma}(\beta_x) f(\zeta_x(s)) \tag{19}$$

with ties broken by choosing the alternative with the smallest index.

The term $Q^{N-1}(s, x) - \max_{x'} \mu_{x'} = \tilde{\sigma}(\beta_x) f(\zeta_x(s))$ is in some sense the expected value of the information that would be obtained by measuring alternative x and is sometimes called the "expected value of information," or EVI, e.g., in [12] and [10].

Computation of the KG policy via (19) scales linearly with the number of alternatives M. This compares well with other policies that might be used on this problem. To compute the KG policy at time n, we must first find the largest and second largest μ_x^n across all alternatives x, which will be used to compute $\zeta_x^n := \zeta_x(S^n)$. This may be implemented either by an initial pass through the alternatives at each time period, or by storing and updating the two values across time periods. Once we have the largest and second largest μ_x^n, we iterate through the alternatives, calculating $\tilde{\sigma}(\beta_x^n) f(\zeta_x^n)$ for each one and returning the alternative with the largest value for this expression. This iteration may be streamlined by recomputing the expression only for those alternatives that changed ζ_x^n or β_x^n from the previous iteration.

The following remark, which is an easily obtained consequence of Theorems 1 and 2 in [16] and may also be obtained directly from (18), may also be used to accelerate the computation of the KG policy by eliminating some alternatives from consideration. It is also useful for proving later results. It states that if an alternative dominates another in both mean and variance, then of the two, KG prefers the dominating alternative.

Remark 4.2. For every $s = (\mu, \beta) \in \mathbb{S}$ such that $\mu_j \geq \mu_i$ and $\beta_j \leq \beta_i$ we have $Q^{N-1}(s, j) \geq Q^{N-1}(s, i)$.

Finally, during computation, we may also use the following remark to eliminate some alternatives from consideration, again improving the speed with which we may compute the KG policy.

Remark 4.3. Take $n = N-1$ in Corollary 3.1. If $\beta_j = \infty$ for some $j \in \{1, \ldots, M\}$ (that is, if the predictive distribution $\mathcal{N}(\mu_j, 1/\beta_j)$ for Y_j is a point mass), then $Q^{N-1}(S, i) \geq Q^{N-1}(S, j)$ for every $i \in \{1, \ldots, M\}$.

Thus, KG will never measure an alternative with zero variance unless every alternative has zero variance. Corollary 3.1 shows that the optimal policy shares this behavior of preferring not to measure any alternative whose true value is known perfectly.

4.3. Behavior. KG balances two considerations when it chooses its measurement decisions. First, it prefers to measure those alternatives about which comparatively little is known. These alternatives x are the ones whose predictive distributions

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

have large variance $(\sigma_x^n)^2$ or, equivalently, have small precision β_x^n. Thus, we have that if KG prefers to measure some alternative i over another alternative j, then it would still prefer to measure alternative i over j if the predictive variance of i were increased.

Second, KG prefers to measure alternatives x with $|\mu_x^n - \max_{x' \neq x} \mu_{x'}^n|$ close to 0. We call $-|\mu_x^n - \max_{x' \neq x} \mu_{x'}^n|$ the *unnormalized influence* and $\zeta_x^n = -|\mu_x^n - \max_{x' \neq x} \mu_{x'}^n| / \tilde{\sigma}(\beta_x^n)$ the *normalized influence*, or simply the *influence*, of alternative x, where $\tilde{\sigma}(\beta_x^n)$ is understood as a normalization term because predictions for different alternatives have different variances and comparison does not make sense unless we standardize these differences. Measurements of alternatives with large influence are more likely to cause a change in the optimal implementation decision; that is, to cause $\arg\max_{x'} \mu_{x'}^n \neq \arg\max_{x'} \mu_{x'}^{n+1}$. KG's preference for small predictive precision and large influence are formalized in Propositions 4.1 and 4.2, but first we calculate the derivative of f, as defined in (17), in a lemma.

LEMMA 4.1. *We have* $f'(z) = \Phi(z) \geq 0$ *for every* $z \in \mathbb{R}$.

Proof. First note that $\frac{d}{dz} e^{-z^2/2} = -z e^{-z^2/2}$, showing that $\varphi'(z) = -z\varphi(z)$. From this we see that f has nonnegative derivative $f'(z) = \Phi(z) + z\varphi(z) - z\varphi(z) = \Phi(z)$, which completes the proof. ∎

PROPOSITION 4.1. *Let states* $s = (\mu, \beta) \in \mathbb{S}$, $s' = (\mu', \beta') \in \mathbb{S}$ *and alternatives* $i, j \in \{1, \ldots, M\}$ *satisfy the following criteria:* $\zeta_i(s') > \zeta_i(s)$, $\zeta_j(s') = \zeta_j(s)$, $\beta_i' < \beta_i$, *and* $\beta_j' = \beta_j$. *If* $Q^{N-1}(s, i) > Q^{N-1}(s, j)$, *then* $Q^{N-1}(s', i) > Q^{N-1}(s', j)$.

Proof. First, $\tilde{\sigma}(\beta_i') \geq \tilde{\sigma}(\beta_i)$ by Remark 3.1 and $f(\zeta_i(s')) \geq f(\zeta_i(s))$ by Lemma 4.1. By (18), $Q^{N-1}(s', i) > Q^{N-1}(s, i)$. Also, the equalities $\tilde{\sigma}(\beta_j') = \tilde{\sigma}(\beta_j)$ and $f(\zeta_j(s')) = f(\zeta_j(s))$ imply through (18) that $Q^{N-1}(s', j) = Q^{N-1}(s, j)$. Thus, if $Q^{N-1}(s, i) > Q^{N-1}(s, j)$, then $Q^{N-1}(s', i) \geq Q^{N-1}(s, i) > Q^{N-1}(s, j) = Q^{N-1}(s', j)$. ∎

PROPOSITION 4.2. *If alternative* i *and state* $s = (\mu, \beta)$ *are such that* $\zeta_i(s) \geq \zeta_j(s)$ *and* $\beta_i < \beta_j$ *for every alternative* $j \neq i$, *then* $X^{KG}(s) = i$.

Proof. Let j be an alternative different from i. Then $\tilde{\sigma}(\beta_i) > \tilde{\sigma}(\beta_j)$ by Remark 3.1 and $f(\zeta_i(s)) \geq f(\zeta_j(s))$ by Lemma 4.1. This implies that $Q^{N-1}(s, i) > Q^{N-1}(s, j)$ by Proposition 4.1. Since this is true for all $j \neq i$, we have that $i = \arg\max_j Q^{N-1}(s, j) = X^{KG}(s)$ where the arg max is unique. ∎

It is also interesting to note that increasing the predictive mean of a single alternative usually, but not universally, encourages KG to measure it. Thus, having a large predictive mean is similar, but not identical, to having a large unnormalized influence. We formalize this in the following proposition.

PROPOSITION 4.3. *If KG prefers alternative* i *in state* (μ, β), *then it also prefers the same alternative* i *in state* $(\mu + ae_i, \beta)$ *for all positive real numbers* a *such that* $\mu_i + a \leq \max_x \mu_x$, *i.e., for* $0 \leq a \leq -\mu_i + \max_x \mu_x$.

We leave the proof until Appendix A.

5. Asymptotic optimality. In this section we show that the KG policy is asymptotically optimal in the limit as the number of measurements N grows large. This means that, given the opportunity to measure infinitely often, KG will discover which alternative is best. In some sense, this is a convergence result because it shows that the policy's estimate of which alternative is best will converge to the alternative that is truly best.

The KG policy is not alone in possessing this property. Indeed, the following well-known policies are all asymptotically optimal: the equal-allocation policy which distributes its measurements in a round-robin fashion equally among the alternatives; the uniform exploration policy which randomly chooses its measurements with equal

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

probability across the alternatives; and the Boltzmann exploration policy discussed in section 8 which randomly chooses its measurements according to exponentially weighted probabilities.

These policies differ from KG in that they explore for exploration's sake and for the long-term benefit it provides, while KG is purely myopic. Moreover, we argue that KG's asymptotic optimality is notable exactly because the policy is entirely myopic, maximizing its single-period expected reward without regard for the long-term. This is not generally the case with myopic policies for other problems. That a myopic policy is also optimal in the long-term shows that this ranking and selection problem has a special structure, and it foreshadows what is further suggested by our numerical experiments: that this myopic policy, KG, performs quite well in many cases which are neither myopic nor asymptotic.

In addition, one policy, interval estimation, performs very well in our numerical experiments but is not asymptotically optimal as in some cases it "sticks," measuring one alternative only and obtaining its true value perfectly without learning about the others [19]. Indeed, one can construct cases in which this policy's performance is arbitrarily bad compared to any asymptotically optimal policy. Although a policy's asymptotic optimality is not evidence of quality by itself, its absence should raise concern among those who might use a policy lacking it. Finally, a natural question is whether other policies, such as those in the OCBA family and those proposed in [12], are asymptotically optimal. This question is currently open as these other policies are more complex and require more care during analysis than does KG. Nevertheless, we believe that the proof techniques applied here may be extended to show that many other Bayesian look-ahead policies are also asymptotically optimal.

To show that KG is asymptotically optimal, we begin by showing in Proposition 5.1 that the asymptotic value of a policy is well defined and bounded above by the value $\mathbb{E}\max_x Y_x$ of learning every alternative exactly. Then we show in Proposition 5.2 that this value is achieved by any stationary policy that measures every alternative infinitely often. Thus, any stationary policy that samples every alternative infinitely often is asymptotically optimal. Finally, we show in Theorem 5.1 that KG is asymptotically optimal. The proof centers on the notion that, as the number of times an alternative is measured increases, the variance of the value of that alternative shrinks toward 0. Eventually, that variance will be so low that KG will prefer to measure another alternative. This argument is used to show that KG samples every alternative infinitely often and thus is asymptotically optimal.

Since we will be varying the number N of measurements allowed, we use the notation $V^0(\,\cdot\,;N)$ to denote the value function at time 0 when the problem's terminal time is N. We then define the *asymptotic value function* $V(\,\cdot\,;\infty)$ by the limit $V(s;\infty) := \lim_{N\to\infty} V^0(s;N)$ for $s \in \mathbb{S}$. Similarly, we denote the *asymptotic value function for stationary policy* π by $V^\pi(\,\cdot\,;\infty)$ and define it by $V^\pi(s;\infty) := \lim_{N\to\infty} V^{\pi,0}(s;N)$ for $s \in \mathbb{S}$. Proposition 5.1 shows that both limits exist.

If $V^\pi(s;\infty)$ is equal to $V(s;\infty)$ for every $s \in \mathbb{S}$, then π is said to be *asymptotically optimal*. In particular, if a stationary policy π achieves the upper bound $U(\,\cdot\,)$ on $V(\,\cdot\,;\infty)$ shown in Proposition 5.1, then π must be asymptotically optimal. We will use this later to show that KG is asymptotically optimal. The proof of Proposition 5.1 is found in Appendix A.

PROPOSITION 5.1. *Let $s \in \mathbb{S}$. Then the limit $V(s;\infty)$ exists and is bounded above by*

$$U(s) := \mathbb{E}\left[\max_x Y_x \mid S^0 = s\right] < \infty, \tag{20}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

where we recall that $\{Y_x\}_{x\in\{1,\dots,M\}}$ *are independent and* $Y_x \sim \mathcal{N}(\mu_x^0, (\beta_x^0)^{-1})$. *Furthermore,* $V^\pi(s;\infty)$ *exists and is finite for every stationary policy* π.

For any finite terminal time N we define the random variable η_x^N as the number of times that alternative x is measured up to but not including the terminal time N. We also define η_x^∞ as the limit of the η_x^N; namely,

$$\eta_x^N := \sum_{k=1}^{N} 1_{\{x^k=x\}} \quad \text{and} \quad \eta_x^\infty := \lim_{N\to\infty} \eta_x^N.$$

The limit η_x^∞ exists because η_x^N is nondecreasing in N a.s. Note that we allow the limit η_x^∞ to be infinite.

Proposition 5.2 formalizes the idea that if we measure every alternative infinitely often, then we eventually learn the true value of every alternative. This implies asymptotic optimality. We then use Proposition 5.2 in the proof of Theorem 5.1 to show that KG is asymptotically optimal. The proofs for both Theorem 5.1 and Proposition 5.2 are found in Appendix A.

PROPOSITION 5.2. *If* π *is a stationary policy under which* $\eta_x^\infty = \infty$ *a.s. for every* x, *then* π *is asymptotically optimal.*

THEOREM 5.1. *The KG policy is asymptotically optimal and has value* $U(S^0)$.

6. Bound on suboptimality. We have shown that KG is optimal when $N = 1$ and in the limit as $N \to \infty$. In this section we address the range of N between these extremes by bounding KG's suboptimality in this region. This bound will be tight for small N and will grow as N increases.

We begin with a theorem that implies our bound as a corollary. This theorem shows that there is a limit on how much we may learn through any single measurement.

THEOREM 6.1. *Let* $s = (\mu, \beta) \in \mathbb{S}$ *and* $c = (2\pi)^{-1/2} \max_x \tilde{\sigma}(\beta_x)$. *Then*

$$V^n(s) \le V^{N-1}(s) + c(N - n - 1).$$

The proof is found in Appendix A. We combine this result with Theorem 3.1 to bound KG's suboptimality. Here, $V^{KG,n}(s)$ is the value of the KG policy at time n when $S^n = s$.

COROLLARY 6.1. *Let* $s = (\mu, \beta) \in \mathbb{S}$ *and* $c = (2\pi)^{-1/2} \max_x \tilde{\sigma}(\beta_x)$. *Then*

$$V^n(s) - V^{KG,n}(s) \le c(N - n - 1).$$

Proof. By Remark 4.1, we have $V^{N-1}(s) = V^{KG,N-1}(s)$. From Theorem 3.1 we have $V^{KG,N-1}(s) \le V^{KG,n}(s)$. Substituting the inequality $V^{N-1}(s) \le V^{KG,n}(s)$ into Theorem 6.1 shows the corollary. □

7. Optimality for finite horizon special cases. We saw in Remark 4.1 that KG is optimal when $N = 1$. We will show that KG is optimal in two other special cases: first, when there are only two alternatives to measure; second, when the measurements are free from noise, $(\sigma^\varepsilon)^2 = 0$, and when the parameters of the time 0 prior can be ordered by $\mu_1^0 \ge \mu_2^0 \ge \cdots \ge \mu_M^0$ and $\sigma_{11}^0 \ge \sigma_{22}^0 \ge \cdots \ge \sigma_{MM}^0$. Before showing optimality under these conditions, we first define and discuss a property called the KG persistence property. This property is useful because it provides a sufficient condition for optimality.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

7.1. Persistence of the knowledge-gradient policy. Proofs of the optimality of the KG policy in these special cases is based on the KG persistence property. A problem setting is said to have the KG persistence property if, operating the problem under some policy other than KG, an alternative preferred by KG will remain preferred until the alternative is measured. Below, in Theorem 7.1, we show that if a problem setting has the KG persistence property, then KG is optimal in that problem setting. Before stating this theorem, we formally define the KG persistence property and an associated term, "covering of the future."

DEFINITION 7.1. *A sequence of subsets of* $\mathbb{S}$, $\{\mathbb{S}^n\}_{n=k}^N$, *is called a covering of the future from* k *if* $T(s,x,Z^{n+1}) \in \mathbb{S}^{n+1}$ *a.s. for every* $s \in \mathbb{S}^n$, $x \in \{1,\dots,M\}$, *and* $n \in \{k,\dots,N-1\}$.

DEFINITION 7.2. *We say that the KG persistence property holds on a covering* $\{\mathbb{S}^n\}_{n=k}^N$ *of the future from* k *if* $X^{KG}(T(s,x,Z^{n+1})) = X^{KG}(s)$ *a.s. for every* $s \in \mathbb{S}^n$, $x \neq X^{KG}(s)$, *and* $n \in \{k,\dots,N-1\}$.

This KG persistence property gives us a sufficient condition for the optimality of the KG policy, as stated in the following theorem.

THEOREM 7.1. *If the KG persistence property holds on a covering* $\{\mathbb{S}^n\}_{n=k}^N$ *of the future from* k *for some* $k \in \{0\dots N-1\}$, *then* $V^{KG,k}(s) = V^k(s)$ *for every* $s \in \mathbb{S}^k$.

We leave the proof until Appendix A, but we give a sketch here. Consider a time $n < N-1$ and the alternative that KG prefers. If the problem setting has the KG persistence property, then, even if we do not measure that alternative now, KG will continue to prefer it until we reach the final measurement $N-1$. At this measurement, KG is optimal by construction and so it is now provably optimal to measure this persistent alternative. Thus, there exists an optimal policy that measures the persistent alternative a.s., and by the temporal symmetry in the model, there exists an optimal policy that measures the persistent alternative immediately at time n. This argument is used with induction to show that there exists an optimal policy making the same measurements as KG.

7.2. Optimality for two alternatives. We use the KG persistence principle to show that KG is optimal when there are exactly two alternatives to consider, i.e., $M = 2$. In this case we will see that the optimal policy is one that, at each decision point, measures the alternative with the largest variance. This policy is actually deterministic, and it was shown in [15] that this policy is optimal among the class of deterministic policies. Theorem 7.2 extends this result to show that this same policy is also optimal among the class of fully sequential policies. It is not generally true that the best deterministic policy is also as good as or better than every sequential policy, but Theorem 7.2 shows that this is exactly the case for this particular problem.

We will see that the policy of measuring the alternative with the largest variance is optimal because knowing the correct implementation decision is the same as knowing the true sign of $Y_1 - Y_2$. Each measurement measures only one of Y_1 or Y_2, and an equal reduction in variance for Y_1 or Y_2 contributes equally to the overall reduction in variance of $Y_1 - Y_2$, regardless of which expected value is bigger. Thus, the best way to learn about the difference between points $Y_1 - Y_2$ is to measure that point about which the least is known.

To show that KG is optimal when $M = 2$, we need to show that KG persistence holds when $M = 2$ and then refer to Theorem 7.1.

LEMMA 7.1. *If* $M = 2$, *then* $X^{KG}(s) \in \arg\min_x \beta_x$ *for each* $s = (\mu,\beta) \in \mathbb{S}$ *with ties broken by choosing the alternative with the smaller index.*

Proof. By (19) from Theorem 4.1, it is enough to show equality between the sets

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$\arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(s))$ and $\arg\min_x \beta_x$. When $M = 2$, $\zeta_x(s) = -|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)$, so $\arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(s)) = \arg\max_x \tilde{\sigma}(\beta_x) f\left(-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)\right)$. The function $\tilde{\sigma}$ is strictly decreasing by Remark 3.1. This fact will be used on its own, and it also implies that $-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)$ is a decreasing function of β_x. The function f is nondecreasing by Lemma 4.1, so the function $\beta_x \mapsto f\left(-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)\right)$ is the composition of a nondecreasing function with a nonincreasing function and is thus itself nonincreasing. Thus, the function $\beta_x \mapsto \tilde{\sigma}(\beta_x) f\left(-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)\right)$ is the product of a strictly decreasing function with a nonincreasing function and is thus itself strictly decreasing. This implies that $\arg\max_x \tilde{\sigma}(\beta_x) f\left(-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)\right) = \arg\min_x \beta_x$. ∎

THEOREM 7.2. *If $M = 2$, then KG is optimal.*

Proof. Let $\mathbb{S}^n = \mathbb{S}$ for all n, and note that $\{\mathbb{S}^n\}_{n=0}^N$ is a covering of the future from 0. We will show that the KG persistence property holds on $\{\mathbb{S}^n\}_{n=0}^N$.

Let $n \in \{0, \ldots, N-1\}$ and $s = (\mu, \beta) \in \mathbb{S}$. First consider the case when $\beta_1 \le \beta_2$. By Lemma 7.1, $X^{KG}(s) = 1$. The precision component of $T(s, 2, Z^{n+1})$ is $(\beta_1, \beta_2 + \beta^\epsilon)$. Since $\beta_1 \le \beta_2 \le \beta_2 + \beta^\epsilon$ and by Lemma 7.1, $X^{KG}(T(s, 2, Z^{n+1})) = 1$ a.s.

Now consider the case when $\beta_1 > \beta_2$. By Lemma 7.1, $X^{KG}(s) = 2$. The precision component of $T(s, 1, Z^{n+1})$ is $(\beta_1 + \beta^\epsilon, \beta_2)$. Since $\beta_1 + \beta^\epsilon \ge \beta_1 > \beta_2$ and by Lemma 7.1, $X^{KG}(T(s, 1, Z^{n+1})) = 2$ a.s.

In both cases, $x \ne X^{KG}(s)$ implies $X^{KG}(T(s, x, Z^{n+1})) = X^{KG}(s)$ a.s., so KG persistence holds. Then, by Theorem 7.1, $V^{KG,0}(s) = V^0(s)$ for every $s \in \mathbb{S}$, and KG is optimal. ∎

This theorem is founded on the intuition that the policy that learns the most is also the one that changes our beliefs the most. This has a comparison in other measurement problems—for example, the problem in which we have a quadratic function with known second derivative and we measure the first derivative to find the maximum of the function. In this case the optimal policy is also the one that maximizes the variance of the change in our final belief with respect to our current belief. In both cases we measure the change between our current and final beliefs by taking the variance. In other problems the variance is likely not the right measure of change, but the same intuition would apply with some other measure of change.

7.3. Optimality when the state space is ordered. The KG policy is also optimal when there is no measurement noise, i.e., $(\sigma^\varepsilon)^2 = 0$, and when the components of S^0 may be ordered in such a way that we have $\mu_1^0 \ge \cdots \ge \mu_M^0$ together with $\beta_1^0 \le \cdots \le \beta_M^0$. In other words, the optimality result requires that we may order the alternatives with increasing means while simultaneously ordering them with increasing variances. With the assumption of no measurement noise, the problem is interesting only if the number of alternatives M is larger than the measurement budget N.

We present this optimality result formally in the theorem below, but first, as these conditions are particularly restrictive, we motivate them with an example. Consider a problem in marketing research in which we have a collection of potential advertising campaigns, some of which are more ambitious than others. The predictive distributions for the value obtained from the ambitious campaigns have larger mean but larger variance as well. We may test a few of these campaigns in test markets before committing to one of them. We will assume that the number of test markets allowed is smaller than the number of potential campaigns. If we are willing to make two additional assumptions—that loss is linear and that test markets give us perfect knowledge of the campaign's true value—then the example meets the conditions of the theorem. These additional assumptions would not be met perfectly satisfied in reality, but it is not too unreasonable to imagine situations in which loss would be

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

approximately linear, and in which the knowledge obtained from a test market would be large enough that one would not wish to perform a second test market. With this marketing application as an illustrative example, we expect that this sort of ordering of means and variances may also occur in financial applications, or wherever greater expected reward brings greater risk along with it.

THEOREM 7.3. *If* $(\sigma^\varepsilon)^2 = 0$ *and* $s = (\mu, \beta) \in \mathbb{S}$ *is such that the implication*

$$(\beta_i \neq \infty \text{ and } \beta_j \neq \infty \text{ and } \beta_i < \beta_j) \implies \mu_i \geq \mu_j$$

holds for every $i, j \in \{1, \ldots, M\}$, *then* $V^0(s) = V^{KG,0}(s)$.

The full proof can be found in Appendix A, but the essential idea is that when this ordering holds, the tension between exploration and exploitation is gone, and KG will simply choose that alternative with the largest variance. This is because the alternative with the largest variance is also the alternative with the largest mean among those which are not yet perfectly known. This ordering by variances is persistent, as it was in the $M = 2$ case. Thus, the KG persistence property holds and KG is optimal.

8. Computational experiments. We compared KG against other sampling policies using Monte Carlo simulation on 100 randomly generated problems and found that it performs competitively. In particular, KG performed best when measured by average performance across all the problems, and the margin by which it outperformed the best competing policies in favorable cases was significantly larger than the margin by which it was outperformed in unfavorable cases. Its comparative performance was particularly good when the measurement budget was not much larger than the number of alternatives to measure, and we would argue that performing well in these cases is particularly important.

The space of problems is parameterized by a number of measurements N, a number of alternatives M, an initial precision $\beta^0 \in (0, \infty]^M$, an initial mean $\mu^0 \in \mathbb{R}^M$, and a measurement noise $(\sigma^\varepsilon)^2 \in [0, \infty)$. We chose a collection of 100 problems randomly generated within this space according to the following distribution: M was integer-valued between 2 and 100. N was chosen by first choosing M and then choosing a ratio N/M uniformly from the set $\{1, 3, 10\}$. Each μ_x was uniformly distributed in the interval $[-1, 1]$, and each β_x was independently chosen as 1 with probability .9 and 1000 with probability .1. The noise variance $(\sigma^\varepsilon)^2$ was set to 1 in all cases.

For each problem, we performed simulations in which true function values were generated independently according to the prior. Rather than collecting the value obtained by the policy in each simulation, we collected the opportunity cost realized, where the opportunity cost is the difference in true value between the best option and the option chosen by the policy. The difference in expected opportunity cost is the same as the difference in policy value, but samples of opportunity cost have less error, and this allowed us to obtain accurate estimates with fewer simulations. We ran 10^5 simulations for each policy.

We compared KG against seven other policies: the OCBA for linear loss of [18], the LL(S) policy of [12], the interval estimation (IE) policy of [19], Boltzmann exploration (see, e.g., [28]), equal allocation, and exploitation. Several of these policies required choosing one or more parameters, which we did by simulating several choices on all 100 problems and taking the parameters whose resulting opportunity cost was smallest when summed over all 100 problems. We briefly describe each policy and its tuning.

- *OCBA.* This policy has three parameters: the number of alternatives to allocate to in each stage, m; the number of measurements to allocate to each

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

alternative in the first stage, n_0; and the number of measurements per chosen alternative to allocate in each stage, τ. We set n_0 to 0 because our prior is informative and thus may be thought of as already providing the results of a first stage. To calibrate m and τ, we ran initial experiments with 5000 samples each with settings of $m = 1, \tau \in \{1, 2, 5, 10\}$, and also with $\tau = 1$, $m \in \{2, 5, 10\}$. We found that $m = 1, \tau = 1$ performed best.

- *LL(S) for known variance.* The LL(S) policy allows normal measurement errors with *unknown* variance and uses a normal-gamma prior for the unknown mean and measurement precision. We adapted this policy to the known-variance case by taking the limit as the gamma prior on the precision becomes a point mass at the known variance. Details can be found in Appendix B. The policy has two parameters, n_0 and τ. We set n_0 to 0 as we did with OCBA. We tested the values $1, 2, 3, 4, 5, 10$ for τ on our collection of 100 problems with 5000 samples for each problem and found that $\tau = 1$ worked best for every problem. This is the value we used in comparison with KG.
- *Interval estimation.* IE is parameterized by $z_{\alpha/2}$. As [19] suggests that values of 2, 2.5, or 3 often work best for $z_{\alpha/2}$, we tested values between 2 and 4 in increments of .1 and found that $z_{\alpha/2} = 3.1$ worked best. Although we found IE worked very well when properly tuned, we also found it to be very sensitive to the choice of tuning parameter.
- *Boltzmann exploration.* Boltzmann exploration chooses its measurements by $\mathbb{P}\{x^n = x \mid \mathcal{F}^n\} = \frac{\exp(\mu_x^n/T^n)}{\sum_{x'=1}^M \exp\left(\mu_{x'}^n/T^n\right)}$, where the policy is parameterized by a decreasing sequence of "temperature" coefficients $(T^n)_{n=0}^{N-1}$. We tuned this temperature sequence within the set of exponentially decreasing sequences defined by $T^{n+1} = \gamma T^n$ for some constant $\gamma \in (0, 1]$. The set of all such sequences is parameterized by γ and T^N. We tested $\gamma \in \{.1, .5, .8, .9, 1\}$ with $T^N \in \{.1, 1, 10\}$ and found that $\gamma = 1$ performed best. We then tested the set of possible T^N between .1 and 10 with γ fixed to 1 and found that $T^N = .55$ performed best.
- *Equal allocation.* The equal-allocation policy is $x^n \in \arg\min_x \beta_x^n$, since we think of the prior as providing the results of some previous first-stage measurements, and we interpret β_x^n/β^ϵ as the number of measurements of alternative x taken by time n. It requires no tuning.
- *Exploitation.* The exploitation policy is $x^n \in \arg\max_x \mu_x$. It requires no tuning.

The work required to tune other policies highlights one practical advantage of KG policy: it requires no tuning.

8.1. Results. On each of the 100 randomly generated problems, we took 10^5 samples of opportunity cost from every policy. The distribution of opportunity cost is not normal, as it is positive a.s. and often equal to 0. We averaged groups of 500 samples to obtain approximately normal samples from which we estimated expected opportunity cost as well as standard errors on these estimates. The difference in value between KG and any other policy on any particular problem was then estimated as the difference in sampled opportunity costs, with standard error equal to the square root of the sum of the squared standard errors. The resulting standard errors of the difference, reporting maximum and averaged values across the 100 problems, were .0018 and .0007 for IE; .0018 and .0007 for OCBA; .0019 and .0007 for LL(S); .0020 and .0009 for Boltzmann exploration; .0024 and .0013 for equal allocation; and .0026 and .0021 for exploitation.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

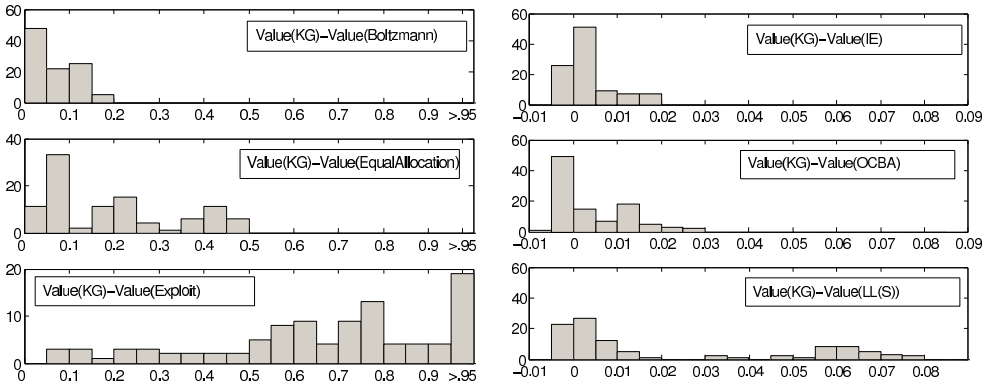

FIG. 1. *Histogram of the sampled difference in value for competing policies aggregated across the* 100 *randomly generated problems.*

We show in Figure 1 the sample estimates of $V^{KG} - V^{\pi}$ aggregated across the randomly generated problems for each of the competing policies π. Bars to the right of 0 indicate that KG outperformed the plotted policy on those problems, and bars to the left indicate the converse. Note that the scale of the histograms in the right-hand plots is much smaller than in the left-hand plots. The histograms show that Boltzmann exploration, equal allocation, and exploitation policies were all outperformed by KG in every problem setting tested, while IE, OCBA for linear loss, and the LL(S) policy performed relatively better. Each of these three better competing policies performed better than KG on some problems and were outperformed on others; however, the tail to the right of 0 is larger than to the left. This indicates that the amount by which KG outperformed the competing policies was significantly larger than the amount by which it was outperformed.

We note a seeming discrepancy between our numerical work and that of Chick, Branke, and Schmidt [10], who tested a variance-unknown version of the KG policy called LL_1. They found that LL_1 performed well in small-sample settings, but poorly elsewhere. In contrast, we found that KG, a very similar policy, performed quite well overall. We believe that the difference lies in the stopping rule used. We simply stopped our sampling policies after a fixed horizon N, but [10] drew many of its conclusions from experiments using the expected opportunity cost Bonferonni (EOC Bonf) stopping rule introduced in [2]. In experiments not pictured here we found that KG also performed poorly with EOC Bonf stopping, but much better when it was stopped using a stopping rule that we introduce now.

This new rule stops as soon as the expected myopic value of the next measurement, as determined by $Q^{N-1}(s,x) - \max_{x'} \mu_{x'} = \tilde{\sigma}(\beta_x) f(\zeta_x(s))$, drops below a threshold c. That is, the number of measurements N to take under this rule is defined by $N = \inf\{n \geq 0 : \tilde{\sigma}(\beta_x^n) f(\zeta_x(S^n)) < c\}$. The threshold c should be interpreted as the cost of one measurement. Since the expected marginal value of each subsequent measurement decreases on average, it is reasonable to stop measuring as soon as the marginal expected value of the next measurement drops below its cost. Replacing EOC Bonf with this new stopping rule may improve the performance of the KG sampling policy enough to make it competitive with LL(S) and other commonly used policies in an adaptive stopping setting. Our initial experiments suggest that this may be the case, but space limitations prevent a thorough discussion of the experimental issues.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

9. Conclusion. The KG measurement policy, as first proposed in [16] and as analyzed here, has several attractive features. Under the assumption of independent normally distributed priors with normal sampling errors of common known variance, we showed that the policy is optimal in both extremes of the number of measurements allowed ($N = 1$ and $N \to \infty$), as well as in other special cases, and has bounded suboptimality in the remaining cases. We showed numerically that it performs competitively with, or significantly better than, several other sequential measurement policies in a broad class of problem settings. In addition, KG is simple in concept, easy to implement, fast to compute, and requires no tuning. This simplicity may make it an attractive alternative to its more complex but similarly performing cousins, the OCBA and the LL(S) policy.

One important limitation of the version of the policy discussed herein is its assumption of common known variance, which often fails to be met in practice. To lift this assumption, it is possible to place a normal-gamma prior on the unknown means and variances, as was done in [12], and recompute the optimal single-step look ahead policy. Indeed, if we begin with a noninformative normal-gamma prior for the true mean Y_x and unknown sampling variance β_x^ϵ of alternative x, and after sampling have vectors of statistics $(\mu, \hat{\sigma}^2, n)$ where $(\mu_x, \hat{\sigma}_x^2, n_x)$ indicate the sample mean, sample variance, and number of samples taken for alternative x, then a calculation similar to that of Theorem 4.1 reveals that the corresponding KG policy is $\arg\max_x \tilde{\sigma}_x f_{n_x-1}(\zeta_x)$, where we must redefine $\tilde{\sigma}_x := \sqrt{\hat{\sigma}^2/n_x(n_x+1)}$, leave ζ_x defined as before, and define $f_n(z) := \frac{\nu+z^2}{\nu-1}\varphi_\nu(z) + z\Phi_\nu(z)$, where φ_ν and Φ_ν are, respectively, the probability density function and cumulative density function of the student-t distribution with ν degrees of freedom. This provides a version of KG for the unknown-variance case. This was derived earlier and independently in [10], and is discussed there in much greater detail, together with a numerical analysis of its properties.

Additionally, the KG policy as described herein has used a fixed number of samples instead of an adaptive stopping rule, while [2] has shown that such rules generally improve the efficiency of budgeted ranking and selection policies. Nevertheless, as implied briefly in section 8 and as discussed in [10], one can certainly use an adaptive stopping rule with the KG sampling policy. Future work is needed to assess the quality of such adaptively stopped policies, and to determine which stopping rules are best to use with KG, but this is by no means an insurmountable obstacle.

Other limitations would seem to present more difficulty. The use of common random numbers has proved immensely beneficial for simulation-based ranking and selection. References [11] and [14] discuss Bayesian ranking and selection policies taking advantage of common random numbers, as does [21] for the frequentist formulation, and it may be possible to extend the KG approach along these lines as well. Indeed, KG's benefits may be overshadowed by its inability to leverage common random numbers in simulation-based ranking and selection unless this extension can be made. In addition, KG assumes the alternatives have a common measurement cost, while in practice it may be more expensive or time consuming to measure some alternatives than others. It may be possible to lift this restriction by dividing the benefit of measurement by the cost so as to obtain a normalized quantity for comparison (a benefit per unit cost), but it may also be that the OCBA approach is more appropriate in such instances.

Despite these limitations, KG has great potential for application. As demonstrated here, it should be considered a reasonable alternative to other measurement policies for those applications that meet its assumptions of a fixed sampling budget

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

and normally distributed errors with common known variance.

Appendix A. Proofs.

Proof of Proposition 3.2. We proceed by induction on n. For $n = N-1$ and $s = (\mu, \beta)$ we have

$$Q^{N-1}(s,x) = \mathbb{E}\left[V^N(T(s,x,Z^N))\right] = \mathbb{E}\left[\left(\mu_x + \tilde{\sigma}(\beta_x)Z^N\right) \vee \max_{x' \neq x} \mu_{x'}\right]$$

$$\geq \mu_x \vee \max_{x' \neq x} \mu_{x'} = V^N(s),$$

where the inequality is justified by Jensen's inequality and the convexity of the max operator. Now we prove the induction step. For $0 \leq n < N$,

$$Q^n(s,x) = \mathbb{E}\left[V^{n+1}(T(s,x,Z^{n+1}))\right] = \mathbb{E}\left[\max_{x' \in \{1,\ldots,M\}} Q^{n+1}(T(s,x,Z^{n+1}),x')\right]$$

$$\geq \max_{x' \in \{1,\ldots,M\}} \mathbb{E}\left[Q^{n+1}(T(s,x,Z^{n+1}),x')\right]$$

$$= \max_{x' \in \{1,\ldots,M\}} \mathbb{E}\left[V^{n+2}(T(T(s,x,Z^{n+1}),x',Z^{n+2}))\right]. \tag{21}$$

In this equation both decisions x and x' are fixed, so the state to which we arrive when we measure x first and x' second, $T(T(s,x,Z^{n+1}),x',Z^{n+2})$, is equal in distribution to the state to which we arrive when we measure x' first and x second, $T(T(s,x',Z^{n+2}),x,Z^{n+1})$. This allows us to exchange the time-order of the decisions x and x' in (21) to write

$$Q^n(s,x) \geq \max_{x' \in \{1,\ldots,M\}} \mathbb{E}\left[V^{n+2}(T(T(s,x',Z^{n+2}),x,Z^{n+1}))\right]$$

$$= \max_{x' \in \{1,\ldots,M\}} \mathbb{E}\left[\mathbb{E}\left[V^{n+2}(T(T(s,x',Z^{n+2}),x,Z^{n+1})) \mid Z^{n+2}\right]\right]$$

$$= \max_{x' \in \{1,\ldots,M\}} \mathbb{E}\left[Q^{n+1}(T(s,x',Z^{n+2}),x)\right].$$

Then the induction hypothesis tells us that

$$Q^{n+1}(T(s,x',Z^{n+2}),x) \geq V^{n+2}(T(s,x',Z^{n+2})) \text{ a.s.},$$

allowing us to write

$$Q^n(s,x) \geq \max_{x' \in \{1,\ldots,M\}} \mathbb{E}\left[V^{n+2}(T(s,x',Z^{n+2}))\right] = \max_{x' \in \{1,\ldots,M\}} Q^{n+1}(s,x') = V^{n+1}(s).$$

Proof of Theorem 3.1. We proceed by induction on n. Consider the base case, which is $n = N-1$. Fix $s = (\mu, \beta) \in \mathbb{S}$. Then $V^N(s) = \max_x \mu_x$ is convex in its arguments, so we can employ Jensen's inequality to write

$$V^{\pi,N-1}(s) = \mathbb{E}\left[V^{\pi,N}(T(s,X^\pi(s),Z^N))\right] \geq V^{\pi,N}\left(\mathbb{E}\left[T(s,X^\pi(s),Z^N)\right]\right)$$

$$= V^{\pi,N}(\mu, \beta + \beta^\epsilon e_{X^\pi(s)}) = V^{\pi,N}(\mu,\beta) = V^{\pi,N}(s).$$

Now consider the induction step. For $n < N-1$,

$$V^{\pi,n}(s) = \mathbb{E}\left[V^{\pi,n+1}(T(s,X^\pi(s),Z^{n+1}))\right] \geq \mathbb{E}\left[V^{\pi,n+2}(T(s,X^\pi(s),Z^{n+1}))\right]$$

by the induction hypothesis. Then, by the definition of $V^{\pi,n+1}$ in terms of $V^{\pi,n+2}$ from (10), we have $V^{\pi,n}(s) \geq V^{\pi,n+1}(s)$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Proof of Theorem 4.1. By (15), computing $X^{KG}(s)$ reduces to computing $Q^{N-1}(s,x)$ for each $x \in \{1,\dots,M\}$. By definition (11) we have, for a generic state s and standard normal random variable Z,

$$Q^{N-1}(s,x) = \mathbb{E}\left[V^N(T(s,x,Z))\right] = \mathbb{E}\left[(\mu_x + \tilde{\sigma}(\beta_x)Z) \vee \max_{x' \neq x} \mu_{x'}\right]. \tag{22}$$

This expectation is the expectation of the maximum of a constant and a normal random variable, for which we have an analytical expression from [13]. Let $a \in \mathbb{R}$ be an arbitrary constant and $W \sim \mathcal{N}(b, c^2)$ an arbitrary normal random variable. Then [13] tells us that

$$\mathbb{E}\left[W \vee a\right] = a\Phi\left(\frac{a-b}{c}\right) + b\Phi\left(\frac{b-a}{c}\right) + c\varphi\left(\frac{a-b}{c}\right), \tag{23}$$

which can be rewritten as

$$\begin{aligned}\mathbb{E}\left[W \vee a\right] &= a\Phi\left(\frac{a-b}{c}\right) + b\left(1 - \Phi\left(\frac{a-b}{c}\right)\right) + c\varphi\left(\frac{a-b}{c}\right)\\ &= b + (a-b)\Phi\left(\frac{a-b}{c}\right) + c\varphi\left(\frac{a-b}{c}\right)\\ &= b + c\left[\left(\frac{a-b}{c}\right)\Phi\left(\frac{a-b}{c}\right) + \varphi\left(\frac{a-b}{c}\right)\right].\end{aligned}$$

Fix x and consider two cases. First, consider the case that $\mu_x > \max_{x'} \mu_{x'}$. This is the case in which we measure an alternative that is uniquely the best according to the prior. Then $\mu_x - \max_{x' \neq x} \mu_{x'}$ is positive and $(\max_{x' \neq x} \mu_{x'} - \mu_x)/\tilde{\sigma}(\beta_x) = \zeta_x(s)$. Substitute $\zeta_x(s)$ for $(a-b)/c$ and write (22) as

$$Q^{N-1}(s,x) = \mu_x + \tilde{\sigma}(\beta_x)\left[\zeta_x(s)\Phi(\zeta_x(s)) + \varphi(\zeta_x(s))\right] = \mu_x + \tilde{\sigma}(\beta_x) f(\zeta_x(s)),$$

which can be rewritten in our case using $\mu_x = \max_{x'} \mu_{x'}$ as

$$Q^{N-1}(s,x) = \max_{x'} \mu_{x'} + \tilde{\sigma}(\beta_x) f(\zeta_x(s)). \tag{24}$$

Now consider the case that $\mu_x \leq \max_{x'} \mu_{x'}$. We rewrite (23) again using the substitution $\Phi(-z) = 1 - \Phi(z)$ and also using the symmetric property of the normal probability density function, $\varphi(-z) = \varphi(z)$, as

$$\mathbb{E}\left[Z \vee a\right] = a + c\left[\left(\frac{b-a}{c}\right)\Phi\left(\frac{b-a}{c}\right) + \varphi\left(\frac{b-a}{c}\right)\right].$$

In the case we are considering, $\mu_x - \max_{x' \neq x} \mu_{x'} \leq 0$ and $(\mu_x - \max_{x' \neq x})/\tilde{\sigma}(\beta_x) = \zeta_x(s)$. Substitute $\zeta_x(s)$ for $(b-a)/c$ and write (22) as

$$\begin{aligned}Q^{N-1}(s,x) &= \max_{x' \neq x} \mu_{x'} + \tilde{\sigma}(\beta_x)\left[\zeta_x(s)\Phi(\zeta_x(s)) + \varphi(\zeta_x(s))\right]\\ &= \max_{x' \neq x} \mu_{x'} + \tilde{\sigma}(\beta_x) f(\zeta_x(s)),\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

which can be rewritten in our case using $\max_{x' \neq x} \mu_{x'} = \max_{x'} \mu_{x'}$ as

$$Q^{N-1}(s,x) = \max_{x'} \mu_{x'} + \tilde{\sigma}(\beta_x) f(\zeta_x(s)). \tag{25}$$

In both cases the expression for $Q^{N-1}(s,x)$ agrees with (18), and we use this expression to rewrite (15) as

$$X^{KG}(s) \in \arg\max_x \max_{x'} \mu_{x'} + \tilde{\sigma}(\beta_x) f(\zeta_x(s)) = \arg\max_{x \in \{1,\ldots,M\}} \tilde{\sigma}(\beta_x) f(\zeta_x(s)),$$

since $\max_{x'} \mu_{x'}$ does not depend on x.

Proof of Proposition 4.3. By Theorem 4.1, KG prefers the alternative with the largest value of $\tilde{\sigma}(\beta_x) f(\zeta_x(S))$. Fix $S = (\mu, \beta)$, and let a be as in the statement of Proposition 4.3. Let i be the alternative preferred by KG, so

$$i = \arg\max_{x \in \{1,\ldots,M\}} \tilde{\sigma}(\beta_x) f(\zeta_x(S)), \tag{26}$$

where we recall that we are breaking ties by choosing the smallest index. Note that the theorem's condition on a trivializes the case when $\mu_i = \max_x \mu_x$ because here the range of a contains only the value 0, for which the theorem is obviously true. Thus, without loss of generality we may assume $\mu_i < \max_x \mu_x$, and let $j \in \arg\max_x \mu_x$. Then $j \neq i$.

Let $S' = (\mu + ae_i, \beta)$. We will first show for all alternatives $x \neq i$ that

$$\tilde{\sigma}(\beta_i) f(\zeta_i(S')) \geq \tilde{\sigma}(\beta_x) f(\zeta_x(S')). \tag{27}$$

This will show that $i \in \arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S'))$. We will then show that the implication

$$\tilde{\sigma}(\beta_x) f(\zeta_x(S)) < \tilde{\sigma}(\beta_i) f(\zeta_i(S)) \implies \tilde{\sigma}(\beta_x) f(\zeta_x(S')) < \tilde{\sigma}(\beta_i) f(\zeta_i(S')) \tag{28}$$

holds for all $x \neq i$. This will suffice to show the proposition because if we choose any $x' \notin \arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S))$, (26) will imply $\tilde{\sigma}(\beta_{x'}) f(\zeta_{x'}(S)) < \tilde{\sigma}(\beta_i) f(\zeta_i(S))$. The implication (28) will then imply that $\tilde{\sigma}(\beta_{x'}) f(\zeta_{x'}(S')) < \tilde{\sigma}(\beta_i) f(\zeta_i(S'))$ and, moreover, that $x' \notin \arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S'))$. Taking the contrapositive of the statement

$$x' \notin \arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S)) \implies x' \notin \arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S'))$$

reveals that

$$x' \in \arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S')) \implies x' \in \arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S)).$$

By this argument, (28) implies that $\arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S')) \subseteq \arg\max_x \tilde{\sigma}(\beta_x) \cdot f(\zeta_x(S))$. Therefore i is the element of $\arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S))$ with the smallest index, and thus i is the alternative that KG prefers in state S'.

We will show (27) and (28) by treating three cases separately, noting in general that $\zeta_i(\mu, \beta) \leq \zeta_i(\mu + ae_i, \beta)$. The first case is when $x \neq i, j$. Then

$$\zeta_x(S') = \zeta_x(\mu + ae_i, \beta) = \zeta_x(\mu, \beta) = \zeta_x(S).$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Thus, (27) is true because

$$\tilde{\sigma}(\beta_i)f(\zeta_i(S')) \geq \tilde{\sigma}(\beta_i)f(\zeta_i(S)) \geq \tilde{\sigma}(\beta_x)f(\zeta_x(S)) = \tilde{\sigma}(\beta_x)f(\zeta_x(S')),$$

and (28) is true because if $\tilde{\sigma}(\beta_x)f(\zeta_x(S)) < \tilde{\sigma}(\beta_i)f(\zeta_i(S))$, then

$$\tilde{\sigma}(\beta_x)f(\zeta_x(S')) = \tilde{\sigma}(\beta_x)f(\zeta_x(S)) < \tilde{\sigma}(\beta_i)f(\zeta_i(S)) \leq \tilde{\sigma}(\beta_i)f(\zeta_i(S')).$$

The second case is when $x = j$ and $\mu_i + a < \max_{x' \neq j} \mu_{x'}$. Then again $\zeta_j(S') = \zeta_j(S)$ because $j \neq i$, and both (27) and (28) hold by the same reasoning as in the first case.

The third case is when $x = j$ and $\mu_i + a \geq \max_{x' \neq j} \mu_{x'}$. Then we have $\zeta_j(\mu + ae_i, \beta) = \frac{-|\mu_i + a - \mu_j|}{\tilde{\sigma}(\beta_j)}$. For $x = j$, KG's preference of alternative i implies that $\beta_i \leq \beta_j$. Otherwise, by Remark 4.2 and because $\mu_j \geq \mu_i$, KG would prefer alternative j. This shows that

$$\zeta_i(\mu + ae_i, \beta) = \frac{-|\mu_i + a - \mu_j|}{\tilde{\sigma}(\beta_i)} \geq \frac{-|\mu_i + a - \mu_j|}{\tilde{\sigma}(\beta_j)} = \zeta_j(\mu + ae_i, \beta).$$

This shows (27). To show (28), assume the antecedent of condition (28). Since $|\mu_j - \max_{x' \neq j} \mu_{x'}| \leq |\mu_j - \mu_i|$ and $\tilde{\sigma}(\beta_j)f(\zeta_j(S)) < \tilde{\sigma}(\beta_i)f(\zeta_i(S))$, it must be that $\tilde{\sigma}(\beta_j) < \tilde{\sigma}(\beta_i)$ since otherwise j would have been KG's choice in state S. Thus,

$$\zeta_i(\mu + ae_i, \beta) = \frac{-|\mu_i + a - \mu_j|}{\tilde{\sigma}(\beta_i)} > \frac{-|\mu_i + a - \mu_j|}{\tilde{\sigma}(\beta_j)} = \zeta_j(\mu + ae_i, \beta).$$

Proof of Proposition 5.1. We will show that $V^0(S^0; N)$ is a nondecreasing function of N bounded from above by $U(S^0)$, which will imply that the limit $V(S^0; \infty)$ exists and is bounded as claimed. To show that $V^0(S^0; N)$ is nondecreasing in N, note that $V^0(S^0; N-1) = V^1(S^0; N)$, and thus

$$V^0(S^0; N) - V^0(S^0; N-1) = V^0(S^0; N) - V^1(S^0; N).$$

This difference is positive by Corollary 3.2.

Now we show that $V^0(S^0; N) \leq U(S^0)$. For every $N \geq 1$ and policy π,

$$\begin{aligned}\mathbb{E}^\pi \left[\max_x \mu_x^N\right] &= \mathbb{E}^\pi \left[\max_x \mathbb{E}_N^\pi [Y_x]\right] \leq \mathbb{E}^\pi \left[\mathbb{E}_N^\pi \left[\max_x Y_x\right]\right] \\ &= \mathbb{E}^\pi \left[\max_x Y_x\right] = \mathbb{E} \left[\max_x Y_x\right].\end{aligned}$$

This value is independent of π and is equal to $U(S^0)$. Thus

$$V^0(S^0; N) := \sup_\pi \mathbb{E}^\pi \left[\max_x \mu_x^N\right] \leq U(S^0)$$

for every $N \geq 1$. Taking the limit as $N \to \infty$ shows $V(S^0; \infty) \leq U(S^0)$.

Finally, we show that the limit $V^\pi(S^0; \infty)$ exists and is finite for every stationary policy π. Fix a stationary policy π. Then Theorem 3.1 implies that $V^{\pi,0}(S^0; N)$ is nondecreasing in N, and $V^{\pi,0}(S^0; N)$ is bounded by $V^0(S^0; N)$, which is itself uniformly bounded in N by $U(S^0)$. Then $V^\pi(S^0; \infty)$ is the limit of a nondecreasing bounded sequence. Hence, it exists.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Proof of Proposition 5.2. We assumed in the formal model in section 3.1 that our measurement-noise variance $(\sigma^\varepsilon)^2$ is finite. This implies via the strong law of large numbers that the sequence of posterior predictive means μ_x^N converges as $\lim_{N\to\infty}\mu_x^N = Y_x$ a.s. for each $x = 1,\dots,M$. Thus $\lim_{N\to\infty}\max_x \mu_x^N$ exists a.s. and in probability. We will show next that the sequence $\left(\max_x \mu_x^N\right)_{N\geq 1}$ is uniformly integrable, and then convergence in probability together with uniform integrability implies convergence in L^1 (see, e.g., [20, Theorem 3.12]). Convergence in L^1 of $\max_x \mu_x^N$ as $N\to\infty$ implies

$$V^\pi(S^0;\infty) = \lim_{N\to\infty}\mathbb{E}^\pi\left[\max_x \mu_x^N\right] = \mathbb{E}^\pi\left[\lim_{N\to\infty}\max_x \mu_x^N\right] = \mathbb{E}^\pi\left[\max_x Y_x\right] = U(S^0).$$

Proposition 5.1 showed that $U(S^0)\geq V(S^0;\infty)$, so $V^\pi(S^0;\infty) = V(S^0;\infty)$ and π must be asymptotically optimal.

To complete the proof we must show uniform integrability of the sequence $\left(\max_x \mu_x^N\right)_{N\geq 1}$. For every fixed K we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\max_x \mu_x^N\right|\ 1_{\{|\max_x \mu_x^N|\geq K\}}\right] \leq \mathbb{E}\left[\max_x \left|\mu_x^N\right|\ 1_{\{\max_x|\mu_x^N|\geq K\}}\right]\\
&= \mathbb{E}\left[\max_x \left|\mathbb{E}_N\left[Y_x\right]\right|\ 1_{\{\max_x|\mathbb{E}_N[Y_x]|\geq K\}}\right] \leq \mathbb{E}\left[\max_x \mathbb{E}_N\left[|Y_x|\right]\ 1_{\{\max_x \mathbb{E}_N[|Y_x|]\geq K\}}\right]\\
&\leq \mathbb{E}\left[\mathbb{E}_N\left[\max_x |Y_x|\right]\ 1_{\{\mathbb{E}_N[\max_x|Y_x|]\geq K\}}\right] = \mathbb{E}\left[\mathbb{E}_N\left[\max_x |Y_x|\ 1_{\{\mathbb{E}_N[\max_x|Y_x|]\geq K\}}\right]\right]\\
&= \mathbb{E}\left[\max_x |Y_x|\ 1_{\{\mathbb{E}_N[\max_x|Y_x|]\geq K\}}\right].
\end{aligned}$$

We assumed in the formal model in section 3.1 that $\max_x |Y_x|$ was integrable. This implies via Markov's inequality that

$$\mathbb{P}\left\{\mathbb{E}_N\left[\max_x |Y_x|\right]\geq K\right\} \leq \frac{\mathbb{E}\left[\mathbb{E}_N\left[\max_x |Y_x|\right]\right]}{K} = \frac{\mathbb{E}\left[\max_x |Y_x|\right]}{K}.$$

This is bounded uniformly in N, and the bound goes to zero as $K\to\infty$.

Proof of Theorem 5.1. First note that KG is stationary. We will show that $\lim_{N\to\infty}\eta_x^N = \infty$ a.s. for all x under KG, and then Proposition 5.2 will complete the proof.

First we show that, for each x, $\{\mu_x^n\}_{n=0}^\infty$ is a uniformly integrable martingale with respect to the filtration $\mathcal{F}$ and hence converges. μ_x^n is defined by $\mu_x^n := \mathbb{E}\left[Y_x \mid \mathcal{F}^n\right]$ and thus is $\mathcal{F}^n$-measurable and, by the tower property of conditional expectation, satisfies the martingale identity. Y_x is a normal random variable with finite variance. Thus, $Y_x\in L^2\subset L^1$, and by the Doob uniform integrability lemma [20, Lemma 5.5], the collection of conditional expectations $\{\mu_x^n\}_n$ is uniformly integrable (and hence each μ_x^n is integrable). Thus, $\{\mu_x^n\}_n$ is a uniformly integrable martingale and hence converges a.s. to an integrable random variable μ_x^∞. In addition, $\lim_{n\to\infty}\beta_x^n \overset{a.s.}{=} \beta_x^0 + \beta^\epsilon\eta_x^\infty$ for each x.

By the computation performed in Theorem 4.1, the Q-factors for each alternative x are continuous functions of their arguments (μ,β), and, hence,

$$\lim_{n\to\infty} Q^{N-1}(S^n;x) \overset{a.s.}{=} \max_{x'}\mu_{x'}^\infty + \tilde\sigma(\beta_x^\infty) f\left(\frac{\mu_x^\infty - \max_{x''\neq x}\mu_{x''}^\infty}{\tilde\sigma(\beta_x^\infty)}\right).$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Define Ω_0 to be the almost sure event on which this convergence holds, and define the event $\mathcal{H}_x$ to be $\mathcal{H}_x := \{\omega : \eta_x^\infty(\omega) < \infty\}$. Then,

$$\lim_{n\to\infty} Q^{N-1}(S^n(\omega);x) > \max_{x'} \mu_{x'}^\infty(\omega) \qquad \text{for all } \omega \in \mathcal{H}_x \cap \Omega_0, \tag{29}$$

$$\lim_{n\to\infty} Q^{N-1}(S^n(\omega);x) = \max_{x'} \mu_{x'}^\infty(\omega) \qquad \text{for all } \omega \in \mathcal{H}_x^c \cap \Omega_0. \tag{30}$$

Let A be any subset of $\{1,\dots,M\}$, and define the event $\mathcal{H}_A$ to be $\mathcal{H}_A := (\cap_{x\in A}\mathcal{H}_x)\cap(\cap_{x\in A^c}\mathcal{H}_x^c)$. We will show that, if $A \neq \emptyset$, then $\mathbb{P}(\mathcal{H}_A) = 0$. This will prove the theorem because $\Omega = \cup_{A\subseteq\{1,\dots,M\}}\mathcal{H}_A$, so if we know that $A \neq \emptyset \implies \mathbb{P}(\mathcal{H}_A) = 0$, then $1 = \mathbb{P}(\mathcal{H}_\emptyset) = \mathbb{P}\{\lim_{n\to\infty}\eta_x^n = \infty \text{ for all } x\}$.

Fix A nonempty and suppose for contradiction that $\mathcal{H}_A \cap \Omega_0$ is nonempty so that we may choose $\omega \in \mathcal{H}_A \cap \Omega_0$ to be an element of this set. By (29) and (30), for all $x \in A$ and all $y \in A^c$,

$$\lim_{n\to\infty} Q^{N-1}(S^n(\omega);x) > \lim_{n\to\infty} Q^{N-1}(S^n(\omega);y),$$

and there exists a finite number K_{xy} such that, for all $n > K_{xy}$,

$$Q^{N-1}(S^n(\omega);x) > Q^{N-1}(S^n(\omega);y).$$

Let $K := \max_{x\in A, y\in A^c} K_{xy}$ if A^c is nonempty, and let $K := 1$ if A^c is empty. Then K is finite and for all $n > K$ and all $x \in A$ and $y \in A^c$,

$$Q^{N-1}(S^n(\omega);x) > Q^{N-1}(S^n(\omega);y).$$

Therefore, KG distributes all measurements $n > K$ only to alternatives in the set A, and $\sum_{x\in A}\eta_x^\infty(\omega) = \infty$. This is a contradiction because $x \in A$ implies $\omega \in \mathcal{H}_x$, which implies $\eta_x^\infty(\omega) < \infty$.

Thus, $\mathbb{P}(\mathcal{H}_\emptyset \cap \Omega_0) = 0$, and since $\mathbb{P}(\Omega_0) = 1$, $\mathbb{P}(\mathcal{H}_\emptyset) = 0$.

Proof of Theorem 6.1. Note that $\varphi(0) = (2\pi)^{-1/2}$, where φ is the normal probability density function. We will use this throughout. We induct backward over n. First, when $n = N-1$, the theorem is trivially true with equality. Now, under the assumption that the theorem is true for some $n+1$,

$$\begin{aligned} V^n(s) &= \max_x \mathbb{E}\left[V^{n+1}(T(s,x,Z^{n+1}))\right] \\ &\le \max_x \mathbb{E}\left[V^{N-1}(T(s,x,Z^{n+1})) + \varphi(0)(N-n-2)\max_x \tilde\sigma\left(\beta_{x'} + \beta^\epsilon 1_{\{x=x'\}}\right)\right]. \end{aligned}$$

Then, since $\tilde\sigma$ is a decreasing function and $\beta_{x'}^n \le \beta_{x'}^n + \beta^\epsilon 1_{\{x=x'\}}$,

$$V^n(s) \le \max_x \mathbb{E}\left[V^{N-1}(T(s,x,Z^{n+1})) + \varphi(0)(N-n-2)\max_{x'} \tilde\sigma(\beta_{x'})\right].$$

Since the last term is a constant and does not depend on x, we may move it outside the maximum and expectation operators, giving

$$V^n(s) \le \max_x \mathbb{E}\left[V^{N-1}(T(s,x,Z^{n+1}))\right] + \varphi(0)(N-n-2)\max_{x'} \tilde\sigma(\beta_{x'}). \tag{31}$$

We will rewrite the first term on the right-hand side of this inequality as a maximum over a set of Q-factors using the definition of V^{N-1} in terms of Q^{N-1}, but before

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

making this substitution, let us bound Q^{N-1}. We rewrite the expression (24) for Q^{N-1} as $Q^{N-1}(s,x') = \max_{x''} \mu_{x''} + \tilde{\sigma}(\beta_{x'}) f(\zeta_{x'}) = V^N(s) + \tilde{\sigma}(\beta_{x'}) f(\zeta_{x'})$. Lemma 4.1 tells us that f is nondecreasing, so $\zeta_{x'} \le 0$ implies that $f(\zeta_{x'}) \le f(0) = \varphi(0)$. Thus,

$$Q^{N-1}(s,x') \le V^N(s) + \varphi(0)\tilde{\sigma}(\beta_{x'}).$$

Using this and the definition of the value function in terms of the Q-factors from (10) and (11), we have

$$\begin{aligned} V^{N-1}(T(s,x,Z^{n+1})) &= \max_{x'} Q^{N-1}(T(s,x,Z^{n+1}),x') \\ &\le \max_{x'} V^N(T(s,x,Z^{n+1})) + \varphi(0)\tilde{\sigma}\left(\beta_{x'} + \beta^\epsilon 1_{\{x=x'\}}\right) \\ &= V^N(T(s,x,Z^{n+1})) + \varphi(0) \max_{x'} \tilde{\sigma}\left(\beta_{x'} + \beta^\epsilon 1_{\{x=x'\}}\right) \\ &\le V^N(T(s,x,Z^{n+1})) + \varphi(0) \max_{x'} \tilde{\sigma}(\beta_{x'}). \end{aligned}$$

Combining this bound with (31) and moving the $\tilde{\sigma}(\beta_x)$ outside the maximization and expectation operators, we obtain

$$\begin{aligned} V^n(s) &\le \max_x \mathbb{E}\left[V^N(T(s,x,Z^{n+1})) + \varphi(0) \max_{x'} \tilde{\sigma}(\beta^n_{x'})\right] + \varphi(0)(N-n-2) \max_{x'} \tilde{\sigma}(\beta_{x'}) \\ &= \max_x \mathbb{E}\left[V^N(T(s,x,Z^{n+1}))\right] + \varphi(0)(N-n-1) \max_{x'} \tilde{\sigma}(\beta_{x'}) \\ &= V^{N-1}(s) + \varphi(0)(N-n-1) \max_{x'} \tilde{\sigma}(\beta_{x'}), \end{aligned}$$

where in the last step we used the definition of V^N in terms of V^{N-1} from (10).

Proof of Theorem 7.1. The proof is by induction backward on k. The theorem holds for the base case, $k = N-1$, by Remark 4.1. Now let $k < N-1$. Let π^* be an optimal policy, with decision function X^{*k} at time k. Let $s = (\mu, \beta) \in \mathbb{S}^k$. Then

$$V^k(s) = \mathbb{E}\left[V^{k+1}(T(s, X^{*k}(s), Z^{k+1}))\right] = \mathbb{E}\left[V^{KG,k+1}(T(s, X^{*k}(s), Z^{k+1}))\right] \tag{32}$$

by the induction hypothesis, since $\{\mathbb{S}^n\}_{n=k+1}^N$ is a covering of the future from $k+1$ on which KG persistence holds, and $T(s, X^{*k}(s), Z^{k+1}) \in \mathbb{S}^{k+1}$ a.s.

Consider two cases. In the first case, suppose $X^{*k}(s) = X^{KG}(s)$. By (32),

$$V^k(s) = \mathbb{E}\left[V^{KG,k+1}(T(s, X^{KG}(s), Z^{k+1}))\right] = V^{KG,k}(s).$$

In the second case, suppose $X^{*k}(s) \ne X^{KG}(s)$. Then, abbreviating the random state at time $k+1$ under the optimal policy by $S^{k+1} = T(s, X^{*k}(s), Z^{k+1})$,

$$\begin{aligned} V^k(s) &= \mathbb{E}\left[V^{KG,k+2}(T(S^{k+1}, X^{KG}(S^{k+1}), Z^{k+2}))\right] \\ &= \mathbb{E}\left[V^{KG,k+2}(T(S^{k+1}, X^{KG}(s), Z^{k+2}))\right], \end{aligned} \tag{33}$$

since $X^{KG}(s) = X^{KG}(S^{k+1})$ a.s. by the KG persistence property. Let $S^{k+2} = T(S^{k+1}, X^{KG}(s), Z^{k+2})$. Then $V^k(s) = \mathbb{E}\left[V^{KG,k+2}(S^{k+2})\right]$.

Note that S^{k+2} is the state to which we arrive when we measure $X^{*k}(s)$ at time k and $X^{KG}(s)$ at time $k+1$. Let $E_x = e_x(e_x)^T$ be a matrix of all zeros except for

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

a single 1 at row x, column x, and let $\stackrel{d}{=}$ denote equality in distribution. Then the definition (8) of the transition function T and $X^{KG}(s) \neq X^{*,k}(s)$ imply

$$\begin{aligned}
S^{k+2} &= T(S^{k+1}, X^{KG}(s), Z^{k+2}) \\
&= T(T(s, X^{*k}(s), Z^{k+1}), X^{KG}(s), Z^{k+2}) \\
&= \mu + \tilde{\sigma}(\beta_{X^{KG}(s)})Z^{k+1} + \tilde{\sigma}(\beta_{X^{*,k}(s)})Z^{k+2} + \beta^{\epsilon}E_{X^{KG}(s)} + \beta^{\epsilon}E_{X^{*,k}(s)} \\
&\stackrel{d}{=} \mu + \tilde{\sigma}(\beta_{X^{KG}(s)})Z^{k+2} + \tilde{\sigma}(\beta_{X^{*,k}(s)})Z^{k+1} + \beta^{\epsilon}E_{X^{KG}(s)} + \beta^{\epsilon}E_{X^{*,k}(s)} \\
&= T(T(s, X^{KG}(s), Z^{k+1}), X^{*k}(s), Z^{k+2}).
\end{aligned}$$

Thus, we have that $V^k(s) = \mathbb{E}\left[V^{KG,k+2}(S^{k+2})\right]$ equals

$$\mathbb{E}\left[V^{KG,k+2}(T(T(s, X^{KG}(s), Z^{k+1}), X^{*k}(s), Z^{k+2}))\right].$$

This quantity is the value of making decisions $X^{KG}(s)$ at time k, $X^{*k}(s)$ at time $k+1$, and then following KG afterward. This value must be less than the value of making the same decision $X^{KG}(s)$ at time k and following the optimal policy afterward. Thus, $V^k(s) \leq \mathbb{E}\left[V^{k+1}(T(s, X^{KG}(s), Z^{k+1}))\right]$. Now, $T(s, X^{KG}(s), Z^{n+1}) \in \mathbb{S}^{n+1}$ a.s., so by the induction hypothesis we may replace the optimal value function with the KG value function when operating on this state. This allows us to write

$$V^k(s) \leq \mathbb{E}\left[V^{KG,k+1}(T(s, X^{KG}(s), Z^{k+1}))\right] = V^{KG,k}(s).$$

Finally, $V^k(s) \geq V^{KG,k}(s)$ implies $V^k(s) = V^{KG,k}(s)$.

Proof of Theorem 7.3. For $n \in \{0, \ldots, N-1\}$, define $\mathbb{S}^n$ to be the set of all $s = (\mu, \beta) \in \mathbb{S}$ satisfying

$$(\beta_i \neq \infty \text{ and } \beta_j \neq \infty \text{ and } \beta_i < \beta_j) \implies \mu_i \geq \mu_j \tag{34}$$

for all $i, j \in \{1, \ldots, M\}$. Note that the sets $\mathbb{S}^n$ are identical for all n. We will show that $\{\mathbb{S}^n\}$ is a covering of the future from 0.

Let $n \in \{0, \ldots, N-2\}$, $x \in \{1, \ldots, M\}$, $s \in \mathbb{S}^n$, and $S^n = s$ a.s. Consider $S^{n+1} := T(S^n, x, Z^{n+1})$. Let $i, j \in \{1, \ldots, M\}$ meet the conditions of the implication (34) for S^{n+1}, so $\beta_i^{n+1} \neq \infty$ and $\beta_j^{n+1} \neq \infty$ and $\beta_i^{n+1} < \beta_j^{n+1}$. We will show that $\mu_i^n \geq \mu_j^n$, which will show that S^{n+1} meets condition (34) and is in $\mathbb{S}^{n+1}$.

First, $\beta^n \leq \beta^{n+1}$ componentwise implies that $\beta_i^n \neq \infty$ and $\beta_j^n \neq \infty$. Also, since $(\sigma^{\varepsilon})^2 = 0$, $\beta_x^{n+1} = \infty$, which implies that $x \neq i, j$, and the measurement between S^n and S^{n+1} alters neither the i component nor the j component. Thus, $\beta_i^n = \beta_i^{n+1} < \beta_j^{n+1} = \beta_j^n$. This shows that i, j meet the conditions of the implication (34) for S^n as well as S^{n+1}. Thus, since $S^n \in \mathbb{S}^n$, $\mu_i^n \geq \mu_j^n$. Then, again because $x \neq i, j$ implies that the means of the i and j components did not change from time n to $n+1$, $\mu_i^{n+1} \geq \mu_j^{n+1}$, showing that S^{n+1} meets the condition (34), and $S^{n+1} \in \mathbb{S}^{n+1}$. Thus, $\{\mathbb{S}^n\}$ is a covering of the future from 0.

Now we will show that KG is persistent on $\{\mathbb{S}^n\}$. Let $s \in \mathbb{S}^n$ and $S^n = s$ a.s. Condition (34) together with Remarks 4.2 and 4.3 implies $X^{KG}(S^n) \in \arg\min_{x'} \beta_{x'}^n$, with ties broken by the smallest index. Let $x \neq X^{KG}(S^n)$. We showed that $S^{n+1} := T(S^n, x, Z^{n+1}) \in \mathbb{S}^{n+1}$ a.s. Thus, again by condition (34) and Remarks 4.2 and 4.3,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$X^{KG}(S^{n+1}) \in \arg\min_{x'} \beta^{n+1}_{x'}$. We use the state transition function for the case with $(\sigma^{\varepsilon}) = 0$, $\beta^{n+1}_{x'} = \beta^n_{x'} + \infty 1_{\{x'=x\}}$, and we consider two cases.

In the first case suppose $\beta^n_{x'} < \infty$ for some $x' \neq x$. Then, since $\beta^{n+1}_x = \infty$, we have $\beta^{n+1}_{x'} = \beta^n_{x'} < \beta^{n+1}_x$. Thus, we may drop x from the argmin set as in

$$\arg\min_{x'} \beta^{n+1}_{x'} = \arg\min_{x' \neq x} \beta^{n+1}_{x'} = \arg\min_{x' \neq x} \beta^n_x.$$

$X^{KG}(S^n)$ is the element of this set with the smallest index, and since $X^{KG}(S^{n+1})$ is also defined to be the element of this set with the smallest index, $X^{KG}(S^{n+1}) = X^{KG}(S^n)$.

In the second case suppose $\beta^n_{x'} = \infty$ for all $x' \neq x$. Then, by $X^{KG} \in \arg\min_{x'} \beta^n_{x'}$, and since $X^{KG}(S^n) \neq x$, we also have that $\beta^n_x = \infty$. The state transition rule for β implies that $\beta^{n+1}_{x'} = \infty$ for all x'. Thus, $\arg\min_{x'} \beta^n_{x'} = \{1, \ldots, M\} = \arg\min_{x'} \beta^{n+1}_{x'}$, and since the tie-breaking rule is fixed to choose the element with the smallest index, $X^{KG}(S^{n+1}) = X^{KG}(S^n)$.

In both cases KG is persistent on $\{\mathbb{S}^n\}$, and Theorem 7.1 shows that $V^{KG,0}(s) = V^0(s)$ for all $s \in \mathbb{S}^0$.

Appendix B. Known variance LL(S) policy. The LL(S) policy was developed for normal measurement errors with *unknown* variance and uses a normal-gamma prior for the unknown mean and measurement precision. To adapt it to the known-variance case, we take both the shape and rate parameters in the gamma prior on the measurement precision to infinity while keeping their ratio fixed to the known measurement precision β^{ϵ}; we obtain a prior in which the measurement precision is known perfectly and the alternative's true value is still normally distributed. Taking this limit in the allocation given by [12, Corollary 1] provides the following policy. The steps below describe how the policy allocates τ measurements for the stage beginning at a generic time n, and should be repeated a total of N/τ times beginning at time 0 and finishing at time N. We use the notation $[i]$ to indicate the alternative whose μ^n component is ith largest. That is, $\mu^n_{[M]} \geq \cdots \geq \mu^n_{[1]}$.

(i) For each alternative calculate $n_i = \beta^n_i/\beta^{\epsilon}$, which may be interpreted as the effective number of times alternative i has been sampled.

(ii) Initialize $\mathcal{S}$, the set of alternatives under consideration for measurement in the current stage, to $\mathcal{S} = \{1, \ldots, M\}$.

(iii) For each $i \in \mathcal{S} \setminus \{[M]\}$ set $\lambda_{i,M}$ as follows. If $[M] \notin \mathcal{S}$, set $\lambda_{i,M} = \beta_{[i]}$. If $[M] \in \mathcal{S}$, set $\lambda_{i,M} = \left((\beta^n_{[M]})^{-1} + (\beta^n_{[i]})^{-1}\right)^{-1}$.

(iv) Calculate a tentative number of samples $r_{[i]}$ to take from alternative $[i]$,

$$r_{[i]} = \frac{\tau + \sum_{j \in \mathcal{S}} n_j}{\sum_{j \in \mathcal{S}} \sqrt{\gamma_j/\gamma_{[i]}}} - n_{[i]},$$

where

$$\gamma_{[i]} = \begin{cases} \sqrt{\lambda_{i,M}}\ \phi\left(\sqrt{\lambda_{i,M}}(\mu^n_{[M]} - \mu^n_{[i]})\right) & \text{if } [i] \neq [M], \\ \sum_{[j] \in \mathcal{S} \setminus \{[M]\}} \gamma_{[j]} & \text{if } [i] = [M]. \end{cases}$$

(v) For each $[i] \in \mathcal{S}$ with $r_{[i]} < 0$, remove $[i]$ from $\mathcal{S}$ and set $r_{[i]} = 0$. If any $[i]$ was removed, then return to step (iii).

(vi) Round the $r_{[i]}$ to integer values so that they still sum to τ.

(vii) Run $r_{[i]}$ additional samples for each alternative $[i]$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

REFERENCES

[1] R. Bechhofer, T. Santner, and D. Goldsman, *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*, John Wiley & Sons, New York, 1995.

[2] J. Branke, S. Chick, and C. Schmidt, *New developments in ranking and selection: An empirical comparison of the three main approaches*, in Proceedings of the 2005 Winter Simulation Conference, M. Kuhl, N. Steiger, F. Armstrong, and J. Joines, eds., IEEE, Piscataway, NJ, 2005, pp. 708–717.

[3] H. Chang, M. Fu, J. Hu, and S. Marcus, *Simulation-Based Algorithms for Markov Decision Processes*, Springer, Berlin, 2007.

[4] C. Chen, *An effective approach to smartly allocate computing budget for discrete event simulation*, in Proceedings of the 34th IEEE Conference on Decision and Control, New Orleans, LA, 1995, pp. 2598–2603.

[5] C. Chen, L. Dai, and H. Chen, *A gradient approach for smartly allocating computing budget for discrete event simulation*, in Proceedings of the 28th Winter Simulation Conference, IEEE Computer Society, Washington, DC, 1996, pp. 398–405.

[6] C. Chen, K. Donohue, E. Yücesan, and J. Lin, *Optimal computing budget allocation for Monte Carlo simulation with application to product design*, Simul. Model. Practice Theory, 11 (2003), pp. 57–74.

[7] C. Chen, J. Lin, E. Yücesan, and S. Chick, *Simulation budget allocation for further enhancing the efficiency of ordinal optimization*, Discrete Event Dyn. Syst., 10 (2000), pp. 251–270.

[8] H. Chen, C. Chen, and E. Yücesan, *Computing efforts allocation for ordinal optimization and discrete event simulation*, IEEE Trans. Automat. Control, 45 (2000), pp. 960–964.

[9] H. Chen, L. Dai, C. Chen, and E. Yücesan, *New development of optimal computing budget allocation for discrete event simulation*, in Proceedings of the 29th Winter Simulation Conference, IEEE Computer Society, Washington, DC, 1997, pp. 334–341.

[10] S. Chick, J. Branke, and C. Schmidt, *New myopic sequential sampling procedures*, INFORMS J. Computing, submitted.

[11] S. Chick and K. Inoue, *New procedures to select the best simulated system using common random numbers*, Manage. Sci., 47 (2001), pp. 1133–1149.

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

[13] C. Clark, *The greatest of a finite set of random variables*, Oper. Res., 9 (1961), pp. 145–163.

[14] M. C. Fu, J.-Q. Hu, C.-H. Chen, and X. Xiong, *Simulation allocation for determining the best design in the presence of correlated sampling*, INFORMS J. Comput., 19 (2007), pp. 101–111.

[15] S. Gupta and K. Miescke, *Bayesian look ahead one stage sampling allocations for selecting the largest normal mean*, Statist. Papers, 35 (1994), pp. 169–177.

[16] S. Gupta and K. Miescke, *Bayesian look ahead one-stage sampling allocations for selection of the best population*, J. Statist. Plann. Inference, 54 (1996), pp. 229–244.

[17] M. Hartmann, *An improvement on Paulson's procedure for selecting the population with the largest mean from k normal populations with a common unknown variance*, Sequential Anal., 10 (1991), pp. 1–16.

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

[19] L. P. Kaelbling, *Learning in Embedded Systems*, MIT Press, Cambridge, MA, 1993.

[20] O. Kallenberg, *Foundations of Modern Probability*, Springer, New York, 1997.

[21] S. Kim and B. Nelson, *Selecting the best system*, in Simulation, Handbooks Oper. Res. Management Sci. 13, North–Holland, Amsterdam, 2006, pp. 501–534.

[22] S. Kim and B. Nelson, *On the asymptotic validity of fully sequential selection procedures for steady-state simulation*, Oper. Res., 54 (2006), pp. 475–488.

[23] S.-H. Kim and B. L. Nelson, *A fully sequential procedure for indifference-zone selection in simulation*, ACM Trans. Model. Comput. Simul., 11 (2001), pp. 251–273.

[24] B. Nelson, J. Swann, D. Goldsman, and W. Song, *Simple procedures for selecting the best simulated system when the number of alternatives is large*, Oper. Res., 49 (2001), pp. 950–963.

[25] E. Paulson, *A sequential procedure for selecting the population with the largest mean from k normal populations*, Ann. Math. Statist., 35 (1964), pp. 174–180.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[26] E. Paulson, *Sequential procedures for selecting the best one of K Koopman-Darmois populations*, Sequential Anal., 13 (1994), pp. 207–220.
[27] Y. Rinott, *On two-stage selection procedures and related probability-inequalities*, Comm. Statist. A—Theory Methods, 7 (1978), pp. 799–811.
[28] S. Singh, T. Jaakkola, M. Littman, and C. Szepesvari, *Convergence results for single-step on-policy reinforcement-learning algorithms*, Mach. Learn., 38 (2000), pp. 287–308.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.