The Knowledge-Gradient Policy for Correlated Normal Beliefs

Peter Frazier, Warren Powell, Savas Dayanik
Department of Operations Research and Financial Engineering,
Princeton University, Princeton, NJ 08544, USA,
{pfrazier@princeton.edu, powell@princeton.edu, sdayanik@princeton.edu}

We consider a Bayesian ranking and selection problem with independent normal rewards and a correlated multivariate normal belief on the mean values of these rewards. Because this formulation of the ranking and selection problem models dependence between alternatives' mean values, algorithms may utilize this dependence to perform efficiently even when the number of alternatives is very large. We propose a fully sequential sampling policy called the knowledge-gradient policy, which is provably optimal in some special cases and has bounded suboptimality in all others. We then demonstrate how this policy may be applied to efficiently maximize a continuous function on a continuous domain while constrained to a fixed number of noisy measurements.

Key words: 763 Simulation : Design of experiments; 055 Decision analysis: Sequential; 793 Statistics, Bayesian

History: Received January 2008; revised July, October 2008.

---

1. Introduction

Consider the following problem: we are confronted with a collection of alternatives, and asked to choose one from among them. It may be convenient to think of these alternatives as possible configurations of an assembly line, different temperature settings for a chemical production process, or different drugs for treating a disease. The chosen alternative will return a reward according to its merit, but these rewards are unknown and so it is unclear which alternative to choose. Before choosing, however, we have the opportunity to measure some of the alternatives. As measurements have a cost, we are only allowed a limited number, and our strategy should allocate these measurements across the alternatives in such a way as to maximize the information gained and the reward obtained. Measurements are typically noisy, and so a single alternative may merit more than one measurement. This problem is known as the ranking and selection (R&S) problem.

The R&S problem is ubiquitous in application. Consider the following short list of examples chosen from the long list that could be enumerated:

- We wish to choose the dosage level for a drug with the median aggregate response in patients. This dosage level is desirable because it achieves the positive effect of the drug while minimizing side effects. Similarly, we might wish to find the dosage level maximizing some utility function which is increasing in positive drug response and decreasing in negative drug response. The set of dosage levels from which to choose is finite because only finitely many amounts of a drug can be easily distributed and administered to patients.

- We wish to select the fastest path through a network subject to traffic delays by sampling travel times through it. This network might be a data network through which we would like to transmit packets of data, or a network of roads through which we would like to route vehicles.

- In the early stages of drug discovery, pharmaceutical companies often perform robotically automated tests in which chemical compounds are screened for effectiveness against a particular disease. These tests, in which surviving diseased and non-diseased cells are counted after exposure to a compound, are performed on a large number of chemical compounds from which a small number of candidates will be selected.

- We wish to measure heat or pollution at discrete points in a continuous medium to ascertain which of the finitely many discrete locations have the highest levels.

Common to these examples is the characteristic of dependence, by which we mean that when we measure one alternative, we learn something about the others. In the drug dosage example, drug response is generally increasing in dosage. In the network example, each congested link slows travel times along all paths that share it. In the drug development example, chemically related compounds often have similar effects. In the pollution example, pollution levels at nearby locations are correlated. In this article, we introduce a fully sequential sampling technique called the correlated knowledge-gradient (KG) policy which takes advantage of this dependence in the prior belief to improve sampling efficiency.

While each of the examples has correlation in the *belief*, we will assume that any measurement errors are independent. This may require additional assumptions in some of the

examples. For example, in the continuous medium and network path examples, we assume that measurements are taken sufficiently far apart from each other in time that measurement noise can be assumed independent. In the drug discovery example, we assume that there are no confounding factors like a time-varying laboratory temperature that would induce correlated measurement noise.

The R&S and experimental design literature has devoted the most attention to our problem class (see Bechhofer et al. (1995) for a comprehensive treatment of R&S, and Fu (2002); Swisher et al. (2003) for a review of R&S within the simulation community). Within this literature, the techniques that most successfully exploit dependence are variance-reduction techniques for simulation (Law and Kelton, 2000), which include control variates (Nelson and Staum, 2006) and common random numbers (Kim and Nelson, 2001). While both variance-reduction techniques and the correlated KG policy we describe here exploit dependence to improve efficiency, the dependencies they exploit are different in kind. Variance-reduction techniques use dependence in the noise, while we use dependence between the true values of different alternatives under a Bayesian prior. Many applications admit one form of dependence without admitting the other. Other applications admit both, and although it is possible to exploit them both simultaneously, we do not treat that case here.

The use of a Bayesian framework for R&S is well-established, beginning with Raiffa and Schlaifer (1968), who consider deterministic designs for maximizing the expected value of the chosen alternative under an independent normally distributed prior. Several approximate sequential and two-stage policies exist for maximizing a quality measure applied to the chosen alternative, beginning with Gupta and Miescke (1996) and continuing with two distinct families of policies: the Optimal Computing Budget Allocation (OCBA) (Chen et al., 1996, 2000; He et al., 2007), and Value of Information Procedures (VIP) (Chick and Inoue, 2001b; Chick et al., 2007). Computational experiments (Inoue et al., 1999; Branke et al., 2007) and theoretical results (Frazier et al., 2008) demonstrate that these policies perform very well, and their sequential nature allows them to achieve even greater efficiency than could a deterministic or two-staged policy (Chen et al., 2006).

While OCBA- and VIP-based policies for exploiting common random numbers have also been introduced in Chick and Inoue (2001a) and Fu et al. (2007), to our knowledge no work has been done within the R&S literature to exploit the dependence inherent in our prior belief about the values of related alternatives. For example, in the drug discovery example described above, we believe that similar chemicals are likely to have similar effects. Our

prior should embody this belief.

Contrasting their rarity within R&S, correlated Bayesian priors have appeared frequently within Bayesian global optimization, modeling belief in the similarity of continuous functions at nearby points. Bayesian global optimization, which began with Kushner (1964) and was recently reviewed in Sasena (2002) and Kleijnen (2009), uses a Gaussian process prior to model belief about an unknown function, and then chooses experiments to most efficiently optimize that function. Algorithms often evaluate the desirability of potential measurements via a one-step Bayesian analysis, and then choose to perform a measurement whose desirability is maximal, or nearly maximal. We will employ a similar approach, but for the general class of multivariate normal priors on a finite number of alternatives.

In this article, we adopt a theoretical framework and one-step analysis introduced for a one-dimensional continuous domain with Wiener process prior by Mockus (1972) (see Mockus et al. (1978) for a description in English), and for the finite domain discrete independent normal means case by Gupta and Miescke (1996). The independent normal means case was analyzed further by Frazier et al. (2008), and extended to the unknown variance case by Chick et al. (2007). In the one-step analysis used, one computes the sampling decision that would be optimal if one were allowed to take only one additional sample, and then samples according to this decision in general. We call this the "knowledge-gradient" approach, and the resulting sampling policy the "knowledge-gradient policy". The resulting policy has also been called a Bayes one-step policy by Mockus et al. (1978), and a myopic policy by Chick et al. (2007).

Such policies operate by greedily acquiring as much information as possible with each measurement, and they work well to the extent to which this greed does not interfere with information acquisition over longer timescales. One may also compare KG policies for R&S to coordinate ascent methods for optimization, since KG policies choose the measurement that would be best under the assumption that no other alternatives will be measured, and coordinate optimization methods optimize each coordinate under the assumption that no other coordinates will be optimized. Both KG and coordinate optimization methods work well when the immediate benefits realized by their decisions are in harmony with long-term progress. While the KG approach does not work well in all information collection problems, and should be applied with care (see Chen et al. (2006) for a perfect information R&S problem for which a myopic policy required modification to perform well), it has been successfully applied to at least two other R&S problems (Frazier et al. (2008); Frazier and

Powell (2008)), and promises to produce a class of principled yet flexible algorithms for information collection.

While previous knowledge-gradient, Bayes one-step, and myopic approaches assumed either an independent normal or one-dimensional Wiener process prior on the alternatives' true means, we assume a general multivariate normal prior on a finite number of alternatives. Different statistical models and priors lead to different KG policies, and although the theoretical foundations leading to this KG policy and its progenitors are similar, the resulting policies are quite different. In comparison with the independent policies, the correlated KG policy is more computationally intensive, requiring $O(M^2 \log(M))$ computations to reach a sampling decision where M is the number of alternatives, while the independent policy requires only $O(M)$, but the correlated KG policy often requires dramatically fewer samples to achieve the same level of accuracy. In comparison with the one-dimensional Wiener process prior policy of Mockus (1972), the correlated KG policy is also more computationally intensive, but can handle more general finite alternative correlation structure, including but not limited to other kinds of one- and multi-dimensional discretized correlation structure.

We begin our discussion of the KG policy in detail in Section 2 by making explicit the correlated prior and associated model, and then, in Section 3, computing the KG policy that results from this prior. We then generalize to the correlated normal case three theoretical results that were first shown for the independent normal case in Frazier et al. (2008): the KG policy is optimal by construction when there is only one measurement left to make; the KG policy is convergent, in the sense that it always eventually discovers the best alternative if allowed enough measurements; and the suboptimality of the KG policy is bounded in the finite sample case. Finally, in Section 4, we apply the correlated KG policy to the maximization of a random function in noisy and noise-free environments, which is a problem previously considered by the Bayesian global optimization literature. We compare the correlated KG policy to two recent Bayesian global optimization methods, Efficient Global Optimization, or EGO (Jones et al., 1998), for use in the noise-free case, and Sequential Kriging Optimization, or SKO, (Huang et al., 2006), for use in the noisy case. We show that KG performs as well or better than the other methods in almost every situation tested, with a small improvement detected in the noise-free (EGO) case, and larger improvements seen in the noisy (SKO) case.

2. Model

Suppose that we have a collection of M distinct alternatives, and that samples from alternative i are normally and independently distributed with unknown mean θ_i and known variance λ_i. We will write θ to indicate the column vector $(\theta_1, \ldots, \theta_M)'$. We will further assume, in accordance with our Bayesian approach, that our belief about θ is distributed according to a multivariate normal prior with mean vector μ^0 and positive semi-definite covariance matrix Σ^0,

$$\theta \sim \mathcal{N}\left(\mu^0, \Sigma^0\right). \tag{1}$$

Consider a sequence of N sampling decisions, x^0,x^1,...,x^{N-1}. The measurement decision x^n selects an alternative to sample at time n from the set $\{1, \ldots, M\}$. The measurement error $\varepsilon^{n+1} \sim \mathcal{N}(0, \lambda_{x^n})$ is independent conditionally on x^n, and the resulting sample observation is $\hat{y}^{n+1} = \theta_{x^n} + \varepsilon^{n+1}$. Conditioned on θ and x^n, the sample has conditional distribution $\hat{y}^{n+1} \sim \mathcal{N}(\theta_{x^n}, \lambda_{x^n})$. Note that our assumption that the errors $\varepsilon^1, \ldots, \varepsilon^N$ are independent differentiates our model from one that would be used for common random numbers. Instead, we introduce correlation by allowing a non-diagonal covariance matrix Σ^0.

We may think of θ as having been chosen randomly at the initial time 0, unknown to the experimenter but according to the prior distribution (1), and then fixed for the duration of the sampling sequence. Through sampling, the experimenter is given the opportunity to better learn what value θ has taken.

We define a filtration $(\mathcal{F}^n)$ wherein $\mathcal{F}^n$ is the sigma-algebra generated by the samples observed by time n and the identities of their originating alternatives. That is, $\mathcal{F}^n$ is the sigma-algebra generated by $x^0, \hat{y}^1, x^1, \hat{y}^2, \ldots, x^{n-1}, \hat{y}^n$. We write $\mathbb{E}_n$ to indicate $\mathbb{E}[\cdot \mid \mathcal{F}^n]$, the conditional expectation taken with respect to $\mathcal{F}^n$, and then define $\mu^n := \mathbb{E}_n[\theta]$ and $\Sigma^n := \mathrm{Cov}[\theta \mid \mathcal{F}^n]$. Conditionally on $\mathcal{F}^n$, our posterior predictive belief for θ is multivariate normal with mean vector μ^n and covariance matrix Σ^n. Further discussion of the way in which μ^n and Σ^n are obtained as functions of μ^{n-1}, Σ^{n-1}, $\hat{y}^n$, and x^{n-1} is left until Section 2.1.

Intuitively we view the learning that occurs from sampling as a narrowing of the conditional predictive distribution $\mathcal{N}(\mu^n, \Sigma^n)$ for θ, and as the tendency of μ^n, the center of the predictive distribution for θ, to move toward θ as n increases. In fact we will later see that, subject to certain conditions, μ^n converges to θ almost surely as n increases to infinity.

After exhausting the allotment of N opportunities to sample, we will suppose that the experimenter will be asked to choose one of the alternatives $1, \ldots, M$ and given a reward

equal to the true mean θ_{i^*} of the chosen alternative i^*. We assume an experimenter who desires maximizing expected reward, and such a risk-neutral decision-maker will choose the alternative with largest expected value according to the posterior predictive distribution $\theta \sim \mathcal{N}(\mu^N, \Sigma^N)$. That is, the experimenter will choose an alternative from the set $\arg\max_i \mu_i^N$, attaining a corresponding conditional expected reward $\max_i \mu_i^N$. Note that a risk-averse experimenter would penalize variance and might make a different choice. We do not consider risk-aversion here.

We assume that the experimenter controls the experimental design, that is, the choice of measurement decisions $x^0, x^1, \ldots, x^{N-1}$. We allow the experimenter to make these decisions sequentially, in that x^n is allowed to depend upon samples observed by time n. We write this requirement as $x^n \in \mathcal{F}^n$. Note that we have chosen our indexing so that random variables measurable with respect to the filtration at time n are indexed by an n in the superscript.

We define Π to be the set of experimental designs, or measurement policies, satisfying our sequential requirement. That is, $\Pi := \left\{(x^0, \ldots, x^{N-1}) : x^n \in \mathcal{F}^n\right\}$. We will often write $\pi = (x^0, \ldots, x^{N-1})$ to be a generic element of Π, and we will write $\mathbb{E}^\pi$ to indicate the expectation taken when the measurement policy is fixed to π. The goal of our experimenter is to choose a measurement policy maximizing expected reward, and this can be written as

$$\sup_{\pi \in \Pi} \mathbb{E}^\pi \left[\max_i \mu_i^N\right]. \tag{2}$$

2.1. Updating equations

Since the prior on θ is multivariate normal and all samples are normally distributed, each of the posteriors on θ will be multivariate normal as well. After each sample is observed, we may obtain a posterior distribution on θ as a function of x^n, $\hat{y}^{n+1}$, and the prior distribution specified by μ^n and Σ^n. The posterior distribution is specified by μ^{n+1} and Σ^{n+1}, so to understand the relationship between the posterior and the prior it is enough to write μ^{n+1} and Σ^{n+1} as functions of x^n, $\hat{y}^{n+1}$, μ^n and Σ^n.

Temporarily supposing that our covariance matrix Σ^n is non-singular, we may use Bayes' law and complete the square (see, e.g., Gelman et al. (2004)) to write

$$\mu^{n+1} = \Sigma^{n+1}\left((\Sigma^n)^{-1}\mu^n + (\lambda_{x^n})^{-1}\hat{y}^{n+1}e_{x^n}\right), \tag{3}$$

$$\Sigma^{n+1} = ((\Sigma^n)^{-1} + (\lambda_{x^n})^{-1}e_{x^n}(e_{x^n})')^{-1}, \tag{4}$$

where e_x is a column M-vector of 0s with a single 1 at index x, and $'$ indicates matrix-transposition. Note that the new mean is found by a weighted sum of the prior mean and the measurement value, where the weighting is done according to the inverse variance. Also note that Σ^{n+1} is measurable with respect to $\mathcal{F}^n$ rather than merely $\mathcal{F}^{n+1}$.

We may rewrite the formula (4) using the Sherman-Woodbury matrix identity (see, e.g., Golub and Loan (1996)) to obtain a recursion for Σ^{n+1} that does not require matrix inversion. We can then substitute this new expression for Σ^{n+1} into (3) to obtain a new recursion for μ^{n+1} as well. Taking $x = x^n$ temporarily to simplify subscripts, the recursions obtained are

$$\mu^{n+1} = \mu^n + \frac{\hat{y}^{n+1} - \mu^n_x}{\lambda_x + \Sigma^n_{xx}} \Sigma^n e_x, \tag{5}$$

$$\Sigma^{n+1} = \Sigma^n - \frac{\Sigma^n e_x e_x' \Sigma^n}{\lambda_x + \Sigma^n_{xx}}. \tag{6}$$

The formulas (5) and (6) hold even when Σ^n is positive semi-definite and not necessarily invertible, even though the formulas (3) and (4) hold only when Σ^n is positive-definite.

We will now obtain a third version of the updating equation for μ^{n+1} which will be useful later when considering the pair (μ^n, Σ^n) as a stochastic process in a dynamic-programming context. Toward this end, let us define a vector-valued function $\tilde{\sigma}$ as

$$\tilde{\sigma}(\Sigma, x) := \frac{\Sigma e_x}{\sqrt{\lambda_x + \Sigma_{xx}}}. \tag{7}$$

We will later write $\tilde{\sigma}_i(\Sigma, x)$ to indicate the component $e_i' \tilde{\sigma}(\Sigma, x)$ of the vector $\tilde{\sigma}(\Sigma, x)$.

By noting that $\mathrm{Var}\left[\hat{y}^{n+1} - \mu^n \mid \mathcal{F}^n\right] = \mathrm{Var}\left[\theta_{x^n} + \varepsilon^{n+1} \mid \mathcal{F}^n\right] = \lambda_{x^n} + \Sigma^n_{x^n x^n}$, and defining random variables $(Z^n)_{n=1}^N$ by $Z^{n+1} := (\hat{y}^{n+1} - \mu^n)/\sqrt{\mathrm{Var}\left[\hat{y}^{n+1} - \mu^n \mid \mathcal{F}^n\right]}$, we can rewrite (5) as

$$\mu^{n+1} = \mu^n + \tilde{\sigma}(\Sigma^n, x^n) Z^{n+1}. \tag{8}$$

The random variable Z^{n+1} is standard normal when conditioned on $\mathcal{F}^n$, and so we can view (μ^{n+1}) as a stochastic process with Gaussian increments given by (8). This implies that, conditioned on $\mathcal{F}^n$, μ^{n+1} is a Gaussian random vector with mean vector μ^n and covariance matrix $\tilde{\sigma}(\Sigma^n, x^n)(\tilde{\sigma}(\Sigma^n, x^n))'$. The expression (8) will be useful when computing conditional expectations of functions of μ^{n+1} conditioned on $\mathcal{F}^n$ because it will allow computing these expectations in terms of the normal distribution.

We conclude this discussion by noting that the update (6) for Σ^{n+1} may also be rewritten in terms of $\tilde{\sigma}$ by

$$\Sigma^{n+1} = \Sigma^n - \tilde{\sigma}(\Sigma^n, x^n)(\tilde{\sigma}(\Sigma^n, x^n))' = \Sigma^n - \mathrm{Cov}\left[\mu^{n+1} \mid \mathcal{F}^n\right].$$

This expression may be interpreted by thinking of the covariance matrix Σ^n as representing our "uncertainty" about θ at time n. The measurement x^n and its result $\hat{y}^{n+1}$ removes some of this uncertainty, and in doing so alters our point estimate of θ from μ^n to μ^{n+1}. The quantity of uncertainty removed from Σ^n, which the expression shows is $\mathrm{Cov}\left[\mu^{n+1} \mid \mathcal{F}^n\right]$, is equal to the amount of uncertainty added to μ^n.

2.2. Dynamic programming formulation

We will analyze this R&S problem within a dynamic programming framework. We begin by defining our state space. As a multivariate random variable, the distribution of θ under our belief at any point in time n is completely described by its mean vector μ^n and its covariance matrix Σ^n. Thus we define our state space $\mathbb{S}$ to be the cross-product of $\mathbb{R}^M$, in which μ^n takes its values, and the space of positive semidefinite matrices, in which Σ^n takes its values. We also define the random variable $S^n := (\mu^n, \Sigma^n)$, and call it our state at time n.

We now define a sequence of value functions $(V^n)_n$, one for each time n. We define $V^n : \mathbb{S} \mapsto \mathbb{R}$,

$$V^n(s) := \sup_{\pi \in \Pi} \mathbb{E}^\pi \left[\max_i \mu_i^N \mid S^n = s \right] \quad \text{for every } s \in \mathbb{S}.$$

The terminal value function V^N may be computed directly from this definition by noting that $\max_i \mu_i^N$ is $\mathcal{F}^N$-measurable, and thus the expectation does not depend on π. The resulting expression is

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

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

$$V^n(s) = \max_{x \in \{1 \ldots M\}} \mathbb{E}\left[V^{n+1}(S^{n+1}) \mid S^n = s, x^n = x \right], \quad \text{for every } s \in \mathbb{S}. \tag{9}$$

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

$$Q^n(s, x) := \mathbb{E}\left[V^{n+1}(S^{n+1}) \mid S^n = s, x^n = x \right], \quad \text{for every } s \in \mathbb{S}.$$

We may think of $Q^n(s, x)$ as giving the value of being in state s at time n, sampling from alternative x, and then behaving optimally afterward. For a Markovian policy π, we denote by $X^{\pi,n} : \mathbb{S} \mapsto \{1 \ldots M\}$ the function that satisfies $X^{\pi,n}(S^n) = x^n$ almost surely under $\mathbb{P}^\pi$, which is the probability measure induced by π, and call this function the decision function for π. A policy is said to be stationary if there exists a single function $X^\pi : \mathbb{S} \mapsto \{1 \ldots M\}$

such that $X^\pi(S^n) = x^n$ almost surely under $\mathbb{P}^\pi$. We define the value of a measurement policy $\pi \in \Pi$ as

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

A policy π is said to be optimal if $V^n(s) = V^{\pi,n}(s)$ for every $s \in \mathbb{S}$ and $n \leq N$. The dynamic programming principle tells us that any policy π^* whose measurement decisions satisfy

$$X^{\pi^*,n}(s) \in \arg\max_{x \in \{1 \dots M\}} Q^n(s,x), \quad \text{for every } s \in \mathbb{S},\ n < N, \text{ and } x \in \{1 \dots M\}, \tag{10}$$

is optimal.

In some cases, when discussing the effect of varying the number N of measurements allowed, we make the dependence on N explicit by using the notation $V^0(\,\cdot\,; N)$ to denote the optimal value function at time 0 when the problem's terminal time is N. Similarly, $V^{\pi,0}(\,\cdot\,; N)$ denotes the value function of policy π at time 0 when the terminal time is N.

3. Knowledge gradient

We define the KG policy π^{KG} to be the stationary policy that chooses its measurement decisions according to

$$X^{KG}(s) \in \arg\max_x \mathbb{E}_n \left[\max_i \mu_i^{n+1} \mid S^n = s, x^n = x \right] - \max_i \mu_i^n \tag{11}$$

with ties broken by choosing the alternative with the smallest index. Note that $\max_i \mu_i^n$ is the value that we would receive were we to stop immediately, and so $(\max_i \mu_i^{n+1}) - (\max_i \mu_i^n)$ is in some sense the incremental random value of the measurement made at time n. Thinking of this incremental change as a gradient, we give the policy described the name "knowledge gradient" because it maximizes the expectation of this gradient. This is the same general form of the knowledge-gradient that appears in Frazier et al. (2008), and may be used together with an independent normal prior to derive the $(R_1, \dots, R_1)$ policy in Gupta and Miescke (1996). It may also be used together with a Wiener process prior to derive the one-step Bayes policy in Mockus et al. (1978).

Note that we write X^{KG} rather than the more cumbersome $X^{\pi^{KG}}$. We will also write $V^{KG,n}$ rather than $V^{\pi^{KG},n}$ to indicate the value function for the KG policy at time n. We immediately note the following remarks concerning the one-step optimality of this policy.

Remark 1. *When $N = 1$, the KG policy satisfies condition* (10) *and is thus optimal.*

Remark 2. *Consider any stationary policy π and suppose that it is optimal when $N = 1$. Then its decision function X^π must satisfy* (10)*, and hence must also satisfy* (11)*. The policy π is then the same as the KG policy, except possibly in the way it breaks ties in* (11)*. In this sense, the KG policy is the only stationary myopically optimal policy.*

The KG policy (11) was calculated in Gupta and Miescke (1996), and again more explicitly in Frazier et al. (2008), under the assumption that Σ^0 is diagonal. In this case the components of θ are independent under the prior, and under all subsequent posteriors. It was shown that in this case,

$$X^{KG}(S^n) \in \arg\max_x \tilde{\sigma}_x(\Sigma^n, x) f\left(\frac{-\left|\mu_x^n - \max_{i \neq x} \mu_i^n\right|}{\tilde{\sigma}_x(\Sigma^n, x)}\right) \quad \text{if } \Sigma^n \text{ is diagonal,} \tag{12}$$

where the function f is given by $f(z) := \varphi(z) + z\Phi(z)$, with φ as the normal probability density function and Φ as the normal cumulative density function. Furthermore, if Σ^n is diagonal then $\tilde{\sigma}_x(\Sigma^n, x) = \Sigma^n_{xx}/\sqrt{\lambda_x + \Sigma^n_{xx}}$.

In general, one may model a problem with a correlated prior, i.e., one in which Σ^0 is not diagonal, but then adjust the model by removing all non-diagonal components, keeping only $\mathrm{diag}(\Sigma^0)$. This allows using the formula (12), which we will see is easier to compute than the general case (11). We will also see, however, that the additional computational complexity incurred by computing (11) for non-diagonal Σ^n is rewarded by increased per-measurement efficiency.

3.1. Computation

We may use our knowledge of the multivariate normal distribution to compute an explicit formula for the KG policy's measurement decisions in the general case that Σ^n is not diagonal. The definition of the KG policy, (11), may be rewritten as

$$\begin{aligned} X^{KG}(S^n) &= \arg\max_x \mathbb{E}\left[\max_i \mu_i^n + \tilde{\sigma}_i(\Sigma^n, x^n) Z^{n+1} \mid S^n, x^n = x\right] - \max_i \mu_i^n \\ &= \arg\max_x h(\mu^n, \tilde{\sigma}(\Sigma^n, x)) \end{aligned} \tag{13}$$

where $h : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ is defined by $h(a, b) = \mathbb{E}\left[\max_i a_i + b_i Z\right] - \max_i a_i$, where a and b are any deterministic vectors and Z is a one-dimensional standard normal random variable. We will provide an algorithm for computing this function h as a generic function of any vectors a and b. This will allow us to compute the KG policy at any time n by substituting

μ^n for a and $\tilde{\sigma}(\Sigma^n, x)$ for b with each possible choice of $x \in \{1, \ldots, M\}$, and then choosing the x that makes $h(\mu^n, \tilde{\sigma}(\Sigma^n, x))$ largest.

Consider the function h with generic vector arguments a and b, and note that $h(a, b)$ does not depend on the ordering of the components, so that $h(\tilde{a}, \tilde{b}) = h(a, b)$ where i and j are two alternatives, $\tilde{a}$ is a but with the components a_i and a_j flipped, and $\tilde{b}$ is b but with the components b_i and b_j flipped. Thus we may assume without loss of generality that the alternatives are ordered so that $b_1 \leq b_2 \leq \ldots \leq b_M$. Furthermore, if there are two alternatives i, j with $b_i = b_j$ and $a_i \leq a_j$, $h(a, b)$ will be unchanged if we remove alternative i from both a and b. Thus we may assume without loss of generality that the ordering in b is strict so $b_1 < b_2 < \ldots < b_M$. This ordering allows us to make several remarks concerning the lines $z \mapsto a_i + b_i z$, of which we have one for each $i = 1, \ldots, M$.

Remark 3. *Let $z < w$ be real numbers and $i < j$ be elements of $\{1, \ldots, M\}$. Then, since $b_j - b_i > 0$ we have*

$$(a_i + b_i w) - (a_j + b_j w) = (a_i - a_j) - w(b_j - b_i) < (a_i - a_j) - z(b_j - b_i) = (a_i + b_i z) - (a_j + b_j z),$$

and thus, if $a_i + b_i z \leq a_j + b_j z$ then $a_i + b_i w < a_j + b_j w$.

This remark shows that the relative ordering of the lines $z \mapsto a_i + b_i z$, $i = 1, \ldots, M$, changes in a particular fashion as z increases. Taking this line of thought further, let us define a function $g : \mathbb{R} \mapsto \{1, \ldots, M\}$ by $g(z) := \max(\arg\max_i a_i + b_i z)$. This function g tells us which component $i \in \{1, \ldots, M\}$ is maximal, in the sense that its corresponding line $a_i + b_i z$ has the largest value of all the lines when evaluated at the particular point $z \in \mathbb{R}$. We break ties by choosing the largest index.

With this definition, if i is an element of $\{1, \ldots, M\}$ and $z < w$ are real numbers such that $i < g(z)$, then the component $g(z)$ satisfies $a_i + b_i z \leq a_{g(z)} + b_{g(z)} z$, and Remark 3 implies that $a_i + b_i w < a_{g(z)} + b_{g(z)} w$. Thus, $i \neq g(w)$. Since this is true for any $i < g(z)$, we have shown that $g(w) \geq g(z)$, and thus g is a non-decreasing function. Additionally, g is obtained by taking the maximum index in the argmax set, and so is itself right-continuous. Combining these facts, that g is non-decreasing and right-continuous, we see that there must exist a non-decreasing sequence $(c_i)_{i=0}^M$ of extended real numbers such that $g(z) = i$ iff $z \in [c_{i-1}, c_i)$. Note that $c_0 = -\infty$ and $c_M = +\infty$.

Observe further that if an alternative i is such that $c_i = c_{i-1}$, then $g(z) = i$ iff $z \in [c_{i-1}, c_i) = \emptyset$ implies that $g(z)$ can never equal i. Such an alternative is always less

than or equal to another alternative, and we say that it is dominated. We define a set A containing only the undominated alternatives, $A := \{i \in \{1, \ldots, M\} : c_i > c_{i-1}\}$. We will call the set A the *acceptance set*.

One algorithm for computing the sequence (c_i) and the set A is Algorithm 1, which has computational complexity $O(M)$. The algorithm may be understood as belonging to the class of scan-line algorithms (see, e.g., Preparata and Shamos (1985)), whose member algorithms all share the characteristic of scanning in one dimension without backtracking and performing operations when certain structures are encountered during the scan. In the case of Algorithm 1, it keeps counters i and j that it increments as it scans, and performs an operation whenever it encounters an intersection between lines $z \mapsto a_j + b_j z$ and $z \mapsto a_{i+1} + b_{i+1} z$. The details of the algorithm's derivation and computational complexity are given in the online supplement.

Algorithm 1 Calculate the vector c and the set A

Require: Inputs a and b, with b in strictly increasing order.
Ensure: c and A are such that $i \in A$ and $z \in [c_{i-1}, c_i) \iff g(z) = i$.

$c_0 \leftarrow -\infty$, $c_1 \leftarrow +\infty$, $A \leftarrow \{1\}$
for $i = 1$ to $M - 1$ **do**
  $c_{i+1} \leftarrow +\infty$,
  repeat
    $j \leftarrow A[end(A)]$
    $c_j \leftarrow (a_j - a_{i+1})/(b_{i+1} - b_j)$.
    if length$(A) \neq 1$ and $c_j \leq c_k$, where $k = A[end(A) - 1]$ **then**
      $A \leftarrow A(1, \ldots, end(A) - 1)$
      loopdone $\leftarrow$ false
    else
      loopdone $\leftarrow$ true
    end if
  until loopdone
  $A \leftarrow (A, i + 1)$
end for

We now compute $h(a, b)$ using the identity $\max_i a_i + b_i z = a_{g(z)} + b_{g(z)} z$, recalling that the function g is fully specified by the sequence (c_i) and the set A as computed by Algorithm 1. Since $\max_i a_i + b_i z = \max_{i \in A} a_i + b_i z$ for all $z \in \mathbb{R}$, alternatives outside A do not affect the computation of $h(a, b)$, and we may suppose without loss of generality that these alternatives have been removed from the vectors a, b, and c. To compute h, we could use the identity $h(a, b) = \sum_{j=1}^{M} a_j \mathbb{P}\{g(Z) = j\} + b_j \mathbb{E}\left[Z \mathbf{1}_{\{g(Z)=j\}}\right]$ and then calculate

$\mathbb{P}\{g(Z) = j\} = \mathbb{P}\{Z \in [c_{j-1}, c_j)\}$ and $\mathbb{E}\left[Z\mathbf{1}_{\{g(Z)=j\}}\right] = \mathbb{E}\left[Z\mathbf{1}_{\{Z\in[c_{j-1},c_j)\}}\right]$, but this leads to an expression that, while correct, is sometimes numerically unstable.

Instead, we write $a_{g(Z)} + b_{g(Z)}Z$ as the telescoping sum

$$a_{g(0)} + b_{g(0)}Z + \left[\sum_{i=g(0)}^{g(Z)-1} (a_{i+1} - a_i) + (b_{i+1} - b_i)Z\right] + \left[\sum_{i=g(Z)}^{g(0)-1} (a_i - a_{i+1}) + (b_i - b_{i+1})Z\right],$$

where only the first sum has terms if $Z \geq 0$ and only the second sum has terms if $Z < 0$.

We then apply the identity $a_{i+1} - a_i = -(b_{i+1} - b_i)c_i$ and alter the sums using indicator functions to rewrite this as,

$$a_{g(0)} + b_{g(0)}Z + \left[\sum_{i=g(0)}^{M-1} (b_{i+1} - b_i)(-c_i + Z)\mathbf{1}_{\{g(Z)>i\}}\right] + \left[\sum_{i=1}^{g(0)-1} (b_{i+1} - b_i)(c_i - Z)\mathbf{1}_{\{g(Z)\leq i\}}\right].$$

Note that $(-c_i + Z)\mathbf{1}_{\{g(Z)>i\}} = (-c_i + Z)^+$ and $(c_i - Z)\mathbf{1}_{\{g(Z)\leq i\}} = (c_i - Z)^+$ with $z^+ = \max(0, z)$ being the positive part of z. Noting that $a_{g(0)} = \max_i a_i$, we can then evaluate $h(a, b)$ as

$$\begin{aligned}
h(a,b) &= \mathbb{E}\left[a_{g(Z)} + b_{g(Z)}Z - a_{g(0)}\right] \\
&= \left[\sum_{i=g(0)}^{M-1} (b_{i+1} - b_i)\mathbb{E}\left[(-c_i + Z)^+\right]\right] + \left[\sum_{i=1}^{g(0)-1} (b_{i+1} - b_i)\mathbb{E}\left[(c_i - Z)^+\right]\right] \\
&= \sum_{i=1}^{M-1} (b_{i+1} - b_i)\mathbb{E}\left[(-|c_i| + Z)^+\right] = \sum_{i=1}^{M-1} (b_{i+1} - b_i) f(-|c_i|), \qquad (14)
\end{aligned}$$

where the function f is given as above in terms of the normal cdf and pdf as $f(z) = \varphi(z) + z\Phi(z)$. In the first equality on the third line we have used that $i \geq g(0)$ implies $c_i \geq 0$ and $i < g(0)$ implies $c_i < 0$, and that Z is equal in distribution to $-Z$. In the second equality on this line we have evaluated the expectation using integration by parts.

For avoiding rounding errors in implementation, the expression (14) has the advantage of being a sum of positive terms, rather than involving subtraction of terms approximately equal in magnitude. Its accuracy can be further improved by evaluating the logarithm of each term as $\log(b_{i+1}-b_i)+\log\varphi(c_i)+\log(1-|c_i|R(|c_i|))$, where $R(s) = \Phi(-s)/\varphi(s)$ is Mills' ratio. One can then evaluate $\log h(a, b)$ from these terms using the identity $\log\sum_i \exp(d_i) = \log(\max_j d_j) + \log\sum_i \exp(d_i - \max_j d_j)$. To evaluate $\log(1 - |c_i|R(|c_i|))$ accurately for large values of $|c_i|$, use the function log1p available in most numerical software

packages, and an asymptotic approximation to Mills' ratio such as $R(|c_i|) \approx |c_i|/(c_i^2+1)$, which is based on the bounds $|c_i|/(c_i^2+1) \leq R(|c_i|) \leq 1/|c_i|$ (Gordon (1941)).

In summary, one computes the KG policy by first computing the sequence (c_i) and the set A using Algorithm 1, then dropping the alternatives not in A and using (14) to compute $h(a,b)$. The complete algorithm for doing so is given in Algorithm 2.

Algorithm 2 KnowledgeGradient(μ^n, Σ^n)

Require: Inputs μ^n and Σ^n.
Ensure: $x^* = X^{KG}(\mu^n, \Sigma^n)$
1: **for** $x = 1$ to M **do**
2: $a \leftarrow \mu^n$, $b \leftarrow \tilde{\sigma}(\Sigma^n, x)$.
3: Sort the sequence of pairs $(a_i, b_i)_{i=1}^M$ so that the b_i are in non-decreasing order and ties in b are broken so that $a_i \leq a_{i+1}$ if $b_i = b_{i+1}$.
4: **for** $i = 1$ to $M-1$ **do**
5: **if** $b_i = b_{i+1}$ **then**
6: Remove entry i from the sequence $(a_i, b_i)_{i=1}^M$.
7: **end if**
8: **end for**
9: Use Algorithm 1 to compute c and A from a and b.
10: $a \leftarrow a[A]$, $b \leftarrow b[A]$, $c \leftarrow (c[A], +\infty)$, $M \leftarrow \text{length}(A)$.
11: $\nu \leftarrow \log\left(\sum_{i=1}^{M-1}(b_{i+1}-b_i)f(-|c_i|)\right)$
12: **if** $x = 1$ or $\nu > \nu^*$ **then**
13: $\nu^* \leftarrow \nu$, $x^* \leftarrow x$
14: **end if**
15: **end for**

To analyze the computational complexity of Algorithm 2, we note that the loop executes M times, and that within that loop, the step with the largest computational complexity is the sort in Step 3 with complexity $O(M \log M)$. Therefore the algorithm has computational complexity $O(M^2 \log M)$.

3.2. Optimality and convergence results

The KG policy exhibits several optimality and convergence properties. We only state and briefly discuss these properties here, leaving proofs and further discussion to the Online Supplement. First, as shown in Remark 1, the KG policy is optimal by construction when $N = 1$. Second, in the limit as $N \to \infty$, the suboptimality gap of the KG policy shrinks to 0. Third, for $1 < N < \infty$, we provide a bound on the suboptimality gap of the KG policy. These results extend optimality results proved in Frazier et al. (2008) for independent normal

priors. Because the prior lacks independence, the proofs of convergence and bounded finite sample suboptimality are more involved, and the statements of the theorems themselves are somewhat different than in the independent case.

The second optimality result, that the suboptimality of the KG policy shrinks to 0 as $N \to \infty$, is given in the following theorem.

Theorem 4. *For each* $s \in \mathbb{S}$, $\lim_{N\to\infty} V^0(s;N) = \lim_{N\to\infty} V^{KG,0}(s;N)$.

We refer to this property as asymptotic optimality of the KG policy, since it shows that the values of KG and optimal policies are asymptotically identical. It should be emphasized that this use of the term "asymptotic optimality" does not refer to the asymptotic rate of convergence, but only to the asymptotic equality between the two value functions. Theorem 4 is essentially a convergence result, since both the KG policy and the optimal policy achieve their asymptotic values $\lim_{N\to\infty} V^0(s;N)$ by exploring often enough to learn perfectly which alternative is best. In other words, our posterior belief about which alternative is best converges to one in which the best alternative is known perfectly.

While convergence and hence asymptotic optimality is generally easy to prove for simple non-adaptive policies like equal allocation, it is usually more difficult to prove for adaptive policies. Since non-adaptive policies like equal allocation usually perform badly in the finite sample case, the value of proving convergencing under the KG policy lies in KG's adaptive nature, and in KG's good finite sample performance in numerical experiments (see Section 4). By itself, convergence is not sufficient evidence to use a particular policy in an application, but when a policy has other good properties, convergence provides extra reassurance that it may be a good choice. We prove and discuss Theorem 4 further in Section A.2 of the Online Supplement.

The third optimality result, which provides a general bound on suboptimality in the cases $1 < N < \infty$ not covered by the first two optimality results, is given by the following theorem. This bound is tight for small N and loosens as N increases. It uses the notation $||\tilde{\sigma}(\Sigma, \cdot)||$ to indicate $\max_{x,i,j} \tilde{\sigma}_i(\Sigma, x) - \tilde{\sigma}_j(\Sigma, x)$.

Theorem 5. $V^n(S^n) - V^{KG,n}(S^n) \leq \frac{1}{\sqrt{2\pi}} \max_{x^n,\ldots,x^{N-2}} \sum_{k=n+1}^{N-1} ||\tilde{\sigma}(\Sigma^k, \cdot)||$.

A proof of this theorem is given in Section A.3 of the Online Supplement.

4. Numerical experiments

To illustrate the application of the KG policy, we consider the problem of maximizing a continuous function over a compact subset of $\mathbb{R}^d$. We will suppose that noisy evaluations of the function may be obtained from some "black box", but that each evaluation has a cost and so we should try to minimize the number of evaluations needed. This problem appears in many applications: finding the optimal dosage of a drug; finding the temperature and pressure that maximize the yield of a chemical process; pricing a product through a limited number of test markets; or finding aircraft design parameters that provide the best performance in a computer simulation. The problem is particularly well-studied in the context in which the function is evaluated by running a time-consuming simulation, as in the last of these examples, where it is known as simulation optimization. When the problem is accompanied by a modeling decision to place a Bayesian prior belief on the unknown function θ, it is further known as Bayesian global optimization.

Bayesian global optimization is a well-developed approach, dating to the seminal work of Kushner (1964). Because it is so well-developed, and contains several well-regarded algorithms, it offers a meaningful and competitive arena for assessing the KG policy's performance. We will compare the KG policy against two recent Bayesian global optimization methods that compare well with other global optimization methods: the Efficient Global Optimization (EGO) policy introduced in Jones et al. (1998), and the Sequential Kriging Optimization (SKO) policy introduced in Huang et al. (2006). Both algorithms were designed for use with a continuous domain, but can be easily adapted to the discretized version of the problem treated here.

The modeling approach generally employed in Bayesian global optimization is to suppose that the unknown function θ is a realization from a Gaussian process. Wiener process priors, a special case of the Gaussian process prior, were common in early work on Bayesian global optimization, being used by techniques introduced in Kushner (1964) and Mockus (1972). The Wiener process in one dimension is computationally convenient both because of an independence property under the posterior probability measure, and because the maximum of the posterior mean is always achieved by a previously measured point. Later work (see Stuckman (1988) as well as Mockus (1989, 1994)) extended these two methods to multiple dimensions while continuing to use the Wiener process prior.

The paths of the Wiener process are nowhere-differentiable with probability 1, which

can cause difficulty when using it as a prior belief for smooth functions. A more general class of Gaussian processes has been used for estimating mineral concentrations within the geostatistics community since the 1960s under the name kriging (see Cressie (1993) for a comprehensive treatment, and Currin et al. (1991); Kennedy and O'Hagan (2001) for a Bayesian interpretation) and it was this more general class of priors that was advocated for use by Sacks et al. (1989) and others. The EGO algorithm against which we compare uses this more general class of priors. EGO assumes the absence of measurement noise, but was extended to the noisy case by Williams et al. (2000), and then later by Huang et al. (2006), which introduced the SKO algorithm. To maintain computational efficiency, EGO and its descendants assume that the point with the largest posterior mean is one of those that was previously measured. While true under the Wiener process prior, this assumption is not true with a general Gaussian process prior.

The class of Gaussian process priors is parameterized by the choice of a mean function with domain $\mathbb{R}^d$, and a covariance function with domain $\mathbb{R}^d \times \mathbb{R}^d$. Under a Gaussian process prior so parameterized, the prior belief on the vector $(\theta(i_1), \ldots, \theta(i_K))$ for any fixed finite collection of points $i_1, \ldots, i_K$ is given by a multivariate normal distribution whose mean vector and covariance matrix are given by evaluating the mean function at each of the K points and the covariance function at each pair of points. If there are known trends in the data then the mean function may be chosen to reflect this, but otherwise it is often taken it to be identically 0, as we do in the experiments described here. The class of Gaussian process priors used in practice is usually restricted further by choosing the covariance function from some finite dimensional family of functions. In our experiments we use the class of power exponential covariance functions, under which, for any two points i and j,

$$\mathrm{Cov}(\theta(i), \theta(j)) = \beta \exp\left\{ -\sum_{k=1}^{d} \alpha_k (i_k - j_k)^2 \right\}, \tag{15}$$

where $\alpha_1, \ldots, \alpha_d > 0$ and $\beta > 0$ are hyperparameters chosen to reflect our belief. Since $\mathrm{Var}(\theta(i)) = \beta$, we choose β to represent our confidence that θ is close overall to our chosen mean function. We may even take the limit as $\beta \to \infty$ to obtain an improper prior that does not depend upon the mean function. The hyperparameter α_k should be chosen to reflect how quickly we believe θ changes as we move in dimension k, with larger values of α_k suggesting more rapid change. This class of covariance functions produces Gaussian process priors whose paths are continuous and differentiable with probability 1, and for this reason is often used for modeling smooth random functions.

In practice, one is often unsure about which hyperparameters are best, and particularly about the smoothness parameters $\alpha_1, \ldots, \alpha_d$. This ambivalence may be accommodated by placing a second-level prior on the hyperparameters. In this hierarchical setting, inference with the full posterior is often computationally intractable, so instead the maximum a posteriori (MAP) estimate of the hyperparameters is used by first maximizing the posterior likelihood of the data across the hyperparameters, and then proceeding as if our posterior belief were concentrated entirely on the values attaining the maximum. If the prior is taken to be uniform on the hyperparameters then the MAP estimate is identical to the MLE. This is the approach we apply here.

While usual approaches to Bayesian global optimization generally assume a continuous domain, the knowledge-gradient approach described herein requires discretizing it. We choose some positive integer L and discretize the domain via a mesh with L pieces in each dimension, obtaining $M = L^d$ total points. Our task is then to discover the point i in this mesh that maximizes $\theta(i)$.

We now describe in greater detail the algorithms against which we will compare KG: EGO and SKO. The EGO algorithm is designed for the case when there is no measurement noise. It proceeds by assigning to each potential measurement point an "expected improvement" (EI) given by

$$\mathrm{EI}(x) = \mathbb{E}_n \left[\max \left(\theta(x^n), \max_{k<n} \theta(x^k) \right) \mid x^n = x \right] - \max_{k<n} \theta(x^k), \tag{16}$$

and then measuring the x with the largest value of $\mathrm{EI}(x)$. In the version of the problem with a continuous domain, the above formula may be used to compute $\mathrm{EI}(x)$ for any given value of x, and then a global optimization algorithm such as the Nelder-Mead simplex search is used to search for the x that maximizes $\mathrm{EI}(x)$. In our discretized version of the problem EGO simply evaluates $\mathrm{EI}(x)$ at each of the finitely many points and measures a point attaining the maximum. If there is more than one point attaining the maximum then EGO chooses uniformly at random among them.

In the calculation (16) of $\mathrm{EI}(x)$, the term $\max_{k<n} \theta(x^k)$ is the value of the best point we have measured by time n, and is $\mathcal{F}^n$-measurable in light of the assumption of no measurement noise. The term $\theta(x^n)$ is the value of the point that we are about to measure, and is $\mathcal{F}^{n+1}$-measurable. Thus $\mathrm{EI}(x)$ is exactly the expected value of measuring at $x^n = x$ and then choosing as implementation decision the best among the points $x^0, \ldots, x^n$. This quantity is quite similar to the factor $Q^{N-1}(S^n; x)$ used by the KG policy to make its decisions, except

that $Q^{N-1}(S^n; x)$ does not restrict its potential implementation decisions to those points measured previously. Generally speaking, the points maximizing $\text{EI}(x)$ and $Q^{N-1}(S^n; x)$ are frequently distinct from one another, but they are also often close together, and so KG and EGO policies often perform similarly in those noise-free cases in which EGO can be used.

SKO is a generalization of the EGO policy to the case of non-zero measurement noise. It operates at time n by first considering a utility function, $u(x) = \mu^n(x) - c\sqrt{\Sigma^n_{xx}}$, and maximizing this over the points already measured to obtain an "effective best point", $x^{**} \in \arg\max_{x^k, k<n} u(x^k)$. Then, when considering whether to measure at some candidate point x, it calculates an augmented expected improvement function,

$$\text{EI}(x) = \mathbb{E}_n \left[\max \left(\mu^{n+1}(x) - \mu^n(x^{**}), 0 \right) \mid x^n = x \right] \cdot \left(1 - \sqrt{\frac{\lambda_x}{\Sigma^n_{xx} + \lambda_x}} \right). \tag{17}$$

The first term is essentially the expected improvement over implementing at x^{**}, and the second term is added to suggest more measurement in unexplored regions of the domain. As λ_x goes to 0, the second term goes to 1 and x^{**} goes to $\arg\max_{x^k, k<n} \mu^n_{x^k}$, and so the augmented expected improvement in (17) goes to the noise-free expected improvement in (16). In this limit, SKO behaves identically to EGO.

KG is similar to EGO and SKO in that all three do some type of one-step analysis considering the change in the expected value of the best implementation decision before and after the measurement, but KG is essentially different from EGO and SKO in its understanding that measuring at a point x^n can cause the best posterior implementation decision to be at some entirely new location not equal to any previously measured point. We illustrate this in Figure 1, where we show two posterior beliefs and the decision process of KG and EGO in each. In the first situation (two left panels), EGO prefers to measure at a point that is very close to previous measurements. EGO prefers this location because it has a large mean in comparison with the unexplored region of the function's domain. The unexplored region also has value to EGO, but not as much as does the region with large mean, as displayed by the plot of expected improvement.

In contrast, KG prefers to measure in the unexplored region. When calculating the value of measuring in this region, both KG and EGO include the potential benefit of learning that the measured point is better than the previous best point. KG, however, also includes a more subtle benefit: measurement in the unexplored region will alter the location of the posterior maximum even if the point measured is not found to be better than the previous best point.

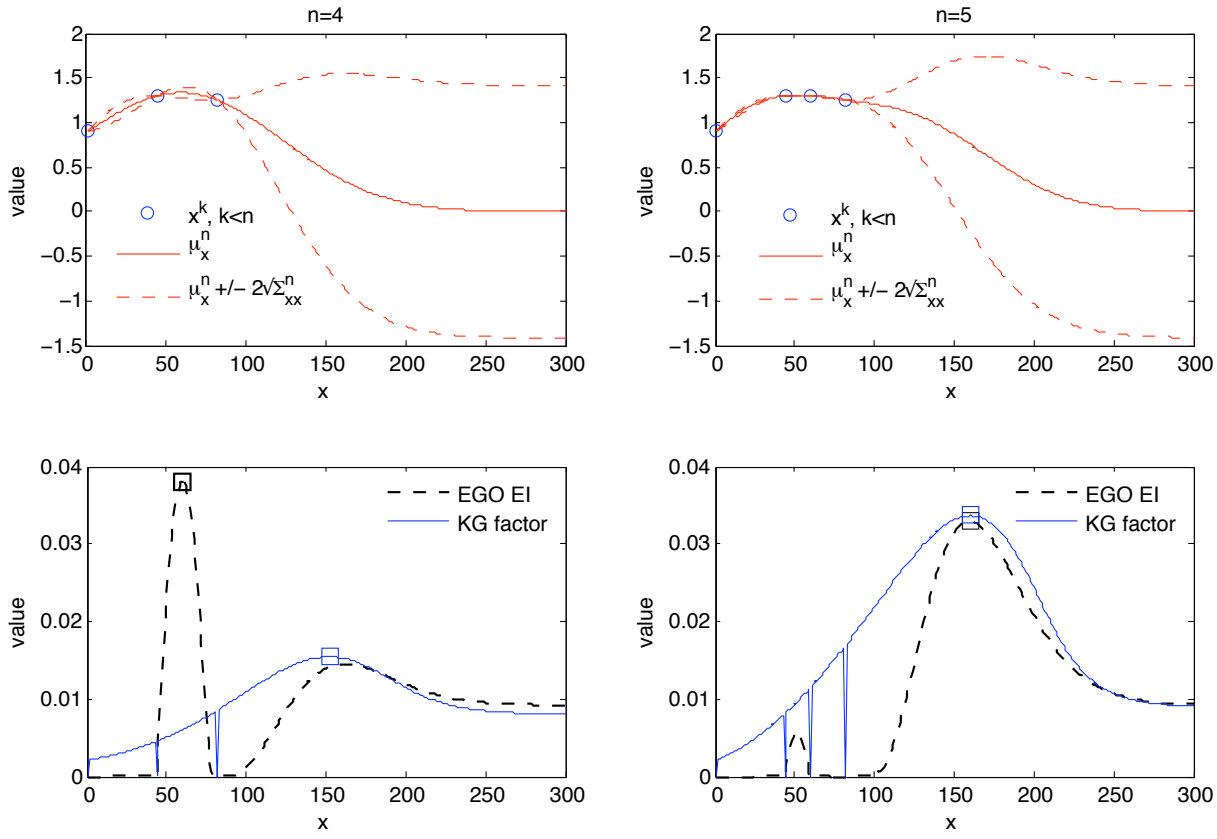

Figure 1: Upper plots display the posterior belief at two different points in time, with time $n = 4$ on the left and $n = 5$ on the right. The prior mean is plotted as a solid line, with two standard deviations above and below plotted as dotted lines. Previous measurements are circles. The $n = 5$ belief is obtained by beginning with $n = 4$ and taking the EGO decision. Lower plots display EGO's Expected Improvement quantity and KG's improvement factor $\mathbb{E}_n \left[\max_i \mu_i^{n+1} \mid x^n = x\right] - \max_i \mu_i^n$ for the corresponding belief above. The alternative that each policy would measure is marked with a square, with disagreement at $n = 4$ but agreement at $n = 5$.

If the measurement reveals the point to be worse than expected, this will shift the maximum to the left of where it was previously, and if the measurement reveals the point to be even a small amount better than expected, this will shift the maximum to the right. This shifting left and right also carries with it shifting up and down, and a positive net benefit. This added benefit is enough to convince KG to measure in the unexplored region.

Such differences in measurement decision between EGO and KG tend to cause relatively small differences in their expected performances, as demonstrated in our second set of experiments to be discussed below, with one reason pictured in the two right panels of Figure 1. Here we see the belief state resulting from the measurement advocated by EGO in the left

panel. Now both KG and EGO agree, with the point of their agreement being close to where KG wanted to measure originally. This situation, with KG and EGO choosing similar measurements, is common.

The differences between SKO and KG have as origin KG's inclusion of extra considerations into its calculation, but they also include SKO's inclusion of an extra exploration term in its calculation. The benefits provided by SKO's explicit exploration term appear to be provided implicitly by KG's full one-step analysis, and their difference in expected performance tends to be greater than is the difference between KG and EGO. This is demonstrated in our experiments below.

When estimating the hyperparameters from previous observations using the MLE and at the same time measuring according to a policy that depends upon the hyperparameters like KG, EGO or SKO, it is necessary to initially sample according to some other design to obtain a reasonable estimate of the hyperparameters, and then to switch over to the hyperparameter-dependent policy. When the measurement noise is zero, Jones et al. (1998) recommends using an initial Latin hypercube design with $2 \times$ number of dimensions measurements. When the measurement noise is unknown, Huang et al. (2006) recommends using the same Latin hypercube design with the same number of measurements followed by two additional measurements at the previously measured locations with the two best outcomes. We followed these recommendations in the experiments described here.

In our first set of experiments, pictured in Figure 2, we generated three one-dimensional random functions labeled a, b and c and discretized them into $M = 80$ points each. The three functions were drawn from Gaussian process priors with mean 0 and power exponential covariance matrices with $\beta = 1/2$ and α_1 equal to $100/(M-1)^2$, $16/(M-1)^2$ and $4/(M-1)^2$ respectively. With each truth, experiments were performed with normally distributed noise with standard deviations of 0.1 and 0.2. We compared KG with SKO, both with correlated priors, and also with KG under an independent noninformative prior.

With both correlated KG and SKO algorithms we used an initial design of 12 points as described above to obtain an initial MAP estimate of the hyperparameters, updating this MAP estimate with each sample taken. With the independent KG algorithm, we began with a noninformative prior in which the prior probability distribution on θ_i was uniform over $\mathbb{R}$, resulting in a first stage of size $M = 80$ in which each alternative was measured once in random order. Each combination of truth, noise variance and policy was replicated between 860 and 1100 times, and the opportunity cost was recorded as a function

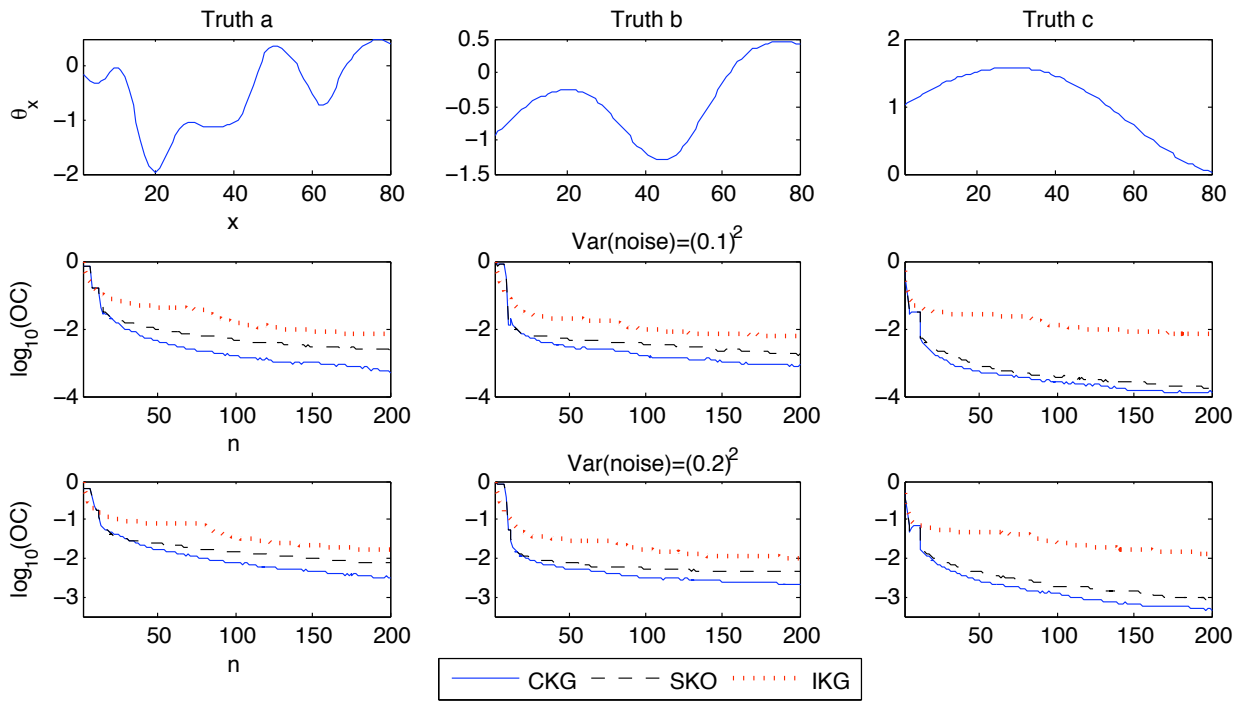

Figure 2: Comparison of correlated KG (CKG), SKO, and independent KG (IKG) on three functions drawn from Gaussian process priors. CKG and SKO policies estimated hyperparameters adaptively with an initial stage of 22 measurements. IKG used an independent noninformative prior. The top row shows the three functions tested, and the middle and bottom rows show policy performance at noise variances $(.1)^2$ and $(.2)^2$ respectively. Each policy performance plot shows the $\log_{10}$ of expected opportunity cost (OC) vs. iteration. Standard errors were too small to be plotted. The maximum in each plot of $|\log_{10}(\text{estimated OC} \pm 2 \times \text{stderr}) - \log_{10}(\text{estimated OC})|$ over all three policies is, from left to right, $.17, .12, .12$ for the second row and $.09, .07, .12$ for the third row.

of iteration n. Opportunity cost is here defined as $(\max_i \theta_i) - \theta_{i^*}$, with i^* being given by $i^* \in \arg\max_{x^k, k<n} Y_{x^k}$ during the first stage when the hyperparameters have not yet been estimated, and $i^* \in \arg\max_x \mu_x^n$ after the first stage. After the first stage, opportunity cost is the difference between the best implementation decision given perfect knowledge and the best implementation decision given the knowledge collected by the policy by time n.

The base-10 logarithm of the sample average of the opportunity costs observed over all the replications is plotted against iteration in Figure 2 for each choice of truth and noise variance. Sampled opportunity costs from batches of 25 replications were averaged together to obtain approximately normally distributed estimates of expected opportunity cost, and their sample deviations were used to estimate the error in the plotted lines, but the resulting error estimates were too small to be graphed. Instead, we state them in the caption to

Figure 2.

The figure shows that correlated KG outperformed SKO in each of the six situations tested, and at the final measurement ($n = 200$) the expected opportunity cost incurred by SKO was as much as 4.4 times larger than that incurred by KG. For the truths a, b and c respectively, the ratio of opportunity costs at the final measurement was 4.4, 2.1 and 1.3 when the measurement variance was $(.1)^2$, and 2.4, 2.0 and 1.9 when the measurement variance was $(.2)^2$.

The figure also shows that both SKO and correlated KG outperformed independent KG, often by a significant margin. The independent KG policy was shown in Frazier et al. (2008) to perform well in comparison with other R&S policies on problems with independent beliefs, and so this relative performance should be seen as a function of the correlation present in the prior, and as likely to be evidenced by other R&S policies assuming independent beliefs like OCBA and VIP. Indeed, these results show that there is often great benefit to using correlations in the prior when the problem encourages it. The margin between independent KG and the other policies is largest for truth c because it has the largest correlation across the domain. Generally, the advantage of including correlation in the prior increases as the underlying function becomes more strongly correlated. In particular, had we chosen a finer discretization level but used the same truths, independent KG would have suffered while the performance of correlated KG and SKO would have been relatively unaffected.

In our second set of experiments, pictured in Figure 3, we compare EGO and CKG. In the previous set of experiments we also examined KG and EGO performance with no measurement noise, but found no statistically significant difference between them with the number of replications we performed. Indeed, without measurement noise, the test problems were easy enough that the best point was discovered during the first stage of measurements where there is no difference between the two policies. This second set of experiments was designed with this similarity in mind to be as sensitive as possible to differences in the measurement policies. Instead of estimating expected opportunity cost for a single true function θ, we generated $26,000$ 1-dimensional functions from a Gaussian process prior, simulated each policy on each function and averaged them together to obtain expected opportunity cost under the prior. The Gaussian process prior had mean identically 0 and power exponential covariance function with $\beta = 1/2$ and $\alpha_1 = 1/64$, and discretization level $L = 200$. Also, instead of using a large first stage to adaptively estimate the hyperparameters, we restricted the first stage to a single uniformly distributed measurement, and we allowed the

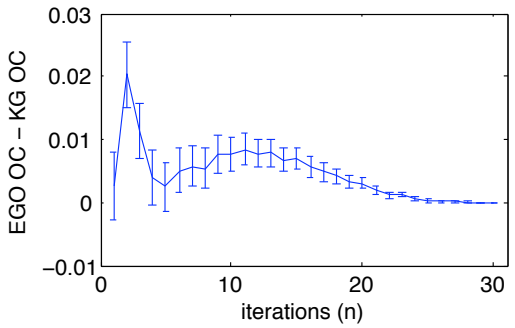

Figure 3: Comparison of KG with EGO on 26,000 one-dimensional functions drawn from a Gaussian process prior with parameters $\beta = 1/2$, $\alpha_1 = 1/64$ and $L = 200$. The plot shows the difference in expected opportunity cost between the two polices, with a positive difference indicating that KG performed better.

measurement policies to use the true hyperparameter values rather than the MAP estimate. The results, in Figure 3, show that the difference between the policies was quite small but still statistically significant, with KG performing better than EGO, and with the biggest improvement in the early iterations.

In our third and final set of experiments, pictured in Figure 4, we compared KG with SKO on several standard test functions with measurement variance $(.1)^2$. These test functions are the six-hump camelback function from Branin (1972), a "tilted" version of the Branin function from Huang et al. (2006), and the Hartman-3 function from Hartman (1973). Their functional forms, discretization levels and domains are given in the table in Figure 4. These functions are traditionally minimized, and we do so in these numerical experiments by maximizing their negative.

In all three tests the algorithms performed similarly in the first stage. Then, on the Tilted Branin and Hartman-3 functions, both KG and SKO rapidly improved their opportunity cost as the first stage ended and both their implementation decision and measurement decisions became free to range across the entire domain. KG was able to maintain this rapid improvement for longer, achieving a lower opportunity cost by approximately iteration 30 in the Tilted Branin example, and by approximately iteration 40 in the Hartman-3 example. KG then maintained this advantage through the increasing iterations.

On the six-hump camelback function, both SKO and KG algorithms suffered an initial increase in opportunity cost after the first stage in which the belief acquired by the Latin hypercube sampling combined with the Gaussian process prior led them to believe that the function was better at a point far from where they have measured previously, when in fact this belief was incorrect. Both policies quickly recovered, but SKO initially recovered more quickly than KG, outperforming it until approximately iteration 45. This may be because SKO has a greater tendency toward measuring the alternative that it would like to implement,

Name	Functional Form, Domain and Discretization Level (L)	Source
Six-hump camelback	$f(x) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4$, with $x \in [-1.6, 2.4] \times [-.8, 1.2]$ and $L = 30$.	Branin (1972)
Tilted Branin	$f(x) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10(1 - \frac{1}{8\pi})\cos(x_1) + 10 + \frac{1}{2}x_1$, with $x \in [-5.10] \times [0, 15]$ and $L = 30$.	Huang et al. (2006), modified from Branin (1972)
Hartman-3	$f(x) = -\sum_{i=1}^4 c_i \exp\left(-\sum_{j=1}^3 \alpha_{ij}(x_j - p_{ij})^2\right)$, where $\alpha = \begin{pmatrix} 3 & 10 & 30 \\ .1 & 10 & 35 \\ 3 & 10 & 30 \\ .1 & 10 & 35 \end{pmatrix}$ $c = \begin{pmatrix} 1 \\ 1.2 \\ 3 \\ 3.2 \end{pmatrix}$ $p = \begin{pmatrix} .3689 & .1170 & .2673 \\ .4699 & .4387 & .7470 \\ .1091 & .8732 & .5547 \\ .03815 & .5743 & .8828 \end{pmatrix}$, with $x \in [0, 1]^3$ and $L = 10$.	Hartman (1973)

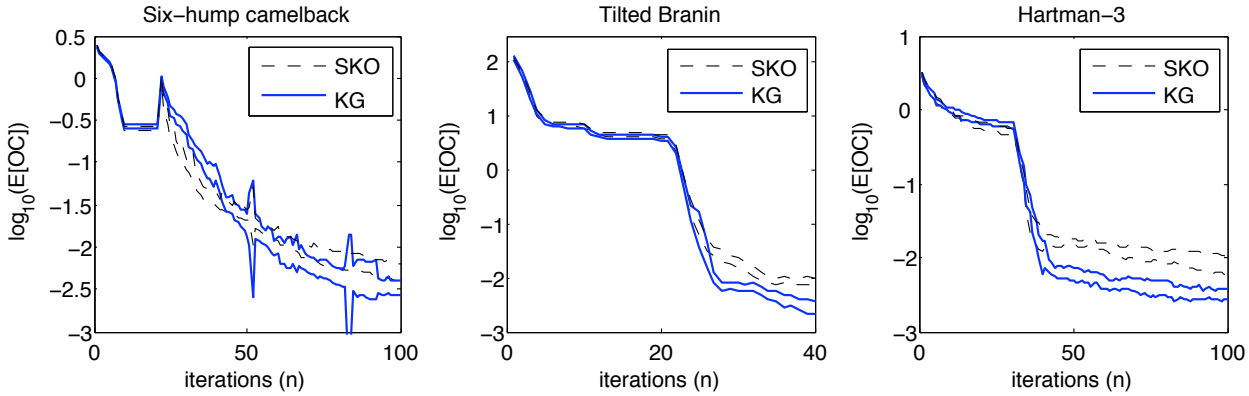

Figure 4: Comparison of KG with SKO on three standard test functions. Both policies estimated hyperparameters adaptively with an initial stage of 2 × dimension + 2 measurements. Plots show the $\log_{10}$ of the expected opportunity cost vs. iteration n at measurement variance $(.1)^2$ for three functions: Six-hump camelback (left); Tilted Branin (center); and Hartman-3 (right). Lines are plotted for each policy at $\log_{10}$(estimated OC $\pm$ 2 × stderr).

i.e., that has the largest posterior mean, and this helps to correct posterior beliefs that are incorrect in the manner described. By iteration 45 KG had recovered completely and was reducing its opportunity cost more rapidly. In the larger iterations, KG outperformed SKO.

Across these three sets of experiments, we found that KG performed well in comparison with SKO and EGO, performing as well or better than these two other policies in every situation tested except on the early iterations of the six-hump camelback test function in the second set of experiments. That KG performed well in comparison to these other Bayesian global optimization methods should not be surprising, since it is derived along similar lines but with a more complete account of the effect of a single measurement. This improved performance comes at the cost of increased complexity, however. KG requires the cross terms

of the correlation matrix, and in its current form requires discretizing the domain. These complications can dramatically increase the computational complexity of the algorithm, particularly if the discretization needs to be fine. Nevertheless, if the cost of each measurement is large enough then the computational cost of computing the KG policy will be dwarfed by the cost of measurement, and any improvement in measurement efficiency will be worthwhile.

5. Conclusion

In this article we presented a policy for sequential correlated multivariate normal Bayesian R&S, generalizing the policy presented in Gupta and Miescke (1996) and Frazier et al. (2008), which required that alternatives be independent under the prior, and generalizing the policy presented in Mockus (1972) that required the prior to be a one-dimensional Wiener process. We proved optimality of the general policy in certain special cases, and proved that it has bounded suboptimality in the remaining cases. The policy may be used effectively in applications with large numbers of alternatives for which the only way to achieve an efficient solution is by utilizing the dependence between alternatives, and its sequential nature allows greater efficiency by concentrating later measurements on alternatives revealed by earlier measurements to be among the best. Its discrete nature allows an exact calculation of the knowledge-gradient, avoiding the approximations used by other Bayesian global optimization techniques like EGO and SKO, and leading to improved performance in the cases tested.

In closing, we would like to suggest that the method we have pursued for solving the general multivariate normal sequential Bayesian R&S problem can also be applied to other sequential Bayesian R&S problems. Once a problem is formulated in the Bayesian framework, the only further requirement for applying a KG approach is that the quantity $\arg\max_x \mathbb{E}_n \left[\max_i \mu_i^{n+1}\right]$, as in (11), should be calculable exactly or approximately in an efficient manner. For example, one could assume a different prior, e.g., a hierachical multivariate normal prior whose variances are themselves random. One might also consider objectives other than the expected value of the selected alternative, such as the expected risk-averse utility of the selected alternative, or square deviation from a desired target level. In addition, an adaptive stopping rule could be used rather than a fixed sampling budget. With these and other variations in mind, we believe that the technique of posing R&S problems within a Bayesian framework and then calculating a KG policy appropriate for that framework promises practical results for a wide variety of applications.

References

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

Branin, F.H. 1972. Widely convergent method for finding multiple solutions of simultaneous nonlinear equations. *IBM J. Res. Develop* **16** 504–522.

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

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

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

Chen, C.H., J. Lin, E. Yücesan, S.E. Chick. 2000. Simulation Budget Allocation for Further Enhancing the Efficiency of Ordinal Optimization. *Discrete Event Dynamic Sys.* **10** 251–270.

Chick, S.E., J. Branke, C. Schmidt. 2007. New myopic sequential sampling procedures. Submitted to INFORMS J. on Comput.

Chick, S.E., K. Inoue. 2001a. New procedures to select the best simulated system using common random numbers. *Management Sci.* **47** 1133–1149.

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

Cressie, N.A.C. 1993. *Statistics for Spatial Data, revised edition*, vol. 605. Wiley Interscience.

Currin, C., T. Mitchell, M. Morris, D. Ylvisaker. 1991. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. of the American Statistical Association* **86** 953–963.

Frazier, P., W. B. Powell, S. Dayanik. 2008. A knowledge gradient policy for sequential information collection. *SIAM J. on Control and Optimization* **47** 2410–2439.

Frazier, P., W.B. Powell. 2008. The knowledge-gradient stopping rule for ranking and selection. *Winter Simul. Conf. Proc., 2008* .

Fu, M.C. 2002. Optimization for simulation: Theory vs. practice. *INFORMS J. on Comput.* **14** 192–215.

Fu, M.C., J.Q. Hu, C.H. Chen, X. Xiong. 2007. Simulation allocation for determining the best design in the presence of correlated sampling. *INFORMS J. on Computing* **19** 101–111.

Gelman, A.B., J.B. Carlin, H.S. Stern, D.B. Rubin. 2004. *Bayesian data analysis*. 2nd ed. CRC Press.

Golub, G. H., C. F. Van Loan. 1996. *Matrix Computations*. John Hopkins Univ. Press, Baltimore, MD.

Gordon, R.D. 1941. Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Ann. Math. Statist* **12** 364–366.

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

Hartman, J.K. 1973. Some experiments in global optimization. *Naval Res. Logistics Quarterly* **20** 569–576.

He, D.H, S.E. Chick, C.H. Chen. 2007. Opportunity cost and OCBA selection procedures in ordinal optimization for a fixed number of alternative systems. *IEEE Trans. on Sys. Man and Cybernetics Part C-Applications and Reviews* **37** 951–961.

Huang, D., T.T. Allen, W.I. Notz, N. Zeng. 2006. Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models. *J. of Global Optimization* **34** 441–466.

Inoue, K., S.E. Chick, C. Chen. 1999. An empirical evaluation of several methods to select the best system. *ACM Trans. on Model. and Comput. Simul.* **9** 381–407.

Jones, D.R., M. Schonlau, W.J. Welch. 1998. Efficient Global Optimization of Expensive Black-Box Functions. *J. of Global Optimization* **13** 455–492.

Kennedy, M.C., A. O'Hagan. 2001. Bayesian calibration of computer models. *J. of the Royal Statistical Society: Series B (Statistical Methodology)* **63** 425–464.

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

Kleijnen, J.P.C. 2009. Kriging metamodeling in simulation: A review. *European J. of*

Operational Res. **192** 707–716.

Kushner, H. J. 1964. A new method of locating the maximum of an arbitrary multipeak curve in the presence of noise. *J. of Basic Engineering* **86** 97–106.

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

Mockus, J. 1972. On bayesian methods for seeking the extremum. *Automatics and Comp.* 53–62(in Russian).

Mockus, J. 1989. *Bayesian approach to global optimization: theory and applications*. Kluwer Academic, Dordrecht.

Mockus, J. 1994. Application of Bayesian approach to numerical methods of global and stochastic optimization. *J. of Global Optimization* **4** 347–365.

Mockus, J., V. Tiesis, A. Zilinskas. 1978. The application of Bayesian methods for seeking the extremum. L.C.W. Dixon, G.P. Szego, eds., *Towards Global Optimisation*, vol. 2. North-Holland Pub. Co., North Holland, Amsterdam, 117–129.

Nelson, B.L., J. Staum. 2006. Control variates for screening, selection, and estimation of the best. *ACM Trans. on Model. and Comput. Simul.* **16** 52–75.

Preparata, F.P., M.I. Shamos. 1985. *Computational Geometry: An Introduction*. Springer.

Raiffa, H., R. Schlaifer. 1968. *Applied Statistical Decision Theory*. M.I.T. Press.

Sacks, J., W.J. Welch, T.J. Mitchell, H.P. Wynn. 1989. Design and analysis of computer experiments. *Statistical Sci.* **4** 409–423.

Sasena, M.J. 2002. Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations. Ph.D. thesis, University of Michigan.

Stuckman, B.E. 1988. A global search method for optimizing nonlinear systems. *IEEE Trans. on Sys., Man and Cybernetics* **18** 965–977.

Swisher, J.R., S.H. Jacobson, E. Yücesan. 2003. Discrete-event simulation optimization using ranking, selection, and multiple comparison procedures: A survey. *ACM Trans. on Model. and Comput. Simul.* **13** 134–154.

Williams, B.J., T.J. Santner, W.I. Notz. 2000. Sequential design of computer experiments to minimize integrated response functions. *Statistica Sinica* **10** 1133–1152.

Online Supplement to "The Knowledge-Gradient Policy for Correlated Normal Beliefs"

Peter Frazier, Warren Powell, Savas Dayanik
Department of Operations Research and Financial Engineering,
Princeton University, Princeton, NJ 08544, USA,
{pfrazier@princeton.edu, powell@princeton.edu, sdayanik@princeton.edu}

A. Optimality and convergence results

As discussed in Section 3 of the main paper, the KG policy posseses several optimality and convergence properties. First, it is optimal by construction when $N = 1$ (Remark 1). Second, the suboptimality gap between the values of the KG and the optimal policies narrows to 0 as $N \to \infty$ (Theorem 4). This is a convergence result, since it shows that when sampling under the KG policy we are guaranteed to eventually discover the alternative that is truly best. Third, the suboptimality gap is bounded for N between these two extremes (Theorem 5). Here, we discuss and prove these latter two results, discussing the convergence result in Section A.2, and the general bound on suboptimality in Section A.3. These results extend those proved in Frazier et al. (2008) for independent normal priors.

A.1. Benefits of measurement

We begin by stating the following preliminary results concerning the benefits of measurement. These results will be used later to show optimality properties of the KG policy. They show that the values of both stationary and optimal policies increase as more measurements are allowed, which is a natural result since allowing more measurements makes R&S easier.

Proposition A.1 shows that if we provide more measurement opportunities to any stationary measurement policy, then it will perform better on average.

Proposition A.1. *For any stationary policy π and state $s \in \mathbb{S}$, $V^{\pi,n}(s) \geq V^{\pi,n+1}(s)$.*

Proposition A.2 states a stronger result holding for the optimal policy, which is that if we allow it a single extra measurement of a fixed alternative, the optimal policy will perform better on average than if allowed no extra measurement at all.

Proposition A.2. *For $s \in \mathbb{S}$ and $x \in \{1, \ldots, M\}$, $Q^n(s, x) \geq V^{n+1}(s)$.*

Propositions A.1 and A.2 are similar to results proved for the independent case in Frazier et al. (2008), and the proofs contained there may be extended to the more general correlated case without undue difficulty. These proofs have been omitted due to their similarity.

Corollary A.3 then uses Proposition A.2 to show the weaker result that if the optimal policy is allowed to decide how to allocate its extra measurement then it will do better on average than if given no extra measurement at all. This is the analog of Proposition A.1, but for the optimal policy. Note that the optimal policy is not generally known to be stationary.

Corollary A.3. *For $s \in \mathbb{S}$, $V^n(s) \geq V^{n+1}(s)$.*

Proof. In Proposition A.2, take the extra measurement x to be the measurement made by an optimal policy in state s. For such an x, $Q^n(s,x) = V^n(s)$. □

A.2. Convergence and asymptotic optimality

In this section we prove Theorem 4, which states that the difference in value between the KG and optimal policies shrinks to 0 as the number of measurements, N, increases to infinity. This may be understood as convergence, in the sense that the KG policy eventually discovers the alternative that is truly the best given enough measurements. This may also be understood as asymptotic optimality, where we use the term "asymptotic optimality" to mean only that the suboptimality gap shrinks to 0 in the limit, and not that it shrinks to 0 at an optimal rate.

On its own, convergence or asymptotic optimality of a policy is little evidence of efficiency in the finite sample case. Indeed, equal allocation or any other policy measuring every alternative infinitely often will also be convergent, and many such policies do not perform particularly well. With this in mind, convergence may then be understood first as a condition we require a candidate measurement policy to possess before being willing to use it, but not one that by itself suggests a candidate policy is worth using. In this way, it is a necessary but not sufficient condition for merit. If we would like to use the KG policy because of its good finite sample performance, the convergence result then reassures us that no pernicious cases exist in which the KG policy becomes stuck measuring a proper subset of the alternatives, never discovering the best no matter how many measurements it makes.

In the case of the KG policy, it is also interesting to consider convergence and asymptotic optimality together with Remark 1, which we recall states that the KG policy is optimal when there is only one measurement left to give. Considering myopic and asymptotic optimality

together, we see that the KG policy is optimal for both immediate and distant horizons. Short- and long-term benefit are usually countervailing concerns, so it is interesting that the KG policy accommodates both simultaneously.

One may construct other policies that are both myopically and asymptotically optimal, for example by measuring according to the KG policy on the first measurement and according to the equal allocation policy on all subsequent measurements. This will be optimal when $N = 1$, and will also converge to the correct answer as $N \to \infty$, but will not necessarily be a good policy for values of N in between. Distinguishing the KG policy from such mixture policies is the fact that the KG policy is *stationary*, applying a myopic rule at each point and nevertheless still guaranteeing convergence, instead of achieving short-term optimality by behaving myopically in the early iterations, and then later switching over to a "far-sighted" rule that guarantees convergence in the limit. Remark 2 shows that, except for differences on how ties are broken, the KG policy is the only stationary policy that is both myopically and asymptotically optimal.

We begin our proof of Theorem 4 by showing in Proposition A.4 that the asymptotic value of a policy is well defined and bounded above by the value $\mathbb{E}[\max_x \theta_x]$ of learning every alternative exactly. Then, we show in Lemma A.6 that those states s for which there is no residual myopic value to be gained through any single measurement are states in which we have already achieved this upper bound on the asymptotic value. Thus any stationary measurement policy under which the limiting state has this property is asymptotically optimal. (The limiting state is shown to exist in Lemma A.5.) We then show in Lemma A.7 that if we measure an alternative infinitely often then the residual myopic value of measuring it under the limiting state vanishes. Finally, in the proof of Theorem 4 we show that the limiting state under the KG policy is one in which there is no residual myopic value in any single measurement, and thus the KG policy is asymptotically optimal. The proof centers on the notion that as an alternative is measured, the marginal value of measuring it in the future decreases to the point that the KG policy will eventually measure some other alternative.

We define the *asymptotic value function* $V(\,\cdot\,;\infty)$ by the limit $V(s;\infty) := \lim_{N\to\infty} V^0(s;N)$ for $s \in \mathbb{S}$. Below, Proposition A.4 shows that this limit exists. 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 A.4 shows that this limit also exists.

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 A.4 below, then π must be asymptotically optimal. This upper bound U corresponds to the value of an "oracle" that always knows which alternative is the best. This oracle always chooses an implementation decision in $\arg\max_i \theta_i$, and under the prior distribution given by S^0 this perfect implementation decision has expected value $U(S^0)$. The bound shown in Proposition A.4 then corresponds with our intuition that no feasible measurement policy can outperform this oracle. We will use Proposition A.4 later to show the asymptotic optimality of the KG policy.

Proposition A.4. *Let $s \in \mathbb{S}$. Then the limit $V(s;\infty)$ exists and is bounded above by*

$$U(s) := \mathbb{E}\left[\max_i \theta_i \mid S^0 = s\right] < \infty, \tag{A.1}$$

where we recall that $\theta \sim \mathcal{N}(\mu^0, \Sigma^0)$. Furthermore, $V^\pi(s;\infty)$ exists and is finite for every stationary policy π.

Proposition A.4 generalizes Proposition 5.1 from Frazier et al. (2008), and the proof found there may be easily extended to include the general correlated case. We therefore omit the proof from this article.

We now present three lemmas leading up to the main result of this section, Theorem 4.

Lemma A.5. *(S^n) converges almost surely to a random variable S^∞ in $\mathbb{S}$.*

Proof. Let $M^n = (\mu^n, \Sigma^n + \mu^n(\mu^n)')$. It is sufficient to show that M^n converges almost surely as $n \to \infty$ since $S^n = (\mu^n, \Sigma^n)$ is a linear transformation of M^n. We may write the components of M^n as the conditional expectation of an integrable random variable with respect to $\mathcal{F}^n$ by $\mu^n = \mathbb{E}_n[\theta]$, $\Sigma^n + \mu^n(\mu^n)' = \mathbb{E}_n[\theta\theta']$. This implies that M^n is a uniformly integrable martingale and hence converges (see, e.g., Kallenberg (1997) Lemma 5.5 and Theorem 3.12). □

Lemma A.5 states that the sequence of posterior distributions converges to a limiting posterior distribution. Our goal in this section is to show that this limiting posterior distribution is one in which the best alternative is known perfectly.

Lemma A.6. *Let $s = (\mu, \Sigma) \in \mathbb{S}$. If $V^N(s) = Q^{N-1}(s;x)\ \forall x$ then $V^N(s) = U(s)$.*

Proof. Fix any x. We will first show that $\tilde{\sigma}_i(\Sigma, x) = \tilde{\sigma}_1(\Sigma, x)$ for every i.

Without loss of generality we may reorder the index set $\{1, \ldots, M\}$ so that $\mu_1 = \max_i \mu_i = V^N(s)$. For a standard univariate normal random variable Z,

$$\begin{aligned} 0 = Q^{N-1}(s;x) - V^N(s) &= \mathbb{E}\left[\max_i \mu_i + \tilde{\sigma}_i(\Sigma, x)Z\right] - \mu_1 \\ &= \mathbb{E}\left[\max_i (\mu_i - \mu_1) + (\tilde{\sigma}_i(\Sigma, x) - \tilde{\sigma}_1(\Sigma, x))Z\right] + \mathbb{E}\left[\tilde{\sigma}_1(\Sigma, x)Z\right] \\ &= \mathbb{E}\left[\max_i (\mu_i - \mu_1) + (\tilde{\sigma}_i(\Sigma, x) - \tilde{\sigma}_1(\Sigma, x))Z\right]. \end{aligned}$$

This is the expectation of a non-negative random variable since the term over which the maximum is taken, $(\mu_i - \mu_1) + (\tilde{\sigma}_i(\Sigma, x) - \tilde{\sigma}_1(\Sigma, x))Z$, is 0 almost surely when $i = 1$. Thus we can write this expectation, which is known to be 0, as the integral,

$$\int_0^\infty \mathbb{P}\left\{\max_i (\mu_i - \mu_1) + (\tilde{\sigma}_i(\Sigma, x) - \tilde{\sigma}_1(\Sigma, x))Z \geq u\right\} du = 0,$$

which implies that $\mathbb{P}\{\max_i (\mu_i - \mu_1) + (\tilde{\sigma}_i(\Sigma, x) - \tilde{\sigma}_1(\Sigma, x))Z \geq u\} = 0$ for almost every u in $[0, \infty)$. Taking the limit as $u \to 0$ and using the bounded convergence theorem,

$$\begin{aligned} 0 &= \lim_{u \to 0} \mathbb{P}\left\{\max_i (\mu_i - \mu_1) + (\tilde{\sigma}_i(\Sigma, x) - \tilde{\sigma}_1(\Sigma, x))Z \geq u\right\} \\ &= \mathbb{P}\left\{\max_i (\mu_i - \mu_1) + (\tilde{\sigma}_i(\Sigma, x) - \tilde{\sigma}_1(\Sigma, x))Z > 0\right\}. \end{aligned}$$

As already noted, the random variable $\max_i (\mu_i - \mu_1) + (\tilde{\sigma}_i(\Sigma, x) - \tilde{\sigma}_1(\Sigma, x))Z$ is non-negative, so this implies that $\max_i (\mu_i - \mu_1) + (\tilde{\sigma}_i(\Sigma, x) - \tilde{\sigma}_1(\Sigma, x))Z = 0$ almost surely, which implies in turn that $\tilde{\sigma}_i(\Sigma, x) = \tilde{\sigma}_1(\Sigma, x)$ for every i.

Now fix x^n to x and define a normal random vector W with components $W_i := \mu_i^{n+1} - \mu_x^{n+1} + \theta_x$. Conditioned on $\mathcal{F}^{n+1}$, it has mean vector μ^{n+1} and covariance matrix with all entries equal to Σ_{xx}^{n+1}. We will show that W is equal in distribution to θ, with the interpretation being that the only variability left in θ is a constant translation term that affects each component equally.

Define a constant c by $c := \left(\lambda_x / \sqrt{\Sigma_{xx}^n + \lambda_x}\right) \tilde{\sigma}_1(\Sigma^n, x)$. Then, regardless of the choice of i, we have

$$\frac{\sqrt{\Sigma_{xx}^n + \lambda_x}}{\lambda_x} c = \tilde{\sigma}_i(\Sigma^n, x) = e_i' \tilde{\sigma}(\Sigma^n, x) = \frac{\sqrt{\Sigma_{xx}^n + \lambda_x}}{\lambda_x} e_i' \Sigma^{n+1} e_x.$$

Cancelling the $\sqrt{\Sigma_{xx}^n + \lambda_x}/\lambda_x$ shows that $\mathrm{Cov}\left[\theta_i, \theta_x \mid \mathcal{F}^{n+1}\right] = e_i' \Sigma^{n+1} e_x = c$, which does not depend on i. Furthermore, by choosing $i = x$ we have $c = \Sigma_{xx}^{n+1}$, and so the conditional covariance matrices of θ and W agree at $\mathcal{F}^{n+1}$. We also have agreement in the mean vectors, which are μ^{n+1} for both W and θ. Thus, since the distribution of a normal random vector

is completely determined by its mean and covariance, we must have equality in distribution between W and θ when conditioned on $\mathcal{F}^{n+1}$. We use this fact to write,

$$\begin{aligned} U(S^{n+1}) &= \mathbb{E}_{n+1}\left[\max_i \theta_i\right] = \mathbb{E}_{n+1}\left[\max_i W_i\right] = \mathbb{E}_{n+1}\left[\max_i \mu_i^{n+1} + \theta_x - \mu_x^{n+1}\right] \\ &= \max_i \mu_i^{n+1} + \mathbb{E}_{n+1}\left[\theta_x - \mu_x^{n+1}\right] = \max_i \mu_i^{n+1} = V^N(S^{n+1}). \end{aligned}$$

Finally, we use that $U(S^{n+1}) = V^N(S^{n+1})$ almost surely, together with the tower property, to complete the proof.

$$\begin{aligned} V^N(s) = Q^{N-1}(s,x) &= \mathbb{E}\left[V^N(S^{n+1}) \mid S^n = s, x^n = x\right] \\ &= \mathbb{E}\left[U(S^{n+1}) \mid S^n = s, x^n = x\right] = \mathbb{E}\left[\mathbb{E}\left[\max_i \theta_i \mid S^{n+1}\right] \mid S^n = s, x^n = x\right] \\ &= \mathbb{E}\left[\max_i \theta_i \mid S^n = s, x^n = x\right] = U(s). \end{aligned}$$ □

Lemma A.6 states that if a posterior distribution given by $s = (\mu, \Sigma)$ is such that there is no benefit gained by taking one more measurement, then the best alternative is known perfectly under this posterior distribution. We may also think of $V^N(s) = Q^{N-1}(s;x)$ as meaning that alternative x is known perfectly, and hence there is no information to be gained by measuring it. This lemma gives us a criterion by which to judge whether the limiting distribution S^∞ shown to exist in Lemma A.5 satisfies asymptotic optimality.

Lemma A.7. *If the policy π measures alternative x infinitely often almost surely, then $Q^{N-1}(S^\infty; x) = V^N(S^\infty)$ almost surely under π.*

Proof. Let $\mathcal{G}$ be the sigma-algebra generated by the collection $\left\{\hat{y}^{n+1}\mathbf{1}_{\{x^n=x\}}\right\}_{n\geq 0}$ random variables. This collection of random variables contains the information learned from the measurements of θ_x, and that information only. Since the collection has infinitely many independent measurements of θ_x with finite variance λ_x, the strong law of large numbers implies $\theta_x \in \mathcal{G}$. Then, since $\mathcal{G} \subseteq \mathcal{F}^\infty$, we have that $\theta_x \in \mathcal{F}^\infty$. Let ε be a scalar random variable equal in distribution to ε^1 but independent of $\mathcal{F}^\infty$. Then

$$Q^{N-1}(S^\infty, x) = \mathbb{E}\left[\max_i \mathbb{E}\left[\theta_i \mid \mathcal{F}^\infty, \theta_x + \varepsilon\right] \mid \mathcal{F}^\infty\right].$$

Since θ_x is measurable with respect to $\mathcal{F}^\infty$ and ε is independent of $\mathcal{F}^\infty$,

$$\mathbb{E}\left[\theta_i \mid \mathcal{F}^\infty, \theta_x + \varepsilon\right] \;=\; \mathbb{E}\left[\theta_i \mid \mathcal{F}^\infty\right].$$

Substituting this relation shows

$$Q^{N-1}(S^\infty, x) = \mathbb{E}\left[\max_i \mathbb{E}\left[\theta_i \mid \mathcal{F}^\infty\right] \mid \mathcal{F}^\infty\right] = \max_i \mathbb{E}\left[\theta_i \mid \mathcal{F}^\infty\right] = V^N(S^\infty).$$ □

Lemma A.7 is a natural consequence of the law of large numbers and shows that, if we have measured an alternative infinitely many times, there is no benefit to measuring it one more time. This will take us closer to showing that the limiting distribution S^∞ satisfies the precondition of Lemma A.6.

We now restate Theorem 4 from Section 3.2.

Theorem 4. *For each* $s \in \mathbb{S}$, $\lim_{N\to\infty} V^0(s;N) = \lim_{N\to\infty} V^{KG,0}(s;N)$.

This theorem, which states that the asymptotic value functions $V^{KG}(\,\cdot\,;\infty)$ and $V(\,\cdot\,;\infty)$ are identical, is equivalent to the statement that the KG policy is asymptotically optimal. It may also be understood primarily as a convergence result because it is equivalent to the statement that, with probability 1, the KG policy eventually learns which alternative is best.

We sketch the proof first, before proving the theorem in detail. The proof's main argument is that there can never be an alternative whose measurement would provide additional useful information under the limiting distribution achieved by the KG policy. This is because if any such alternative were to exist, it would satisfy $Q^{N-1}(S^\infty;x) < V^N(S^\infty)$ and the KG policy would prefer to measure it over some other alternative x' for which $Q^{N-1}(S^\infty;x') = V^N(S^\infty)$. Thus, among those alternatives satisfying $Q^{N-1}(S^\infty;x) < V^N(S^\infty)$, at least one gets measured infinitely often. This is a contradiction because measuring an alternative x infinitely often causes $Q^{N-1}(S^\infty;x) = V^N(S^\infty)$. We now give the full proof.

Proof of Theorem 4. Lemma A.5 shows that S^∞ exists. We will show that, under the KG policy, $V^N(S^\infty) = U(S^\infty)$ almost surely. This will imply

$$V^{KG}(S^0;\infty) = \mathbb{E}^{KG}\left[V^N(S^\infty)\right] = \mathbb{E}^{KG}\left[U(S^\infty)\right] = \mathbb{E}\left[\mathbb{E}\left[\max_i \theta_i \mid \mathcal{F}^\infty\right]\right] = \mathbb{E}\left[\max_i \theta_i\right] = U(S^0),$$

and $U(S^0) \geq V(S^0;\infty)$ by Proposition A.4. Since we also know $V^{KG}(S^0;\infty) \leq V(S^0;\infty)$, this shows $V^{KG}(S^0;\infty) = V(S^0;\infty)$ and the KG policy is asymptotically optimal.

Consider the event $H_x := \left\{Q^{N-1}(S^\infty;x) > V^N(S^\infty)\right\}$ where $x \in \{1,\ldots,M\}$. Let A be a subset of $\{1,\ldots,M\}$ and define

$$H_A := \left[\cap_{x\in A} H_x\right] \cap \left[\cap_{x\notin A} H_x^C\right],$$

where H_x^C is the complement of H_x. Since Proposition A.2 implies $Q^{N-1}(\cdot;x) \geq V^N(\cdot)$, H_A is the event that $Q^{N-1}(S^\infty;x) > V^N(S^\infty)$ for $x \in A$ and $Q^{N-1}(S^\infty;x) = V^N(S^\infty)$ for $x \notin A$. We will show that $\mathbb{P}\{H_A\} = 0$ when A is nonempty, which will imply $\mathbb{P}\{H_A\} = 1$ when A is the empty set.

Choose $A \neq \emptyset$ and let $\omega \in H_A \cap \{S^n \to S^\infty\}$. By the contrapositive of Lemma A.7 there exists a finite number $K_x(\omega)$ for each $x \in A$ such that the KG policy does not sample x for $n > K_x(\omega)$. Let $K(\omega) := \max_x K_x(\omega)$. Thus, the KG policy samples no x in A for any $n > K(\omega)$. That is,

$$x^n(\omega) \notin A \quad \forall n > K(\omega). \tag{A.2}$$

But the fact that $Q^{N-1}(S^\infty(\omega); x) > V^N(S^\infty(\omega)) = Q^{N-1}(S^\infty(\omega); y)$ for all $x \in A$, $y \notin A$, together with $S^n(\omega) \to S^\infty(\omega)$, implies that there exists $\tilde{n}(\omega) > K(\omega)$ such that

$$\min_{x \in A} Q^{N-1}(S^{\tilde{n}(\omega)}(\omega); x) > \max_{y \neq A} Q^{N-1}(S^{\tilde{n}(\omega)}(\omega); y).$$

Thus the KG policy must sample from $x \in A$ at time $\tilde{n}(\omega)$. That is, $x^{\tilde{n}(\omega)} \in A$. This contradicts our statement (A.2) that the KG policy never samples from A for $n > K_x(\omega)$ This contradiction implies that the event $H_A \cap \{S^n \to S^\infty\}$ is empty and, since $\mathbb{P}\{S^n \to S^\infty\} = 1$, we have $\mathbb{P}\{H_A\} = 0$ for our nonempty A. Therefore $\mathbb{P}\{H_\emptyset\} = 1$ and $Q^{N-1}(S^\infty; x) = V^N(S^\infty)$ almost surely for all x. Finally, by Lemma A.6, $V^N(S^\infty) = U(S^\infty)$ almost surely. □

In practice, the KG policy will begin by distributing measurements to those alternatives that early samples suggest are better. Eventually, as the variance of these better alternatives shrinks small enough, measurements will flow again to those alternatives with smaller μ_x but much larger Σ_{xx}. Measurements will flow in this fashion such that every alternative is either known perfectly in finite time through a perfect measurement or a zero variance prior, or in the limit through an infinite number of measurements.

Note that the correlated multivariate prior allows a policy to achieve asymptotic optimality without measuring an initially unknown alternative infinitely often because one may learn θ_x perfectly without measuring x if θ_x is perfectly correlated with the values of other alternatives. This is the essential difference between asymptotic optimality for the independent and correlated cases, and is the reason why the proof in Frazier et al. (2008) of the asymptotic optimality of the KG policy under an independent prior cannot be simply extended to the correlated case.

A.3. Bound on suboptimality

We have shown that the KG policy is optimal when $N = 1$ and in the limit as $N \to \infty$. In this section we prove Theorem 5, which bounds the suboptimality of the KG policy in the

intermediate region. The heart of Theorem 5 is contained in the following lemma, which bounds the marginal value of the last measurement, x^{N-1}.

The proof uses a norm $||\cdot||$ on $\mathbb{R}^M$ defined by $||u|| := \max_i u_i - \min_j u_j$. Note that this defines an operator on vectors, while the same notation $||\cdot||$ applied to the function $\tilde{\sigma}(\Sigma, \cdot)$ (a function that maps measurement decisions in $\{1, \ldots, M\}$ to vectors in $\mathbb{R}^M$) was defined in Section 3.2 by $||\tilde{\sigma}(\Sigma, \cdot)|| = \max_{x,i,j} \tilde{\sigma}_i(\Sigma, x) - \tilde{\sigma}_j(\Sigma, x)$. This previously defined notation can be written in terms of the newly defined norm on $\mathbb{R}^M$ as $||\tilde{\sigma}(\Sigma, \cdot)|| = \max_x ||\tilde{\sigma}(\Sigma, x)||$.

Lemma A.8. *Let $s = (\mu, \Sigma) \in \mathbb{S}$. Then $V^{N-1}(s) \leq V^N(s) + ||\tilde{\sigma}(\Sigma, \cdot)||/\sqrt{2\pi}$.*

Proof. Bellman's equation implies $V^{N-1}(s) = \max_{x^{N-1}} \mathbb{E}\left[V^N(S^N) \mid S^{N-1} = s\right]$. We may bound the inner term $V^N(S^N)$ by

$$\begin{aligned}
V^N(S^N) &= \max_i \mu_i^N = \max_i \left(\mu_i^{N-1} + \tilde{\sigma}_i(\Sigma^{N-1}, x^{N-1}) Z^N\right) \\
&= \left(\max_j \mu_j^{N-1}\right) + \max_i \left(\underbrace{\mu_i^{N-1} - \left(\max_j \mu_j^{N-1}\right)}_{\text{term is} \leq 0.} + \tilde{\sigma}_i(\Sigma^{N-1}, x^{N-1}) Z^N \right) \\
&\leq \left(\max_j \mu_j^{N-1}\right) + \max_i \tilde{\sigma}_i(\Sigma^{N-1}, x^{N-1}) Z^N \\
&= V^N(S^{N-1}) + \max_i \tilde{\sigma}_i(\Sigma^{N-1}, x^{N-1}) Z^N.
\end{aligned}$$

Thus, we may bound the whole expression by

$$\begin{aligned}
V^{N-1}(s) &\leq \max_{x^{N-1}} \mathbb{E}\left[V^N(S^{N-1}) + \max_i \tilde{\sigma}_i(\Sigma^{N-1}, x^N) Z^N \mid S^{N-1} = s\right] \\
&\leq V^N(s) + \max_{x^{N-1}} \mathbb{E}\left[\max_i \tilde{\sigma}_i(\Sigma^{N-1}, x^{N-1}) Z^N \mid S^{N-1} = s\right].
\end{aligned}$$

The term $\mathbb{E}\left[\max_i \tilde{\sigma}_i(\Sigma^{N-1}, x^{N-1}) Z^N \mid S^{N-1} = s\right]$ is of the form $\mathbb{E}\left[\max_i b_i' Z\right]$ where $b = \tilde{\sigma}(\Sigma^{N-1}, x^{N-1})$ and Z is a one-dimensional standard normal random variable. We have $\max_i b_i Z = (\max_i b_i)\, Z \mathbf{1}_{\{Z \geq 0\}} + (\min_i b_i)\, Z \mathbf{1}_{\{Z < 0\}}$. Thus

$$\mathbb{E}\left[\max_i b_i Z\right] = \left(\max_i b_i\right) \mathbb{E}\left[Z \mathbf{1}_{\{Z \geq 0\}}\right] + \left(\min_i b_i\right) \mathbb{E}\left[Z \mathbf{1}_{\{Z < 0\}}\right] = ||b|| \mathbb{E}\left[Z^+\right]$$

where Z^+ indicates the positive part of Z. Since $\mathbb{E}[Z^+] = 1/\sqrt{2\pi}$ we may write $V^{N-1}(s) \leq V^N(s) + \max_x ||\tilde{\sigma}(\Sigma, x)||/\sqrt{2\pi}$, completing the proof. □

The following proposition extends the bound shown in Lemma A.8 to hold when there is any number of measurements remaining.

Proposition A.9.

$$V^n(S^n) \le V^{N-1}(S^n) + \frac{1}{\sqrt{2\pi}} \max_{x^n,\ldots,x^{N-2}} \sum_{k=n+1}^{N-1} ||\tilde{\sigma}(\Sigma^k, \cdot)||$$

Proof. The proof is by induction. The base case, $n = N-1$, follows trivially. Now consider any $n < N-1$. By Bellman's equation and the induction hypothesis,

$$\begin{aligned} V^n(s) &= \max_{x^n} \mathbb{E}\left[V^{n+1}(S^{n+1}) \mid S^n = s\right] \\ &\le \max_{x^n} \mathbb{E}\left[V^{N-1}(S^{n+1}) + \max_{x^{n+1},\ldots,x^{N-2}} \sum_{k=n+2}^{N-1} ||\tilde{\sigma}(\Sigma^k, \cdot)||/\sqrt{2\pi} \mid S^n = s\right]. \end{aligned}$$

Applying Lemma A.8 to $V^{N-1}(S^{n+1})$ on the right-hand side,

$$\begin{aligned} V^n(S^n) &\le \max_{x^n} \mathbb{E}\left[V^N(S^{n+1}) + \max_{x^{n+1},\ldots,x^{N-2}} \sum_{k=n+1}^{N-1} ||\tilde{\sigma}(\Sigma^k, \cdot)||/\sqrt{2\pi} \mid S^n\right] \\ &\le \max_{x^n} \mathbb{E}\left[V^N(S^{n+1}) \mid S^n\right] + \max_{x^n,x^{n+1},\ldots,x^{N-2}} \sum_{k=n+1}^{N-1} ||\tilde{\sigma}(\Sigma^k, \cdot)||/\sqrt{2\pi}. \end{aligned}$$

Finally, noting that the first term on the right-hand side can be written as $\max_{x^n} \mathbb{E}\left[V^N(S^{n+1}) \mid S^n\right] = V^{N-1}(S^n)$ shows the result. □

We now combine this result with Proposition A.1 to bound the suboptimality of the KG policy in Theorem 5. We restate Theorem 5 here for convenience before the proof.

Theorem 5.

$$V^n(S^n) - V^{KG,n}(S^n) \le \frac{1}{\sqrt{2\pi}} \max_{x^n,\ldots,x^{N-2}} \sum_{k=n+1}^{N-1} ||\tilde{\sigma}(\Sigma^k, \cdot)||$$

Proof. Since the KG policy is optimal when $N = 1$, we have $V^{N-1}(S^n) = V^{KG,N-1}(S^n)$. Furthermore, from Proposition A.1 we have $V^{KG,N-1}(S^n) \le V^{KG,n}(S^n)$. Substituting the resulting inequality $V^{N-1}(S^n) \le V^{KG,n}(S^n)$ into Proposition A.9 shows the result. □

B. Discussion of Algorithm 1

Section 3.1 presented Algorithm 1 (reprinted here for reference) for computing the sequence (c_i) and acceptance set A needed in Algorithm 2 to compute the KG policy, but did not give the details of its derivation or its computational complexity. We present those details here.

Algorithm 1 Calculate the vector c and the set A

Require: Inputs a and b, with b in strictly increasing order.
Ensure: c and A are such that $i \in A$ and $z \in [c_{i-1}, c_i) \iff g(z) = i$.

$c_0 \leftarrow -\infty$, $c_1 \leftarrow +\infty$, $A \leftarrow \{1\}$
for $i = 1$ to $M - 1$ **do**
  $c_{i+1} \leftarrow +\infty$,
  repeat
    $j \leftarrow A[end(A)]$
    $c_j \leftarrow (a_j - a_{i+1})/(b_{i+1} - b_j)$.
    if length$(A) \neq 1$ and $c_j \leq c_k$, where $k = A[end(A) - 1]$ **then**
      $A \leftarrow A(1, \ldots, end(A) - 1)$
      loopdone $\leftarrow$ false
    else
      loopdone $\leftarrow$ true
    end if
  until loopdone
  $A \leftarrow (A, i + 1)$
end for

For ease of presentation, we first consider the case that every alternative is acceptable, so $A = \{1, \ldots, M\}$. We then have the situation illustrated in Figure B.1, and c_i (where $i \in \{1, \ldots, M - 1\}$) is simply the point where the line $a_i + b_i z$ crosses the next line in the sequence, $a_{i+1} + b_{i+1} z$. This point is $c_i = \frac{a_i - a_{i+1}}{b_{i+1} - b_i}$. Note that c_i is finite since $b_{i+1} \neq b_i$. The interior portion of the sequence (c_i), that is the portion $i = 1, \ldots, M - 1$, may be computed with a single pass through the alternatives. To complete the calculation, we set $c_0 = -\infty$ and $c_M = +\infty$.

In general, however, some alternatives will be completely dominated by others and A will not contain the full set of alternatives. This is illustrated in Figure B.2. In this more general case, if we were to calculate each c_i as simply the point where $a_i + b_i z$ crosses $a_{i+1} + b_{i+1} z$, our sequence (c_i) would occasionally decrease. To remedy the situation, we need to remove those lines that are dominated from the set A and then, for $i + 1 \in A$, compute c_j as the point at which the line $a_j + b_j z$ crosses $a_{i+1} + b_{i+1} z$, where j is the first acceptable (undominated) alternative smaller than $i + 1$. If A were the full set of alternatives, j would equal i, giving us the special case above.

Algorithm 1 accomplishes this calculation in general. In support of its analysis, we introduce a function g^i for each $i = 1, \ldots, M$ which is defined by,

$$g^i(z) = \max(\arg\max_{j \leq i} a_j + b_j z).$$

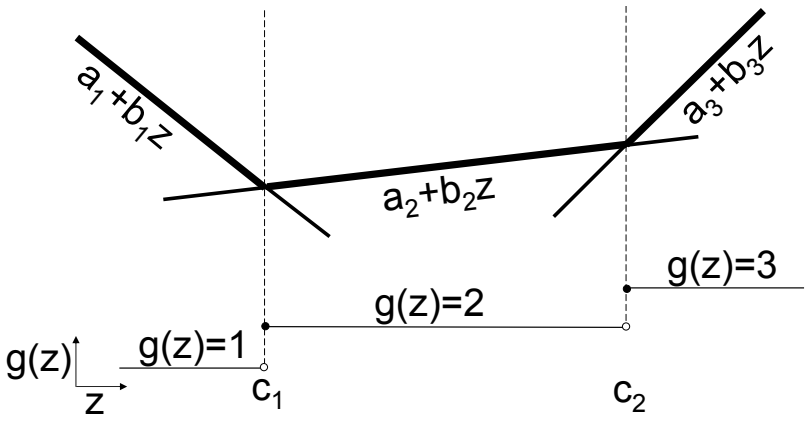

Figure B.1: Illustration of the case when $M = 3$ and no alternatives are dominated. The upper part of the illustration shows the three lines $a_i + b_i z$ for $i = 1, 2, 3$, with z ranging along the horizontal axis. The thicker portions of the lines constitute $\max_i a_i + b_i z$. The lower part of the figure shares the same horizontal z-axis, with the special points c_1 and c_2 annotated, and shows the value of $g(z)$.

At Step 2 in Algorithm 1, the vector c and the set A contain what would be the correct values if M were equal to i. That is, $g^i(z) = l \iff l \in A$ and $z \in [c_{l-1}, c_l)$. Note in particular that i is always an element of A and c_i is always equal to $+\infty$. This is because b_i is strictly the largest component of b with index less than or equal to i, and so as z becomes large enough, $g^i(z)$ will equal i.

In Steps 3 through 14 the algorithm considers adding to A the line defined by $a_{i+1} + b_{i+1} z$. It computes where this line intersects the line indexed by j, which is the undominated line with the largest index among the previously considered lines (that is, among lines with indices $\leq i$). This intersection point is c_j, and if the intersection is to the left of where line j intersects the next undominated line to the left, then line j is now dominated in this larger set of lines that now includes $i+1$. If this happens, we remove j from A in Step 8, reset j to the next undominated line to the left of $i+1$ in Step 5, and recompute where $i+1$ intersects this new j in Step 6. On the other hand, if j is still undominated even under the larger set of lines, then all previously undominated lines to the left of j also remain undominated. We add $i+1$ to the set A and loop back to Step 2.

In this way, the algorithm maintains the post-condition on Step 2 that $g^i(z) = l \iff l \in A$ and $z \in [c_{l-1}, c_l)$. Since $g^M(z) = g(z)$ and $i = M$ when the algorithm terminates, we see that $g(z) = l \iff l \in A$ and $z \in [c_{l-1}, c_l)$ at this termination time. Therefore the algorithm is correct.

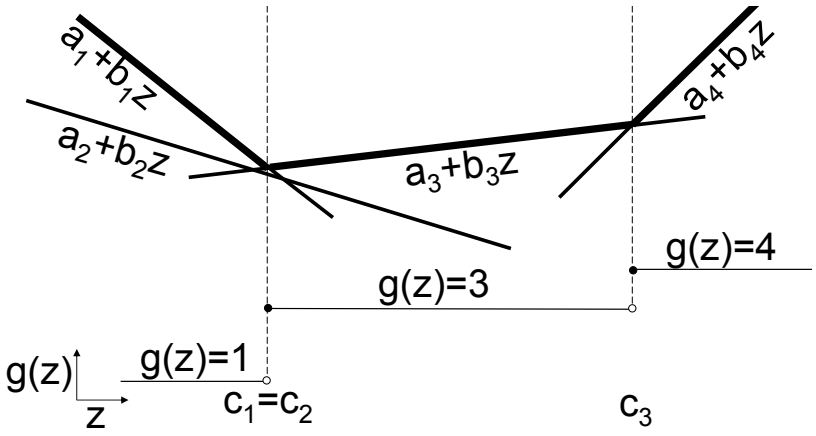

Figure B.2: Illustration of the case when $M = 4$ and alternative 2 is dominated. As in Figure B.1, the upper part of the illustration shows the lines $a_i + b_i z$ for $i = 1, 2, 3, 4$ and the lower part of the figure shows the value of $g(z)$ as a function of z. Alternative 2 is dominated because $a_2 + b_2 z$ is lower than another line for all z, which causes c_2 to be equal to c_1 and $g(z) \neq 2$ for all z.

To analyze this computational complexity of this algorithm, first note that it contains an outer loop at Step 2, and an inner loop beginning at Step 5 that optionally repeats at Step 7. Each time the inner loop repeats it removes an element from A. A total of M elements are added to A in Steps 1 and 14, and A finishes with at least one element, so the inner loop can repeat at most $M-1$ times through the course of the entire algorithm. Note that this $O(M)$ bound on inner-loop iterations is a bound on the number that take place over the course of the *entire algorithm*, and not just a bound on the number per outer loop. The outer loop clearly executes $M - 1$ times, so the maximum number of times that any statement may be executed is $2(M - 1)$. Thus, this algorithm has computational complexity $O(M)$.

References

Frazier, P., W. B. Powell, S. Dayanik. 2008. A knowledge gradient policy for sequential information collection. *SIAM J. on Control and Optimization* **47** 2410–2439.

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