Knowledge-Gradient Methods for Statistical Learning

Peter Frazier

A Dissertation

Presented to the Faculty

of Princeton University

in Candidacy for the Degree

of Doctor of Philosophy

Recommended for Acceptance

by the Department of

Operations Research and Financial Engineering

Advisor: Warren B. Powell

June 2009

© Copyright 2009 by Peter Frazier.

All rights reserved.

Abstract

We consider the class of fully sequential Bayesian information collection problems, a class that includes ranking and selection problems, multi-armed bandit problems, and many others. Although optimal policies for such problems are generally known to exist and to satisfy Bellman's recursion, the curses of dimensionality prevent us from actually computing them except in a few very special cases. Motivated by this difficulty, we develop a general class of practical and theoretically well-founded information collection policies known as *knowledge-gradient* (KG) policies. KG policies have several attractive qualities: they are myopically optimal in general; they are asymptotically optimal in a broad class of problems; they are flexible and may be computed easily in a broad class of problems; and they perform well numerically in several well-studied ranking and selection problems compared with other state-of-the-art policies designed specifically for these problems.

Acknowledgements

I am grateful to many people for their help in completing my PhD.

First, I would like to thank my advisor, Professor Warren Powell, for his ability to choose problems, his untiring availability for questions, his high expectations, and the wonderful opportunities he has created. His ability to discern the important from the unimportant continues to amaze me, and challenges me to find this level of clarity on my own.

I would also like to thank everyone at Princeton in general, and particularly within my department of Operations Research & Financial Engineering, for the marvelous atmosphere of intellectual stimulation and openness they create. I would particularly like to thank: Tim Leung, Ronny Luss, Evan Papageorgiou, Ilya Ryzhov, Vijay Krishnamurthy, and all my other fellow graduate students for their friendship and advice; Diana Negoescu for sharing her prodigious and growing mathematical abilities; and Erhan Çinlar, Savas Dayanik, Patrick Cheridito, Dave Blei, Andreas Hamel, and all the other faculty for unveiling the beautiful panoply of ideas that populate this strange but wonderful amalgam of disciplines we call operations research.

Although completing a PhD may appear to be primarily an intellectual effort, I found that the emotional challenges were equal if not greater. In am deeply indebted to a few people in particular for their help in difficult times. Most significantly, my writing this thesis, and doing the research described herein, would have been utterly impossible without my wife, Emilee. She was strong when I was weak and tired. She had confidence when I had none. Her love is the glue that binds together the moments in which I live.

I am also grateful to my parents, Mike and Cecily Frazier, for their love and untiring support, for their belief in the value of learning, and for the examples they offer through the way they live their lives. I am continually inspired by my father's integrity and energy,

my mother's wisdom and kindness, and their peaceful way of moving through the world.

Finally, I thank my children, Jack, Sophia and Olivia. Research is filled with false starts, failed attempts, and uncertainty, but caring for a child is an essentially confident effort: when wiping a tear or giving a bath, one can be sure that the labor is fruitful and necessary. I thank them for their beauty, and for the stability that the rhythms of childhood supply. These advantages are only a shadow of the great but ultimately indescribable gift that my children have given me.

To my wife, Emilee.
This work is possible because of her strength, wisdom and love.

Contents

List of Figures

Chapter 1

Introduction

Beginning my studies, the first steps
pleas'd me so much...
Walt Whitman, Leaves of Grass

This thesis considers a class of problems known collectively as *information collection problems*. By an information collection problem, we mean any problem in which one makes decisions about how much and of what type of information to collect about some unknown. We often suppose these decisions are made sequentially. An example is the problem a doctor faces in adjusting the dosages of diabetes medications for a patient. At each patient visit, the doctor collects information about the effectiveness of the currently prescribed medications and decides which mix of drugs and dosages the patient should take until the next visit, and when this visit should occur. Information collection problems appear in almost every aspect of engineering, business and science, including medicine, market research, manufacturing optimization, emergency response, neuroscience. Several example information collection problems are collected in Section 1.1.

In making information collection decisions one implicitly trades the cost of collecting the information, whether in terms of money, time, or an opportunity cost, against the benefit of the information obtained, which is the ability to make better decisions in the future. We seek to find an optimal tradeoff between this cost and benefit, and we call collecting information in a way that achieves this optimal tradeoff *optimal learning*.

Under a Bayesian framework, a sequential information collection problem can be formulated as a dynamic program and its solution given, at least in principle, by Bellman's

optimality equation (Bellman 1954). The curses of dimensionality (Powell 2007), however, prevent solving Bellman's equation except in certain special cases, and so we are encouraged to look for good sub-optimal policies. Motivated by this difficulty, this thesis develops a general class of suboptimal but practical and theoretically well-founded information collection policies known as *knowledge-gradient* (KG) policies, all originating from a simple general method that we call the knowledge-gradient method.

Because of their ubiquity in application, information collection problems have been considered by a startlingly diverse collection of distinct research communities and literatures. We now briefly review these literatures. Although the work in this thesis grew from just one of these literatures, the KG method applies to all sequential Bayesian information collection problems, and so spans all of these literatures.

The literature from which the work in this thesis originates is the ranking and selection (R&S) literature. This literature considers the R&S problem, in which we have a collection of alternatives from which we would like to select the one with the largest value. The values are initially unknown, but we have the ability to measure them with noise. Our goal in R&S is to allocate these measurements as efficiently as possible across the alternatives in order to be able to accurately choose which alternative is best with as few measurements as possible. For a fixed measurement allocation it is straightforward say which alternative should be selected, but the difficulty comes in saying which measurement allocation is best. The measurement allocation may either be made all at once, in batches (called *two-stage* or *multi-stage* R&S), or individually (called *fully sequential* R&S). The R&S problem is commonly considered for application to the optimization of Monte Carlo simulations, with a classic example (see, e.g., (Law & Kelton 2000)) being the allocation of computer time to the simulation of different (s, S) inventory policies in order to discover which gives the best performance.

The R&S literature begins with (Bechhofer 1954), with its early work summarized by the research monograph (Bechhofer et al. 1968). The predominance of the first 40 years of this research considers the problem from a non-Bayesian perspective, formulating the problem under the so-called indifference zone (IZ) framework. In this formulation of the problem, we seek policies that satisfy certain worst-case guarantees on the probability of failing to

find the best alternative, across a broad class of possible true sampling distributions. A sequence of IZ policies were proposed in (Paulson 1964, Rinott 1978, Hartmann 1991, Nelson et al. 2001). Recently, the R&S problem has been considered with greater frequency under the Bayesian formulation, in which we seek to maximize average-case performance with respect to a prior probability distribution (see the survey (Swisher et al. 2003)).

Another literature considers a problem, called the multi-armed bandit problem, that is related to the fully sequential R&S problem but that has one critical difference. In the multi-armed bandit problem we again have a collection of alternatives from which we must select one at each point in time, but now the resulting observation is more than just an observation — it is also a reward. The goal in the multi-armed bandit problem is to maximize the sum of the rewards, which may be discounted. The Bayes-optimal policy for the discounted infinite horizon version of the this problem with independent prior was discovered in (Gittins & Jones 1974) to be an index policy that takes careful advantage of the separability of the value function across alternatives. Also see (Gittins 1989, Berry & Fristedt 1985). Although index policies are generally suboptimal when the problem is altered, for example to include a finite horizon or a prior correlated across alternatives, index policies nevertheless have been found to be a useful class of policies (see, e.g., (Whittle 1988, Nino-Mora 2001)). In addition to the Bayesian approach, a substantial portion of this research has considered the problem from the non-Bayesian, and even adversarial perspectives (Auer et al. 1995).

Another literature considering information collection problems is the sequential analysis literature, which generally considers problems in which our main information collection decision is the number of samples observed, and this decision is made adaptively. An influential problem in this literature, studied in the seminal work (Wald & Wolfowitz 1948), is to adaptively choose the number of samples to take in order to decide a simple hypothesis about the underlying distribution of a sequence of independent and identically distributed samples. This problem originated from the authors' work during World War II on the problem of when to raise an alarm in the use of radar for detection of approaching planes. The optimal solution, in both Bayesian and non-Bayesian (Neyman-Pearson) formulations is the sequential probability ratio test, which adopts the simple form of a pair of thresholds on the odds-ratio between the two hypotheses. Later work expanded the set

of problems investigated to include multiple-hypothesis testing, changepoint detection, and others (Siegmund 1985), and considered applications including clinical trials (Jennison & Turnbull 2000) and statistical process control (Wetherill & Brown 1991). See (Lai 2001) for a recent review.

Closely related to the sequential analysis is the literature on sequential design of experiments, which considers problems in which we choose not only the number of observations, but also the experiments from which these observations derive. This literature is also the literature from which the multi-armed bandit problem originates. Early work in sequential design of experiments includes the dual problems of root-finding through noisy measurements of a function's value and finding the maximum of a function through noisy measurements of its gradient, first considered in (Kiefer & Wolfowitz 1952) and considered frequently since then (see (Spall 2003) for an overview). See (Wetherill & Glazebrook 1986) for a review of sequential design of experiments.

Information collection problems have been considered in still other literatures, including: design of experiments (Box et al. 1978); decision analysis (Howard 1966); economics (Weitzman 1979); team decision theory (Marschak & Radner 1972); global optimization (Jones 2001); computer vision (Renninger et al. 2007); search theory (Stone 1975); active learning (Cohn et al. 1996); and medical diagnosis within artificial intelligence (Kapoor & Greiner 2005).

An unfortunate consequence of this diversity of literatures is the difficulty in providing a comprehensive review, creating a tendency to reinvent and rediscover ideas in multiple literatures. Indeed, individual KG policies have been proposed in the context of specific problems by other authors, including (Gupta & Miescke 1996) in the context of the independent normal R&S problem with known sampling variance, and (Mockus 1972) in the context of Bayesian global optimization with a Wiener process prior. By proposing and analyzing the KG method as a general purpose method, we hope to tie together the broad spectrum of information collection problems, and perhaps to induce cross-fertilization between essentially separate literatures, bringing techniques and insights from one community to bear on the often closely related problems considered by others.

In addition to their generality, KG policies have several attractive qualities that make

them worthy of analysis and application: they are myopically optimal in general; they are asymptotically optimal in a broad class of problems; they are flexible and may be computed easily in a broad class of problems; and they perform well numerically in several well-studied ranking and selection problems compared with other state-of-the-art policies designed specifically for these problems.

1.1 Examples of Information Collection Problems

Information collection problems are incredibly common in application. In this section we give examples, classifying them according to the type of the information collection decisions involved. The difficulty of the problem depends in part on the size of the decision space, but also on the complexity of the underlying statistical model. Some of these examples will be considered more fully later, and others are included to underscore the broad scope encompassed by the class of information collection problems.

1.1.1 Stopping decisions

In this first set of problems, the information collection decision is a stopping time that adaptively chooses how much information to collect.

1. **Early warning systems:** When using radar during wartime, given that we observe a suspicious signal, we would like to decide between raising an alarm or waiting and collecting more information through observation over a longer period of time. See (Wald & Wolfowitz 1948).

2. **Statistical process control:** When operating an industrial chemical process and periodically observing the quality of samples pulled from the production line, we would like to stop the process if poor sample quality indicates a production problem. We would like to stop the process as soon as possible when there is a problem, while waiting long enough to control the probability of false alarm under normal fluctuations. See (Wetherill & Brown 1991).

3. **Group sequential clinical trials:** We would like to adaptively choose the length of time to run a clinical trial. Adaptively choosing the duration allows us to stop the

trial early and immediately make the treatment available to the general public if the treatment being tested performs very well on the initial set of patients, saving lives. If the initial results are equivocal we can run the trial longer instead, and if the initial results are very poor we can stop the trial early and declare the drug a failure, saving the cost of a longer trial. See (Jennison & Turnbull 2000).

4. **Drug recall:** While monitoring the incidence of side effects from a drug approved for use in the population, upon observing a few suspicious events we would like to decide whether to recall the drug. By waiting, we collect more information about the incidence of side effects, allowing us to know with greater accuracy whether the incidents observed were due to chance or indicate a real problem, but we also risk more danger to the population.

5. **Technology adoption:** When operating a business, we would like to decide whether to adopt a new technology (such as a new computer system or manufacturing technology) now, or to wait and gather more information about how well it works by observing the experience of other early-adopters.

1.1.2 Decisions from a finite set

In the following information collection problems, the information collection decision at each time period comes from a finite set.

1. **Screening of shipping containers:** We would like to screen shipping containers entering a port for dangerous nuclear materials. We have a number of examination techniques at our disposal, including various types of passive imaging (e.g., observation of gamma rays) and active interrogation (e.g., with a neutron beam), as well as manual inspection of containers. Each examination technique has its own costs and unique detection abilities. We would like to design a protocol to guide our application of these techniques to maximize the probability of detecting dangerous containers while minimizing cost and delay of containers traveling through the port. See (Madigan et al. 2007).

2. **Outsourcing:** When making the decision of whether or not to outsource a project to another firm, we would like to account for the fact that, by keeping the project in-house, employees in our firm will learn and develop expertise that will be useful in future projects. See (Anderson Jr & Parker 2002).

3. **Manufacturing optimization:** We would like to choose the best assembly line configuration from among several related choices to maximize manufacturing efficiency. We have the ability to collect information about the quality of any given configuration from a discrete event simulation of the manufacturing operation, or from using the configuration in practice.

4. **Drug screening:** Upon having identified a target pathway important in the function of a disease, e.g., a protein interaction, we would like to screen a large number of unrelated compounds (usually between 10^3 and 10^6) using automated high-throughput in vitro experiments in order to find a relatively small number of compounds with the ability to disrupt this pathway.

5. **Collaborative filtering:** We would like to use collaborative filtering to suggest books, movies, or music to users according to their underlying but unknown preferences. After viewing an item, the user reports his or her enjoyment of that item with some probability. We may suggest an item because it possesses a high probability of being enjoyed by the user, but we may also suggest an item to learn more about the user's preferences.

6. **Emergency vehicle routing:** We would like route ambulances from their origin to patients and then to the hospital in a minimal amount of time. Before an ambulance leaves, or while it is in the early part of its route, we have the option of asking a traffic helicopter to observe traffic flows along links in the transportation network to choose a route with the smallest possible travel time. See (Ryzhov & Powell 2009).

1.1.3 Continuous univariate decisions

In the following information collection problems, the information collection decision is chosen from a one-dimensional continuous or discrete space. Because the space has only a

single dimension, it can generally be discretized if it was continuous and then truncated to approximate the problem with a finite decision set of manageable size.

1. **Product pricing:** We would like to dynamically vary the price of a product to maximize revenue while learning about demand at different prices. Similarly, we might like to vary the price in test and target markets in order to learn about customer demand and better choose a final market price at which to sell the product.

2. **Venture capital:** We learn about the profitability of startup companies with various characteristics by observing how other companies have fared in the past. We may learn from startup companies funded by other firms, but this may not be possible if our firm provides a large fraction of the funding in our sector, or if the information about companies funded by other firms is delayed in becoming available. At each point in time we must decide how much money to allocate to each company requesting funding, in order to maximize both short-term profit and our information about the marketplace, which will increase future profits.

3. **Sequential planning:** We would like to discover whether a new experimental technique for growing samples in a biology lab produces better results than a standard technique. We have the option of growing samples under the new and standard techniques in batches. There is a large cost paid per batch due to the time required to grow the samples, and a smaller cost paid per sample included due to the experimentalist's time preparing it. The decision at the completion of each batch is whether to continue with another batch, and if so, how many samples to include. See (Schmitz 1993, Schmegner & Baron 2004).

4. **Supply chain management:** We would like to make inventory decisions in a supply chain, for which the distribution of customer demand is unknown and changing over time. Our observations of demand come only through the amount of product we sell, and demand lost due to stockout is unobserved. By stocking more product, we collect more information about demand because stockouts are less frequent, but we may pay more in holding costs or unused product. See (Lariviere & Porteus 1999, Negoescu et al. 2008).

5. **Fishery regulation:** We would like to set limits on fishing low enough to maintain the fish population, without unnecessarily reducing profits in the fishing industry. The current fish population is only partially known, as are the dynamics governing it, and one of the main methods with which to learn about them is through the returns from the catch. Each year we set limits on the catch, keeping in mind that we will have less information later about the fish population if we set the limits low now. See (Tomberlin 2008).

6. **Drug dosing:** In clinical practice, we would like to escalate the dosage of a drug to a therapeutic level as quickly as possible without causing undue side effects. Similarly, in a clinical trial, we would like to find the maximum tolerated dose of a chemotherapy drug as quickly as possible, while minimizing toxic effects to participants. See (Eichhorn & Zacks 1973, Babb et al. 1998).

1.1.4 Continuous multivariate decisions

In the following information collection problems, the information collection decision is chosen from a multidimensional continuous space. The dimension of the space is often large enough that discretization is computationally infeasible.

1. **Oil exploration:** We would like to discover the best place at which to drill a commercial oil well by drilling a sequence of exploratory test wells. We would like to find a good location with as few exploratory wells as possible. See (Bickel & Smith 2006).

2. **Chemical process optimization:** We would like to choose the inputs to an industrial chemical process to maximize the quality of the output. For example, we might be choosing the temperature, pressure, and mixture of gases to use when etching silicon wafers. See (Myers & Montgomery 2002).

3. **Visual search in computer vision:** We would like to design a computer vision algorithm that adaptively chooses where to point a video camera in order to quickly and effectively find a particular object. See (Renninger et al. 2007).

4. **Physical search:** We would like to direct one or more search teams in order to

find a lost person, e.g., a hiker, or a lost object, e.g., a sunken submarine. See (Stone 1975, Chudnovsky 1988).

5. **Emergency response:** Following the release of radioactive material in a major city, we would like to identify contaminated areas using an efficient sequential method for testing contamination at points within the city. At each point in time our decision contains the areas to which our inspectors will travel and investigate next.

6. **X-ray crystallography:** We would like to crystallize a protein in order to interrogate its 3-dimensional structure using X-ray crystallography. Many proteins crystallize only under a delicate and unique set of experimental conditions. Given a particular protein, we would like to search for appropriate experimental conditions, trying these conditions sequentially until we find one that works.

7. **Model calibration:** We have a time-consuming computational model for which we would like to find parameters that best fit observed data. We would like to search through the space of input parameters to the computational model, running the model on parameters of our choosing and learning how well their outputs fit data. See (Frazier & Powell 2009*a*).

1.1.5 Decisions with combinatorial structure

In the following information collection problems, each information collection decision has combinatorial structure, often because it is drawn from the set of subsets of some larger finite set.

1. **Drug combination therapy:** Given a particular patient and his or her medical characteristics, a doctor would like to decide what drug or combination of drugs to prescribe. See (Ryzhov et al. 2008*a*) for an example motivated by diabetes treatment.

2. **Drug design:** We would like to search among chemically similar derivatives of a molecule to find the one that is best at binding to a target protein. These molecules are specified by the chemical functional groups chosen from a larger set and placed at

fixed substituent locations. At each point in time, we decide which molecule to test, and then collect information about its effectiveness.

3. **Product feature selection:** We would like to choose the best set of features to include in a product, e.g., a cell phone, to maximize profitability. To determine consumer preferences for features we may conduct a sequence of focus groups in which we ask participants to judge the quality of feature subsets of our choosing.

4. **Assortment planning:** We would like to choose the best composition of products in a product line to simultaneously satisfy demand from several market segments. We may learn about demand from focus groups or from purchases of products currently offered in the product line.

5. **Pricing of credit default swaps:** We have a mathematical model of credit default swap prices parameterized by a clustering of firms into categories (Papageorgiou & Sircar 2009), under which prices of liquidly traded assets are predicted by taking an expectation via many replications of a Monte Carlo simulation. In order to price more exotic credit derivatives, we would like to calibrate the model by finding a clustering of firms that accurately replicates market prices.

6. **Research and development portfolio selection:** We would like to allocate research funding among the set of proposed research projects to maximize the expected value to society of the resulting research. See (Hannah et al. 2009).

7. **Network travel times:** We would like to find the path with the shortest expected travel time through a transportation or communications network by measuring travel times along paths of our choosing. We might have the ability to observe individual travel times along links in the path, or perhaps only the total travel time.

1.2 Offline and online information collection problems

In addition to classifying problems according to the form of the information collection decision, we may also classify them according to whether the rewards are earned throughout time, in which case we call the problem *online*, or entirely at the final time, in which case

we call the problem *offline*. To be more exact, an offline problem is one in which the total reward earned may be written entirely as a function of the decision made at the final time and the true state of the world, with decisions at earlier times affecting only the information available to the decision-maker at this final time. An online problem is any problem not meeting this offline definition.

In an offline information collection problem, there is a clear distinction between learning and implementation phases. In the learning phase, one performs a sequence of tests, collecting information about the results. Then, after the learning phase is complete, a single decision is made based on the results of the tests and no more learning occurs. This final decision is called the implementation decision. Note that R&S problems with fixed time-horizons are offline problems. As another example, both the drug screening and drug design problems (item 4 in Section 1.1.2 and item 2 in Section 1.1.5) are offline problems, where the learning phase is the period of time we spend doing laboratory tests on molecules, and the implementation decision is the decision of which molecule to send to the next stage of drug development.

In an online problem there is no clear distinction between learning and implementation phases. Instead, they occur simultaneously, and one has the opportunity to learn about the state of the world while simultaneously acting within it. For example, the supply chain management problem (item 4 in Section 1.1.3) is an online problem.

Chapters 2, 4, and 5 consider offline problems, while Chapter 3 considers an online problem. Although the problem considered in Chapter 3 is technically an online problem, it shares much with the offline problems considered because the time before stopping is similar to an offline learning phase, with the difference being that we pay a cost (or negative reward) for each measurement we take and thus rewards depend upon decisions made before the final time.

1.3 KG Policies

In general, both for offline and online problems, KG policies are derived by first considering a stochastic optimization problem in which we are acting in the world while simultaneously

collecting information about it. The KG policy is constructed by supposing that observations resulting from actions at the current point in time will constitute our last opportunity to learn, and that we will continue to act later without any additional learning. The KG policy is defined to be the policy that is optimal under this supposition. This construction balances the need to explicitly account for the role of information in the problem against the computational burden of doing so.

In an offline context this general construction simplifies, and the KG policy is computed by considering the myopic gain in information due to a sampling decision. This myopic gain is computed under the assumption that the current measurement will be the last, and the implementation decision must be made immediately afterward. The value of each particular measurement is given by the expected difference in value between the best final implementation decisions that could be made before and after the measurement.

To state this construction of the KG policy for offline problems more formally, we give the following general formulation of a sequential offline Bayesian information collection problem:

1. Begin with a prior distribution on some unknown truth θ.

2. Choose a sequence of measurements $x^0, x^1, \ldots$ and obtain the corresponding sequence of observations, choosing each new measurement based on all previous observations.

3. After N measurements, choose an implementation decision i and earn a reward $R(\theta, i)$, where R is some known function.

The KG policy for any problem within this general offline framework chooses its next decision x^n according to

$$\arg\max_x \mathbb{E}_n \left[\max_i \mu_i^{n+1} \mid x^n = x \right] - \max_i \mu_i^n,$$

where $\mu_i^n = \mathbb{E}_n \left[R(\theta; i) \right]$ is the expected value of implementation decision i given what we know at time n, and μ^{n+1} is defined similarly. With this definition, $\max_i \mu_i^n$ is the expected value of the best implementation decision we can make given our current (time-n) information, $\mathbb{E}_n \left[\max_i \mu_i^{n+1} \mid x^n = x \right]$ is the expected value of the best implementation decision we can make given the new information provided by x, and their difference approximates

the value of the information. This use of the difference in value between the best decision that can be made with and without a new piece of knowledge is the reason for the name knowledge-gradient.

KG policies for particular information collection problems have been introduced in the past. In particular, (Gupta & Miescke 1996) proposed the KG policy for the independent normal R&S problem described in Chapter 2. This policy is discussed in much greater detail in that chapter. For the multivariate normal R&S problem described in Chapter 4 with the special case of a Wiener process prior, the KG policy was introduced by (Mockus 1972). For the independent normal R&S problem with unknown sampling variance, the KG policy was introduced by (Branke et al. 2007) and called the myopic policy. Although KG policies have been previously introduced for specific problems, the contribution of this thesis is to tie these policies together into a collective framework that includes a broad and cohesive class of methods.

1.4 Thesis Organization

We now review the broader organization of the rest of this thesis, and summarize the contents of each chapter and appendix.

Chapter 2

Chapter 2 considers the KG policy for the most well-studied Bayesian R&S problem, in which measurements are normally distributed with known variance around the alternatives' true values, and the prior on the alternatives' true values is normal with independent components. This policy was proposed by (Gupta & Miescke 1996) as the simplest of a broad collection of policies but has been largely ignored. In its place, a number of more complicated policies had been introduced for the same and similar problems. Chapter 2 demonstrates with theoretical results and numerical examples the quality of the KG policy for this problem. In particular, the KG policy is optimal by construction when a single measurement opportunity remains, and satisfies a bound on suboptimality that grows linearly with the number of measurements remaining. This would seem to indicate that the KG

policy is worth using when the number of measurements remaining is small. In addition, Chapter 2 proves that the KG policy is asymptotically optimal in the limit as the number of measurements remaining grows large. Thus we have optimality results in both extremes of the number of measurements remaining. In numerical examples, the KG policy also performs as well or better than the other policies designed for this well-studied problem in almost all scenarios. This chapter is based on (Frazier et al. 2008*b*).

Chapter 3

Chapter 3 shows how the KG method may be extended to the case of normally distributed measurements with unknown variance, where in addition one pays a fixed cost for each measurement and stops when one sees fit, rather than only being a fixed number of allowed measurements. (Branke et al. 2007) considered how the KG method could be applied to choose the alternative to measure in this context, but had used a different stopping rule. Chapter 3 shows that by stopping according to the KG method one improves performance. This chapter is based on (Frazier & Powell 2008*a*).

Chapter 4

Chapter 4 considers R&S problems in which beliefs about alternatives' values are correlated with each other (the measurement noise is assumed uncorrelated). By including correlation of beliefs into the policy, we can dramatically increase efficiency over traditional R&S methods. For example, when testing a collection of chemical compounds to find the one that is best at treating a disease, we can use the fact that chemically similar compounds usually behave similarly to dramatically improve efficiency. With these correlations we can consider millions of different chemical compounds even if we can only test a small fraction of these. This would be completely impossible using a R&S method with independent beliefs. In such situations, correlated beliefs are critical to successful applications. We show how the KG policy can be computed for this problem and that it is both myopically and asymptotically optimal. We then specialize to a particular correlated problem for which other policies have been developed, Bayesian global optimization of continuous functions, and show that the KG policy performs very well compared with these other policies. This chapter is based on

(Frazier et al. 2008*a*).

Chapter 5

Chapter 5 generalizes to consider all offline sequential Bayesian information collection problems. Under the assumptions that the sampling distribution is from an exponential family and that the number of distinct measurement types is finite, we give sufficient conditions for an adaptive sampling policy to achieve asymptotic optimality. Here, asymptotic optimality is understood to mean that the limit of the expected loss under the given sampling policy as the number of measurements allowed grows to infinity attains the minimum over all possible sampling policies. This property is important because it ensures convergence in the limit for sophisticated policies designed to maximize performance over the short-term. We then apply these sufficient conditions to show asymptotic optimality of three previously proposed R&S policies. We also show how this sufficient condition may be generally applied to a broad class of KG policies. This chapter is based on (Frazier & Powell 2008*b*).

Chapter 6

Chapter 6 summarizes the contributions of this thesis and describes ongoing work in KG methods by this author and others. In particular, it describes new applied problems for which KG methods are being used, new methodological work underway to apply them to online learning problems including variations of the multi-armed bandit problem, and new theoretical work that attempts to characterize their limitations.

Chapter 2

The Knowledge-Gradient Policy for Independent Normal Beliefs

> I returned, and saw under the sun, that the race is not to the swift, nor the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding, nor yet favour to men of skill; but time and chance happeneth to them all.
>
> Ecclesiastes 9:11

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, and was first discussed in Chapter 1.

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

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

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

(iii) Selecting the best of several policies applied to a stochastic Markov decision process. The policies may only be evaluated through Monte Carlo simulation so a method of R&S is needed to determine which policy is best. This selection may be as part of a larger algorithm for finding the optimal policy as in Evolutionary Policy Iteration (Chang et al. 2007).

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

In this chapter we study a measurement policy introduced in (Gupta & Miescke 1996) under the name of the $(R_1, \ldots, R_1)$ policy. This policy is the KG policy for this independent normal R&S problem. We briefly describe this policy and leave further description for section 2.3.1. Let μ_x^n and $(\sigma_x^n)^2$ denote the mean and variance of the posterior predictive distribution for the unknown value of alternative x after the first n measurements. Then the KG policy is the policy that chooses its $(n+1)$st measurement $X^{KG}((\mu_1^n, \sigma_1^n), \ldots, (\mu_M^n, \sigma_M^n))$ from within $\{1, \ldots, M\}$ to maximize the single-period ex-

pected increase in value, $\mathbb{E}_n\left[(\max_{x'}\mu_{x'}^{n+1})-(\max_{x'}\mu_{x'}^{n})\right]$ where $\mathbb{E}_n$ indicates the conditional expectation with respect to what is known after the first n measurements. That is,

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

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

In this chapter we continue the analysis of (Gupta & Miescke 1996). We demonstrate that the KG policy, introduced there as the most rudimentary of a collection of potential policies and studied for its simplicity but neglected thereafter, is actually a powerful and efficient tool for R&S that should be considered for application alongside current state-of-the-art policies. As discussed in detail in section 2.1 below, a number of other sequential Bayesian look-ahead policies have been derived in the years by solving a sequence of single-stage optimization problems just as the KG policy does, and, among these, the optimal computing budget allocation for linear loss of (He et al. 2007) and the LL(S) policy of (Chick & Inoue 2001*b*) assume situations most similar to the one assumed here. The KG policy differs, however, from these other policies in that it solves its single-stage problem exactly while the other policies must use approximations. We believe that solving the look-ahead problem exactly offers an advantage.

After formulating the problem in section 2.2 and defining the policy in section 2.3, we show in section 2.4 that the KG policy is optimal in the limit as $N\to\infty$ in the sense that the policy incurs no opportunity cost in the limit as infinitely many measurements are allowed. Also, by its construction and as noted in (Gupta & Miescke 1996), KG is optimal when there is only one measurement remaining. This provides optimality guarantees at two extremes: N large and N small. While many mediocre policies are optimal in one extreme but not the other, we feel that KG's optimality in both extremes is evidence of quality. For example, the equal-allocation policy guarantees asymptotic optimality but sacrifices short-

term performance. While the equal-allocation policy, and indeed any policy, is optimal for one measurement remaining when the alternative's current predictive distributions are identical, it does not perform nearly as well in other cases. Seldom does a myopic policy also perform optimally in the long-run, and so we feel that the KG policy is notable.

In accord with our belief that optimality at two extremes suggests good performance in the region between, we provide a bound on the policy's suboptimality for finite N in section 2.5. In section 2.6 we introduce the KG persistence property and use it to show both optimality for the case when $M = 2$ and for a further special case in which the means and variances are ordered. Our proof that KG is optimal when $M = 2$ confirms a claim made by (Gupta & Miescke 1994), who showed its optimality among deterministic policies for $M = 2$, but did not offer a formal proof for optimality among sequential policies. Finally, in section 2.7, we demonstrate in numerical experiments that KG performs competitively against the other policies discussed here. In particular, the KG policy is best according to the measure of average performance across a number of randomly generated problems, and the margin by which it outperforms the best competing policies on the most favorable problems is significantly larger than the margin by which it is outperformed on the most unfavorable problems.

2.1 Literature Review

The KG policy was introduced in (Gupta & Miescke 1996) as the simplest of a collection of look-ahead policies, and was studied because its simplicity provided tractability, but this simple policy has been seldom studied or applied in the years since. Instead, a number of more complex Bayesian look-ahead policies have been introduced. A series of researches beginning with (Chen 1995) and continuing with (Chen et al. 1996, Chen et al. 1997, Chen, Chen & Yucesan 2000, Chen, Lin, Yücesan & Chick 2000, Chen et al. 2003) proposed and then refined a family of policies known as the Optimal Computing Budget Allocation (OCBA). These policies are derived by formulating a single-stage optimization problem in which one chooses the second stage measurements to maximize the expected probability of later correctly selecting the best alternative. OCBA policies solve this optimization

problem by approximating the objective function with various bounds, relaxations, and by assuming that the predictive mean will remain unchanged by measurement. They then solve the approximate problem using gradient ascent, greedy heuristics, or with an asymptotic solution that is exact in the limit as the number of measurements in the second stage is large. All OCBA policies assume normal samples with known sampling variance but in practice one may estimate this variance through sampling.

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

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

Recently (He et al. 2007) introduced an OCBA policy designed to minimize expected linear loss. Although more similar to KG than other OCBA policies, it differs in that it uses the Bonferroni inequality to approximate the linear loss objective function for a single stage, and then solves the approximate problem using a second approximation which is accurate in the limit as the second stage is large. This is in contrast to KG, which solves the single-stage problem exactly. The (He et al. 2007) OCBA policy does not assume like the other OCBA approaches that the posterior predictive mean is equal to the prior predictive mean, and in this regard is more similar to the approach of (Chick & Inoue 2001*b*) discussed below.

A set of Bayesian look-ahead R&S policies distinct from OCBA were introduced in

(Chick & Inoue 2001*b*). They differ by not assuming the predictive means equal through time, and by allowing the sampling variance to be unknown. This causes the posterior predictive mean to be student-t distributed, inducing an optimization problem governing the second stage allocation with a somewhat different objective function than that in OCBA formulations. This objective function, corresponding to expected loss, is bounded below, and this lower bound is then approximately minimized. The resulting solution minimizes the lower bound exactly in the limit as sampling costs are small, or as the number second-stage measurements is large.

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

In (Chick et al. 2009) an unknown-variance version of the KG policy was developed under the name LL_1. The authors compared LL_1 to LL(S) using Monte Carlo simulations and found that LL_1 performed well for a small sampling budget, but degraded in performance as the sampling budget increased. We briefly discuss how these results relate to our own in Section 2.7.

In addition to the Bayesian approaches to sequentially ranking and selecting normal populations described thus far, a substantial amount of progress has been made using a frequentist approach. We do not review this literature in detail, and only state that an overview may be found in (Bechhofer et al. 1995), and that a more recent policy which performs quite well in the multi-stage setting with normal rewards is given in (Kim & Nelson 2001, Kim & Nelson 2006*b*). Other sequential and staged policies for independent normal rewards with frequetist guarantees include those in (Paulson 1964), (Rinott 1978), (Hartmann 1991), (Paulson 1994), and (Nelson et al. 2001).

Sequential tests also exist which choose measurements based upon confidence bounds for the value Y_x. Such tests include interval estimation (Kaelbling 1993), which was developed for on-line bandit-style learning in a reinforcement learning setting, and upper confidence bound estimation (Chang et al. 2007), which was developed for estimating value functions for Markov decision processes. Both tests form frequentist confidence intervals for each

Y_x and then select the alternative with the largest upper bound on its confidence interval for measurement. Such policies have general applicability beyond the independent normal setting discussed here.

2.2 Problem formulation

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

2.2.1 A formal model

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\{1, \ldots, M\}$ be the set of alternatives. For each $x \in \{1, \ldots, M\}$ define a random variable Y_x to be the true underlying value of alternative x. We assume a Bayesian setting for the problem in which we have a multivariate normal prior predictive distribution for the random vector Y, and we further assume that the components of Y are independent under the prior, and that $\max_{x=1,\ldots,M} |Y_x|$ is integrable. We will be allotted exactly N measurements, and time will be indexed using n with the first measurement decision made at time 0. At each time $0 \leq n < N$, we choose an alternative x^n to measure. Let ε^{n+1} be the measurement error, which we assume is normally distributed with mean 0 and a finite known variance $(\sigma^\varepsilon)^2$ that is the same across all alternatives. We also assume that errors are independent of each other and of the random vector Y. Then define $\hat{y}^{n+1} = Y_x + \varepsilon^{n+1}$ to be the measurement value observed. At time N, we choose an implementation decision x^N based on the measurements recorded, and we receive an implementation reward $\hat{y}^{N+1}$. We assume that the reward is unbiased, so that $\hat{y}^{N+1}$ satisfies $\mathbb{E}\left[\hat{y}^{N+1} | Y, x^N\right] = Y_{x^N}$. Define the filtration $(\mathcal{F}^n)_{n=0}^N$ by letting $\mathcal{F}^n$ be the sigma-algebra generated by $x^0, \hat{y}^1, x^1, \ldots x^{n-1}, \hat{y}^n$. We will use the notation $\mathbb{E}_n[\cdot]$ to indicate $\mathbb{E}[\,\cdot \mid \mathcal{F}^n]$, the conditional expectation taken with respect to $\mathcal{F}^n$. Measurement and implementation decisions x^n are restricted to be $\mathcal{F}^n$-measurable so that decisions may only depend on measurements observed and decisions made in the past.

Let $\mu^0 := \mathbb{E}[Y]$ and $\Sigma^0 := \mathrm{Cov}[Y]$ be the mean and covariance of the predictive distribution for Y so that Y has prior predictive distribution $\mathcal{N}(\mu^0, \Sigma^0)$ and Σ^0 is a diagonal

covariance matrix. Note that our assumed integrability of $\max_x |Y_x|$ is equivalent to assuming integrability of every Y_x because $|Y_{x'}| \leq \max_x |Y_x|$ and $\max_x |Y_x| \leq |Y_1| + \ldots + |Y_M|$, which is equivalent to assuming Σ^0_{xx} finite for every x.

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

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

$$\sup_{\pi \in \Pi} \mathbb{E}^\pi \left[\max_x \mu^N_x \right]. \tag{2.1}$$

2.2.2 State space and transition function

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

Fix a time n. We use Bayes' rule to update the predictive distribution of Y_x conditioned on $\mathcal{F}^n$ to reflect the observation $\hat{y}^{n+1} = Y_x + \varepsilon^{n+1}$, obtaining a posterior predictive distribution conditioned on $\mathcal{F}^{n+1}$. Since ε^{n+1} is an independent normal random variable,

and the family of normal distributions is closed under sampling, the posterior predictive distribution is also normal. Thus our posterior predictive distribution for Y_x is $\mathcal{N}(\mu_x^n, 1/\beta_x^n)$, and writing it as a function of the prior and the observation reduces to writing μ^{n+1} and β^{n+1} as functions of μ^n, β^n, and $\hat{y}^{n+1}$. Bayes' rule tells us these functions are

$$\mu_x^{n+1} = \begin{cases} \left[\beta_x^n \mu_x^n + \beta^\epsilon \hat{y}^{n+1}\right] / \beta_x^{n+1}, & \text{if } x^n = x, \\ \mu_x^n, & \text{otherwise,} \end{cases} \tag{2.2}$$

$$\beta_x^{n+1} = \begin{cases} \beta_x^n + \beta^\epsilon, & \text{if } x^n = x, \\ \beta_x^n, & \text{otherwise.} \end{cases} \tag{2.3}$$

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

$$\tilde{\sigma}_x^n := \sqrt{\mathrm{Var}\left[\mu_x^{n+1} \mid \mathcal{F}^n, x^n = x\right]}, \tag{2.4}$$

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

Proposition 2.2.1. *For every $x = 1, \ldots, M$, we have $(\tilde{\sigma}_x^n)^2 = (\sigma_x^n)^2 - (\sigma_x^{n+1})^2$.*

Proof. We begin with the relation

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

Squaring both sides, taking the expectation with respect to $\mathcal{F}^{n+1}$, and noting that $(\sigma_x^{n+1})^2 =$

$\mathbb{E}_{n+1}\left[(Y_x - \mu_x^{n+1})^2\right]$ gives

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

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

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

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

$$\tilde{\sigma}(\beta_x) = \sqrt{(\beta_x)^{-1} - (\beta_x + \beta^\epsilon)^{-1}}. \tag{2.5}$$

Then we have that $\tilde{\sigma}_x^n = \tilde{\sigma}(\beta_x^n)$ by Proposition 2.2.1 applied to the identities $(\sigma_x^{n+1})^2 = (\beta_x^{n+1})^{-1} = (\beta_x^n + \beta^\epsilon)^{-1}$ and $(\sigma_x^n)^2 = (\beta_x^n)^{-1}$.

Remark 2.2.2. *For $\beta_x \in (0, \infty)$, we have that $(\tilde{\sigma}(\beta_x))^2 = \beta^\epsilon/[(\beta_x + \beta^\epsilon)\beta_x]$ is strictly decreasing in β_x, and thus so is $\tilde{\sigma}(\beta_x)$.*

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

$$\mu^{n+1} = \mu^n + \tilde{\sigma}(\beta_{x^n}^n) Z^{n+1} e_{x^n}, \tag{2.6}$$

$$\beta^{n+1} = \beta^n + \beta^\epsilon e_{x^n}, \tag{2.7}$$

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

$$T((\mu, \beta), x, z) := (\mu + \tilde{\sigma}(\beta_x) z e_x, \beta + \beta^\epsilon e_x), \tag{2.8}$$

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

We briefly recall and summarize the random variables which play a role in the measurement process. The underlying and unknown value of alternative x is notated Y_x, and

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

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

2.2.3 Dynamic program

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

The terminal value function $V^N : \mathbb{S} \mapsto \mathbb{R}$ is given by (2.1) as

$$V^N(s) := \max_{x \in \{1,\ldots,M\}} \mu_x \quad \text{for every } s = (\mu, \beta) \in \mathbb{S}. \tag{2.9}$$

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

$$V^n(s) = \max_{x \in \{1,\ldots,M\}} \mathbb{E}\left[V^{n+1}(T(s, x, Z^{n+1}))\right], \ s \in \mathbb{S}. \tag{2.10}$$

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

$$Q^n(s,x) := \mathbb{E}\left[V^{n+1}(T(s,x,Z^{n+1}))\right], \ s \in \mathbb{S}, \tag{2.11}$$

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

$$X^{*n}(s) \in \underset{x \in \{1,\ldots,M\}}{\arg\max}\ Q^n(s,x), \ s \in \mathbb{S} \tag{2.12}$$

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

$$V^{n,\pi}(s) := \mathbb{E}^{\pi}\left[V^N(S^N) \mid S^n = s\right], \ s \in \mathbb{S}. \tag{2.13}$$

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

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

Proposition 2.2.3. *$Q^n(s,x) \geq V^{n+1}(s)$ for every $0 \leq n < N$, $s \in \mathbb{S}$, and $x \in \{1, \ldots, M\}$.*

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

Corollary 2.2.4. *Let $i, j \in \{1, \ldots, M\}$, $n < N$, and $s = (\mu, \beta) \in \mathbb{S}$. If $\beta_j = \infty$, then $Q^n(s,i) \geq Q^n(s,j)$.*

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

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

Then, by Proposition 2.2.3,

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

We also have a second corollary to the proposition. Proposition 2.2.3 allowed arbitrarily specifying the alternative to which the extra measurement would be applied, while this corollary points out that the extra measurement may be made according to the optimal policy, in which case $Q^n(s,x)$ is equal to $V^n(s)$. We will use this corollary in section 2.5 to bound the suboptimality of KG.

Corollary 2.2.5. *$V^{n+1}(s) \leq V^n(s)$ for all states $s \in \mathbb{S}$.*

Proof. In Proposition 2.2.3, take the extra measurement x to be the measurement made by the optimal policy in state s. □

Let us say that a policy π is *stationary* if $X^{\pi,n}(s) = X^{\pi,0}(s)$ for all $s \in \mathbb{S}$ and all $n = 1, \ldots, N-1$. In this case we denote $X^{\pi,n}$ simply by X^{π}. Corollary 2.2.5 showed that the value of the optimal policy increases as more measurements are allowed, and we will see in Theorem 2.2.6 below that this monotonicity also holds for stationary policies.

Theorem 2.2.6. *$V^{\pi,n}(s) \geq V^{\pi,n+1}(s)$ for every stationary policy π and every state $s \in \mathbb{S}$.*

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

2.3 The knowledge-gradient policy

In our problem, the entire reward is received after the final measurement. We may formulate an equivalent problem in which the reward is given in pieces over time, but the total reward given is identical. We define the KG policy as that policy which maximizes the single period reward under this alternate formulation. We will see later that this KG policy is optimal in several cases and has bounded suboptimality in all others. This policy was first introduced in (Gupta & Miescke 1996) under the name of the $(R_1, \ldots, R_1)$ policy.

2.3.1 Definition

The problem given by (2.1) has a terminal reward $V^N(S^N) := \max_x \mu_x^N$, but no rewards at any other times. We restructure these rewards by writing $V^N(S^N)$ as a telescoping

sequence,

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

Thus, the problem that provides single period reward $V^N(S^n)$ at time n and $V^N(S^k) - V^N(S^{k-1})$ at times $k = n+1, \ldots, N$ is equivalent to problem (2.1) because the total reward provided is the same in each case. The KG policy π^{KG} is defined as the policy that chooses its measurements to maximize the expectation of the single period reward provided under this alternate formulation, $\mathbb{E}_n\left[V^N(T(S^n, x, Z^{n+1})) - V^N(S^n)\right]$. Since the $(Z^n)_{n=1}^N$ are i.i.d. normal random variables, we may take Z to be a generic standard normal random variable and write the decision function of the KG policy $X^{KG} : \mathbb{S} \mapsto \{1, \ldots, M\}$ as

$$X^{KG}(s) \in \underset{x\in\{1,\ldots,M\}}{\arg\max}\ \mathbb{E}\left[V^N(T(s, x, Z)) - V^N(s)\right] \quad \text{for every } s \in \mathbb{S}, \tag{2.14}$$

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

$$X^{KG}(s) \in \underset{x\in\{1,\ldots,M\}}{\arg\max}\ \mathbb{E}\left[V^N(T(s, x, Z))\right] = \underset{x\in\{1,\ldots,M\}}{\arg\max}\ Q^{N-1}(s, x). \tag{2.15}$$

Remark 2.3.1. *As noted in (Gupta & Miescke 1996), KG is optimal by construction when* $N = 1$. *This is because* $V^{N-1} = V^{KG,N-1}$ *by* (2.12) *and* (2.15), *where* $V^{KG,n}$ *denotes the value of the KG policy at time* n *and is defined according to* (2.13) *with the policy* π *fixed to KG.*

If we think of $V^N(\cdot)$ as a utility function, or as a measure of the amount of "knowledge" contained in a state, we see from (2.14) that the knowledge-gradient policy chooses its decisions in the direction of steepest expected ascent of this measure. This is the reason behind the name *knowledge-gradient policy.*

2.3.2 Computation

It was already known in (Gupta & Miescke 1996) that an exact and computationally tractable expression exists for X^{KG}. We present it here.

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

$$\zeta_x(s) := -\left| \frac{\mu_x - \max_{x' \neq x} \mu_{x'}}{\tilde{\sigma}(\beta_x)} \right|. \tag{2.16}$$

Except for the sign, $\zeta_x(S^n)$ is the minimum distance, in terms of the number of standard deviations $\tilde{\sigma}(\beta_x^n)$, that a measurement of alternative x must alter μ_x^{n+1} from its pre-measurement value of μ_x^n to make $\arg\max_{x'} \mu_{x'}^{n+1} \neq \arg\max_{x'} \mu_{x'}^n$ — that is, to change the identity of the alternative with the largest conditional expected value. In addition, define the function $f : \mathbb{R} \mapsto \mathbb{R}$ as

$$f(z) := z\Phi(z) + \varphi(z), \tag{2.17}$$

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

Theorem 2.3.2. *For every $s = (\mu, \beta) \in \mathbb{S}$, we have*

$$Q^{N-1}(s, x) = \max_{x'} \mu_{x'} + \tilde{\sigma}(\beta_x) f(\zeta_x(s)), \tag{2.18}$$

$$X^{KG}(s) \in \underset{x \in \{1,\ldots,M\}}{\arg\max} \tilde{\sigma}(\beta_x) f(\zeta_x(s)). \tag{2.19}$$

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

The term $Q^{N-1}(s, x) - \max_{x'} \mu_{x'} = \tilde{\sigma}(\beta_x) f(\zeta_x(s))$ is in some sense the expected value of the information that would be obtained by measuring alternative x, and is sometimes called the "expected value of information" or EVI, e.g., in (Chick & Inoue 2001*b*) and (Chick et al. 2009).

Computation of the KG policy via (2.19) scales linearly with the number of alternatives M. This compares well with other policies that might be used on this problem. To compute the KG policy at time n, we must first find the largest and second largest μ_x^n across all alternatives x, which will be used to compute $\zeta_x^n := \zeta_x(S^n)$. This may be implemented either by an initial pass through the alternatives at each time period, or by storing and updating the two values across time periods. Once we have the largest and second largest μ_x^n, we iterate through the alternatives, calculating $\tilde{\sigma}(\beta_x^n) f(\zeta_x^n)$ for each one, and returning the alternative with the largest value for this expression. This iteration may be streamlined

by recomputing the expression only for those alternatives that changed ζ_x^n or β_x^n from the previous iteration.

The following remark, which is an easily obtained consequence of Theorems 1 and 2 in (Gupta & Miescke 1996) and may also be obtained directly from (2.18), may also be used to accelerate the computation of the KG policy by eliminating some alternatives from consideration. It is also useful for proving later results. It states that if an alternative dominates another in both mean and variance, then of the two, KG prefers the dominating alternative.

Remark 2.3.3. *For every $s = (\mu, \beta) \in \mathbb{S}$ such that $\mu_j \geq \mu_i$ and $\beta_j \leq \beta_i$ we have $Q^{N-1}(s, j) \geq Q^{N-1}(s, i)$.*

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

Remark 2.3.4. *Take $n = N - 1$ in Corollary 2.2.4. If $\beta_j = \infty$ for some $j \in \{1, \ldots, M\}$ (that is, if the predictive distribution $\mathcal{N}(\mu_j, 1/\beta_j)$ for Y_j is a point mass), then $Q^{N-1}(S, i) \geq Q^{N-1}(S, j)$ for every $i \in \{1, \ldots, M\}$.*

Thus, KG will never measure an alternative with zero variance unless every alternative has zero variance. Corollary 2.2.4 shows that the optimal policy shares this behavior of preferring not to measure any alternative whose true value is known perfectly.

2.3.3 Behavior

KG balances two considerations when it chooses its measurement decisions. First, it prefers to measure those alternatives about which comparatively little is known. These alternatives x are the ones whose predictive distributions have large variance $(\sigma_x^n)^2$ or equivalently with small precision β_x^n. Thus, we have that if KG prefers to measure some alternative i over another alternative j, then it would still prefer to measure alternative i over j if the predictive variance of i were increased.

Second, KG prefers to measure alternatives x with $|\mu_x^n - \max_{x' \neq x} \mu_{x'}^n|$ close to 0. We call $-|\mu_x^n - \max_{x' \neq x} \mu_{x'}^n|$ the *unnormalized influence* and $\zeta_x^n = -|\mu_x^n - \max_{x' \neq x} \mu_{x'}^n| / \tilde{\sigma}(\beta_x^n)$ the

normalized influence, or simply the *influence*, of alternative x, where $\tilde{\sigma}(\beta_x^n)$ is understood as a normalization term because predictions for different alternatives have different variances and comparison does not make sense unless we standardize these differences. Measurements of alternatives with large influence are more likely to cause a change in the optimal implementation decision; that is, to cause $\arg\max_{x'} \mu_{x'}^n \neq \arg\max_{x'} \mu_{x'}^{n+1}$. KG's preference for small predictive precision and large influence are formalized in Propositions 2.3.6 and 2.3.7, but first we calculate the derivative of f, as defined in (2.17), in a lemma.

Lemma 2.3.5. *We have $f'(z) = \Phi(z) \geq 0$ for every $z \in \mathbb{R}$.*

Proof. First note that $\frac{d}{dz} e^{-z^2/2} = -z e^{-z^2/2}$, showing that $\varphi'(z) = -z\varphi(z)$. From this we see that f has non-negative derivative $f'(z) = \Phi(z) + z\varphi(z) - z\varphi(z) = \Phi(z)$. □

Proposition 2.3.6. *Let states $s = (\mu, \beta) \in \mathbb{S}$, $s' = (\mu', \beta') \in \mathbb{S}$, and alternatives $i, j \in \{1, \ldots, M\}$ satisfy the following criteria: $\zeta_i(s') > \zeta_i(s)$, $\zeta_j(s') = \zeta_j(s)$, $\beta_i' < \beta_i$, and $\beta_j' = \beta_j$. If $Q^{N-1}(s, i) > Q^{N-1}(s, j)$, then $Q^{N-1}(s', i) > Q^{N-1}(s', j)$.*

Proof. First, $\tilde{\sigma}(\beta_i') \geq \tilde{\sigma}(\beta_i)$ by Remark 2.2.2 and $f(\zeta_i(s')) \geq f(\zeta_i(s))$ by Lemma 2.3.5. By (2.18), $Q^{N-1}(s', i) > Q^{N-1}(s, i)$. Also, the equalities $\tilde{\sigma}(\beta_j') = \tilde{\sigma}(\beta_j)$ and $f(\zeta_j(s')) = f(\zeta_j(s))$ imply through (2.18) that $Q^{N-1}(s', j) = Q^{N-1}(s, j)$. Thus, if $Q^{N-1}(s, i) > Q^{N-1}(s, j)$, then $Q^{N-1}(s', i) \geq Q^{N-1}(s, i) > Q^{N-1}(s, j) = Q^{N-1}(s', j)$. □

Proposition 2.3.7. *If alternative i and state $s = (\mu, \beta)$ are such that $\zeta_i(s) \geq \zeta_j(s)$ and $\beta_i < \beta_j$ for every alternative $j \neq i$, then $X^{KG}(s) = i$.*

Proof. Let j be an alternative different from i. Then $\tilde{\sigma}(\beta_i) > \tilde{\sigma}(\beta_j)$ by Remark 2.2.2 and $f(\zeta_i(s)) \geq f(\zeta_j(s))$ by Lemma 2.3.5. This implies that $Q^{N-1}(s, i) > Q^{N-1}(s, j)$ by Proposition 2.3.6. Since this is true for all $j \neq i$, we have that $i = \arg\max_j Q^{N-1}(s, j) = X^{KG}(s)$ where the arg max is unique. □

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

Proposition 2.3.8. *If KG prefers alternative i in state (μ, β), then it also prefers the same alternative i in state $(\mu + ae_i, \beta)$ for all positive real numbers a such that $\mu_i + a \leq \max_x \mu_x$, i.e., for $0 \leq a \leq -\mu_i + \max_x \mu_x$.*

We leave the proof until the appendix.

2.4 Asymptotic optimality

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

The KG policy is not alone in possessing this property. Indeed, the following well-known policies are all asymptotically optimal: the equal-allocation policy which distributes its measurements in a round-robin fashion equally among the alternatives; the uniform exploration policy which randomly chooses its measurements with equal probability across the alternatives; and the Boltzmann exploration policy discussed later in section 2.7 which randomly chooses its measurements according to exponentially weighted probabilities.

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

In addition, one policy, interval estimation, performs very well in our numerical experiments but is not asymptotically optimal as in some cases it "sticks", measuring one alternative only and obtaining its true value perfectly without learning about the others (Kaelbling 1993). Indeed, one can construct cases in which this policy's performance is arbi-

trarily bad compared to any asymptotically optimal policy. Although a policy's asymptotic optimality is not evidence of quality by itself, its absence should raise concern among those who might use a policy lacking it. Finally, a natural question is whether other policies, such as those in the OCBA family and those proposed in (Chick & Inoue 2001*b*), are asymptotically optimal. This question is currently open as these other policies are more complex and require more care during analysis than does KG. Nevertheless, we believe that the proof techniques applied here may be extended to show that many other Bayesian look-ahead policies are also asymptotically optimal.

To show that KG is asymptotically optimal, we begin by showing in Proposition 2.4.1 that the asymptotic value of a policy is well defined and bounded above by the value $\mathbb{E}\max_x Y_x$ of learning every alternative exactly. Then we show in Proposition 2.4.2 that this value is achieved by any stationary policy that measures every alternative infinitely often. Thus, any stationary policy that samples every alternative infinitely often is asymptotically optimal. Finally, we show in Theorem 2.4.3 that KG is asymptotically optimal. The proof centers on the notion that, as the number of times an alternative is measured increases, the variance of the value of that alternative shrinks toward 0. Eventually, that variance will be so low that KG will prefer to measure another alternative. This argument is used to show that KG samples every alternative infinitely often, and thus is asymptotically optimal.

Since we will be varying the number N of measurements allowed, we use the notation $V^0(\,\cdot\,;N)$ to denote the value function at time 0 when the problem's terminal time is N. We then define the *asymptotic value function* $V(\,\cdot\,;\infty)$ by the limit $V(s;\infty) := \lim_{N\to\infty} V^0(s;N)$ for $s \in \mathbb{S}$. Similarly, we denote the *asymptotic value function for stationary policy* π by $V^\pi(\,\cdot\,;\infty)$ and define it by $V^\pi(s;\infty) := \lim_{N\to\infty} V^{\pi,0}(s;N)$ for $s \in \mathbb{S}$. Proposition 2.4.1 shows that both limits exist.

If $V^\pi(s;\infty)$ is equal to $V(s;\infty)$ for every $s \in \mathbb{S}$, then π is said to be *asymptotically optimal*. In particular, if a stationary policy π achieves the upper bound $U(\,\cdot\,)$ on $V(\,\cdot\,;\infty)$ shown in Proposition 2.4.1, then π must be asymptotically optimal. We will use this later to show that KG is asymptotically optimal. The proof of Proposition 2.4.1 may be found in the appendix.

Proposition 2.4.1. *Let $s \in \mathbb{S}$. Then the limit $V(s;\infty)$ exists and is bounded above by*

$$U(s) := \mathbb{E}\left[\max_x Y_x \mid S^0 = s\right] < \infty, \tag{2.20}$$

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

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

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

The limit η_x^∞ exists because η_x^N is non-decreasing in N almost surely. Note that we allow the limit η_x^∞ to be infinite.

Proposition 2.4.2 formalizes the idea that if we measure every alternative infinitely often, then we eventually learn the true value of every alternative. This implies asymptotic optimality. We then use Proposition 2.4.2 in the proof of Theorem 2.4.3 to show that KG is asymptotically optimal. The proofs for both Theorem 2.4.3 and Proposition 2.4.2 may be found in the appendix.

Proposition 2.4.2. *If π is a stationary policy under which $\eta_x^\infty = \infty$ almost surely for every x, then π is asymptotically optimal.*

Theorem 2.4.3. *The KG policy is asymptotically optimal and has value $U(S^0)$.*

2.5 Bound on suboptimality

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

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

Theorem 2.5.1. *Let $s = (\mu, \beta) \in \mathbb{S}$ and $c = (2\pi)^{-1/2} \max_x \tilde{\sigma}(\beta_x)$. Then*

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

The proof may be found in the appendix. We combine this result with Theorem 2.2.6 to bound KG's suboptimality. Here, $V^{KG,n}(s)$ is the value of the KG policy at time n when $S^n = s$.

Corollary 2.5.2. *Let $s = (\mu, \beta) \in \mathbb{S}$ and $c = (2\pi)^{-1/2} \max_x \tilde{\sigma}(\beta_x)$. Then*

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

Proof. By Remark 2.3.1, we have $V^{N-1}(s) = V^{KG,N-1}(s)$. From Theorem 2.2.6 we have $V^{KG,N-1}(s) \leq V^{KG,n}(s)$. Substituting the inequality $V^{N-1}(s) \leq V^{KG,n}(s)$ into Theorem 2.5.1 shows the corollary. □

2.6 Optimality for finite horizon special cases

We saw in Remark 2.3.1 that KG is optimal when $N = 1$. We will show that KG is optimal in two other special cases: first, when there are only two alternatives to measure; second, when the measurements are free from noise, $(\sigma^\varepsilon)^2 = 0$, and when the parameters of the time 0 prior can be ordered by $\mu_1^0 \geq \mu_2^0 \geq \ldots \geq \mu_M^0$ and $\sigma_{11}^0 \geq \sigma_{22}^0 \geq \ldots \geq \sigma_{MM}^0$. Before showing optimality under these conditions, we first define and discuss a property called the KG persistence property. This property is useful because it provides a sufficient condition for optimality.

2.6.1 Persistence of the knowledge-gradient policy

Proofs of the optimality of the KG policy in these special cases is based on the KG persistence property. A problem setting is said to have the KG persistence property if, operating the problem under some policy other than KG, an alternative preferred by KG will remain preferred until the alternative is measured. Below, in Theorem 2.6.3, we show that if a problem setting has the KG persistence property, then KG is optimal in that problem setting. Before stating this theorem, we formally define the KG persistence property and an associated term, "covering of the future."

Definition 2.6.1. *A sequence of subsets of $\mathbb{S}$, $\{\mathbb{S}^n\}_{n=k}^N$, is called a covering of the future from k if $T(s, x, Z^{n+1}) \in \mathbb{S}^{n+1}$ almost surely for every $s \in \mathbb{S}^n$, $x \in \{1, \ldots, M\}$, and $n \in \{k, \ldots, N-1\}$.*

Definition 2.6.2. *We say that the KG persistence property holds on a covering* $\{\mathbb{S}^n\}_{n=k}^N$ *of the future from* k *if* $X^{KG}(T(s,x,Z^{n+1})) = X^{KG}(s)$ *almost surely for every* $s \in \mathbb{S}^n$, $x \neq X^{KG}(s)$, *and* $n \in \{k,\ldots,N-1\}$.

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

Theorem 2.6.3. *If the KG persistence property holds on a covering* $\{\mathbb{S}^n\}_{n=k}^N$ *of the future from* k *for some* $k \in \{0\ldots N-1\}$, *then* $V^{KG,k}(s) = V^k(s)$ *for every* $s \in \mathbb{S}^k$.

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

2.6.2 Optimality for two alternatives

We use the KG persistence principle to show that KG is optimal when there are exactly two alternatives to consider, i.e., $M = 2$. In this case we will see that the optimal policy is one that, at each decision point, measures the alternative with the largest variance. This policy is actually deterministic, and it was shown in (Gupta & Miescke 1994) that this policy is optimal among the class of deterministic policies. Theorem 2.6.5 extends this result to show that this same policy is also optimal among the class of fully sequential policies. It is not generally true that the best deterministic policy is also as good or better than every sequential policy, but Theorem 2.6.5 shows that this is exactly the case for this particular problem.

We will see that the policy of measuring the alternative with the largest variance is optimal because knowing the correct implementation decision is the same as knowing the

true sign of $Y_1 - Y_2$. Each measurement measures only one of Y_1 or Y_2, and an equal reduction in variance for Y_1 or Y_2 contributes equally to the overall reduction in variance of $Y_1 - Y_2$, regardless of which expected value is bigger. Thus, the best way to learn about the difference between points $Y_1 - Y_2$ is to measure that point about which the least is known.

To show that KG is optimal when $M = 2$, we need to show that KG persistence holds when $M = 2$ and then refer to Theorem 2.6.3.

Lemma 2.6.4. *If $M = 2$, then $X^{KG}(s) \in \arg\min_x \beta_x$ for each $s = (\mu, \beta) \in \mathbb{S}$ with ties broken by choosing the alternative with the smaller index.*

Proof. By (2.19) from Theorem 2.3.2, it is enough to show equality between the sets $\arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(s))$ and $\arg\min_x \beta_x$. When $M = 2$, $\zeta_x(s) = -|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)$, so $\arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(s)) = \arg\max_x \tilde{\sigma}(\beta_x) f\left(-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)\right)$. The function $\tilde{\sigma}$ is strictly decreasing by Remark 2.2.2. This fact will be used on its own, and it also implies that $-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)$ is a decreasing function of β_x. The function f is non-decreasing by Lemma 2.3.5, so the function $\beta_x \mapsto f\left(-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)\right)$ is the composition of a non-decreasing function with a non-increasing function, and is thus itself non-increasing. Thus, the function $\beta_x \mapsto \tilde{\sigma}(\beta_x) f\left(-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)\right)$ is the product of a strictly decreasing function with a non-increasing function, and is thus itself strictly decreasing. This implies that $\arg\max_x \tilde{\sigma}(\beta_x) f\left(-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)\right) = \arg\min_x \beta_x^n$. □

Theorem 2.6.5. *If $M = 2$, then KG is optimal.*

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

Let $n \in \{0, \ldots, N-1\}$ and $s = (\mu, \beta) \in \mathbb{S}$. First consider the case when $\beta_1 \leq \beta_2$. By Lemma 2.6.4, $X^{KG}(s) = 1$. The precision component of $T(s, 2, Z^{n+1})$ is $(\beta_1, \beta_2 + \beta^\epsilon)$. Since $\beta_1 \leq \beta_2 \leq \beta_2 + \beta^\epsilon$ and by Lemma 2.6.4, $X^{KG}(T(s, 2, Z^{n+1})) = 1$ a.s.

Now consider the case when $\beta_1 > \beta_2$. By Lemma 2.6.4, $X^{KG}(s) = 2$. The precision component of $T(s, 1, Z^{n+1})$ is $(\beta_1 + \beta^\epsilon, \beta_2)$. Since $\beta_1 + \beta^\epsilon \geq \beta_1 > \beta_2$ and by Lemma 2.6.4, $X^{KG}(T(s, 1, Z^{n+1})) = 2$ a.s.

In both cases, $x \neq X^{KG}(s)$ implies $X^{KG}(T(s, x, Z^{n+1})) = X^{KG}(s)$ a.s., so KG persistence holds. Then, by Theorem 2.6.3, $V^{KG,0}(s) = V^0(s)$ for every $s \in \mathbb{S}$, and KG is

optimal. □

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

2.6.3 Optimality when the state space is ordered

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

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

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

Theorem 2.6.6. *If* $(\sigma^\varepsilon)^2 = 0$ *and* $s = (\mu, \beta) \in \mathbb{S}$ *is such that the implication*

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

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

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

2.7 Computational experiments

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

The space of problems is parameterized by a number of measurements N, a number of alternatives M, an initial precision $\beta^0 \in (0, \infty]^M$, an initial mean $\mu^0 \in \mathbb{R}^M$, and a measurement noise $(\sigma^\varepsilon)^2 \in [0, \infty)$. We chose a collection of 100 problems randomly generated within this space according to the following distribution: M was integer-valued between

2 and 100. N was chosen by first choosing M and then choosing a ratio N/M uniformly from the set $\{1, 3, 10\}$. Each μ_x was uniformly distributed in the interval $[-1, 1]$, and each β_x was independently chosen as 1 with probability .9 and 1000 with probability .1. The noise variance $(\sigma^\varepsilon)^2$ was set to 1 in all cases. The space of problems is quite large, and with only 100 randomly generated problems we have necessarily failed to represent many problem settings that would be met in applications. Nevertheless, we believe this collection of problems includes problems with structure characteristic to a large number of applications.

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

We compared KG against seven other policies: the Optimal Computing Budget Allocation (OCBA) for linear loss of (He et al. 2007), the LL(S) policy of (Chick & Inoue 2001*b*), the interval estimation (IE) policy of (Kaelbling 1993), Boltzmann exploration (see, e.g., (Singh et al. 2000)), equal allocation, and exploitation. Several of these policies required choosing one or more parameters, which we did by simulating several choices on all 100 problems and taking the parameters whose resulting opportunity cost was smallest when summed over all 100 problems. We briefly describe each policy and its tuning:

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

- (LL(S) for known variance) The LL(S) policy allows normal measurement errors with

unknown variance and uses a normal-gamma prior for the unknown mean and measurement precision. We adapted this policy to the known-variance case by taking the limit as the gamma prior on the precision becomes a point mass at the known variance. Details may be found in the appendix. The policy has two parameters, n_0 and τ. We set n_0 to 0 as we did with OCBA. We tested the values $1, 2, 3, 4, 5, 10$ for τ on our collection of 100 problems with 5000 samples for each problem and found that $\tau = 1$ worked best for every problem. This is the value we used in comparison with KG.

- (Interval Estimation) IE is parameterized by $z_{\alpha/2}$. As (Kaelbling 1993) suggests that values of 2, 2.5 or 3 often work best for $z_{\alpha/2}$, we tested values between 2 and 4 in increments of .1 and found that $z_{\alpha/2} = 3.1$ worked best. Although we found IE worked very well when properly tuned, we also found it to be very sensitive to the choice of tuning parameter.

- (Boltzmann exploration) Boltzmann exploration chooses its measurements by $\mathbb{P}\left\{x^n = x \mid \mathcal{F}^n\right\} = \frac{\exp(\mu_x^n/T^n)}{\sum_{x'=1}^{M} \exp\left(\mu_{x'}^n/T^n\right)}$, where the policy is parameterized by a decreasing sequence of "temperature" coefficients $(T^n)_{n=0}^{N-1}$. We tuned this temperature sequence within the set of exponentially decreasing sequences defined by $T^{n+1} = \gamma T^n$ for some constant $\gamma \in (0, 1]$. The set of all such sequences is parameterized by γ and T^N. We tested $\gamma \in \{.1, .5, .8, .9, 1\}$ with $T^N \in \{.1, 1, 10\}$ and found that $\gamma = 1$ performed best. We then tested the set of possible T^N between .1 and 10 with γ fixed to 1 and found that $T^N = .55$ performed best.

- (Equal allocation) The equal-allocation policy is $x^n \in \arg\min_x \beta_x^n$, since we think of the prior as providing the results of some previous first stage measurements, and we interpret β_x^n/β^ϵ as the number of measurements of alternative x taken by time n. It requires no tuning.

- (Exploitation) The exploitation policy is $x^n \in \arg\max_x \mu_x$. It requires no tuning.

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

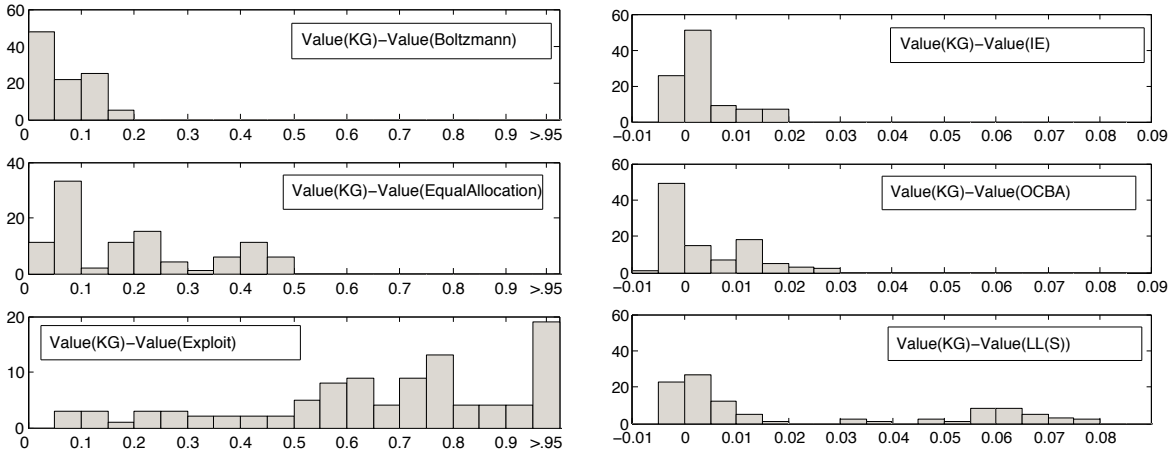

Figure 2.1: Histogram of the sampled difference in value for competing policies aggregated across the 100 randomly generated problems.

2.7.1 Results

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

We show in Figure 2.1 the sample estimates of $V^{KG} - V^{\pi}$ aggregated across the randomly generated problems for each of the competing policies π. Bars to the right of 0 indicate that KG outperformed the plotted policy on those problems and bars to the left indicate the converse. Note that the scale of the histograms in the right-hand plots is much smaller than in the left-hand plots. The histograms show that Boltzmann exploration, equal allocation, and exploitation policies were all outperformed by KG in every problem setting tested, while IE, OCBA for linear loss, and the LL(S) policy performed relatively better. Each

of these three better competing policies performed better than KG on some problems, and were outperformed on others; however, the tail to the right of 0 is larger than to the left. This indicates that the amount by which KG outperformed the competing policies was significantly larger than the amount by which it was outperformed.

We note a seeming discrepancy between our numerical work and that in (Chick et al. 2009), who tested a variance-unknown version of the KG policy called LL_1. They found that LL_1 performed well in small-sample settings, but poorly elsewhere. In contrast, we found that KG, a very similar policy, performed quite well overall. We believe that the difference lies in the stopping rule used. We simply stopped our sampling policies after a fixed horizon N, but (Chick et al. 2009) drew many of its conclusions from experiments using the EOC Bonf stopping rule introduced in (Branke et al. 2005). In experiments not pictured here we found that KG also performed poorly with EOC Bonf stopping, but much better when it was stopped using a stopping rule that we introduce now.

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

2.8 Conclusion

The KG measurement policy, as first proposed by (Gupta & Miescke 1996) and as analyzed here, has several attractive features. Under the assumption of independent normally distributed priors with normal sampling errors of common known variance, we showed that

the policy is optimal in both extremes of the number of measurements allowed ($N = 1$ and $N \to \infty$), as well as in other special cases, and has bounded suboptimality in the remaining cases. We showed numerically that it performs competitively with, or significantly better than, several other sequential measurement policies in a broad class of problem settings. In addition, KG is simple in concept, easy to implement, fast to compute, and requires no tuning. This simplicity may make it an attractive alternative to its more complex but similarly performant cousins, the optimal computing budget allocation and the LL(S) policy.

One important limitation of the version of the policy discussed herein is its assumption of common known variance, which often fails to be met in practice. To lift this assumption, it is possible to place a normal-gamma prior on the unknown means and variances, as was done in (Chick & Inoue 2001*b*), and recompute the optimal single-step-lookahead policy. Indeed, if we begin with a non-informative normal-gamma prior for the true mean Y_x and unknown sampling variance β_x^ϵ of alternative x, and after sampling have vectors of statistics $(\mu, \hat{\sigma}^2, n)$ where $(\mu_x, \hat{\sigma}_x^2, n_x)$ indicate the sample mean, sample variance, and number of samples taken for alternative x, then a calculation similar to that of Theorem 2.3.2 reveals that the corresponding KG policy is $\arg\max_x \tilde{\sigma}_x f_{n_x-1}(\zeta_x)$, where we must redefine $\tilde{\sigma}_x := \sqrt{\hat{\sigma}^2/n_x(n_x+1)}$, leave ζ_x defined as before, and define $f_n(z) := \frac{\nu+z^2}{\nu-1}\varphi_\nu(z) + z\Phi_\nu(z)$ where φ_ν and Φ_ν are respectively the pdf and cdf of the student-t distribution with ν degrees of freedom. This provides a version of KG for the unknown-variance case. This was derived earlier and independently in (Chick et al. 2009), and is discussed there in much greater detail, together with a numerical analysis of its properties.

Additionally, the KG policy as described herein has used a fixed number of samples instead of an adaptive stopping rule, while (Branke et al. 2005) has shown that such rules generally improve the efficiency of budgeted R&S policies. Nevertheless, as implied briefly in section 2.7 and as discussed in (Chick et al. 2009), one can certainly use an adaptive stopping rule with the KG sampling policy. Future work is needed to assess the quality of such adaptively-stopped policies, and to determine which stopping rules are best to use with KG, but this is by no means an insurmountable obstacle.

Other limitations would seem to present more difficulty. The use of common random numbers has proved immensely beneficial for simulation-based R&S. (Chick & Inoue 2001*a*)

and (Fu et al. 2007) discuss Bayesian R&S policies taking advantage of common random numbers, as does (Kim & Nelson 2006*a*) for the frequentist formulation, and it may be possible to extend the KG approach along these lines as well. Indeed, KG's benefits may be overshadowed by its inability to leverage common random numbers in simulation-based R&S unless this extension can be made. In addition, KG assumes the alternatives have a common measurement cost while in practice it may be more expensive or time consuming to measure some alternatives than others. It may be possible to lift this restriction by dividing the benefit of measurement by the cost so as to obtain a normalized quantity for comparison (a benefit per unit cost), but it may also be that the OCBA approach is more appropriate in such instances.

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

Chapter 3

The Knowledge-Gradient Stopping Rule

In this chapter continue to consider the fully sequential independent normal R&S problem, as we did in Chapter 2, but we now allow the policy to adaptively choose how many samples to take. We also allow for uncertainty in the sampling variance by placing a prior on it and updating it to obtain a posterior. By considering the sampling and stopping problems jointly rather than separately, we derive a composite KG stopping/sampling rule.

The sampling component of the composite KG rule we derive is the same as the LL1 sampling rule introduced in (Chick et al. 2007) and (Chick et al. 2009). This LL1 sampling rule is the same as the KG rule for R&S with independent normal rewards of unknown mean and variance, *with a fixed sampling horizon.* Although the sampling portion of the KG rule for this problem was previously introduced, the stopping rule is new. This new stopping rule significantly improves the performance of LL1 as compared to its performance under the best other generally known adaptive stopping rule, EOC Bonf (Branke et al. 2005), outperforming it in every case tested.

As in Chapters 2 and 4, the rule is derived using the linear loss function, which penalizes according to the difference in value between the chosen option and the best, contrasting it with another common choice, $0-1$ loss, which penalizes a constant 1 for failing to find the best alternative. Other algorithms designed for the linear loss objective function under independent normal sampling with unknown means and variances include LL(S) (Chick &

Inoue 2001*b*) and OCBA for linear loss (He et al. 2007).

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

Extensive numerical comparisons were made in (Chick et al. 2009) with these two procedures under two different stopping rules: the naive or fixed stopping rule, which simply stops after a fixed number of samples have been taken; and the EOC Bonf stopping rule, introduced in (Branke et al. 2005), which more intelligently decides when to stop based on an approximation to the expected loss that would occur due to remaining uncertainty. As shown in (Branke et al. 2005) and again in (Branke et al. 2007), using the EOC Bonf stopping rule instead of the fixed one generally improves the performance of R&S procedures.

These numerical comparisons in (Chick et al. 2009) between LL(S) and LL1 revealed that LL(S) outperforms LL1 in a broad class of problem configurations when both operate under the EOC Bonf stopping rule. In other situations, for example with the fixed stopping rule taking a small number of samples, LL(S) performed better but the great benefit provided by using an adaptive stopping rule like EOC Bonf led the authors to conclude that LL1 may not be broadly applicable. Further, they supposed that its myopic assumption was the culprit behind its poor performance.

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

3.1 Bayesian Formulation

We briefly review the Bayesian formulation of the R&S problem with normal means and unknown mean and variance, as well as the decisions made by the LL(S) and LL1 rules.

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

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

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

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

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

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

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

compute the new parameters via

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

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

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

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

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

$$\sup_\pi \mathbb{E}^\pi \left[\max_x \mu_x^\tau\right], \tag{3.1}$$

where again the supremum is over all policies adapted to the filtration. Note that maximizing this reward is equivalent to minimizing the expected opportunity cost, where opportunity cost is defined to be $\mu_{x^*} - \mu_{i^*}$, with $x^* \in \arg\max_x \mu_x$ and $i^* \in \arg\max_x \mu_x^\tau$. This

opportunity cost is the difference in value between the best alternative and the one that we have chosen. This corresponds to the linear loss function discussed above.

We have assumed in (3.1) that τ is given. Indeed, the derivations of most existing sampling rules do not explicitly consider the role that the choice of stopping rule plays in overall performance. A common assumption for the purposes of analysis (this assumption was used in Chapter 2) is that τ is some fixed constant. Our goal in this chapter, however, is to show that there is value in deriving the sampling and stopping rules together, and so later, in Section 3.2, we will consider the optimization problem in which we control both the sampling rule π and the stopping rule τ.

Although (3.1) can in principle be solved through dynamic programming, the computational challenges are prohibitive. Instead, a number of heuristic approaches have been presented. We briefly review two of these approaches: LL1, and LL(S).

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

$$\nu_x^n := \mathbb{E}\left[\max_{x'} \mu_{x'}^{n+1} \mid \mathcal{F}^n, x^n = x\right] - \max_{x'} \mu_{x'}^n = \lambda_{\{x\cdot\}}^{-1/2} \Psi_{\rho_x^n}\left(\lambda_{\{x\cdot\}}^{1/2} |\mu_x^n - \max_{x' \neq x} \mu_{x'}^n|\right). \quad (3.2)$$

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

$$\lambda_{\{x\cdot\}} := \rho_x^n(\rho_x^n + 1)a_x^n/b_x^n,$$
$$\Psi_d(s) := \int_{u=s}^{\infty} \phi_d(u)\, du = \frac{d+s^2}{d-1}\phi_d(s) - s\Phi_d(-s),$$

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

The LL1 policy is then given by

$$x^n \in \arg\max_x \nu_x^n \quad (3.3)$$

The LL(S) rule, introduced in (Chick & Inoue 2001*b*), considers the effect of blocks of measurements. It is parameterized by the block size, which is commonly denoted by τ,

but which we will refer to as B to avoid confusing it with our stopping time τ. At the beginning of each stage, the LL(S) allocation considers the marginal benefit of the next B measurements,

$$\mathbb{E}\left[\max_{x'} \mu_{x'}^{n+B} \mid \mathcal{F}^n, x^n, \ldots, x^{n+B-1}\right] - \max_{x'} \mu_{x'}^n, \tag{3.4}$$

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

Ideally, the LL(S) algorithm would like to optimize (3.4) over $x^n, \ldots, x^{n+B-1}$, but since computing (3.4) is computationally intensive and optimizing over it is even more so, LL(S) uses the Bonferonni inequality to approximate the optimal allocation and allocates according to that approximation. A full description of the LL(S) algorithm may be found in (Chick & Inoue 2001*b*).

Note that in this formulation the decision of which alternatives to measure between times n and $n + B - 1$ may depend only with the information available at time n, while under LL1 each measurement decision is made with the full information available. Also note that when $B = 1$ the objective function (3.4) from which LL(S)'s allocation is derived is identical to (3.1), but LL1 optimizes this expression exactly while LL(S)'s use of the Bonferonni inequality results in an approximation to the optimal.

3.2 Optimal Stopping Rule

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

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

Our objective function is then,

$$\sup_{\pi,\tau} \mathbb{E}^{\pi} \left[\max_{x} \mu_x^{\tau} - C(\tau) \right]. \tag{3.5}$$

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

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

3.3 Knowledge-Gradient Stopping Rule

Just as solving the Bayesian R&S problem with a given stopping rule τ is computationally intractable, so is solving the more difficult problem (3.5) in which we also optimize over τ. This justifies the introduction of a heuristic, which we derive using a method that we refer to as the knowledge-gradient (KG) method.

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

and measurement cost paid on each measurement. We argue later that, in this problem, this myopia does not hinder the efficient acquisition of information over longer periods, justifying the KG policy as a reasonable and interpretable heuristic.

We now apply the KG method by computing what the optimal decision would be if the KG assumption were true. The optimal decision is the best among either stopping now at n, or measuring any alternative x, incurring the measurement cost, and stopping at time $n+1$. Stopping now by taking $\tau = n$ has value $\max_{x'} \mu^n_{x'} - C(n)$, while measuring alternative x and then stopping has value

$$\mathbb{E}\left[\max_{x'} \mu^{n+1}_{x'} - C(n+1) \mid \mathcal{F}^n, x^n = x\right] = \left(\max_{x'} \mu^n_{x'}\right) + \nu^n_x - C(n+1), \tag{3.6}$$

as can be seen directly from (3.2). The x that maximizes (3.6) is exactly the x^n maximizing (3.2), and is thus the same as the decision of the LL1 sampling rule. Thus, the decision we face in our sampling and stopping problem is between sampling the alternative suggested by the LL1 sampling rule, or stopping. Furthermore, the difference in value between sampling this best x and stopping now is equal to

$$-\left(C(n+1) - C(n)\right) + \max_x \nu^n_x, \tag{3.7}$$

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

(i) If $C(n+1) - C(n) \geq \max_x \nu^n_x$, then stop sampling.

(ii) Otherwise, sample $x^n \in \arg\max_x \nu^n_x$.

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

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

Proposition 3.3.1. *Let τ^{KG} be the stopping rule defined by the KG stopping rule and let τ^* be a stopping rule that is optimal for the problem*

$$\sup_{\tau} \mathbb{E}^{\pi=\mathrm{LL1}} \left[\max_x \mu_x^\tau - C(\tau) \right]. \tag{3.8}$$

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

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

Although the KG stopping rule can stop too soon, there is numerical evidence to suggest that the cost of this early stopping is low. We present this evidence in Section 3.4. Further evidence comes from the tendency of the net value of continued measurement to decrease. Consider the case when the expected net marginal value of continuing given by (3.7) is negative. This is the situation in which the KG stopping rule stops, and is the only case in which it can err. The only reason an optimal stopping rule τ^* would continue in this situation would be an expectation that net marginal values of continuing will be positive in the future, compensating for the net loss incurred by the current measurement. Thus continuing in this situation incurs an immediate loss with the *possibility* of future profit. But if we expect future net marginal values of continuing to be even more negative than they are now, there is little possibility of future profit and the KG assumption is reasonable.

In Figure 3.1, the marginal value $(\max_x \mu_x^{n+1}) - (\max_x \mu_x^n)$ of the information obtained from the sample taken at time n is plotted against n for one particular simulation from the slippage configuration described in Section 3.4.1. Samples are taken according to the LL1 sampling rule. In this simulation the marginal value indeed tended to decrease. Other sampling rules also display this tendency.

Finally, we note the KG stopping rule will better approximate the optimal stopping rule τ^* of Proposition 3.3.1 as the function C becomes more strictly convex. This is because the possibility of continuing after $n+1$ becomes increasingly remote as the marginal cost of

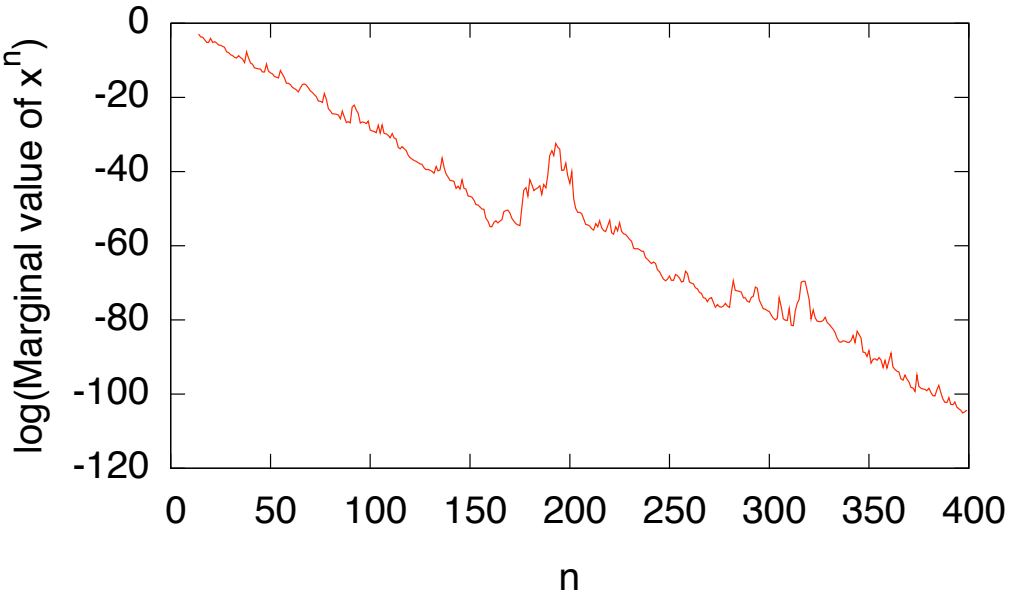

Figure 3.1: The logarithm of the marginal value of measurement x^n, $\log((\max_x \mu_x^{n+1}) - (\max_x \mu_x^n))$, plotted vs. n. Sampling decisions were made by LL1, and the slippage configuration described in Section 3.4.1 was used.

continued measurement increases, better justifying the heuristic's single-sample assumption. The KG stopping rule is perfect, for example, in the case when $C(n) = \infty \, \mathbf{1}_{\{n \geq N\}}$ because it correctly continues for $n < N$, and stops at time N.

3.4 Numerical Results

We now explore the relative quality of KG and other stopping rules through numerical simulation on several test cases. The selection of test cases owes a great deal to the work in (Chick et al. 2007) and (Chick et al. 2009). We concentrated our effort on the test cases used in that work, since that is where the LL1 policy was first presented, and the test cases presented there show that LL1 with the EOC Bonf stopping rule underperforms. In particular, we provide results here for LL1 and LL(S) sampling rules under KG, EOC Bonf, and fixed stopping rules. This makes for a total of six choices of sampling and stopping rules. We explore these choices on three configurations: the slippage configuration (SC); the monotone decreasing means configuration (MDM); and random problem instances (RPI). These configurations are described in more detail below.

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

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

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

The results for LL1 and LL(S) on EOC Bonf and fixed stopping rules replicate results originally presented in (Chick et al. 2009), while the results for the KG stopping rule are new. We will see from these experiments that the KG stopping rule improves the performance of

LL1 to the point where it is comparable to LL(S) with the EOC Bonf stopping rule.

3.4.1 Slippage Configuration

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

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

We picture the relative performance of LL(S) and LL1 under KG, EOC Bonf, and fixed stopping rules in Figure 3.2. We see in these results that LL1 performs better under the KG sampling rule than it does under EOC Bonf, and that both adaptive stopping rules perform better than their fixed counterparts under both sampling rules. We also see that LL1 under KG stopping performs better than does LL(S) with EOC Bonf stopping in this problem setting.

3.4.2 Monotone Decreasing Means

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

We picture the relative performance of our sampling and stopping rules for the MDM configuration in Figure 3.3. We see again in these results that LL1 performs better under the KG sampling rule than it does under EOC Bonf, and that both adaptive stopping rules outperform better their fixed counterparts. Unlike in the SC configuration, however, we see in this configuration that LL1 under KG stopping performs is outperformed by LL(S) with

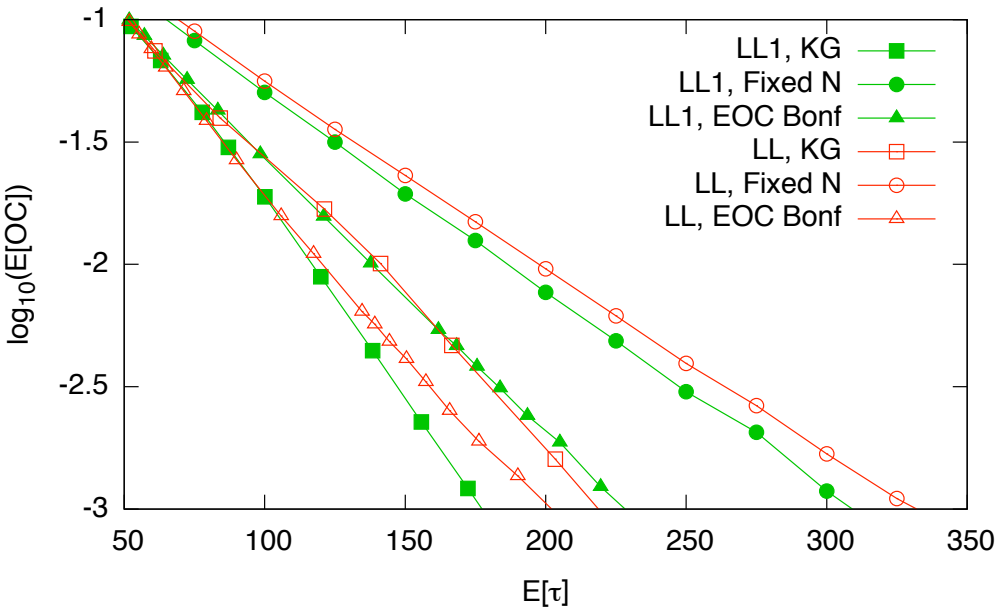

Figure 3.2: Slippage configuration with $\delta = 0.5$, 5 alternatives, and sampling variance 1.

EOC Bonf stopping.

3.4.3 Random Problem Instances

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

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

In Figure 3.4 we see results from a typical randomly generated problem configuration.

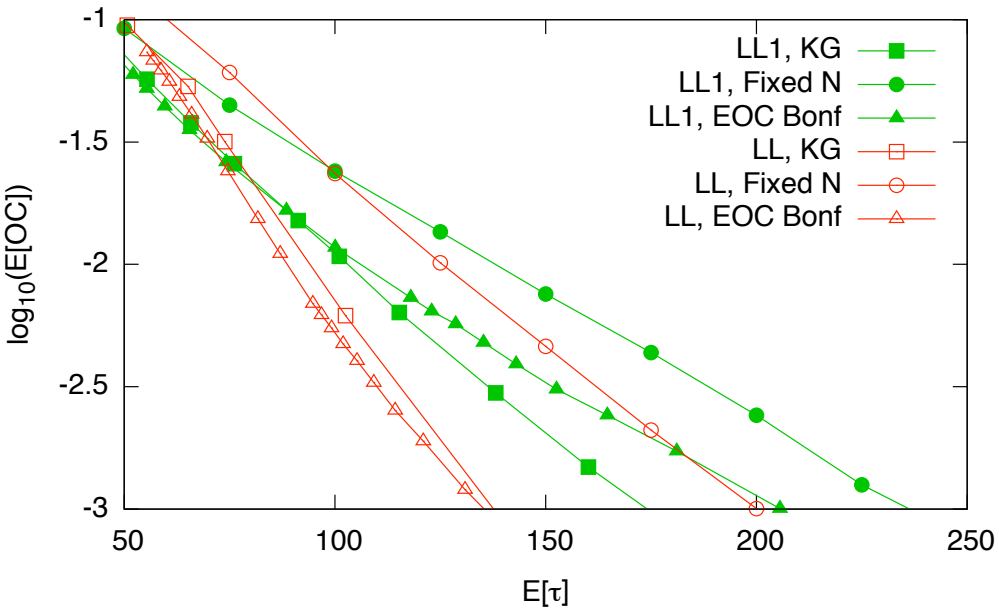

Figure 3.3: Monotone decreasing means configuration with $\delta = 0.5$, 10 alternatives, and sampling variance 1.

Again, these results are typical in the particular sense that LL1 performed better with KG stopping than with EOC Bonf stopping, and that LL1 with KG stopping performed similarly to LL(S) with EOC Bonf stopping. In this particular case LL1 with KG stopping performed better, but this advantage was reversed in other configurations not pictured.

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

3.5 Conclusion

We have shown that the LL1 procedure, introduced as a sampling rule that can be derived exactly under the knowledge-gradient assumption, can be understood in a broader context as the sampling portion of a composite sampling and stopping rule that can again be derived exactly under the same knowledge-gradient assumption. Furthermore, we can obtain

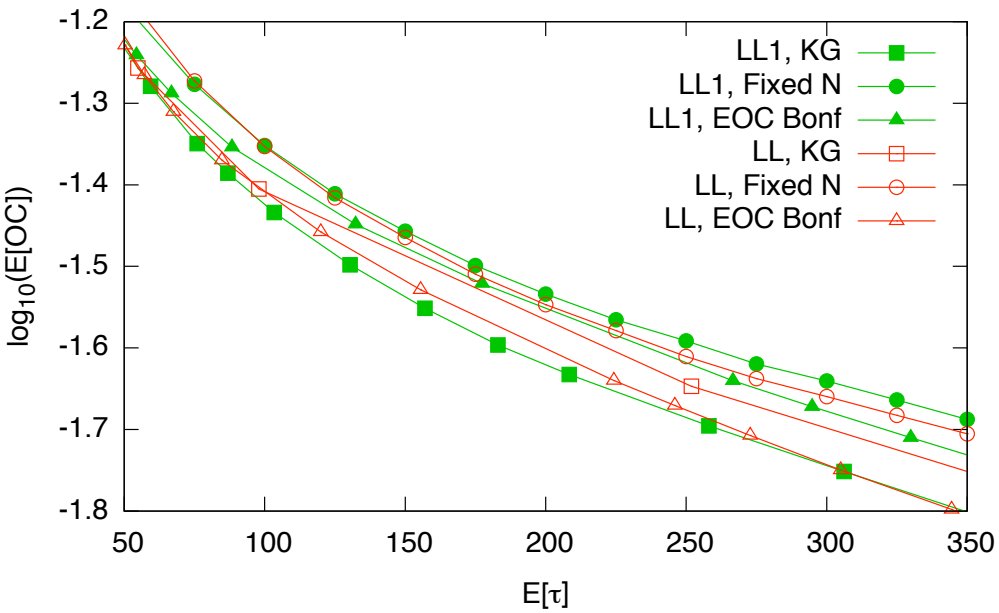

Figure 3.4: Random problem instance generated according to a normal-gamma prior with 5 alternatives. The sampling distribution had means $\mu = [0.16, 0.21, -1.40, -1.20, -0.16]$ and precisions $\beta = [1.07, 1.04, 0.89, 1.05, 0.97]$.

significantly better efficiency by sampling and stopping according to this composite rule as compared to sampling with LL1 but using another adaptive stopping rule. The resulting performance is commensurate with other well-regarded sampling/stopping rules like LL(S) with EOC Bonf stopping.

This is valuable first because it provides a new sampling/stopping rule that works well overall and is likely to work even better in small sample situations. Secondly, and we believe more importantly, it provides a general framework under which composite sampling/stopping rules may be derived. The knowledge-gradient assumption, which is that the current time-period is our last opportunity to sample, depended in no way on the normality of the sample distributions, the particular form of the prior, or on the independence of the samples through time. As long as we can evaluate the one-dimensional integral needed to compute the marginal value of a single measurement, we can create a knowledge-gradient based heuristic for the problem at hand. The quality of the resulting heuristic must of course be evaluated in each new situation to which it is applied, but the results described here add to the evidence accumulated in other problem settings (see Chapters 2 and 4) suggesting that the heuristic performs well in many important problems.

Chapter 4

The Knowledge-Gradient Policy for Correlated Normal Beliefs

In this chapter we consider an R&S problem, as we did in Chapters 2 and 4, but in our statistical model we explicitly account for correlation or dependence between alternatives. In many problems, this explicit accounting for dependence allows the resulting KG policy to perform dramatically better than policies (including the KG policy) derived using the independent and more traditional Bayesian formulation of the R&S problem given in Chapter 2. We say that there is dependence between alternatives in an R&S problem if, when we measure one alternative, we learn something about the others. To further understand what we mean by dependence between alternatives in an R&S problem, consider the following short list of example dependent R&S problems 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.

We see dependence between alternatives in these examples as follows. 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 chapter, we introduce a fully sequential sampling technique called the correlated 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.

In this chapter, we develop a KG policy for the Bayesian R&S problem with correlated normal beliefs. We call the KG policy for this problem the *correlated KG policy* to distinguish it from KG policies for other problems. Special cases of the correlated KG policy were introduced for a one-dimensional continuous domain with a Wiener process prior by

(Mockus 1972) (see (Mockus et al. 1978) for a description in English), and, as discussed in Chapter 2, for the finite domain discrete independent normal means case by (Gupta & Miescke 1996). While these previous 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 procedures 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 correlated KG policy in detail in Section 4.2 by making explicit the correlated prior and associated model, and then, in Section 4.3, 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 Chapter 2. The first, Remark 4.3.1 in Section 4.3, follows directly from the policy's construction and states that the policy is optimal when there is only one measurement left to make. We then show in Section 4.4 that the correlated KG policy always eventually discovers the best alternative if allowed enough measurement opportunities. This convergence result in turn implies asymptotic optimality of the policy. While many inefficient policies are asymptotically optimal, e.g., the equal allocation policy, we argue that asymptotic and myopic optimality taken together are more suggestive of quality than each would be individually. We note further that the correlated KG policy is unique in being the only stationary policy that is both myopically and asymptotically optimal. In Section 4.5 we then show a general bound on suboptimality for the finite sample case. Because the prior lacks independence, the proofs of asymptotic optimality and bounded finite sample suboptimality are more in-

volved, and the statements of the theorems themselves are somewhat different than in the independent case.

Finally, in Section 4.6, we apply the correlated KG policy to the maximization of a random function in noisy and noise-free environments, which are problems previously considered by Bayesian global optimization methods. 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.

4.1 Literature Review

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 & Kelton 2000), which include control variates (Nelson & Staum 2006) and common random numbers (Kim & 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 likely 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 & 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 procedures exist for maximizing a quality measure applied to

the chosen alternative, beginning with (Gupta & Miescke 1996) and continuing with two distinct families of procedures: the Optimal Computing Budget Allocation (OCBA) (Chen et al. 1996, Chen, Lin, Yücesan & Chick 2000, He et al. 2007), and Value of Information Procedures (VIP) (Chick & Inoue 2001*b*, Chick et al. 2009). Computational experiments (Inoue et al. 1999, Branke et al. 2007) demonstrate that these procedures perform very well, and their sequential nature allows them to achieve even greater efficiency than could a deterministic or two-staged procedure (Chen et al. 2006).

While OCBA- and VIP-based procedures for exploiting common random numbers have also been introduced in (Chick & Inoue 2001*a*) 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.

4.2 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,

$$\theta \sim \mathcal{N}\left(\mu^0, \Sigma^0\right). \tag{4.1}$$

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 (4.1), 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 $\Sigma^n := \mathrm{Cov}\left[\theta \mid \mathcal{F}^n\right]$. 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 4.2.1.

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

$$\sup_{\pi \in \Pi} \mathbb{E}^\pi \max_i \mu_i^N. \tag{4.2}$$

4.2.1 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

$$\mu^{n+1} = \Sigma^{n+1}\left((\Sigma^n)^{-1}\mu^n + (\lambda_{x^n})^{-1}\hat{y}^{n+1}e_{x^n}\right), \tag{4.3}$$

$$\Sigma^{n+1} = ((\Sigma^n)^{-1} + (\lambda_{x^n})^{-1}e_{x^n}(e_{x^n})')^{-1}, \tag{4.4}$$

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 (4.4) using the Sherman-Woodbury matrix identity (see, e.g., (Golub & Van 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 (4.3) to obtain a new recursion for μ^{n+1} as well. Taking $x = x^n$ temporarily to simplify subscripts, the recursions obtained are

$$\mu^{n+1} = \mu^n + \frac{\hat{y}^{n+1} - \mu^n_x}{\lambda_x + \Sigma^n_{xx}} \Sigma^n e_x, \tag{4.5}$$

$$\Sigma^{n+1} = \Sigma^n - \frac{\Sigma^n e_x e_x' \Sigma^n}{\lambda_x + \Sigma^n_{xx}}. \tag{4.6}$$

The formulas (4.5) and (4.6) hold even when Σ^n is positive semi-definite and not necessarily invertible, even though the formulas (4.3) and (4.4) 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

$$\tilde{\sigma}(\Sigma, x) := \frac{\Sigma e_x}{\sqrt{\lambda_x + \Sigma_{xx}}}. \tag{4.7}$$

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 (4.5) as

$$\mu^{n+1} = \mu^n + \tilde{\sigma}(\Sigma^n, x^n) Z^{n+1}. \tag{4.8}$$

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 (4.8). 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 (4.8) 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 (4.6) 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.

4.2.2 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 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

$$V^n(s) = \max_{x \in \{1,\ldots,M\}} \mathbb{E}\left[V^{n+1}(S^{n+1}) \mid S^n = s, x^n = x\right], \quad \text{for every } s \in \mathbb{S}. \tag{4.9}$$

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

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

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

$$X^{\pi^*,n}(s) \in \underset{x \in \{1,\ldots,M\}}{\arg\max}\ Q^n(s,x), \quad \text{for every } s \in \mathbb{S},\ n < N, \text{ and } x \in \{1, \ldots, M\}, \tag{4.10}$$

is optimal.

4.2.3 Benefits of measurement

We state the following preliminary results concerning the benefits of measurement. These results will be used later to show various 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 4.2.1 shows that if we provide more measurement opportunities to any stationary measurement policy, then it will perform better on average.

Proposition 4.2.1. *For any stationary policy π and state $s \in \mathbb{S}$, $V^{\pi,n}(s) \geq V^{\pi,n+1}(s)$.*

Proposition 4.2.2 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 4.2.2. *For $s \in \mathbb{S}$ and $x \in \{1, \ldots, M\}$, $Q^n(s,x) \geq V^{n+1}(s)$.*

Propositions 4.2.1 and 4.2.2 are similar to results proved for the independent case in Chapter 2, 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 4.2.3 then uses Proposition 4.2.2 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 4.2.1, but for the optimal policy. Note that the optimal policy is not generally known to be stationary.

Corollary 4.2.3. *For $s \in \mathbb{S}$, $V^n(s) \geq V^{n+1}(s)$.*

Proof. In Proposition 4.2.2, take the extra measurement x to be the measurement made by an optimal policy in state s. For such an x, $Q^n(s,x) = V^n(s)$. □

4.3 Knowledge Gradient

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

$$X^{KG}(s) \in \arg\max_x \mathbb{E}_n \left[\max_i \mu_i^{n+1} \mid S^n = s, x^n = x \right] - \max_i \mu_i^n \tag{4.11}$$

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 Chapter 2, and may be used together with an independent normal prior to derive the $(R_1, \ldots, R_1)$ procedure in (Gupta & Miescke 1996).

It may also be used together with a Wiener process prior to derive the one-step Bayes procedure 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 4.3.1. *When $N = 1$, the KG policy satisfies condition* (4.10) *and is thus optimal.*

Remark 4.3.2. *Consider any stationary policy π and suppose that it is optimal when $N = 1$. Then its decision function X^π must satisfy* (4.10)*, and hence must also satisfy* (4.11)*. The policy π is then the same as the KG policy, except possibly in the way it breaks ties in* (4.11)*. In this sense, the KG policy is the only stationary myopically optimal policy.*

The KG policy (4.11) was calculated in (Gupta & Miescke 1996), and again more explicitly in Chapter 2, 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,

$$X^{KG}(S^n) \in \arg\max_x \tilde{\sigma}_x(\Sigma^n, x) f\left(\frac{-|\mu_x^n - \max_{i \neq x} \mu_i^n|}{\tilde{\sigma}_x(\Sigma^n, x)}\right) \quad \text{if } \Sigma^n \text{ is diagonal,} \tag{4.12}$$

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 $\text{diag}(\Sigma^0)$. This allows using the formula (4.12), which we will see is easier to compute than the general case (4.11). We will also see, however, that the additional computational complexity incurred by computing (4.11) for non-diagonal Σ^n is rewarded by increased per-measurement efficiency.

4.3.1 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, (4.11), may be rewritten as

$$X^{KG}(S^n) = \arg\max_x \mathbb{E}\left[\max_i \mu_i^n + \tilde{\sigma}_i(\Sigma^n, x^n)Z^{n+1} \mid S^n, x^n = x\right] - \max_i \mu_i^n$$
$$= \arg\max_x h(\mu^n, \tilde{\sigma}(\Sigma^n, x)) \tag{4.13}$$

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 4.3.3. *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 4.3.3 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 & 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 Appendix B.

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.

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 ← −∞, c_1 ← +∞, A ← {1}
for i = 1 to M − 1 do
    c_{i+1} ← +∞,
    repeat
        j ← A[end(A)]
        c_j ← (a_j − a_{i+1})/(b_{i+1} − b_j).
        if length(A) ≠ 1 and c_j ≤ c_k, where k = A[end(A) − 1] then
            A ← A(1, . . . , end(A) − 1)
            loopdone ← false
        else
            loopdone ← true
        end if
    until loopdone
    A ← (A, i + 1)
end for
```

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,

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

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

$$\begin{aligned} h(a,b) &= \mathbb{E}\left[a_{g(Z)} + b_{g(Z)}Z - a_{g(0)}\right] \\ &= \left[\sum_{i=g(0)}^{M-1} (b_{i+1} - b_i)\mathbb{E}\left[(-c_i + Z)^+\right]\right] + \left[\sum_{i=1}^{g(0)-1} (b_i - b_{i+1})\mathbb{E}\left[(c_i - Z)^+\right]\right] \\ &= \sum_{i=1}^{M-1} (b_{i+1} - b_i)\mathbb{E}\left[(-|c_i| + Z)^+\right] = \sum_{i=1}^{M-1} (b_{i+1} - b_i)f(-|c_i|), \end{aligned} \tag{4.14}$$

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 (4.14) 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 $\log(b_{i+1}-b_i)+\log\phi(c_i)+\log(1-|c_i|R(-|c_i|))$, 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 $1/|c_i| \leq R(-|c_i|) \leq |c_i|/(c_i^2+1)$ (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 (4.14) 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 $O(M^2 \log M)$.

4.3.2 Summary of optimality results

The KG policy exhibits several optimality properties. First, it is optimal by construction when $N = 1$ (Remark 4.3.1). Second, it is optimal in the limit as $N \to \infty$ (Section 4.5). Third, the suboptimality gap between the optimal policy and the KG policy is bounded (Section 4.5). These optimality results extend optimality results proved in Chapter 2 for independent normal priors.

4.4 Convergence and asymptotic optimality

In this section we show that the KG policy is asymptotically optimal in the limit as the number of measurements N grows large. This may be understood as a convergence result – that given enough opportunities to measure, the policy eventually discovers the alternative that is truly the best.

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 carry this kind of asymptotic optimality, 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.

In the case of the KG policy, however, asymptotic optimality is more suggestive when understood together with Remark 4.3.1, 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 any policy that maximizes both simultaneously is worth consideration.

One may construct other policies that are both myopically and asymptotically optimal, for example by measuring according to the knowledge-gradient policy on the first measure-

ment 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 4.3.2 shows that, except for differences on how ties are broken in (4.11), the KG policy is the only stationary policy that is both myopically and asymptotically optimal.

We begin our proof of the asymptotic optimality of the KG policy by showing in Proposition 4.4.1 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 4.4.4 that this value is achieved by any stationary policy that measures every alternative infinitely often. Thus, any stationary policy that samples every alternative infinitely often is asymptotically optimal. Finally, we show in Theorem 4.4.5 with the use of Lemmas 4.4.2 and 4.4.3 that 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 measure some other alternative.

Since we will be varying the number N of measurements allowed, we use the notation $V^0(\,\cdot\,; N)$ to denote the value function at time 0 when the problem's terminal time is N. We then define the *asymptotic value function* $V(\,\cdot\,; \infty)$ by the limit $V(s; \infty) := \lim_{N\to\infty} V^0(s; N)$ for $s \in \mathbb{S}$. Below, Proposition 4.4.1 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 4.4.1 shows that the limit $V^\pi(s; \infty)$ exists for every $s \in \mathbb{S}$.

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 4.4.1 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 4.4.1 then corresponds with our intuition that no feasible measurement policy can outperform this oracle. We will use Proposition 4.4.1 later to show the asymptotic optimality of the KG policy.

Proposition 4.4.1. *Let $s \in \mathbb{S}$. Then the limit $V(s;\infty)$ exists and is bounded above by*

$$U(s) := \mathbb{E}\left[\max_i \theta_i \mid S^0 = s\right] < \infty, \tag{4.15}$$

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 4.4.1 generalizes Proposition 5.1 from Chapter 2, and the proof found there may be easily extended to include the general correlated case. We therefore omit the proof.

We now present three lemmas leading up to the main result of this section, Theorem 4.4.5. The proofs of all three lemmas may be found in the appendix.

Lemma 4.4.2. *(S^n) converges almost surely to a random variable S^∞ in $\mathbb{S}$.*

Lemma 4.4.2 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 4.4.3. *Let $s = (\mu, \Sigma) \in \mathbb{S}$. If $V^N(s) = Q^{N-1}(s;x)\ \forall x$ then $V^N(s) = U(s)$.*

Lemma 4.4.3 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 4.4.2 satisfies asymptotic optimality.

Lemma 4.4.4. *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.*

Lemma 4.4.4 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 4.4.3.

We now state the main theorem of this section.

Theorem 4.4.5. *The KG policy is asymptotically optimal.*

This theorem shows that, given the opportunity to measure infinitely often, the KG policy will discover which alternative is best. In a sense, this theorem is a convergence result because it shows that the policy's estimate of the best alternative will always converge to the correct result.

The proof of Theorem 4.4.5 may be found in the appendix, but we provide a sketch here. The main argument of the proof 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 $Q^{N-1}(S^\infty; x) = V^N(S^\infty)$.

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 Chapter 2 of the asymptotic optimality of the KG policy under an independent prior cannot be simply extended to the correlated case.

4.5 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 bound the suboptimality of the KG policy in the region in between. This bound will be tight for small N and will loosen as N increases. It will involve a norm $||\cdot||$ on $\mathbb{R}^M$ defined by $||u|| := \max_i u_i - \min_j u_j$. We will also write $||\tilde{\sigma}(\Sigma, \cdot)||$ to indicate $\max_x ||\tilde{\sigma}(\Sigma, x)||$.

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

Lemma 4.5.1. *Let $s = (\mu, \Sigma) \in \mathbb{S}$. Then $V^{N-1}(s) \le V^N(s) + ||\tilde{\sigma}(\Sigma, \cdot)||/\sqrt{2\pi}$.*

Proof. By (4.9), we have $V^{N-1}(s) = \max_{x^{N-1}} \mathbb{E}\left[V^N(S^N) \mid S^{N-1} = s\right]$. We may bound the inner term $V^N(S^N)$ by

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

Thus, we may bound the whole expression by

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

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 $\max_i b_i Z = (\max_i b_i)\, Z\mathbf{1}_{\{Z \ge 0\}} + (\min_i b_i)\, Z\mathbf{1}_{\{Z<0\}}$. Thus

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

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

The following theorem extends the bound shown in Lemma 4.5.1 to hold when there are any number of measurements remaining.

Theorem 4.5.2.

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

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,

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

Applying Lemma 4.5.1 to $V^{N-1}(S^{n+1})$ on the right-hand side,

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

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 4.2.1 to bound the suboptimality of the KG policy in the following corollary.

Corollary 4.5.3.

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

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 4.2.1 we have $V^{KG,N-1}(S^n) \leq V^{KG,n}(S^n)$. Substituting the resulting inequality $V^{N-1}(S^n) \leq V^{KG,n}(S^n)$ into Theorem 4.5.2 shows the corollary. □

4.6 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, Mockus 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 & 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 assumed 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,

$$\mathrm{Cov}(\theta(i), \theta(j)) = \beta \exp\{-\sum_{k=1}^{d} \alpha_k (i_k - j_k)^2\}, \tag{4.16}$$

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 are 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 calculations with the full posterior is often 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, our knowledge-gradient approach as 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

$$\mathrm{EI}(x) = \mathbb{E}_n \left[\max \left(\theta(x^n), \max_{k<n} \theta(x^k) \right) \mid x^n = x \right] - \max_{k<n} \theta(x^k), \tag{4.17}$$

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 like the Nelder-Mead simplex search procedure 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 (4.17) 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 $\theta(x^k)$ 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 $\mathrm{EI}(x)$ 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,

$$\mathrm{EI}(x) = \mathbb{E}_n \left[\max \left(\mu^n(x^{**}) - \mu^{n+1}(x), 0 \right) \mid x^n = x \right] \cdot \left(1 - \sqrt{\frac{\lambda_x}{\Sigma^n_{xx} + \lambda_x}} \right). \tag{4.18}$$

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 (4.18) goes to the noise-free expected improvement in (4.17). 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 4.1, 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. 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 4.1. 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

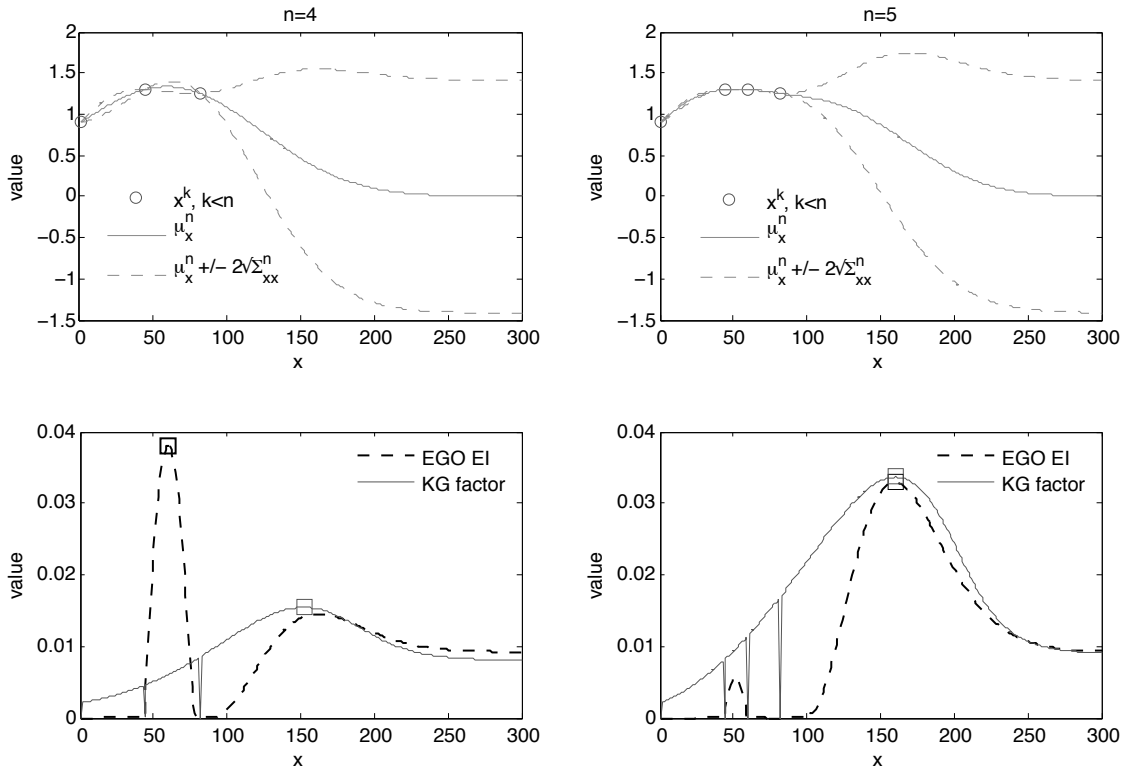

Figure 4.1: Upper plots display the posterior belief at two different points in time, with time $n = 4$ on the left and $n = 5$ on the right. The prior mean is plotted as a solid line, with two standard deviations above and below plotted as dotted lines. Previous measurements are circles. The $n = 5$ 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 $n = 5$.

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 follow these recommendations in the experiments described here.

In our first set of experiments, pictured in Figure 4.2, 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 $100/(M-1)^2$, $16/(M-1)^2$, $4/(M-1)^2$ respectively. With each truth, experiments were performed with normally distributed noise with standard deviations of 0.1 and 0.2. In this experiment we compare KG with SKO, both with correlated priors, and we also picture the performance of KG with 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 is uniform over $\mathbb{R}$, resulting in a first stage of size $M = 80$ in which each alternative is measured once in random order. Each combination of truth, noise variance and policy was replicated between 1000 and 2000 times, and the opportunity cost was recorded as a function 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 the 500 replications is plotted against iteration in Figure 4.2 for each choice of truth and noise variance. Sampled opportunity costs from batches of 25 replications were averaged

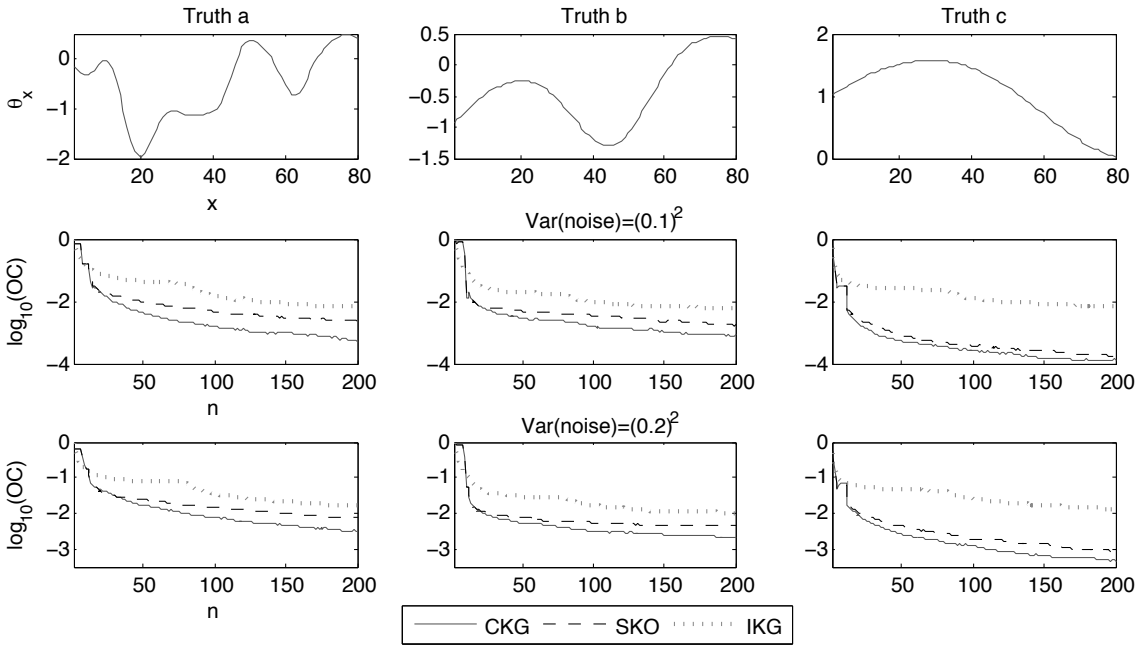

Figure 4.2: Comparison of SKO, correlated KG (CKG), and independent KG (IKG) on three functions drawn randomly from Gaussian process priors. CKG and SKO policies estimated hyperparameters adaptively with an initial stage of 22 measurements. IKG used an independent noninformative prior. The top row shows the three functions tested, the middle row shows policy performance at noise variance $(.1)^2$, and the bottom row shows policy performance at noise variance $(.2)^2$. Each policy performance plot shows the $\log_{10}$ of expected opportunity cost vs. iteration for the two algorithms. Standard errors were estimated from batches of 25 replications, but 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.

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 4.2.

The figure shows that correlated KG outperforms SKO in each of the six situations tested, and at the final measurement ($n = 200$) the expected opportunity cost incurred by SKO is 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 is 4.4, 2.1 and 1.3 when the measurement variance is $(.1)^2$, and 2.4, 2.0 and 1.9 when the measurement variance is $(.2)^2$.

The figure also shows that both SKO and correlated KG outperform independent KG,

often by a significant margin. The independent KG policy was shown in Chapter 2 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 4.3, 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 the discretization level was $L = 200$. 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 measurement policies to use the true hyperparameter values rather than the MAP estimate. The results, in Figure 4.3, show that the difference between the policies is 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 4.4, we compare KG

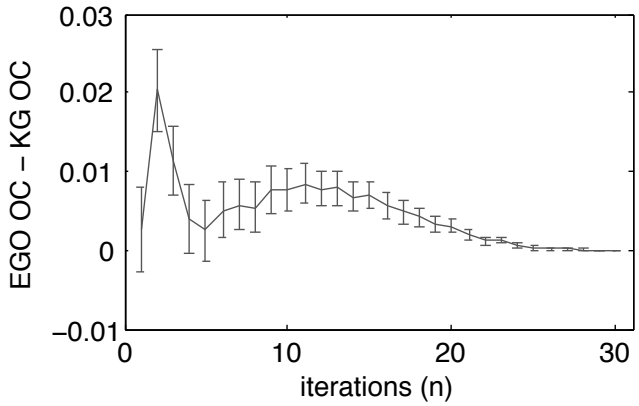

Figure 4.3: Comparison of KG with EGO on 26,000 one-dimensional functions drawn randomly from a Gaussian process prior with parameters $\beta = 1/2$, $\alpha_1 = 1/64$ and $L = 200$. Measurements were error-free, both policies were given the correct hyperparameters, and the initial stage had one measurement. The plot shows the difference in expected opportunity cost between the two polices, with a positive difference indicating that KG performed better than EGO. Standard errors were calculated from batches of 100 replications, and the error bars represent a single standard error above and below the estimated difference.

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 and the discretization levels and domains are given in the table in Figure 4.4. These functions are traditionally minimized, and we do so in these numerical experiments by maximizing their negative.

In all three tests we see the initial stage in which both algorithms perform similarly. On the Tilted Branin and Hartman-3 functions, both KG and SKO rapidly improve their opportunity cost as the first stage ends and both their implementation decision and measurement decisions are free to range across the entire domain. KG is 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 maintains this advantage through the increasing iterations.

On the six-hump camelback function, both SKO and KG algorithms suffer 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 leads them to believe that the function is better at a point far from where they have measured previously, when in fact

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)

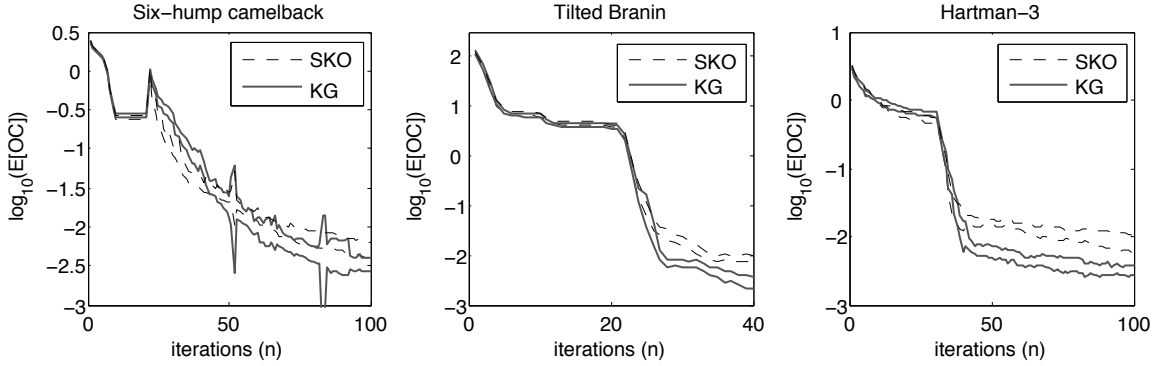

Figure 4.4: Comparison of KG with SKO on three standard test functions. Both policies estimated hyperparameters adaptively with an initial stage of 2 × dimension + 2 measurements. Plots show the $\log_{10}$ of the expected opportunity cost vs. iteration n at measurement variance $(.1)^2$ for three functions: 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). Standard errors were calculated by taking batches of 25 replications.

this belief is incorrect. Both policies quickly recover, but SKO initially recovers 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, 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 has recovered completely and is reducing its opportunity cost more rapidly. By the larger iterations, KG is outperforming 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 performs 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.

4.7 Conclusion

In this chapter we presented a policy for sequential correlated multivariate normal Bayesian R&S, generalizing the policy presented in (Gupta & Miescke 1996) and Chapter 2, 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 (4.11), 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.

Chapter 5

Asymptotic Optimality of Sequential Sampling Policies

We consider the general class of offline sequential Bayesian information collection problems. This class of problems is distinguished by: the assumption of a prior distribution on some underlying and unknown truth; the opportunity to adaptively perform a sequence of measurements whose results are observed with noise; an "implementation decision" made after all measurements are complete; and a loss assessed as a function of the implementation decision chosen and the underlying truth. Our goal when studying such problems is to design a measurement strategy that will best allow making a good implementation decision and, by doing so, minimize the expected loss.

A measurement strategy is a rule for selecting which of several types of measurement to make at each point in time. If the number of measurement types is finite, then one simple strategy is to choose the measurement types in round-robin fashion. With such a strategy, as the total number of measurements grows to infinity, each measurement type is chosen infinitely often, guaranteeing that we learn the full probability distribution of the observations resulting from each measurement type, and that the expected loss shrinks to its minimal attainable value. We call this property "asymptotic optimality."

It is often true that finite sample performance can be improved dramatically by employing more sophisticated adaptive measurement strategies in place of the round-robin or equal-allocation policy. When using such adaptive strategies, we would like to guarantee

that we retain asymptotic optimality, but checking this is seldom straightforward. Indeed, one adaptive policy lacking asymptotic optimality is interval estimation (Kaelbling 1993). When applied to R&S this policy usually works quite well, but occasionally fails to disastrous effect due to its lack of asymptotic optimality (see Chapter 2).

In this chapter, we provide an easy-to-check sufficient condition for asymptotic optimality, and use it to show asymptotic optimality of three measurement policies proposed previously in the literature: the OCBA policy for linear loss proposed in (He et al. 2007), the LL(S) policy proposed in (Chick & Inoue 2001*b*), and the LL(1) policy proposed in (Branke et al. 2007). We also offer a new proof of asymptotic optimality within this general framework for the $(R_1, \ldots, R_1)$ policy from (Gupta & Miescke 1996), for which asymptotic optimality was shown in Chapter 2.

We also show asymptotic optimality of a broad class of those sampling policies known as knowledge-gradient policies. A knowledge-gradient policy is any policy that chooses the measurement type that would be Bayes-optimal were it the last one allowed. Asymptotic optimality, which is optimality in the limit as the number of measurements remaining grows to infinity, is particularly interesting for knowledge-gradient policies because these policies are constructed to be optimal in the other extreme situation in which only one measurement remains, without any concern for the large-sample limit.

- We would like to discover which of several chemical compounds would make the most promising candidate for drug development against a particular target disease. We estimate a compound's potential by using it to treat diseased and non-diseased cells in a laboratory setting. Measurement decisions are made regarding which set of chemicals to test on a given day, observations are noisy, and the implementation decision is which subset of the chemicals to pass on to more advanced stages of drug development.

- We would like to discover which combination of pressure, temperature, and ingredient concentrations maximizes the yield of a chemical production process. Batches of chemical are produced, each with a particular choice of these factors, and the yields are measured. After a finite measurement period, a single choice of input factors is

fixed and the chemical is produced over a long period of time with these input factors.

- We would like to identify those sections of the water supply system that have dangerous chemical or biological contamination. Contamination may either be the result of industrial pollution or intentional contamination. At each decision period, the measurement policy chooses one or more chemical tests to run on water drawn from one or more several discrete locations in the water supply. After a measurement period, a message is relayed to the public, the content of which is decided by the implementation decision.

Different types of offline sequential Bayesian information collection problems have been considered previously by distinct areas of the literature. The first example above is representative of the Bayesian R&S literature (see, e.g., (Berger & Deely 1988), (Chick 2000), (Branke et al. 2007)). Ranking and selection problems involve a finite number of alternatives (in the example, drug compounds) whose true values are unknown but can be measured with noise. The goal is to find which alternatives have the largest values, with numerous variations possible in the loss function, assumptions about sampling noise, choice of prior distribution on the unknown truths, and type of implementation decision (e.g., whether a single alternative believed to be best is selected, or a set of alternatives that likely contains the best). The examples to which we apply our sufficient condition for asymptotic optimality are all policies for R&S.

The second example shares with R&S a desire to discover which alternatives are best, but it differs in that the number of alternatives is now infinite. Such problems are considered from a Bayesian perspective in the Bayesian global optimization literature (see, e.g., (O'Hagan & Kingman 1978, Currin et al. 1991, Kennedy & O'Hagan 2001, Kleijnen 2009, van Beers & Kleijnen 2008)).

The third example differs from the previous two in that the goal of measurement is to find all locations with true values above a given level rather than finding that location with the largest level. It would be considered outside the scope of either the R&S or Bayesian global optimization. Indeed, offline sequential Bayesian information collection problems have been considered often outside both the R&S and Bayesian global optimization literatures, in, for

example, medical diagnosis (Kapoor & Greiner 2005); oil exploration (Bickel & Smith 2006); computer vision (Rimey & Brown 1994); and drug dosage-response estimation (Eichhorn & Zacks 1973). Although we focus here on applications to R&S, these other examples of offline sequential Bayesian information collection problems illustrate the broad importance of these problems.

Although the class of offline sequential Bayesian information collection problems is broad, we stress here the assumed distinctness of measurement and implementation decisions. In particular, we assume that all rewards are collected at the final time as a function only of the implementation decision, and depend only *indirectly* on measurements. This excludes a large class of problems such as multi-armed bandit problems (Gittins 1989, Berry & Fristedt 1985, Lai 1987, Brezzi & Lai 2002), in which actions provide both information and a direct cost or reward.

Asymptotic optimality for specific policies and specific offline sequential Bayesian information collection problems has been previously shown in the literature. Within the context of R&S, Chapters 2 and 4 show asymptotic optimality for the knowledge-gradient policy for the R&S problem with linear loss and independent normally distributed rewards, with Chapter 2 showing it for the case of an independent normal prior and 4 showing it for the more general case of a multivariate normal prior. Asymptotic optimality of the Expected Improvement policy (Mockus et al. 1978, Schonlau & Welch 1996), which is the knowledge-gradient policy for a Wiener-process prior on continuous functions on $[0, 1]$, is shown in (Locatelli 1997).

Asymptotic optimality is also similar to *consistency*, which is the property that the posterior belief eventually becomes concentrated on the true parameter values, either for any set of possible true parameters that is almost sure under the prior (first type), or for a specific set of potential truths (second type), where the specific set is often the entire space of potential truths. If we consider a special offline sequential Bayesian information collection problem with a single measurement type, then there is only one measurement policy possible, which is to measure with the only measurement type repeatedly. In this special case, the asymptotic optimality property is the same as the first type of consistency. This type of consistency was shown to hold almost surely in (Doob 1949). Subsequent work

on consistency (see, e.g., (Freedman 1963, Diaconis & Freedman 1986, Walker 2004) as well as the review in (Ghosh & Ramamoorthi 2003)) has concentrated on the second form of consistency, which is a generalization of the first. The results we present here may be understood as a generalization of the work in (Doob 1949) in a different direction, to the case of multiple measurement types.

We begin in Section 5.1 with a general description of the offline sequential Bayesian information collection problem. Then, in Section 5.2, we present the main result, which is a sufficient condition for asymptotic optimality of a measurement policy. In Section 5.3 we apply this result to the case of a R&S problem with normally distributed rewards and known variance, showing asymptotic optimality of the OCBA for linear loss policy from (He et al. 2007). In Section 5.4 we apply the main result again to the case of a R&S problem with normally distributed rewards, but now with unknown variance. We then show asymptotic optimality of the LL(S) policy from (Chick & Inoue 2001*b*). Finally, in Section 5.5, we show that the main result may be applied to a broad class of knowledge-gradient policies, and show in particular that it may be applied to the $(R_1, \ldots, R_1)$ policy from (Gupta & Miescke 1996), and to the LL_1 policy from (Chick et al. 2007, Chick et al. 2009). We conclude in Section 5.6.

5.1 Problem Description

We are interested in whether a particular sequential sampling rule is consistent. We examine this question in the setting of conjugate sampling within a parametric family.

5.1.1 Sampling Model

We begin by supposing we have an observation space $\mathcal{Y} \subseteq \mathbb{R}^{d'}$, a parameter space $\Theta \subseteq \mathbb{R}^d$, and a finite set $\mathcal{X}$ of measurement types. We also adopt a probability measure Q_0 on $(\Theta, \mathcal{B}(\Theta))$, which is our Bayesian prior, quantifying our beliefs on which underlying truths are most likely. Here $\mathcal{B}(\Theta)$ denotes the Borel σ-algebra on Θ, and similarly for sets other than Θ. We use the notation $M(\Theta)$ to denote the space of probability measures on Θ.

Corresponding to each $u \in \Theta$ and $x \in \mathcal{X}$ is a probability measure $P(\,\cdot\,; x, u)$ on the

observation space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ governing the likelihoods of our observations. We assume that this probability measure has a density $p(\,\cdot\,; x, u) : \mathcal{Y} \mapsto \mathbb{R}_+$ with respect to some σ-finite measure $\nu(\,\cdot\,; x)$ on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ given by

$$p(y; x, u) = \exp\left(\alpha(u; x)^T \gamma(y; x) - \zeta(u; x)\right), \tag{5.1}$$

where ^{T} denotes matrix transposition, $\alpha : \mathbb{R}^d \times \mathcal{X} \mapsto \mathbb{R}^{l'}$ and $\gamma : \mathcal{Y} \times \mathcal{X} \mapsto \mathbb{R}^{l'}$ are given functions with l' an integer, and the normalizing function $\zeta : \mathbb{R}^d \times \mathcal{X} \mapsto \mathbb{R}$ is defined by

$$\zeta(u; x) = \log \int_{\mathcal{Y}} \exp(\alpha(u; x)^T \gamma(y; x)) \nu(dy; x).$$

We further assume $\Theta = \left\{u \in \mathbb{R}^d : |\zeta(u; x)| < \infty \ \forall x\right\}$. This assumption is not restrictive, since we can place all our prior's probability mass on a proper subset of Θ if we desire.

We now construct the over-arching probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which we define the random variable θ and sequences of random variables $(X_t)_{t\geq 1}$ and $(Y_t)_{t\geq 1}$. We define a filtration $(\mathcal{F}^t)$ by letting $\mathcal{F}^t$ be the σ-algebra generated by $\{(X_{t'}, Y_{t'}) \mid t' \leq t\}$, and we define $\mathbb{P}$ by requiring

$$\begin{aligned}
&\mathbb{P}\{\theta \in A\} = Q_0(A) \quad \text{for } A \in \mathcal{B}(\Theta), \\
&\mathbb{P}\{Y_t \in C \mid X_t, \theta\} = P(C; X_t, \theta) \quad \text{for } C \in \mathcal{B}(\mathcal{Y}).
\end{aligned}$$

We also require the Y_t to be conditionally independent of the other random variables given θ and X_t, and for the X_t to be conditionally independent of θ given $\mathcal{F}^t$. We use the notation $\mathbb{P}_t$ to denote the conditional probability measure $\mathbb{P}\left\{\,\cdot \mid \mathcal{F}^t\right\}$, and the notation $Q_t := \mathbb{P}_t\{\theta \in \cdot\,\}$ to represent the posterior distribution on θ.

The distributions $\mathbb{P}_t\{X_{t+1} \in \cdot\,\}$ are chosen by the experimenter and may either be chosen to be concentrated on a single point in $\mathcal{X}$, or may be more general, allowing randomized measurement decisions. We assume that the sequential sampling rule depends only upon t and the information collected up to t on θ. That is, we assume the existence of some measurable function $\Pi : \mathbb{N} \times M(\Theta) \times \mathcal{B}(\mathcal{X}) \mapsto [0, 1]$ giving

$$\mathbb{P}_t\{X_{t+1} \in C\} = \Pi(t, Q_t, C) \text{ a.s.}$$

This choice of Π is the choice of experimental design or sequential sampling rule, and should be chosen to make the inference about θ as accurate as possible. This construction

may be understood by supposing that nature chooses and fixes θ according to Q_0 sight-unseen. The experimenter then conducts a sequence of experiments indexed by t, choosing for each experiment an experiment type X_t based only on the previous experiment types $X_{t'}$ and results $Y_{t'}$ (where $t' < t$) and possibly on some external source of randomization, and then observing the result Y_t whose distribution is determined by X_t and θ. From these experiments we then infer the value of θ.

With these definitions, we have specified that the sampling distribution for each measurement type in $\mathcal{X}$ comes from an exponential family, with the functions α and γ and their dependence on the common parameter θ allowing the possibility for dependence between the sampling distributions at each measurement type, with information from samples of one measurement type allowing inference about the sampling distributions for other measurement types.

5.1.2 Posterior Distribution on θ

In our Bayesian setting, our belief about θ adjusted by our observations up to time t may be described as the posterior distribution Q_t. In this exponential family setting, the posterior distribution on θ takes a convenient form. First, for each $t \in \mathbb{N}$ and $x \in \mathcal{X}$ define

$$S_{tx} = \sum_{t' \leq t} \mathbf{1}_{\{X_{t'}=x\}} \gamma(Y_{t'}; x), \qquad N_{tx} = \sum_{t' \leq t} \mathbf{1}_{\{X_{t'}=x\}}.$$

Also define larger random vectors containing them, $S_t = [S_{tx}]_{x \in \mathcal{X}}$ and $N_t = [N_{tx}]_{x \in \mathcal{X}}$. Here, S_{tx} takes values in $\mathbb{R}^{l'}$, S_t in $\mathbb{R}^{l'|\mathcal{X}|}$, N_{tx} in $\mathbb{R}$, and N_t in $\mathbb{R}^{|\mathcal{X}|}$. It will also be useful to define for each $x \in \mathcal{X}$ a random variable $N_{\infty,x} := \sum t' \mathbf{1}_{\{X_{t'}=x\}}$ taking values in $\mathbb{N} \cup \{\infty\}$. With these definitions, we write the posterior density on θ as

$$\frac{dQ_t}{dQ_0}(u) = \exp\left(\sum_{x \in \mathcal{X}} \left[\alpha(u;x)^T S_{tx} - \zeta(u;x) N_{tx} \right] - \Gamma(S_t, N_t) \right),$$

where the normalizer $\Gamma : \mathbb{R}^{l'|\mathcal{X}|} \times \mathbb{R}^{|\mathcal{X}|} \mapsto \mathbb{R}$ is defined by

$$\Gamma(s,n) := \log \int_\Theta \exp\left(\sum_{x \in \mathcal{X}} \alpha(u;x)^T s_x - \zeta(u;x) n_x \right) Q_0(du),$$

and has domain $\mathrm{dom}(\Gamma)$ given by

$$\mathrm{dom}(\Gamma) = \left\{ (s,n) \in \mathbb{R}^{l'|\mathcal{X}|} \times \mathbb{R}^{|\mathcal{X}|} : |\Gamma(s,n)| < \infty \right\}.$$

We assume that $\mathrm{dom}(\Gamma)$ is open.

This is an exponential family for the posterior distribution on θ parameterized by S_t and N_t. As written, the family has $(l'+1)|\mathcal{X}|$ parameters, but its rank l may be smaller. We now re-parameterize the family into its minimal-rank parameterization, and from there into its mean-value parameterization. The mean-value parameterization is then used in Section 5.2.

To write the family in minimal-rank form, we define linear functions $\beta : \Theta \mapsto \mathbb{R}^l$ and $\tau : \mathbb{R}^{l'|\mathcal{X}|} \times \mathbb{R}^{|\mathcal{X}|} \mapsto \mathbb{R}^l$ so that

$$\sum_{x\in\mathcal{X}} \alpha(u;x)^T s_x - \zeta(u;x) n_x = \beta(u)^T \tau(s,n) \quad \text{for all } u \in \Theta, s \in \mathbb{R}^{l'|\mathcal{X}|}, n \in \mathbb{R}^{|\mathcal{X}|}.$$

If the original family was of full rank then $l = (l'+1)|\mathcal{X}|$ and $\beta(u)$ is formed by concatenating the vectors $[\alpha(u;x), -\zeta(u;x)]$, $x \in \mathcal{X}$, and $\tau(s,n)$ is formed by concatenating $[s_x, n_x]$, $x \in \mathcal{X}$. If not, and $l < (l'+1)|\mathcal{X}|$, then $\beta(u)$ contains a subset of l linearly independent rows from $[\alpha(u;x), -\zeta(u;x)]_{x\in\mathcal{X}}$, and $\tau(s,n)$ is a linear combination of $[s_x, n_x]_{x\in\mathcal{X}}$ in which the linearly dependent rows have been combined away. With β and τ defined in this way,

$$\frac{dQ_t}{dQ_0}(u) = \exp\left(\beta(u)^T \tau(S_t, N_t) - \Lambda(\tau(S_t, N_t))\right), \tag{5.2}$$

where Λ is defined by

$$\Lambda(\tau) := \log \int_\Theta \exp\left(\beta(u)^T \tau\right) Q_0(du),$$

so that $\Lambda(\tau(S_t, N_t)) = \Gamma(S_t, N_t)$ and Λ has domain

$$\mathrm{dom}(\Lambda) = \left\{\tau \in \mathbb{R}^l : |\Lambda(\tau)| < \infty\right\} = \tau(\mathrm{dom}(\Gamma)).$$

Our assumption that $\mathrm{dom}(\Gamma)$ is open, together with the linearity of τ, implies that the natural parameter space $\mathrm{dom}(\Lambda) = \tau(\mathrm{dom}(\Gamma))$ is open. We then recall a fundamental result on exponential families (see, e.g., (Bickel & Doksum 2007) Theorem 1.6.4). The openness of $\mathrm{dom}(\Lambda)$ implies that $\nabla\Lambda(\tau(s,n)) = \mathbb{E}\left[\beta(\theta) \mid S_t = s, N_t = n\right]$ for all $s, n \in \mathrm{dom}(\Gamma)$, and $\nabla\Lambda$ is a bijection between the natural parameter space $\mathrm{dom}(\Lambda)$ and $\mathcal{K} := \nabla\Lambda(\mathrm{dom}(\Lambda))$. Thus we are free to parameterize our posterior at time t in terms of $K_t := \nabla\Lambda(\tau(S_t, N_t))$, which is a random variable taking values in $\mathcal{K}$. This is called the mean-value parameterization. Since K_t completely determines the posterior distribution Q_t on θ, our policy's sampling

decision can then be written entirely in terms of K_t, and we write $\Pi(t, k, C)$ to indicate $\mathbb{P}\{X_{t+1} \in C \mid K_t = k\}$.

We also use the notation $\mathbb{P}^{(k)}$ for $k \in \mathcal{K}$ to denote the measure on $(\Theta, \mathcal{B}(\Theta))$ uniquely determined by k. In particular, we then have $Q_t = \mathbb{P}^{(K_t)}$. We also write $\mathbb{E}^{(k)}$ to indicate the expectation taken under $\mathbb{P}^{(k)}$.

The following lemma is needed later when we work with the asymptotic behavior of the sampling policy, and concerns the limit of the stochastic process $(K_t)_{t\in\mathbb{N}}$. It uses the notation $\mathrm{cl}(\mathcal{K})$ to denote the closure of $\mathcal{K}$ in $\mathbb{R}^l$.

Lemma 5.1.1. *The limit $K_\infty := \lim_{t\to\infty} K_t$ exists almost surely, is integrable, and takes values in* $\mathrm{cl}(\mathcal{K})$.

Its proof may be found in the appendix.

5.1.3 Risk Function and Asymptotic Optimality

The sampling model and posterior distribution from Sections 5.1.1 and 5.1.2, and the inference on θ that they support, exist together with a loss function. The loss function quantifies a sampling policy's performance, and defines an optimization problem over the space of potential sampling policies.

Let $\mathcal{I}$ be a finite set of terminal or "implementation" decisions that could be employed, and $R : \Theta \times \mathcal{I} \mapsto \mathbb{R}_+$ a function that specifies the (non-negative) loss $R(\theta; i)$ incurred when taking implementation decision i while the true sampling distribution is given by θ. We assume that $u \mapsto R(u; i)$ is a measurable function for each i, and that $R(\theta; i)$ is integrable under Q_0 for all $i \in \mathcal{I}$.

Given this loss function and a fixed and finite number of samples T that we may take, we define the expected loss of the sampling policy as $\mathbb{E}\left[\min_{i\in\mathcal{I}} \mathbb{E}_T\left[R(\theta; i)\right]\right]$. Our concern is with the asymptotic expected loss of a policy, and so we define the asymptotic risk of a policy as

$$\lim_{T\to\infty} \mathbb{E}\left[\min_{i\in\mathcal{I}} \mathbb{E}_T\left[R(\theta; i)\right]\right].$$

This limit always exists since Jensen's inequality and the tower property of conditional expectation shows that $\mathbb{E}\left[\min_{i\in\mathcal{I}} \mathbb{E}_T\left[R(\theta; i)\right]\right]$ is nondecreasing in T, and the non-negativity

of R show that it is non-negative. We seek conditions on the sampling policy under which this asymptotic loss is as small as possible.

The best asymptotic loss that we can achieve through sampling is to perfectly learn the sampling distribution of each measurement type $x \in \mathcal{X}$. With this in mind, we call the sampling policy *asymptotically optimal* if

$$\lim_{T\to\infty} \mathbb{E}\left[\min_{i\in\mathcal{I}} \mathbb{E}_T\left[R(\theta;i)\right]\right] = \mathbb{E}\left[\min_i \mathbb{E}\left[R(\theta;i) \mid P(\,\cdot\,;x,\theta)\ \forall x \in \mathcal{X}\right]\right]. \tag{5.3}$$

If knowing the sampling distribution for all $x \in \mathcal{X}$ completely determines θ, i.e. if θ is identifiable, then the right hand side is simply $\mathbb{E}\left[\min_i R(\theta;i)\right]$, which is the minimal loss achievable given perfect information about θ. Thus, we may think of the asymptotic optimality condition as indicating convergence to perfect knowledge and a perfect implementation decision in the limit as our sampling policy is allowed infinitely many measurements.

These conditions, introduced in Section 5.2, use a function g defined for $k \in \mathcal{K}$ and $C \subseteq \mathcal{X}$ by

$$g(k;C) := \min_i \mathbb{E}^{(k)}\left[R(\theta;i)\right] - \mathbb{E}^{(k)}\left[\min_i \mathbb{E}^{(k)}\left[R(\theta;i) \mid P(\,\cdot\,;x,\theta)\ \forall x \in C\right]\right].$$

The quantity $g(k;C)$ gives the expected incremental value of learning the true sampling distribution of all measurement types $x \in C$ when we already have the posterior distribution given by k. Seen another way, $g(k;C)$ tells us the incremental value of measuring all $x \in C$ an infinite number of times.

The non-negativity of R implies that both outer expectations are non-negative and possibly equal to $+\infty$, and that g is well-defined when both expectations are finite. To ensure g is well-defined, we take its domain to be

$$\mathrm{dom}(g) := \left\{k \in \mathcal{K} : \mathbb{E}^{(k)}\left[R(\theta;i)\right] < \infty\ \forall i \in \mathcal{I}\right\}.$$

Then $g(k;C)$ is well-defined and finite for all $k \in \mathrm{dom}(g)$ because Jensen's inequality and the tower property of conditional expectation imply

$$\mathbb{E}^{(k)}\left[\min_i \mathbb{E}^{(k)}\left[R(\theta;i) \mid P(\,\cdot\,;x,\theta)\ \forall x \in C\right]\right] \le \min_i \mathbb{E}^{(k)}\left[R(\theta;i)\right].$$

This also shows that $g(k;C)$ is non-negative.

Additionally, fixing any $i \in \mathcal{I}$ and any $t \in \mathbb{N}$, our assumption that $R(\theta; i)$ is integrable under Q_0 implies that $\infty > \mathbb{E}\left[R(\theta; i)\right] = \mathbb{E}\left[\mathbb{E}\left[R(\theta; i) \mid K_t\right]\right]$, and hence $R(\theta; i)$ is almost-surely integrable under Q_t. Since this is true for each $i \in \mathcal{I}$, we have that $K_t \in \mathrm{dom}(g)$ almost surely.

We now have the relationship between the function g and asymptotic optimality given by the following lemma. The proof of this lemma may be found in the appendix.

Lemma 5.1.2. *The sampling policy is asymptotically optimal iff* $\lim_{t\to\infty} g(K_t; \mathcal{X}) = 0$ *almost surely.*

We now make the following assumption on the structure of the sampling policy, which may be understood loosely as assuming that, if the expected loss cannot be reduced by learning perfectly the result of any *single* measurement type, then there is also nothing to be gained from learning perfectly the results of *all* the measurement types.

Assumption 5.1.3. *If* $(k_t)_{t\in\mathbb{N}}$ *is a sequence converging in* $\mathrm{cl}(\mathcal{K})$ *and satisfying* $\lim_{t\to\infty} g(k_t; \{x\}) = 0$ *for all* $x \in \mathcal{X}$*, then* $\lim_{t\to\infty} g(k_t; \mathcal{X}) = 0$.

If this assumption holds, asymptotic optimality is equivalent to the condition $\lim_{t\to\infty} g(K_t; \{x\}) = 0$ almost surely for all $x \in \mathcal{X}$. This assumption is not restrictive, since in cases for which it does not hold we may expand the set $\mathcal{X}$ of allowed measurement types to a larger set $\mathcal{X}' \subseteq 2^{\mathcal{X}}$ for which it does hold. This is done by considering a block of measurements of formerly distinct types as a new composite measurement type. Nevertheless, for the discussion that follows, we must check that our sampling model satisfies Assumption 5.1.3 and, if it does not, expand $\mathcal{X}$ until it does.

5.2 Main Result: Sufficient Conditions for Asymptotic Optimality

Before stating the main result, we introduce two lemmas. The first lemma tells us that no further information is to be obtained about $R(\theta; i)$ from $P(\,\cdot\,; x, \theta)$ if we have already measured x infinitely often. It is essentially a restatement of the strong law of large numbers, and shares much with the Glivenko-Cantelli theorem (see, e.g., (Kallenberg 1997)). The

second lemma builds upon the first and tells us that the quantity $g(K_t; x)$, which is the incremental value beyond what is known at t of being told the sampling distribution of x, goes to zero if we measure x infinitely often and include increasing knowledge of its sampling distribution into K_t. The proofs for both lemmas may be found in the appendix.

Lemma 5.2.1. *For each $x \in \mathcal{X}$ and $i \in \mathcal{I}$, $\mathbb{E}_\infty [R(\theta; i) \mid P(\,\cdot\,; x, \theta)] = \mathbb{E}_\infty [R(\theta; i)]$ almost surely on $\{N_{\infty,x} < \infty\}$.*

Lemma 5.2.2. *For each $x \in \mathcal{X}$, the event $\{N_{\infty,x} < \infty\} \cup \{\lim_{t\to\infty} g(K_t; \{x\}) = 0\}$ is almost sure.*

For each $x \in \mathcal{X}$, define a special subset M_x of $\mathrm{cl}(\mathcal{K})$ by

$$M_x := \left\{ k \in \mathrm{cl}(\mathcal{K}) : \exists (k_t) \subseteq \mathrm{dom}(g) \text{ converging to } k \text{ with } \lim_{t\to\infty} g(k_t; \{x\}) = 0 \right\}.$$

In addition, define

$$M_* := \left\{ k \in \mathrm{cl}(\mathcal{K}) : \forall (k_t) \subseteq \mathrm{dom}(g) \text{ converging to } k, \lim_{t\to\infty} g(k_t; \{x\}) = 0 \ \forall x \in \mathcal{X} \right\}.$$

If, for each $x \in \mathcal{X}$, the function $k \mapsto g(k; \{x\})$ is continuous and can be extended continuously onto the closure $\mathrm{cl}(\mathrm{dom}(g))$ of its domain with a new range $\mathbb{R}_+ \cup \{\infty\}$, then these definitions may be simplified to

$$M_x = \{k \in \mathrm{cl}(\mathrm{dom}(g)) : g(k; \{x\}) = 0\},$$
$$M_* = \{k \in \mathrm{cl}(\mathrm{dom}(g)) : g(k; \{x\}) = 0, \ \forall x \in \mathcal{X}\}.$$

If Assumption 5.1.3 also holds then we have $M_* = \{k \in \mathrm{cl}(\mathrm{dom}(g)) : g(k; \mathcal{X}) = 0\}$.

In this case we may understand the set M_x as containing those knowledge states in which learning more about the sampling distribution of x will not improve the expected loss. This includes knowledge states in which the sampling distribution of x is already known. Because M_x may not be entirely contained within $\mathcal{K}$, some of its elements may not correspond to members of the exponential family. These elements may instead be understood as limits of posteriors from the exponential family, in the sense that they are in the closure $\mathrm{cl}(\mathcal{K})$ of $\mathcal{K}$. For example, in the normal setting discussed in Section 5.3, $\mathrm{cl}(\mathcal{K})$ includes knowledge states corresponding to measures concentrated entirely on a single point in Θ, and although such

a concentrated measure is not a normal distribution, it is a limit of normal distributions. The set M_* may be understood as the set of knowledge states in which further sampling has no value. This is the set to which we would like our sampling policy to push us, since it is the set in which we have learned as much as possible by sampling.

Since M_x contains those knowledge states for which measuring x has no incremental value, the sets $M_x \setminus M_*$ are the "x-sticking sets" in which a sampling policy measuring x will not its expected loss. If a policy measures a stuck measurement type when other measurement types are not stuck and offer improvement in the expected loss, it will not be asymptotically optimal.

For each $k \in \mathrm{cl}(\mathcal{K})$, we define $A_k := \{x \in \mathcal{X} : k \in M_x\}$ to be the set of measurements that are stuck when we are in knowledge state k. To avoid sticking, a sampling policy should avoid measuring x when in such states. In fact, to guarantee asymptotic optimality, a policy need only do a little more. It need only maintain an open buffer region around these sticking sets in which it avoids measuring the stuck alternatives. This is the essential content of Theorem 5.2.3 below.

The open buffer regions required by Theorem 5.2.3 are illustrated in Figure 5.1, where we picture the knowledge state that results from a R&S problem with two alternatives, samples from which are normally distributed with unknown mean and known variance. This problem is discussed in greater detail in Section 5.3. In it, if we take an independent normally distributed prior distribution on these two unknown means then the posterior distribution on each alternative is parameterized by its mean and variance, causing knowledge states to have four dimensions.

In the figure, we picture two of these four dimensions: the variances of our belief about each of the two unknown means. When the variance of belief about an unknown mean is zero then that mean is known perfectly, and so the sticking set for an alternative is this set of points. These are pictured as thick lines along the vertical axis for alternative 2 and along the horizontal axis for alternative 1. The point at the origin is M_*, and is where both alternatives are known perfectly. To ensure asymptotic optimality by Theorem 5.2.3, a sampling policy needs an open buffer region around the vertical axis, excluding the origin, in which it only measures alternative 2, and a buffer region around the horizontal axis,

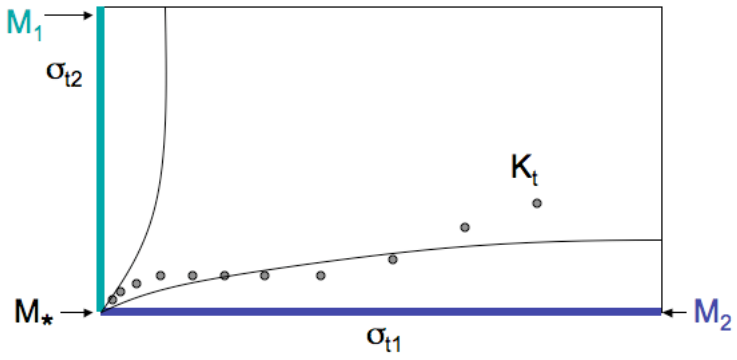

Figure 5.1: Two dimensional projection of $\mathcal{K}$ for a R&S problem with two alternatives and normally distributed rewards of unknown mean and known variance. The two dimensions pictured are the variance of the posterior belief on the unknown mean. The thick line on the vertical axis is M_2, where the variance of the posterior belief on alternative 2 is zero and this alternative's mean is known perfectly. Similarly, the thick line on the horizontal axis is M_1. The thin lines illustrate buffer regions around M_1 and M_2 in which the stuck alternatives are not measured, and the dots labeled K_t illustrate a sequence of knowledge states converging to M_*, where both alternatives' means are known perfectly.

excluding the origin, in which it only measures alternative 1. If this condition is met, then the sequence of knowledge states K_t will converge to the origin, attaining asymptotic optimality.

In this example, we see the necessity of the open buffer regions around the sticking sets. The variance of our belief about an alternative's unknown mean is always strictly positive after finitely many measurements, and can only approach 0 in the limit. Thus, a policy can never reach M_1 or M_2 with finitely many measurements but can only come arbitrarily close. If a policy merely guarantees that it never measures alternative 1 when in M_1, but measures 1 in all other knowledge states, then its knowledge state will converge to a point outside M_*, and the policy will not be asymptotically optimal.

We now state the theorem.

Theorem 5.2.3. *Suppose that the sampling model satisfies Assumption 5.1.3, and let* $\widetilde{\mathcal{K}}$ *be a measurable subset of the closure* $\mathrm{cl}(\mathcal{K})$ *of* $\mathcal{K}$ *such that the event* $\left\{K_t \in \widetilde{\mathcal{K}},\ \forall t \in \mathbb{N} \cup \{\infty\}\right\}$ *is almost sure. If each* $k \in \widetilde{\mathcal{K}} \setminus M_*$ *has an open neighborhood* U *in* $\mathrm{cl}(\mathcal{K})$ *satisfying*

$$\limsup_{t\to\infty} \sup_{k' \in U \cap \widetilde{\mathcal{K}} \cap \mathcal{K}} \Pi(t, k', A_k) < 1, \tag{5.4}$$

then the sampling policy given by Π *is asymptotically optimal.*

Although our discussion of Theorem 5.2.3 has assumed that g can be extended continuously, note that the theorem itself does not make this assumption. The theorem can be applied in numerous settings to show asymptotic optimality of specific policies. In the following sections we apply it to two different R&S problems, showing asymptotic optimality of two previously proposed policies. We also use it show asymptotic optimality of a broad class of policies known as knowledge-gradient policies, and to two specific knowledge-gradient policies previously proposed in the literature.

5.3 Application to Ranking and Selection with Normal Rewards and Known Variance

We apply Theorem 5.2.3 from the previous section to the normal R&S problem, using it to show convergence of a policy previously proposed in the literature. In the R&S problem, we have a collection of alternatives from which we would like to choose the one that is best in some sense. One may make a series of measurements before making the final selection. We suppose that measurements of an alternative are normally distributed, and the best alternative is the one with the largest mean.

This problem is formulated as follows. We call the set of alternatives $\mathcal{I}$ and its cardinality $|\mathcal{I}| = d > 1$. Then we take $\Theta = \mathbb{R}^d$, and samples of alternative $i \in \mathcal{I}$ are normally distributed with mean θ_i and known precision $\lambda_i > 0$. We suppose that measurements are taken in batches of some fixed integer size $B \geq 1$. We allow $B = 1$, which corresponds to taking the measurements one-at-a-time. The possible measurement types are then $\mathcal{X} = \mathcal{I}^B$, with $\sum_{b=1}^{B} \mathbf{1}_{\{x_b=i\}}$ being the number of samples measurement type x takes from alternative i. We also have $\mathcal{Y} = \mathbb{R}^B$, with y_b being an observation of alternative x_b. We take the *linear loss function* given by

$$R(\theta; i) = \max_{j \in \mathcal{I}} \theta_j - \theta_i. \tag{5.5}$$

The sampling density with respect to the Lebesgue measure is

$$p(y; x, \theta) = \prod_{b=1}^{B} \sqrt{\frac{\lambda_{x_b}}{2\pi}} \exp\left(-\frac{1}{2}\lambda_{x_b}(y_b - \theta_{x_b})^2\right).$$

To put this in the form of (5.1), we take $\alpha(u; x) = (u_{x_1}\lambda_{x_1}, \ldots, u_{x_b}\lambda_{x_b})$, $\gamma(y; x) = y$,

$\zeta(u;x) = \sum_{b=1}^{B} u_{x_b}^2 \lambda_{x_b}/2$, and $\nu(dy;x) = \prod_{b=1}^{B} \sqrt{\frac{\lambda_{x_b}}{2\pi}} \exp(-\frac{1}{2}\lambda_{x_b} y_{x_b}^2) dy$. Then the density of the posterior is given by (5.2) with

$$\beta(\theta) = \left[\theta_i \lambda_i, -\frac{1}{2}\theta_i^2 \lambda_i\right]_{i\in\mathcal{I}}, \qquad \tau(S_t, N_t) = \left[\tilde{S}_{ti}, \widetilde{N}_{ti}\right]_{i\in\mathcal{I}},$$

where $\tilde{S}_{ti} = \sum_{t'\le t} \sum_{b=1}^{B} \mathbf{1}_{\{X_{t'b}=i\}} Y_{t'b} = \sum_{x\in\mathcal{X}} S_{tx} \sum_{b=1}^{B} \mathbf{1}_{\{x_b=i\}}$ is the sum of all observations of alternative i, and $\widetilde{N}_{ti} = \sum_{t'\le t} \sum_{b=1}^{B} \mathbf{1}_{\{X_{t'b}=i\}} = \sum_{x\in\mathcal{X}} N_{tx} \sum_{b=1}^{B} \mathbf{1}_{\{x_b=i\}}$ is the number of times we have observed alternative i.

We suppose that the prior Q_0 on θ is a multivariate normal distribution with independent components and with θ_i having mean $\hat{\mu}_{0i}$ and variance $\sigma_{0i}^2 > 0$. With this assumption, the posterior Q_t is again normal with independent components, but now the mean and variance of θ_i, which we denote $\hat{\mu}_{ti}$ and σ_{ti}^2 respectively, are given by

$$\hat{\mu}_{ti} = \left(\frac{\hat{\mu}_{0i}}{\lambda_i \sigma_{0i}^2} + \tilde{S}_{ti}\right) \Big/ \left(\frac{1}{\lambda_i \sigma_{0i}^2} + \widetilde{N}_{ti}\right), \qquad \sigma_{ti}^2 = \left(\frac{1}{\lambda_i}\right) \Big/ \left(\frac{1}{\lambda_i \sigma_{0i}^2} + \widetilde{N}_{ti}\right).$$

With this, we compute K_t as $K_t = \mathbb{E}_t\left[\beta(\theta)\right] = \left[\hat{\mu}_{ti}\lambda_i, -\lambda_i(\hat{\mu}_{ti}^2 + \sigma_{ti}^2)/2\right]_{i\in\mathcal{I}}$, and $\mathcal{K}$ and its closure are given by

$$\mathcal{K} = \left\{ \left[u_i\lambda_i, -\lambda_i(u_i^2 + v_i)/2\right]_{i\in\mathcal{I}} : u \in \mathbb{R}^d, v \in \mathbb{R}_{++}^d \right\},$$
$$\mathrm{cl}(\mathcal{K}) = \left\{ \left[u_i\lambda_i, -\lambda_i(u_i^2 + v_i)/2\right]_{i\in\mathcal{I}} : u \in \mathbb{R}^d, v \in \mathbb{R}_{+}^d \right\},$$

with $\mathbb{R}_{++}$ being the set of strictly positive real numbers.

We may recover $\hat{\mu}_{ti}$ and σ_{ti}^2 from K_t, and, more generally, recover the mean and variance of θ_i under $\mathbb{P}^{(k)}$ from any $k \in \mathcal{K}$. To do so, we define functions $\hat{\mu}_i : \mathrm{cl}(\mathcal{K}) \mapsto \mathbb{R}$ and $\sigma_i^2 : \mathrm{cl}(\mathcal{K}) \mapsto \mathbb{R}_+$ by

$$\hat{\mu}_i(k) = k_{i1}/\lambda_i, \qquad \sigma_i^2(k) = -(k_{i1}/\lambda_i)^2 - 2k_{i2}/\lambda_i.$$

Then $\hat{\mu}_{ti} = \hat{\mu}_i(K_t)$, and $\sigma_{ti}^2 = \sigma_i^2(K_t)$. Note that we define $\hat{\mu}_i$ and σ_i^2 on $\mathrm{cl}(\mathcal{K})$, rather than on just $\mathcal{K}$. This is used in the two subsequent sections.

5.3.1 Continuity of g, and the sets M_x and M_*

We now discuss the continuity of the function g in this sampling model, and describe the sets M_x and M_*.

For $C \subseteq \mathcal{X}$, define $\mathcal{I}_C = \left\{ i \in \mathcal{I} : \sum_{x \in C} \sum_{b=1}^{B} \mathbf{1}_{\{x_b = i\}} > 0 \right\}$ and, for $k \in \mathrm{cl}(\mathcal{K})$, define $\mathcal{I}(k) = \left\{ i \in \mathcal{I} : \sigma_i^2(k) = 0 \right\}$. For $k \in \mathrm{cl}(\mathcal{K}) \setminus \mathcal{K}$, let $\mathbb{P}^{(k)}$ denote the measure on Θ under which μ_i is independent of $(\mu_j)_{j \neq i}$, and is distributed according to $\mu_i = \hat{\mu}_i(k)$ almost surely if $\sigma_i^2(k) = 0$, or $\mu_i \sim \mathrm{Normal}(\hat{\mu}_i(k), \sigma_i^2(k))$ if $\sigma_i^2(k) > 0$, This definition of $\mathbb{P}^{(k)}$ for $k \notin \mathcal{K}$ is a natural extension to the earlier definition for $k \in \mathcal{K}$ because we can see by checking convergence on elementary sets that if $(k_t) \subseteq \mathrm{cl}(\mathcal{K})$ converges to k_* in $\mathrm{cl}(\mathcal{K})$, then $\mathbb{P}^{(k_t)}$ converges weakly to $\mathbb{P}^{(k_*)}$. Let $\mathbb{E}^{(k)}$ denote the expectation under $\mathbb{P}^{(k)}$.

We now have the following pair of lemmas, whose proofs may be found in the appendix. In the first lemma, and elsewhere throughout this chapter, max over the empty set is understood to be $-\infty$, and min over the empty set is understood to by $+\infty$.

Lemma 5.3.1. *We have* $\mathrm{dom}(g) = \mathcal{K}$*, and, for each* $C \subseteq \mathcal{X}$*,* $k \mapsto g(k; C)$ *is continuous on* $\mathrm{dom}(g)$ *and can be extended continuously onto* $\mathrm{cl}(\mathrm{dom}(g)) = \mathrm{cl}(\mathcal{K})$ *by*

$$g(k; C) = \mathbb{E}^{(k)} \left[\max \left(\max_{i \in \mathcal{I}_C} \mu_i, \max_{i \notin \mathcal{I}_C} \hat{\mu}_i(k) \right) \right] - \max_{i \in \mathcal{I}} \hat{\mu}_i(k). \tag{5.6}$$

Lemma 5.3.2. *For* $k \in \mathrm{cl}(\mathrm{dom}(g))$ *and* $C \subseteq \mathcal{X}$*,* $g(k; C) = 0$ *iff* $\mathcal{I}_C \subseteq \mathcal{I}(k)$*.*

From these lemmas, we can exactly describe the sets M_x and M_*.

$$\begin{aligned} M_x &= \{k \in \mathrm{cl}(\mathrm{dom}(g)) : g(k; \{x\}) = 0\} = \left\{ k \in \mathrm{cl}(\mathcal{K}) : \mathcal{I}_{\{x\}} \subseteq \mathcal{I}(k) \right\} \\ &= \left\{ k \in \mathrm{cl}(\mathcal{K}) : \sigma_i^2(k) = 0 \ \forall i \in \mathcal{I}_{\{x\}} \right\}, \\ M_* &= \left\{ k \in \mathrm{cl}(\mathcal{K}) : \sigma_i^2(k) = 0 \ \forall i \in \mathcal{I} \right\}. \end{aligned}$$

We can also conclude that the sampling model satisfies Assumption 5.1.3, since if $k \in \mathrm{cl}(\mathcal{K})$ has $g(k; \{x\}) = 0$ for all $x \in \mathcal{X}$, then $\sigma_i^2(k) = 0$ for all $i \in \mathcal{I}$, and $\mathcal{I}(k) = \mathcal{I} = \mathcal{I}_{\mathcal{X}}$ implying through Lemma 5.3.2 that $g(k; \mathcal{X}) = 0$.

5.3.2 OCBA for linear loss

We now show asymptotic optimality of the optimal computing budget allocation (OCBA) for linear loss, which is a R&S policy that operates within this known-variance normal sampling model, and was first proposed by (He et al. 2007). It is derived by first supposing that

the current batch of B samples to be taken will be the last, and then choosing a measurement allocation that approximates the measurement allocation that would be optimal were this single-batch assumption true. The policy bases its measurement decisions upon an approximation to the reduction in expected linear loss achieved by a batch of measurements. It is described fully in Algorithm 3.

Algorithm 3 OCBA for linear loss

Require: Input knowledge state $k \in \mathrm{cl}(\mathcal{K})$ satisfying $|\arg\max_i \hat{\mu}_i(k)| = 1$, the batch size B, and an integer parameter m dividing B.

1: Choose $i^* \in \arg\max_i \hat{\mu}_i(k)$. This choice is unique by assumption.

2: For $i \in \mathcal{I} \setminus \{i^*\}$, define

$$\delta_i := \hat{\mu}_{i^*}(k) - \hat{\mu}_i(k), \qquad \tilde{\sigma}_{i,i^*} := \sqrt{\sigma_i^2(k) + (B/m + 1/\sigma_{i^*}^2(k))^{-1}},$$
$$\tilde{\sigma}_i := \sqrt{\sigma_i^2(k) + \sigma_{i^*}^2(k)}, \qquad \tilde{\sigma}_{i^*,i} := \sqrt{\sigma_{i^*}^2(k) + (B/m + 1/\sigma_i^2(k))^{-1}},$$

with $\tilde{\sigma}_{i,i^*} = \sigma_i(k)$ if $\sigma_{i^*}(k) = 0$, and $\tilde{\sigma}_{i^*,i} = \sigma_{i^*}(k)$ if $\sigma_i(k) = 0$.

3: Define

$$D_i(k) := \begin{cases} \tilde{\sigma}_{i^*,i} f(\delta_i/\tilde{\sigma}_{i^*,i}) - \tilde{\sigma}_i f(\delta_i/\tilde{\sigma}_i), & \text{if } i \neq i^*, \\ \sum_{i \neq i^*} \tilde{\sigma}_{i,i^*} f(\delta_i/\tilde{\sigma}_{i,i^*}) - \tilde{\sigma}_i f(\delta_i/\tilde{\sigma}_i), & \text{if } i = i^*, \end{cases}$$

where $f(z) := \varphi(z) + z\Phi(z)$, φ is the normal pdf, and Φ is the normal cdf. For any $\delta \in \mathbb{R}$, we take $0 \cdot f(\delta/0) = \lim_{\sigma \to 0} \sigma f(\delta/\sigma) = 0$.

4: Define $S(m) := \{i \in \mathcal{I} : D_i(k) \text{ is among the } m \text{ lowest values}\}$.

5: Allocate B/m samples to each alternative in $S(m)$, and no samples to the other alternatives.

Note that Algorithm 3 assumes of its input that $|\arg\max_i \hat{\mu}_i(k)| = 1$. When using the OCBA for linear loss policy as described in (He et al. 2007), one usually begins with a noninformative prior, then takes a fixed number of measurements from each alternative to obtain an informative posterior belief, and only afterward uses the OCBA policy. The belief succeeding the fixed first stage will almost surely satisfy the assumption, as will all subsequent beliefs, and so it is reasonable to assume this of the inputs to Algorithm 3. In such a setting, our asymptotic optimality result may then be understood as beginning where sampling under OCBA for linear loss begins – immediately after the first stage completes – and Theorem 5.3.3 below shows that OCBA for linear loss is asymptotically optimal under the belief held at this time.

We now use the result of Section 5.2 to show asymptotic optimality of OCBA for linear loss in the following theorem.

Theorem 5.3.3. *Suppose that* $|\arg\max_i \hat{\mu}_{0i}| = 1$. *Then the OCBA for linear loss policy defined in Algorithm 3 is asymptotically optimal for the sampling model and loss function in Section 5.3.*

Proof. Let $\widetilde{\mathcal{K}} = \{k \in \mathrm{cl}(\mathcal{K}) : |\arg\max_i \hat{\mu}_i(k)| = 1\}$. The assumed uniqueness of $\arg\max_i \hat{\mu}_{0i}$ implies $K_t \in \widetilde{\mathcal{K}}$ almost surely for each $t \in \mathbb{N}$, and this together with the fact that $\mathbb{P}\{\mu_i = \mu_j,\ i \neq j\} = 0$ implies $K_\infty \in \widetilde{\mathcal{K}}$ almost surely.

Now choose $\tilde{k} \in \widetilde{\mathcal{K}} \setminus M_*$, let $i^* \in \arg\max_i \hat{\mu}_i(\tilde{k})$, and define U by

$$U := \left\{ k \in \mathrm{cl}(\mathcal{K}) : \hat{\mu}_{i^*}(k) > \max_{i \neq i^*} \hat{\mu}_i(k),\ \min_{i \in \mathcal{I}(\tilde{k})} D_i(k) > \max_{i \notin \mathcal{I}(\tilde{k})} D_i(k) \right\}.$$

We first note that if $i \in \mathcal{I}(\tilde{k})$ then $\sigma_i^2(\tilde{k}) = 0$, implying $D_i(\tilde{k}) = 0$. Also, if $i \notin \mathcal{I}(\tilde{k})$ then $\sigma_i^2(\tilde{k}) > 0$, implying $D_i(\tilde{k}) < 0$. Thus $\min_{i \in \mathcal{I}(\tilde{k})} D_i(\tilde{k}) > \max_{i \notin \mathcal{I}(\tilde{k})} D_i(\tilde{k})$, including in the cases $\mathcal{I}(\tilde{k}) = \emptyset$ and $\mathcal{I}(\tilde{k}) = \mathcal{X}$. Thus $\tilde{k} \in U$.

Second, for each $k \in U \cap \widetilde{\mathcal{K}} \cap \mathcal{K}$, $\min_{i \in \mathcal{I}(\tilde{k})} D_i(k) > \max_{i \notin \mathcal{I}(\tilde{k})} D_i(k)$ implies that the OCBA for linear loss policy assigns at least B/m measurements to an alternative outside $\mathcal{I}(\tilde{k})$. Since $A_{\tilde{k}}$ consists of those measurement types allocating all B samples within $\mathcal{I}(\tilde{k})$, we have for any $t \in \mathbb{N}$ that $\Pi(t, k, A_{\tilde{k}}) = 0$.

Third, since both D_i and $\hat{\mu}_i$ are continuous for each $i \in \mathcal{I}$, U is open in $\mathrm{cl}(\mathcal{K})$. Thus the conditions of Theorem 5.2.3 are met and asymptotic optimality follows. □

5.4 Application to Ranking and Selection with Normal Rewards and Unknown Variance

We now apply Theorem 5.2.3 to the R&S problem, where the rewards are normally distributed as in Section 5.3, but where the variances of these rewards are *unknown*. Although the unknown variances complicate the sampling model, variances are generally unknown in practice and so the resulting model matches reality more closely. Several sequential policies have been developed for use on this problem. We will examine the LL(S)policy proposed in (Chick & Inoue 2001*b*).

As in Section 5.3, we have a finite set $\mathcal{I}$ of alternatives with $|\mathcal{I}| > 1$, but we take $\Theta = \mathbb{R}^{|\mathcal{I}|} \times \mathbb{R}_{++}^{|\mathcal{I}|}$ to be a bigger space, with $\theta \in \Theta$ decomposable as $\theta = (\mu, \lambda)$, where

$\mu \in \mathbb{R}^{|\mathcal{I}|}$ is the vector of means and $\lambda \in \mathbb{R}^{|\mathcal{I}|}_{++}$ is the vector of precisions. Samples from alternative i are then normally distributed with mean μ_i and precision λ_i.

We again suppose measurements are taken in batches of size B, so $\mathcal{X} = \mathcal{I}^B$, $\mathcal{Y} = \mathbb{R}^B$, and y_b is a sample from alternative $x_b \in \mathcal{I}$, for $b = 1, \ldots, B$. We also retain the use of the linear loss function (5.5). The sampling density with respect to the Lebesgue measure is then

$$p(y; x, (\mu, \lambda)) = \prod_{b=1}^{B} \sqrt{\frac{\lambda_{x_b}}{2\pi}} \exp\left(-\frac{1}{2}\lambda_{x_b}(y_b - \mu_{x_b})^2\right).$$

To put this in the form of (5.1), we take $\alpha((\mu, \lambda); x) = \left[\mu_{x_b}\lambda_{x_b}, -\frac{1}{2}\lambda_{x_b}\right]_{b=1}^{B}$, $\gamma(y; x) = \left[y_b, y_b^2\right]_{b=1}^{B}$, $\zeta((\mu, \lambda); x) = \sum_{b=1}^{B} \frac{1}{2}\left(\mu_{x_b}^2\lambda_{x_b} - \log\lambda_{x_b}\right)$, and $\nu(dy; x) = (2\pi)^{-B/2}dy$. Then the density of the posterior is given by (5.2) with

$$\beta(\mu, \lambda) = \left[\mu_i\lambda_i, -\frac{1}{2}\lambda_i, -\frac{1}{2}\left(\mu_i^2\lambda_i - \log\lambda_i\right)\right]_{i\in\mathcal{I}}, \quad \tau(S_t, N_t) = \left[\tilde{S}_{ti1}, \tilde{S}_{ti2}, \widetilde{N}_{ti}\right]_{i\in\mathcal{I}},$$

where $\tilde{S}_{ti1} = \sum_{t'\le t}\sum_{b=1}^{B} \mathbf{1}_{\{X_{t'b}=i\}} Y_{t'b}$ is the sum of all observations of alternative i, $\tilde{S}_{ti2} = \sum_{t'\le t}\sum_{b=1}^{B} \mathbf{1}_{\{X_{t'b}=i\}} Y_{t'b}^2$ is the sum of all squared observations of alternative i, and $\widetilde{N}_i = \sum_{t'\le t}\sum_{b=1}^{B} \mathbf{1}_{\{X_{t'b}=i\}}$ is the number of times we have observed alternative i.

We suppose that the prior Q_0 on θ is normal-gamma with independent components. In particular, we take $\hat{\mu}_0 \in \mathbb{R}^{|\mathcal{I}|}$, $\hat{\lambda}_0 \in \mathbb{R}^{|\mathcal{I}|}_{++}$, $a_0 \in \mathbb{R}^{|\mathcal{I}|}_{++}$, and $\rho_0 \in \mathbb{R}^{|\mathcal{I}|}_{++}$ to be parameters. Then, under Q_0, $\lambda_i \sim \text{Gamma}(a_{0i}, a_{0i}/\hat{\lambda}_{0i})$, $\mu_i|\lambda_i \sim \text{Normal}(\hat{\mu}_{0i}, 1/(\rho_{0i}\lambda_i))$, with the random vectors (μ_i, λ_i), $i \in \mathcal{I}$, independent of one another. We must also assume $a_{0i} > 1/2$ for each i to obtain $Q_0 \in \text{dom}(g)$.

With this assumption, the posterior Q_t is again normal-gamma with independent components. In particular, (DeGroot 1970) Section 9.6 shows that under Q_t, $\lambda_i \sim \text{Gamma}(a_{ti}, a_{ti}/\hat{\lambda}_{ti})$ and $\mu_i|\lambda_i \sim \text{Normal}(\hat{\mu}_{ti}, 1/(\rho_{ti}\lambda_i))$ with $\hat{\lambda}_{ti} := \hat{\lambda}_i(\tau(S_t, N_t))$, $\hat{\mu}_{ti} := \hat{\mu}_i(\tau(S_t, N_t))$, $a_{ti} := a_i(\tau(S_t, N_t))$, $\rho_{ti} := \rho_i(\tau(S_t, N_t))$, where functions $\hat{\mu}_i$, $\hat{\lambda}_i$, a_i, and ρ_i are defined for $i \in \mathcal{I}$ and argument $\tau = [\tilde{s}_{j1}, \tilde{s}_{j2}, \tilde{n}_j]_{j\in\mathcal{I}} \in \mathbb{R}^l$ by

$$\begin{aligned}
\hat{\mu}_i(\tau) &:= \left(\rho_{0i}\hat{\mu}_{0i} + \tilde{s}_{i1}\right)/(\rho_{0i} + \tilde{n}_i), \\
\hat{\lambda}_i(\tau) &:= (2a_{0i} + \tilde{n}_i) \Bigg/ \left(\frac{2a_{0i}}{\hat{\lambda}_{0i}} + \tilde{s}_{i2} + \rho_{0i}\hat{\mu}_{0i}^2 - \frac{(\rho_{0i}\hat{\mu}_{0i} + \tilde{s}_{i1})^2}{\rho_{0i} + \tilde{n}_i}\right), \\
a_i(\tau) &:= a_{0i} + \tilde{n}_i/2, \\
\rho_i(\tau) &:= \rho_{0i} + \tilde{n}_i.
\end{aligned}$$

We may compute the components of $K_t = \mathbb{E}_t\left[\beta(\theta)\right]$ as

$$\mathbb{E}_t\left[\mu_i\lambda_i\right] = \hat{\mu}_{ti}\hat{\lambda}_{ti}, \qquad \mathbb{E}_t\left[-\frac{1}{2}\lambda_i\right] = -\frac{1}{2}\hat{\lambda}_{ti},$$
$$\mathbb{E}_t\left[-\frac{1}{2}\left(\mu_i^2\lambda_i - \log\lambda_i\right)\right] = -\frac{1}{2}\left(\hat{\mu}_{ti}^2\hat{\lambda}_{ti} - \log(\hat{\lambda}_{ti}) + \log(a_{ti}) + \frac{1}{\rho_{ti}} - \Psi_0(a_{ti})\right).$$

where Ψ_0 is the digamma function. Define a function $f_i : (\max(-\rho_{0i}, -2a_{0i}), \infty) \mapsto \mathbb{R}_+$ for each $i \in \mathcal{I}$ by $f_i(\tilde{n}_i) = (\rho_{0i}+\tilde{n}_i)^{-1} + \log\left(a_{0i} + \tilde{n}_i/2\right) - \Psi_0(a_{0i}+\tilde{n}_i/2)$, so the last expectation becomes $\mathbb{E}_t\left[-\frac{1}{2}\left(\mu_i^2\lambda_i - \log\lambda_i\right)\right] = -\frac{1}{2}\left(\hat{\mu}_{ti}^2\hat{\lambda}_{ti} - \log(\hat{\lambda}_{ti}) + f_i(\widetilde{N}_{ti})\right)$.

For arbitrary $\tau = [\tilde{s}_{i1}, \tilde{s}_{i2}, \tilde{n}_i]_{i\in\mathcal{I}} \in \mathbb{R}^l$, $u \mapsto \exp(\beta(u)'\tau)Q_0(du)$ is the kernel of the distribution on (μ, λ) under which the pairs (μ_i, λ_i) are independent and $\lambda_i \sim \text{Gamma}(a_i(\tau), \hat{\lambda}_i(\tau)a_i(\tau))$ with $\mu_i|\lambda_i \sim \text{Normal}(\hat{\mu}(\tau), 1/(\rho_i(\tau)\lambda_i))$. To achieve $|\Gamma(\tau)| < \infty$, it is necessary and sufficient to have $\hat{\lambda}_i(\tau) > 0$, $a_i(\tau) > 0$, and $\rho_i(\tau) > 0$ for each $i \in \mathcal{I}$. Thus $\text{dom}(\Gamma) = \left\{\tau \in \mathbb{R}^l : \hat{\lambda}_i(\tau) > 0, a_i(\tau) > 0, \rho_i(\tau) > 0,\ \forall i \in \mathcal{I}\right\}$. Also,

$$\left\{\left[\hat{\mu}_i(\tau), \hat{\lambda}_i(\tau), a_i(\tau), \rho_i(\tau)\right]_{i\in\mathcal{I}} : \tau \in \mathbb{R}^l\right\} = \left\{\left[\hat{\mu}_i, \hat{\lambda}_i, a_{0i} + \tilde{n}_i/2, \rho_{0i} + \tilde{n}_i\right]_{i\in\mathcal{I}} :\right.$$
$$\left.\hat{\mu} \in \mathbb{R}^{|\mathcal{I}|}, \hat{\lambda} \in \mathbb{R}_{++}^{|\mathcal{I}|}, \tilde{n}_i \in (\max(-2a_{0i}, -\rho_{0i}), \infty)\ \forall i \in \mathcal{I}\right\}.$$

From this and the calculation of K_t, we may compute $\mathcal{K}$ as

$$\mathcal{K} = \left\{\left[\hat{\lambda}_i\hat{\mu}_i, \frac{-\hat{\lambda}_i}{2}, -\frac{1}{2}\left(\hat{\mu}_i^2\hat{\lambda}_i - \log(\hat{\lambda}_i) + f_i(\tilde{n}_i)\right)\right]_{i\in\mathcal{I}} :\right.$$
$$\left.\hat{\mu} \in \mathbb{R}^{|\mathcal{I}|}, \hat{\lambda} \in \mathbb{R}_{++}^{|\mathcal{I}|}, \tilde{n}_i \in (\max(-2a_{0i}, -\rho_{0i}), +\infty)\ \forall i \in \mathcal{I}\right\},$$

The function f_i is strictly decreasing, with $f_i(\tilde{n}_i)$ limited at 0 as $\tilde{n}_i \to \infty$, and $f_i(\tilde{n}_i)$ increasing to ∞ as $\tilde{n}_i$ decreases to $\max(-2a_{0i}, -\rho_{0i})$. Thus the closure $\text{cl}(\mathcal{K})$ of $\mathcal{K}$ is

$$\text{cl}(\mathcal{K}) = \left\{\left[\hat{\lambda}_i\hat{\mu}_i, \frac{-\hat{\lambda}_i}{2}, -\frac{1}{2}\left(\hat{\mu}_i^2\hat{\lambda}_i - \log(\hat{\lambda}_i) + f_i(\tilde{n}_i)\right)\right]_{i\in\mathcal{I}} :\right.$$
$$\left.\hat{\mu} \in \mathbb{R}^{|\mathcal{I}|}, \hat{\lambda} \in \mathbb{R}_{++}^{|\mathcal{I}|}, \tilde{n}_i \in (\max(-2a_{0i}, -\rho_{0i}), +\infty],\ \forall i \in \mathcal{I}\right\},$$

where we have needed to add to $\mathcal{K}$ only those points with $\tilde{n}_i = +\infty$ for at least one $i \in \mathcal{I}$. These points correspond to distributions under which μ_i and λ_i are known perfectly.

Finally, for each $i \in \mathcal{I}$, we introduce functions $\hat{\mu}_i : \text{cl}(\mathcal{K}) \mapsto \mathbb{R}$, $\hat{\lambda}_i : \text{cl}(\mathcal{K}) \mapsto \mathbb{R}_{++}$, $\rho_i : \text{cl}(\mathcal{K}) \mapsto \mathbb{R}_{++} \cup \{\infty\}$ and $a_i : \text{cl}(\mathcal{K}) \mapsto \mathbb{R}_{++} \cup \{\infty\}$. With argument $k = [k_{i1}, k_{i2}, k_{i3}]_{i\in\mathcal{I}}$,

these functions are defined by

$$\begin{aligned}
\hat{\mu}_i(k) &= k_{i1}/\hat{\lambda}_i(k), \\
\hat{\lambda}_i(k) &= -2k_{i2}, \\
a_i(k) &= a_{0i} + \frac{1}{2} f_i^{-1}(-2k_{i3} - \hat{\mu}_i(k)^2 \hat{\lambda}_i(k)), \\
\rho_i(k) &= \rho_{0i} + f_i^{-1}(-2k_{i3} - \hat{\mu}_i(k)^2 \hat{\lambda}_i(k)),
\end{aligned}$$

where f_i^{-1} is the inverse of f_i, which exists since f_i is strictly decreasing. We extend the domain $(0, \infty)$ of f_i^{-1} by taking $f_i^{-1}(0) = \infty$. With these definitions, we have $\hat{\mu}_{ti} = \hat{\mu}_i(K_t)$, $\hat{\lambda}_{ti} = \hat{\lambda}_i(K_t)$, $\rho_{ti} = \rho_i(K_t)$, $a_{ti} = a_i(K_t)$.

5.4.1 Continuity of g, and the sets M_x and M_*

We now discuss the continuity of the function g in this sampling model, and describe the sets M_x and M_*.

For $C \subseteq \mathcal{X}$, define $\mathcal{I}_C = \left\{ i \in \mathcal{I} : \sum_{x \in C} \sum_{b=1}^{B} \mathbf{1}_{\{x_b = i\}} > 0 \right\}$ and, for $k \in \mathcal{K}$, define $\mathcal{I}(k) = \{i \in \mathcal{I} : a_i(k) = \infty\}$. For $k \in \mathrm{cl}(\mathcal{K}) \setminus \mathcal{K}$, let $\mathbb{P}^{(k)}$ denote the measure on Θ under which the pairs (μ_i, λ_i) are independent of one another and have distribution

$$\begin{aligned}
&\mu_i \mid \lambda_i \sim \mathrm{Normal}(\hat{\mu}_i(k), 1/(\rho_i(k)\lambda_i)), \quad \lambda_i \sim \mathrm{Gamma}(a_i(k), \hat{\lambda}_i(k) a_i(k)), && i \notin \mathcal{I}(k), \\
&\mu_i = \hat{\mu}_i(k) \text{ a.s.}, \quad \lambda_i = \hat{\lambda}_i(k) \text{ a.s.}, && i \in \mathcal{I}(k).
\end{aligned}$$

This definition of $\mathbb{P}^{(k)}$ for $k \notin \mathcal{K}$ is a natural extension of the earlier definition for $k \in \mathcal{K}$ because we can see by checking convergence on elementary sets that if $(k_t) \subseteq \mathrm{cl}(\mathcal{K})$ converges to k_* in $\mathrm{cl}(\mathcal{K})$, then $\mathbb{P}^{(k_t)}$ converges weakly to $\mathbb{P}^{(k_*)}$. Let $\mathbb{E}^{(k)}$ denote the expectation under $\mathbb{P}^{(k)}$.

We now have the following pair of lemmas, whose proofs may be found in the appendix.

Lemma 5.4.1. *The sets* $\mathrm{dom}(g)$ *and* $\mathrm{cl}(\mathrm{dom}(g))$ *are given by* $\mathrm{dom}(g) = \left\{ k \in \mathcal{K} : a_i(k) > \frac{1}{2} \ \forall i \right\}$ *and* $\mathrm{cl}(\mathrm{dom}(g)) = \left\{ k \in \mathrm{cl}(\mathcal{K}) : a_i(k) > \frac{1}{2} \right\}$. *Furthermore, for each* $C \subseteq \mathcal{X}$, $k \mapsto g(k; C)$ *is continuous on* $\mathrm{dom}(g)$ *and can be extended continuously onto* $\mathrm{cl}(\mathrm{dom}(g))$ *by*

$$g(k; C) = \begin{cases} \mathbb{E}^{(k)} \left[\max \left(\max_{i \in \mathcal{I}_C} \mu_i, \max_{i \notin \mathcal{I}_C} \hat{\mu}_i(k) \right) \right] - \max_{i \in \mathcal{I}} \hat{\mu}_i(k), & \text{if } a_i(k) > \frac{1}{2} \ \forall i \in \mathcal{I}_C, \\ +\infty, & \text{if not.} \end{cases} \tag{5.7}$$

Lemma 5.4.2. *For $k \in \mathrm{cl}(\mathrm{dom}(g))$ and $C \subseteq \mathcal{X}$, $g(k;C) = 0$ iff $\mathcal{I}_C \subseteq \mathcal{I}(k)$.*

From these lemmas, we can describe the sets M_x and M_* as

$$\begin{aligned} M_x &= \{k \in \mathrm{cl}(\mathrm{dom}(g)) : g(k;\{x\}) = 0\} = \left\{k \in \mathrm{cl}(\mathcal{K}) : \mathcal{I}_{\{x\}} \subseteq \mathcal{I}(k)\right\} \\ &= \left\{k \in \mathrm{cl}(\mathcal{K}) : a_i(k) = \infty \ \forall i \in \mathcal{I}_{\{x\}}\right\}, \\ M_* &= \{k \in \mathrm{cl}(\mathcal{K}) : a_i(k) = \infty \ \forall i \in \mathcal{I}\}. \end{aligned}$$

We can also conclude that the sampling model satisfies Assumption 5.1.3, since if $k \in \mathrm{cl}(\mathcal{K})$ has $g(k;\{x\}) = 0$ for all $x \in \mathcal{X}$, then $a_i(k) = \infty$ for all $i \in \mathcal{I}$, and $\mathcal{I}(k) = \mathcal{I} = \mathcal{I}_{\mathcal{X}}$ implying through Lemma 5.4.2 that $g(k;\mathcal{X}) = 0$.

5.4.2 LL(S) policy

We now use the Theorem 5.2.3 to show asymptotic optimality of the LL(S)policy proposed by (Chick & Inoue 2001*b*). This policy operates in a R&S framework in which samples are normally distributed with unknown mean and variance. The decisions made by the policy may be computed via Algorithm 4. Note that the LL(S)policy as described in (Chick & Inoue 2001*b*) begins with a noninformative prior in which $a_i = -1/2$ and $\rho_i = 0$, samples equally from the alternatives in the first stage, and arrives at a posterior k in which $\rho_i(k)$ is the number of samples taken from alternative i and is equal to $1 + 2a_i(k)$. Substituting this into Algorithm 4 provides the form of the LL(S)policy found in (Chick & Inoue 2001*b*).

We now show asymptotic optimality of the LL(S)policy in Theorem 5.4.3. Note that this theorem assumes that $|\arg\max_i \hat{\mu}_{0i}| = 1$. As with OCBA for linear loss from Section 5.4, this is reasonable because one generally uses LL(S)after a fixed first stage, and the belief that results will almost surely satisfy this assumption. Also note that Algorithm 4 leaves the method of rounding in step 7 unspecified. (Chick & Inoue 2001*b*) does not completely specify the rounding method to be used and leaves some freedom to the practitioner but to guarantee asymptotic optimality we only require that the allocation in step 7 goes entirely to those alternatives in $\mathcal{S}$.

Theorem 5.4.3 also assumes that $a_{0i} > 1$. When beginning with a noninformative prior with the parameter a initially set to $-1/2$, this assumption will be met as long as there are at least 4 samples in the first stage. This assumption is usually met in practice. For

Algorithm 4 LL(S)

Require: Input knowledge state $k \in \mathrm{cl}(\mathrm{dom}(g))$ and a batch size B.

1: Let $i^* \in \arg\max_i \hat{\mu}_i(k)$ and $\mathcal{S} = \mathcal{I}$.

2: For $i \neq i^*$, let

$$\tilde{\lambda}_i = \begin{cases} \left((\hat{\lambda}_i(k)\rho_i(k))^{-1} + (\hat{\lambda}_{i^*}(k)\rho_{i^*}(k))^{-1} \right)^{-1}, & \text{if } i, i^* \in \mathcal{S}, \\ \hat{\lambda}_{i^*}(k)\rho_{i^*}(k), & \text{if } i \notin \mathcal{S},\, i^* \in \mathcal{S}, \\ \hat{\lambda}_i(k)\rho_i(k), & \text{if } i^* \notin \mathcal{S},\, i \in \mathcal{S}. \end{cases}$$

3: For $i \neq i^*$, let

$$\tilde{\nu}_i = \begin{cases} \frac{\left((\hat{\lambda}_i(k)\rho_i(k))^{-1} + (\hat{\lambda}_{i^*}(k)\rho_{i^*}(k))^{-1} \right)^2}{\left((\hat{\lambda}_i(k)\rho_i(k))^2 2a_i(k) \right)^{-1} + \left((\hat{\lambda}_{i^*}(k)\rho_{i^*}(k))^2 2a_{i^*}(k) \right)^{-1}}, & \text{if } i, i^* \in \mathcal{S}, \\ 2a_{i^*}(k), & \text{if } i \notin \mathcal{S}, i^* \in \mathcal{S}, \\ 2a_i(k), & \text{if } i \in \mathcal{S}, i^* \notin \mathcal{S}. \end{cases}$$

4: For $i \in \mathcal{S}$,

$$\eta_i = \begin{cases} \tilde{\lambda}_i^{1/2} \frac{\tilde{\nu}_i + \tilde{\lambda}_i (\hat{\mu}_{i^*}(k) - \hat{\mu}_i(k))^2}{\tilde{\nu}_i - 1} \varphi\left(\tilde{\lambda}_i^{1/2} (\hat{\mu}_{i^*}(k) - \hat{\mu}_i(k));\, \tilde{\nu}_i \right), & \text{if } i \neq i^*, \\ \sum_{j \in \mathcal{S} \setminus \{i^*\}} \eta_j, & \text{if } i = i^*, \end{cases}$$

where $\varphi(\,\cdot\,;\, \tilde{\nu})$ is the density of the student-t distribution with $\tilde{\nu}$ degrees of freedom.

5: For $i \in \mathcal{I}$,

$$r_i = \frac{B + \sum_{j \in \mathcal{S}} \rho_j(k)}{\sum_{j \in \mathcal{S}} \sqrt{\frac{\eta_j / \hat{\lambda}_j}{\eta_i / \hat{\lambda}_i}}} - \rho_i(k).$$

6: If $r_i \geq 0$ for all $i \in \mathcal{S}$, continue to Step 7. Otherwise, remove from $\mathcal{S}$ each $i \in \mathcal{S}$ with $r_i < 0$, and return to Step 2.

7: Round each r_i to an integer, and allocate this number of samples to alternative i.

example, the numerical experiments performed (Chick & Inoue 2001*b*) all use 5 or more samples in the first stage.

Theorem 5.4.3. *Assume* $|\arg\max_i \hat{\mu}_{0i}| = 1$ *and* $\min_i a_{0i} > 1$. *Then the* LL(S)*policy defined in Algorithm 4 is asymptotically optimal for the sampling model and loss function in Section 5.4.*

Proof. Let $\widetilde{\mathcal{K}} = \{k \in \mathrm{cl}(\mathrm{dom}(g)) : |\arg\max_i \hat{\mu}_i(k)| = 1, \min_i a_i(k) > 1\}$. Since $K_0 \in \mathrm{dom}(g)$, $|\arg\max_i \hat{\mu}_{0i}| = 1$, and $\min_i a_{0i} > 1$, the event $\left\{ K_t \in \widetilde{\mathcal{K}},\ \forall t \in \mathbb{N} \cup \{\infty\} \right\}$ is almost sure. Now choose $\tilde{k} \in \widetilde{\mathcal{K}} \setminus M_*$, and we will find a set U open in $\mathrm{cl}(\mathcal{K})$ that contains $\tilde{k}$ and satisfies the condition (5.4) of Theorem 5.2.3.

If $A_{\tilde{k}} = \emptyset$, then we may take U to be any set that is open in $\mathrm{cl}(\mathcal{K})$ and contains $\tilde{k}$, since $\Pi(t, k', \emptyset) = 0$ for all $t \in \mathbb{N}$ and all $k' \in \mathcal{K}$, trivially satisfying (5.4). In particular, we can take $U = \mathrm{cl}(\mathcal{K})$, which is clearly open in itself.

Now consider the case $A_{\tilde{k}} \neq \emptyset$. We have $1 \leq |A_{\tilde{k}}| \leq |\mathcal{X}| - 1$ since $\tilde{k} \notin M_*$ and $M_* = \cap_{x \in \mathcal{X}} M_x$ implies $A_{\tilde{k}} \neq \mathcal{X}$. Let $i^* \in \arg\max_i \hat{\mu}_i(\tilde{k})$, which is unique by $\tilde{k} \in \widetilde{\mathcal{K}}$, and let $\epsilon = \min((\hat{\mu}_{i^*}(\tilde{k}) - \max_{i \neq i^*} \hat{\mu}_i(\tilde{k}))/3, \min_i \hat{\lambda}_i(\tilde{k})/2, \min_i (a_i(\tilde{k}) - 1)/2, \min_i \rho_i(\tilde{k})/2)$. Note $\epsilon > 0$ since $\tilde{k} \in \widetilde{\mathcal{K}}$.

Define for each $c \in [1, \infty)$ a set $U(c)$ by

$$U(c) = \Bigg\{ k \in \mathrm{cl}(\mathrm{dom}(g)) : \max_{i \in \mathcal{I}} |\hat{\mu}_i(k) - \hat{\mu}_i(\tilde{k})| < \epsilon,\ \max_{i \in \mathcal{I}} |\hat{\lambda}_i(k) - \hat{\lambda}_i(\tilde{k})| < \epsilon,$$
$$\max_{i \notin \mathcal{I}(\tilde{k})} |\rho_i(k) - \rho_i(\tilde{k})| < \epsilon,\ \max_{i \notin \mathcal{I}(\tilde{k})} |a_i(k) - a_i(\tilde{k})| < \epsilon,\ \min_{i \in \mathcal{I}(\tilde{k})} \rho_i(k) > c,\ \min_{i \in \mathcal{I}(\tilde{k})} a_i(k) > c \Bigg\}.$$

We have $\tilde{k} \in U(c)$ for each $c \geq 1$. Also, $U(1)$ contains $U(c)$ for all $c \geq 1$, and the following statements hold for each $k \in U(1)$: $\min_{i \notin \mathcal{I}(\tilde{k})} \rho_i(k) > \epsilon$, $\min_{i \notin \mathcal{I}(\tilde{k})} a_i(k) > 1 + \epsilon$, $\min_{i \in \mathcal{I}} \hat{\lambda}_i(k) > \epsilon$, $\arg\max_i \hat{\mu}_i(k) = \{i^*\}$, and $\hat{\mu}_{i^*}(k) - \max_{i \neq i^*} \hat{\mu}_i(k) > \epsilon$.

We now show that, for a sufficiently large value of c and given $k \in U(c)$ as input, Algorithm 4 allocates all its measurements outside $\mathcal{I}(\tilde{k})$, satisfying (5.4) and implying the asymptotic optimality of LL(S)through Theorem 5.2.3. To do so, it is enough to show that on the first visit to step 6, the algorithm eliminates every alternative in $\mathcal{I}(k)$ from $\mathcal{S}$ and retains every alternative in $\mathcal{I} \setminus \mathcal{I}(k)$.

Our argument is based on the boundedness over $U(1)$ of several functions of k used in Algorithm 4. When referring to Algorithm 4, we use the notation $\tilde{\lambda}_i(k)$, $\tilde{\nu}_i(k)$, $\eta_i(k)$, and $r_i(k)$ to indicate the quantities computed within the algorithm with $\mathcal{S} = \mathcal{I}$ and the given value of k.

Let $q := \inf_{k \in U(1)} \min_{i \notin \mathcal{I}(\tilde{k})} \sqrt{\eta_i(k)/\hat{\lambda}_i(k)} \Big/ \sum_{j \in \mathcal{I}} \sqrt{\eta_j(k)/\hat{\lambda}_j(k)}$. We show $q > 0$ by considering two cases. In the first case, suppose $\mathcal{I} \setminus \mathcal{I}(\tilde{k}) = \{i^*\}$. Then, $\eta_{i^*}(k) \geq \eta_j(k)$ for each $j \in \mathcal{I}$ implies

$$q = \inf_{k \in U(1)} \frac{\sqrt{\eta_{i^*}(k)/\hat{\lambda}_{i^*}(k)}}{\sum_{j \in \mathcal{I}} \sqrt{\eta_j(k)/\hat{\lambda}_j(k)}} \geq \frac{1}{|\mathcal{I}|} \sqrt{\frac{\inf_{k \in U(1)} \min_{j \in \mathcal{I}} \hat{\lambda}_j(k)}{\sup_{k \in U(1)} \hat{\lambda}_{i^*}(k)}} > 0.$$

In the second case, suppose $\mathcal{I} \setminus \mathcal{I}(\tilde{k}) \neq \{i^*\}$. Choose $i \notin \mathcal{I}(\tilde{k})$ with $i \neq i^*$. Then,

$$\begin{aligned}\inf_{k\in U(1)} \eta_i(k) &\geq \inf_{k\in U(1)} \tilde{\lambda}_i(k)^{1/2} \frac{\tilde{\nu}_i(k)}{\tilde{\nu}_i(k)-1} \varphi\left(\tilde{\lambda}_i(k)^{1/2}(\hat{\mu}_{i^*}(k)-\hat{\mu}_i(k));\, \tilde{\nu}_i(k)\right) \\ &\geq \left(\inf_{k\in U(1)} \tilde{\lambda}_i(k)^{1/2}\right)\left(\inf_{k\in U(1)} \varphi\left(u;\, \tilde{\nu}_i(k)\right)\right) > 0,\end{aligned}$$

where

$$u = \sup_{k\in U(1)} \tilde{\lambda}_i(k)^{1/2}(\hat{\mu}_{i^*}(k)-\hat{\mu}_i(k)) \leq \sup_{k\in U(1)} \hat{\lambda}_i(k)\rho_i(k)^{1/2}(\hat{\mu}_{i^*}(k)-\hat{\mu}_i(k)) < \infty.$$

Since $\eta_{i^*} \geq \eta_i(k)$, $\inf_{k\in U(1)} \eta_{i^*}(k) > 0$. Thus $\underline{\eta} := \inf_{k\in U(1)} \min_{i\notin\mathcal{I}(\tilde{k})} \eta_i(k) > 0$.

Now we show $\overline{\eta} := \sup_{k\in U(1)} \max_{i\in\mathcal{I}\setminus\{i^*\}} \eta_i(k) < \infty$. Let $i \neq i^*$, and let $u = \sup_{k\in U(1)}(\hat{\mu}_{i^*}(k)-\hat{\mu}_i(k)) < \infty$. Since $\inf_{k\in U(1)}(\hat{\mu}_{i^*}(k)-\hat{\mu}_i(k)) \geq \epsilon$, we have

$$\sup_{k\in U(1)} \eta_i(k) \leq \sup_{k\in U(1)} \tilde{\lambda}_i(k)^{1/2} \frac{\tilde{\nu}_i(k)+u^2\tilde{\lambda}_i(k)}{\tilde{\nu}_i(k)-1} \varphi\left(\epsilon\tilde{\lambda}(k)^{1/2};\, \tilde{\nu}_i(k)\right).$$

Since $\frac{\tilde{\nu}_i(k)+u^2\tilde{\lambda}_i(k)}{\tilde{\nu}_i(k)-1}$ is decreasing in $\tilde{\nu}_i(k)$ and $\tilde{\nu}_i(k) \geq \min(2a_i(k), 2a_{i^*}(k)) \geq 2$,

$$\sup_{k\in U(1)} \eta_i(k) \leq \sup_{k\in U(1)} \tilde{\lambda}_i(k)^{1/2} \left(2+u^2\tilde{\lambda}_i(k)\right) \varphi\left(\epsilon\tilde{\lambda}(k)^{1/2};\, \tilde{\nu}_i(k)\right).$$

The facts $\tilde{\nu}_i(k) \geq 2$ and $\sup_{\tilde{\nu}\geq 2, z\in\mathbb{R}} |z|^j \varphi(z;\tilde{\nu}) < \infty$ for $0 \leq j \leq 3$ together imply that $\sup_{k\in U(1)} \eta_i(k) < \infty$. Then, since $\sup_{k\in U(1)} \eta_{i^*}(k) \leq \sum_{i\neq i^*} \sup_{k\in U(1)} \eta_i(k) < \infty$, we have $\overline{\eta} < \infty$.

These bounds imply

$$q \geq \left(\inf_{k\in U(1)} \min_{i,j\in\mathcal{I}} \frac{1}{|\mathcal{I}|} \sqrt{\frac{\hat{\lambda}_j(k)\underline{\eta}}{\hat{\lambda}_i(k)\overline{\eta}}}\right) > 0,$$

which implies in turn that

$$\inf_{k\in U(c)} \min_{i\notin\mathcal{I}(\tilde{k})} r_i(k) \geq \inf_{k\in U(c)} \min_{i\notin\mathcal{I}(\tilde{k})} \left(B + \sum_{j\in\mathcal{I}} \rho_j(k)\right) q - \rho_i(k) \geq (B+c)q - \max_{i\notin\mathcal{I}(\tilde{k})} (\rho_i(\tilde{k})+\epsilon).$$

Thus, choosing $c > -B + (B + \max_{i\notin\mathcal{I}(\tilde{k})} \rho_i(\tilde{k}) + \epsilon)/q$, $k \in U(c)$ guarantees that $r_i(k) > B$ for every $i \notin \mathcal{I}(\tilde{k})$. Since $\sum_{i\in\mathcal{S}} r_i(k) = B$, this implies that $r_i(k) < 0$ for every $i \in \mathcal{I}(\tilde{k})$. Thus, LL(S)eliminates all of $\mathcal{I}(\tilde{k})$ from $\mathcal{S}$ before deciding on an allocation.

□

5.5 Application to Knowledge-Gradient Policies

We return to the general exponential family setting developed in Section 5.1, and we show asymptotic optimality of a class of policies known as knowledge-gradient (KG) policies. A sampling policy Π is called a *knowledge-gradient policy* if it satisfies

$$\Pi(t, k, \arg\max_{x \in \mathcal{X}} h(k, x)) = 1 \tag{5.8}$$

for all $t \in \mathbb{N}$ and all $k \in \mathcal{K}$. Here, the function $h : \mathcal{K} \times \mathcal{X} \mapsto \mathbb{R}_+ \cup \{\infty\}$ is defined by

$$h(k, x) := \min_{i \in \mathcal{I}} \mathbb{E}\left[R(\theta; i) \mid K_t = k\right] - \mathbb{E}\left[\min_{i \in \mathcal{I}} \mathbb{E}_{t+1}\left[R(\theta; i)\right] \mid K_t = k, X_{t+1} = x\right], \tag{5.9}$$

and may be understood as a myopic benefit of the single measurement x. By construction, KG policies are optimal for a single measurement, in the sense that they achieve the minimum value of $\mathbb{E}\left[\min_{i \in \mathcal{I}} \mathbb{E}_T\left[R(\theta; i)\right]\right]$ over all sampling policies for $T = 1$.

A number of policies existing in the literature satisfy this condition and are KG policies. For example, the $(R_1, \ldots, R_1)$ policy of (Gupta & Miescke 1996), which we called the KG policy for independent normal rewards in Chapter 2, is the KG policy that results from the model given in Section 5.3 with a batch size of 1. Another example is the LL_1policy introduced in (Chick et al. 2007, Chick et al. 2009), which is the KG policy that results from the model given in Section 5.4, again with a batch size of 1. A third example is the policy introduced in Chapter 4, which is the KG policy that results from a linear loss function and an independent normal sampling model with a correlated multivariate normal prior. Note that the definition (5.8) does not include the "online" variety of KG policy considered by (Ryzhov et al. 2008*b*). Such online policies are outside the scope of this chapter.

Given certain extra assumptions on the sampling problem, KG policies satisfy the conditions of Theorem 5.2.3, and hence are asymptotically optimal. We give these conditions below in Theorem 5.5.2, whose proof requires the following lemma. The proof of this lemma may be found in the appendix.

Lemma 5.5.1. *For all $x \in \mathcal{X}$ and all $k \in \mathrm{dom}(g)$, $g(k; \{x\}) \geq h(k, x) \geq 0$.*

We now present sufficient conditions for the asymptotic optimality of a KG policy.

Theorem 5.5.2. *If a sampling model satisfies Assumption 5.1.3, and the functions g and h satisfy the following conditions, then every KG policy for that model is asymptotically optimal.*

- *For each $x \in \mathcal{X}$, both $k \mapsto g(k; \{x\})$ and $k \mapsto h(k, x)$ are continuous on* $\mathrm{dom}(g)$ *and may be extended continuously onto* $\mathrm{cl}(\mathrm{dom}(g))$.
- *For each $k \in \mathrm{cl}(\mathrm{dom}(g))$ and $x \in \mathcal{X}$, if $h(k, x) = 0$ then $g(k; \{x\}) = 0$.*

Proof. We begin by extending Lemma 5.5.1 from $\mathrm{dom}(g)$ to $\mathrm{cl}(\mathrm{dom}(g))$. Let k be an element of $\mathrm{cl}(\mathrm{dom}(g)) \setminus \mathrm{dom}(g)$ and let $(k_t) \subseteq \mathrm{dom}(g)$ be a sequence converging to k. Since Lemma 5.5.1 implies $g(k_t; \{x\}) \geq h(k_t, x) \geq 0$ for each $t \in \mathbb{N}$, and the continuity of the extended versions of g and h imply $\lim_{t\to\infty} g(k_t; \{x\}) = g(k; \{x\})$ and $\lim_{t\to\infty} h(k_t, x) = h(k, x)$, we have $g(k; \{x\}) \geq h(k, x) \geq 0$. With this extension of Lemma 5.5.1, together with the second assumption of the theorem, we see that $g(k; \{x\}) = 0$ if and only if $h(k, x) = 0$, over all $k \in \mathrm{cl}(\mathrm{dom}(g))$ and $x \in \mathcal{X}$.

Choose $k \in \mathrm{cl}(\mathcal{K}) \setminus M_*$. If $k \notin \mathrm{cl}(\mathrm{dom}(g))$ then $A_k = \emptyset$ and we simply take $U = \mathrm{cl}(\mathcal{K})$, which is open in itself and contains k, and we have satisfied the conditions of Theorem 5.2.3.

Now suppose instead that $k \in \mathrm{cl}(\mathrm{dom}(g))$. As noted in Section 5.2, the continuity of g implies $M_x = \{k' \in \mathrm{cl}(\mathcal{K}) : g(k'; \{x\}) = 0\}$. Thus $A_k = \{x \in \mathcal{X} : g(k; \{x\}) = 0\}$, which is equal to $\{x \in \mathcal{X} : h(k; x) = 0\}$. Define

$$U = \left\{ k' \in \mathrm{cl}(\mathrm{dom}(g)) : \max_{x \in A_k} h(k', x) < \min_{x \notin A_k} h(k', x) \right\},$$

and we check that this set satisfies the conditions of Theorem 5.2.3. First, U is open by the continuity of $h(\,\cdot\,, x) : \mathrm{cl}(\mathrm{dom}(g)) \mapsto \mathbb{R}_+$ for each $x \in \mathcal{X}$. Second, $k \in U$ because $h(k, x) = 0$ for all $x \in A_k$ and $h(k, x) > 0$ for all $x \notin A_k$. Third, for each $k' \in U$, $k \notin M_*$ implies $A_k \neq \mathcal{X}$, which implies $\arg\max_{x \in \mathcal{X}} h(k', x) \subseteq \mathcal{X} \setminus A_k$, which together with (5.8), and the fact that Π is a KG policy, implies $\Pi(t, k', A_k) = 1 - \Pi(t, k', \mathcal{X} \setminus A_k) = 0$ for all $t \in \mathbb{N}$.

Thus the conditions of Theorem 5.2.3 are satisfied, and the sampling policy is asymptotically optimal. □

We have as an immediate consequence of this theorem that two of the KG policies mentioned above are asymptotically optimal. We state this consequence as two corollaries.

Corollary 5.5.3. *The KG policy that results from the model in Section 5.3 with $B = 1$ is asymptotically optimal.*

Proof. For the given model, Lemma 5.3.1 shows that $g(\,\cdot\,;\{x\})$ is continuous on $\mathrm{dom}(g) = \mathcal{K}$ and can be extended continuously onto $\mathrm{cl}(\mathrm{dom}(g)) = \mathrm{cl}(\mathcal{K})$. Also, Section 5.3.1 stated as a consequence of Lemmas 5.3.1 and 5.3.2 that the sampling model satisfies Assumption 5.1.3.

Turning our attention to the function h, and noting $\mathcal{X} = \mathcal{I}$ since $B = 1$, h is derived in (Gupta & Miescke 1996), with an alternate derivation in Chapter 2, as

$$h(k,x) = \tilde{\sigma}_x(k) f\left(\frac{-|\hat{\mu}_x(k) - \max_{i \neq x} \hat{\mu}_i(k)|}{\tilde{\sigma}_x(k)}\right), \tag{5.10}$$

where $\tilde{\sigma}_x(k) = \sqrt{\lambda_x \sigma_x^2(k) / \left(\frac{1}{\sigma_x^2(k)} + \lambda_x\right)}$, and $f(z) = z\Phi(z) + \varphi(z)$, with Φ the standard normal cumulative distribution function and φ the standard normal probability density function.

This is a continuous function on $\mathrm{dom}(g)$ and can be extended continuously onto $\mathrm{cl}(\mathrm{dom}(g)) = \mathrm{cl}(\mathcal{K})$ by taking $h(k,x) = 0$ if $\sigma_x^2(k) = 0$ and taking $h(k,x)$ to be defined by (5.10) if $\sigma_x^2(k) > 0$. Thus the first condition of Theorem 5.5.2 is satisfied. Furthermore, with this extension of $h(k,x)$, we then have for $k \in \mathrm{cl}(\mathcal{K})$ that $h(k,x) = 0$ iff $\sigma_x^2(k) = 0$. Since $\sigma_x^2(k) = 0$ implies $g(k;\{x\}) = 0$ as stated in Section 5.3.1, the second condition of Theorem 5.5.2 is satisfied. Thus, by Theorem 5.5.2, this KG policy is asymptotically optimal. □

Corollary 5.5.4. *The KG policy that results from the model in Section 5.4 with $B = 1$ is asymptotically optimal.*

Proof. For the given model, Lemma 5.4.1 shows that $g(\,\cdot\,;\{x\})$ is continuous on $\mathrm{dom}(g)$ and can be extended continuously onto $\mathrm{cl}(\mathrm{dom}(g))$. Also, Section 5.4.1 states as a consequence of Lemmas 5.4.1 and 5.4.2 that the sampling model satisfies Assumption 5.1.3.

Turning our attention to the function h, and noting $\mathcal{X} = \mathcal{I}$ since $B = 1$, h is derived in (Chick et al. 2009) as

$$h(k,x) := \tilde{\lambda}_x(k)^{-1/2} f\left(\tilde{\lambda}_x(k)^{1/2} \left|\hat{\mu}_x(k) - \max_{i \neq x} \hat{\mu}_i(k)\right|; \rho_x(k)\right), \tag{5.11}$$

where $\tilde{\lambda}_x(k) := \rho_x(k)(\rho_x(k)+1)\hat{\lambda}_x(k)$, $f(z;\tilde{\nu}) := \frac{\tilde{\nu}+z^2}{\tilde{\nu}-1}\varphi_{\tilde{\nu}}(z) - z\Phi_{\tilde{\nu}}(-z)$, and where $\Phi_{\tilde{\nu}}$ and $\varphi_{\tilde{\nu}}$ are respectively the cumulative distribution function and probability density function for the student-t distribution with $\tilde{\nu}$ degrees of freedom.

The function $k \mapsto h(k,x)$ is continuous on $\mathrm{dom}(g)$ and can be extended continuously onto $\mathrm{cl}(\mathrm{dom}(g))$ by taking $h(k,x) = 0$ if $a_x(k) = \infty$ and taking $h(k,x)$ to be defined by (5.11) if $a_x(k) < \infty$. Thus the first condition of Theorem 5.5.2 is satisfied. Furthermore, with this extension of $h(k,x)$, we have for $k \in \mathrm{cl}(\mathrm{dom}(g))$ that $h(k,x) = 0$ iff $a_x(k) = \infty$. Since $a_x(k) = \infty$ implies $g(k;\{x\}) = 0$ as stated in Section 5.4.1, the second condition of Theorem 5.5.2 is satisfied. Thus, by Theorem 5.5.2, this KG policy is asymptotically optimal. □

Of these two corollaries, the KG policy for which asymptotic optimality is shown in Corollary 5.5.3 was already shown to be asymptotically optimal in Chapter 2 in a framework specific to that policy. The KG policy for which asymptotic optimality is shown in Corollary 5.5.4, called LL_1, was proposed in (Chick et al. 2007, Chick et al. 2009), and its asymptotic optimality has not been shown previously. With these two corollaries serving as examples, we see that Theorem 5.5.2 applies to a broad class of KG policies and offers a relatively simple check that one can perform when considering the use of a new KG policy.

5.6 Conclusion

We have presented a powerful and general sufficient condition for asymptotic optimality of a sequential sampling policy. We have demonstrated the applicability of this sufficient condition by using it to show asymptotic optimality of the OCBA for linear loss and LL(S)policies, and also of any knowledge-gradient policy meeting some relatively mild extra conditions, including the $(R_1,\ldots,R_1)$ and LL_1policies.

Although asymptotic optimality by itself is no guarantee that a policy performs well in the finite-sample case, lack of asymptotic optimality is a dangerous signal and suggests that a policy may perform extremely badly in some cases. Indeed, if a policy lacks asymptotic optimality then the wise practitioner should be extremely careful when using it, or not use it at all. For this reason, it is our hope that the work we present here may be used to more

easily check whether a policy under consideration is asymptotically optimal, thus warning those who might otherwise employ a bad policy, and reassuring those who would use a good policy.

Chapter 6

Conclusion

> . . . I hereupon offer my own poor endeavors. I promise nothing complete; because any human thing supposed to be complete, must for that very reason infallibly be faulty.
>
> Herman Melville, Moby Dick

In this thesis we have considered the class of information collection problems and developed knowledge-gradient methods as a general class of methods for making decisions in these problems. These methods are myopically optimal in general, perform well numerically in the particular problems considered in Chapters 2, 3, and 4, and are asymptotically optimal in a broad class of offline problems described in Chapter 5. The main value, however, of KG methods are their flexibility. Given *any* information collection problem, we may write down the KG policy, and there is reasonable hope that this policy may be efficiently computed and that it will perform reasonably well. This flexibility gives the researcher applying KG methods to a particular information collection application the freedom to use a complex and high-fidelity mathematical model to represent the application at hand.

With this stated, there is a great deal more that could be done to guide the applied researcher in using KG methods. First, in many applications it will prove challenging to actually compute expectations necessary for KG decisions, and so computational work will be needed in these cases. Second, although we know KG policies are myopically optimal and we have the theoretical results from Chapter 5 on asymptotic optimality, as well as strong positive empirical results on the performance of KG policies for particular R&S problems, an important open question is to characterize the performance of KG policies in general. In

particular, we would like to be able to say in which situations KG policies will perform well and in which situations they will not. KG policies operate by greedily acquiring as much information as possible with each measurement, and so they should work well to the extent to which this greed does not interfere with information acquisition over longer timescales. Fully characterizing these cases would provide a valuable guide to the researcher who might want to use a KG policy in a particular problem.

During the completion of this thesis, work by this author and others have extended the ideas in this thesis to other problems. We give a short list here.

- (Frazier et al. 2009) considers applications to active text classification and rapid emergency response.

- (Negoescu et al. 2008) applies the KG method to a version of the newsvendor problem in which the demand distribution is unknown but can be learned from sales.

- (Ryzhov et al. 2008*a*) consider KG policies for generalizations of the multi-armed bandit problem, including those in which beliefs are correlated. As in the ranking and selection problem, the inclusion of correlated beliefs into multi-armed bandit problems promises to dramatically improve performance in many applications.

- (Ryzhov & Powell 2009) applies the KG method to the problem of learning on a graph, which is an offline problem in which there is an underlying graph whose edges have associated with them unknown costs. At implementation time we would like to find a path through the graph for which the sum of the costs of the edges traveled is as small as possible, and during the learning phase we have the ability to measure the costs on individual edges of our choosing.

- (Frazier & Powell 2009*a*) applies the KG method to a model calibration problem encountered while using approximate dynamic programming to make strategic decisions for a logistics company.

- (Negoescu et al. 2009) considers the application of KG methods to drug discovery for Ewing's sarcoma. The task of drug discovery can be formulated as a Bayesian

R&S problem in which the prior belief is obtained from existing predictive models for protein and small molecule interaction.

- (Frazier & Powell 2009*b*) considers the lack of concavity in the value of information as a function of the measurement allocation. In some problems, e.g., the Bernoulli R&S problem, a measurement may have significant value when taken together with a group of other measurements, but is essentially worthless when taken by itself. We may understand this phenomenon as a lack of concavity in the value of information as a function of the number of observations taken. KG policies may perform poorly in such problems because they only consider the value of a single measurement.

The author sincerely hopes that the work in this thesis will be of value in applied information collection problems, and that the attempt to provide a simple and flexible class of methods will bear fruit in the ability to attack complex problems.

Appendix A

Known Variance LL(S) Policy

The LL(S) policy was developed for normal measurement errors with *unknown* variance and uses a normal-gamma prior for the unknown mean and measurement precision. To adapt it to the known-variance case, we take both the shape and rate parameter in the gamma prior on the measurement precision to infinity while keeping their ratio fixed to the known measurement precision β^ϵ, we obtain a prior in which the measurement precision is known perfectly and the alternative's true value is still normally distributed. Taking this limit in the allocation given by (Chick & Inoue 2001*b*) in Corollary 1 provides the following policy. The steps below describe how the policy allocates τ measurements for the stage beginning at a generic time n, and should be repeated a total of N/τ times beginning at time 0 and finishing at time N. We use the notation $[i]$ to indicate the alternative whose μ^n component is ith largest. That is, $\mu^n_{[M]} \geq \cdots \geq \mu^n_{[1]}$.

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

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

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

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

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

where

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

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

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

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

Appendix B

Derivation and Computational Complexity of Algorithm 1

Section 4.3.1 presented Algorithm 1 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.

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.

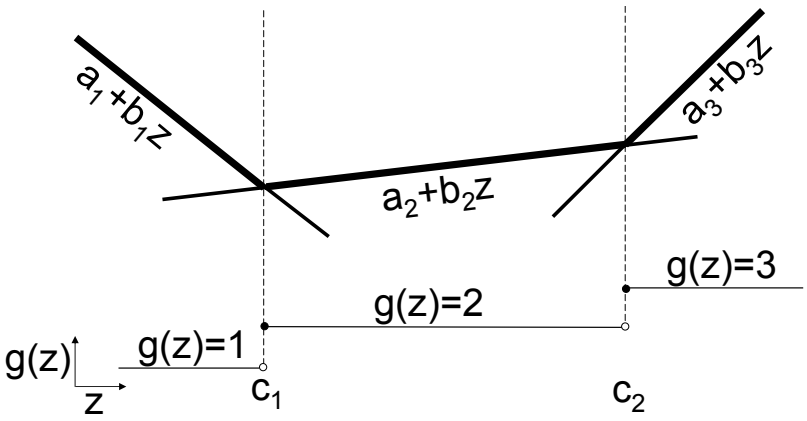

Figure B.1: Illustration of the case when $M = 3$ and no alternatives are dominated. The upper part of the illustration shows the three lines $a_i + b_i z$ for $i = 1, 2, 3$, with z ranging along the horizontal axis. The thicker portions of the lines comprise $\max_i a_i + b_i z$. The lower part of the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

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,

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

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 $\le i$). 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.

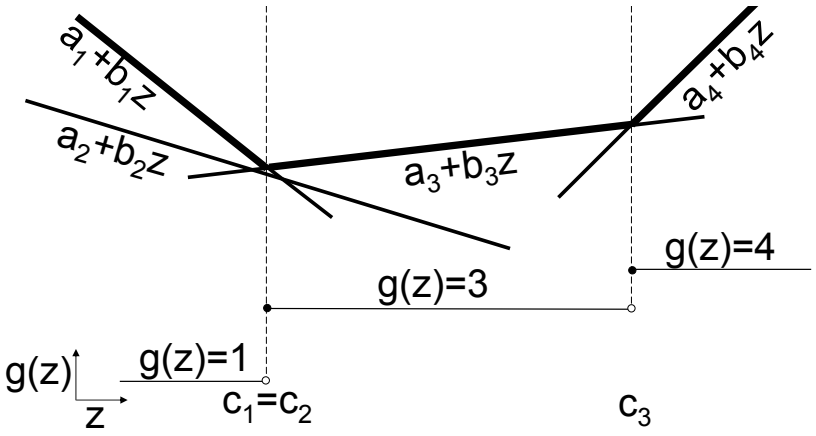

Figure B.2: Illustration of the case when $M = 4$ and alternative 2 is dominated. As in Figure B.1, the upper part of the illustration shows the lines $a_i + b_i z$ for $i = 1, 2, 3, 4$ and the lower part of the figure shows the value of $g(z)$ as a function of z. Alternative 2 is dominated because $a_2 + b_2 z$ 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.

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.

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 $O(M)$.

Appendix C

Proofs

Proof of Proposition 2.2.3

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

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

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

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

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

write

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

Then the induction hypothesis tells us that

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

allowing us to write

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

Proof of Theorem 2.2.6

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

$$\begin{aligned} V^{\pi,N-1}(s) &= \mathbb{E}\left[V^{\pi,N}(T(s,X^\pi(s),Z^N))\right] \geq V^{\pi,N}\left(\mathbb{E}\left[T(s,X^\pi(s),Z^N)\right]\right) \\ &= V^{\pi,N}(\mu, \beta + \beta^\epsilon e_{X^\pi(s)}) = V^{\pi,N}(\mu,\beta) = V^{\pi,N}(s). \end{aligned}$$

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

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

by the induction hypothesis. Then, by the definition of $V^{\pi,n+1}$ in terms of $V^{\pi,n+2}$ from (2.10), we have $V^{\pi,n}(s) \geq V^{\pi,n+1}(s)$.

Proof of Theorem 2.3.2

By (2.15), computing $X^{KG}(s)$ reduces to computing $Q^{N-1}(s,x)$ for each $x \in \{1,\ldots,M\}$. By definition (2.11) we have, for a generic state s and standard normal random variable Z,

$$Q^{N-1}(s,x) = \mathbb{E}\left[V^N(T(s,x,Z))\right] = \mathbb{E}\left[(\mu_x + \tilde{\sigma}(\beta_x)Z) \vee \max_{x' \neq x} \mu_{x'}\right]. \tag{C.2}$$

This expectation is the expectation of the maximum of a constant and a normal random variable, for which we have an analytical expression from (Clark 1961). Let $a \in \mathbb{R}$ be an arbitrary constant and $W \sim \mathcal{N}(b, c^2)$ an arbitrary normal random variable. Then (Clark 1961) tells us,

$$\mathbb{E}\left[W \vee a\right] = a\Phi\left(\frac{a-b}{c}\right) + b\Phi\left(\frac{b-a}{c}\right) + c\varphi\left(\frac{a-b}{c}\right), \tag{C.3}$$

which can be rewritten as

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

Fix x and consider two cases. First, consider the case that $\mu_x > \max_{x'} \mu_{x'}$. This is the case in which we measure an alternative that is uniquely the best according to the prior. Then $\mu_x - \max_{x' \neq x} \mu_{x'}$ is positive and $(\max_{x' \neq x} \mu_{x'} - \mu_x)/\tilde{\sigma}(\beta_x) = \zeta_x(s)$. Substitute $\zeta_x(s)$ for $(a-b)/c$ and write (C.2) as

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

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

$$Q^{N-1}(s,x) = \max_{x'} \mu_{x'} + \tilde{\sigma}(\beta_x) f(\zeta_x(s)). \tag{C.4}$$

Now consider the case that $\mu_x \leq \max_{x'} \mu_{x'}$. We rewrite (C.3) again using the substitution $\Phi(-z) = 1 - \Phi(z)$, and also using the symmetric property of the normal probability density function, $\varphi(-z) = \varphi(z)$, as

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

In the case we are considering, $\mu_x - \max_{x' \neq x} \mu_{x'} \leq 0$ and $(\mu_x - \max_{x' \neq x})/\tilde{\sigma}(\beta_x) = \zeta_x(s)$. Substitute $\zeta_x(s)$ for $(b-a)/c$ and write (C.2) as

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

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

$$Q^{N-1}(s,x) = \max_{x'} \mu_{x'} + \tilde{\sigma}(\beta_x) f(\zeta_x(s)). \tag{C.5}$$

In both cases the expression for $Q^{N-1}(s,x)$ agrees with (2.18), and we use this expression to rewrite (2.15) as

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

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

Proof of Proposition 2.3.8

By Theorem 2.3.2, KG prefers the alternative with the largest value of $\tilde{\sigma}(\beta_x) f(\zeta_x(S))$. Fix $S = (\mu, \beta)$, and let a be as in the statement of Proposition 2.3.8. Let i be the alternative preferred by KG, so

$$i = \arg\max_{x \in \{1,\ldots,M\}} \tilde{\sigma}(\beta_x) f(\zeta_x(S)), \tag{C.6}$$

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

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

$$\tilde{\sigma}(\beta_i) f(\zeta_i(S')) \geq \tilde{\sigma}(\beta_x) f(\zeta_x(S')). \tag{C.7}$$

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

$$\tilde{\sigma}(\beta_x) f(\zeta_x(S)) < \tilde{\sigma}(\beta_i) f(\zeta_i(S)) \implies \tilde{\sigma}(\beta_x) f(\zeta_x(S')) < \tilde{\sigma}(\beta_i) f(\zeta_i(S')) \tag{C.8}$$

holds for all $x \neq i$. This will suffice to show the proposition because if we choose any $x' \notin \arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S))$, (C.6) will imply $\tilde{\sigma}(\beta_{x'}) f(\zeta_{x'}(S)) < \tilde{\sigma}(\beta_i) f(\zeta_i(S))$. The implication (C.8) will then imply that $\tilde{\sigma}(\beta_{x'}) f(\zeta_{x'}(S')) < \tilde{\sigma}(\beta_i) f(\zeta_i(S'))$ and, moreover, that $x' \notin \arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S'))$. Taking the contrapositive of the statement

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

reveals that

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

By this argument, (C.8) implies that $\arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S')) \subseteq \arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S))$. Therefore i is the element of $\arg\max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S))$ with the smallest index, and thus i is the alternative that KG prefers in state S'.

We will show (C.7) and (C.8) by treating three cases separately, noting in general that $\zeta_i(\mu, \beta) \leq \zeta_i(\mu + ae_i, \beta)$. The first case is when $x \neq i, j$. Then

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

Thus, (C.7) is true because

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

and (C.8) is true because if $\tilde{\sigma}(\beta_x) f(\zeta_x(S)) < \tilde{\sigma}(\beta_i) f(\zeta_i(S))$ then

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

The second case is when $x = j$ and $\mu_i + a < \max_{x' \neq j} \mu_{x'}$. Then again $\zeta_j(S') = \zeta_j(S)$ because $j \neq i$, and both (C.7) and (C.8) hold by the same reasoning as in the first case.

The third case is when $x = j$ and $\mu_i + a \geq \max_{x' \neq j} \mu_{x'}$. Then we have $\zeta_j(\mu + ae_i, \beta) = \frac{-|\mu_i + a - \mu_j|}{\tilde{\sigma}(\beta_j)}$. For $x = j$, KG's preference of alternative i implies that $\beta_i \leq \beta_j$. Otherwise, by Remark 2.3.3 and because $\mu_j \geq \mu_i$, KG would prefer alternative j. This shows that

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

This shows (C.7). To show (C.8), assume the antecedent of condition (C.8). Since $|\mu_j - \max_{x' \neq j} \mu_{x'}| \leq |\mu_j - \mu_i|$, and $\tilde{\sigma}(\beta_j) f(\zeta_j(S)) < \tilde{\sigma}(\beta_i) f(\zeta_i(S))$, it must be that $\tilde{\sigma}(\beta_j) < \tilde{\sigma}(\beta_i)$ since otherwise j would have been KG's choice in state S. Thus,

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

Proof of Proposition 2.4.1

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

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

This difference is positive by Corollary 2.2.5.

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

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

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

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

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

Finally, we show that the limit $V^\pi(S^0;\infty)$ exists and is finite for every stationary policy π. Fix a stationary policy π. Then Theorem 2.2.6 implies that $V^{\pi,0}(S^0;N)$ is non-decreasing in N, and $V^{\pi,0}(S^0;N)$ is bounded by $V^0(S^0;N)$, which is itself uniformly bounded in N by $U(S^0)$. Then $V^\pi(S^0;\infty)$ is the limit of a non-decreasing bounded sequence. Hence, it exists.

Proof of Proposition 2.4.2

We assumed in the formal model in section 2.2.1 that our measurement-noise variance $(\sigma^\varepsilon)^2$ is finite. This implies via the strong law of large numbers that the sequence of posterior predictive means μ_x^N converges as $\lim_{N\to\infty} \mu_x^N = Y_x$ almost surely for each $x = 1,\ldots,M$. Thus $\lim_{N\to\infty} \max_x \mu_x^N$ exists almost surely and in probability. We will show next that the sequence $\left(\max_x \mu_x^N\right)_{N\geq 1}$ is uniformly integrable and then convergence in probability together with uniform integrability imply convergence in L^1 (see, e.g., (Kallenberg 1997) Theorem 3.12). Convergence in L^1 of $\max_x \mu_x^N$ as $N \to \infty$ implies

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

Proposition 2.4.1 showed that $U(S^0) \geq V(S^0;\infty)$, so $V^\pi(S^0;\infty) = V(S^0;\infty)$ and π must be asymptotically optimal.

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

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

We assumed in the formal model in section 2.2.1 that $\max_x |Y_x|$. This implies via Markov's inequality that

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

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

Proof of Theorem 2.4.3

First note that KG is stationary. We will show that $\lim_{N\to\infty} \eta_x^N = \infty$ almost surely for all x under KG, and then Proposition 2.4.2 will complete the proof.

First we show that, for each x, $\{\mu_x^n\}_{n=0}^\infty$ is a uniformly integrable martingale with respect to the filtration $\mathcal{F}$ and hence converges. μ_x^n is defined by $\mu_x^n := \mathbb{E}\left[Y_x \mid \mathcal{F}^n\right]$ and thus is $\mathcal{F}^n$-measurable and, by the tower property of conditional expectation, satisfies the martingale identity. Y_x is a normal random variable with finite variance. Thus, $Y_x \in L^2 \subset L^1$, and by the Doob uniform integrability lemma ((Kallenberg 1997) Lemma 5.5), the collection of conditional expectations $\{\mu_x^n\}_n$ is uniformly integrable (and hence each μ_x^n is integrable). Thus, $\{\mu_x^n\}_n$ is a uniformly integrable martingale and hence converges almost surely to an integrable random variable μ_x^∞. In addition, $\lim_{n\to\infty} \beta_x^n \overset{a.s.}{=} \beta_x^0 + \beta^\epsilon \eta_x^\infty$ for each x.

By the computation performed in Theorem 2.3.2, the Q-factors for each alternative x are continuous functions of their arguments (μ, β) and hence,

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

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

$$\lim_{n\to\infty} Q^{N-1}(S^n(\omega); x) > \max_{x'} \mu_{x'}^\infty(\omega) \qquad \text{for all } \omega \in \mathcal{H}_x \cap \Omega_0, \tag{C.9}$$

$$\lim_{n\to\infty} Q^{N-1}(S^n(\omega); x) = \max_{x'} \mu_{x'}^\infty(\omega) \qquad \text{for all } \omega \in \mathcal{H}_x^c \cap \Omega_0. \tag{C.10}$$

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

Fix A nonempty and suppose for contradiction that $\mathcal{H}_A \cap \Omega_0$ is nonempty so that we may choose $\omega \in \mathcal{H}_A \cap \Omega_0$ to be an element of this set. By (C.9) and (C.10), for all $x \in A$ and all $y \in A^c$,

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

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

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

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

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

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

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

Proof of Theorem 2.5.1

Note that $\varphi(0) = (2\pi)^{-1/2}$, where φ is the normal pdf. We will use this throughout. We induct backwards over n. First, when $n = N-1$, the theorem is trivially true with equality.

Now, under the assumption that the theorem is true for some $n+1$,

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

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

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

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

$$V^n(s) \leq \max_x \mathbb{E}\left[V^{N-1}(T(s,x,Z^{n+1}))\right] + \varphi(0)(N-n-2)\max_{x'} \tilde{\sigma}(\beta_{x'}). \tag{C.11}$$

We will rewrite the first term on the right hand side of this inequality as a maximum over a set of Q-factors using the definition of V^{N-1} in terms of Q^{N-1}, but before making this substitution, let us bound Q^{N-1}. We rewrite the expression (C.4) for Q^{N-1} as $Q^{N-1}(s,x') = \max_{x''} \mu_{x''} + \tilde{\sigma}(\beta_{x'})f(\zeta_{x'}) = V^N(s) + \tilde{\sigma}(\beta_{x'})f(\zeta_{x'})$. Lemma 2.3.5 tells us that f is non-decreasing, so $\zeta_{x'} \leq 0$ implies that $f(\zeta_{x'}) \leq f(0) = \varphi(0)$. Thus,

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

Using this and the definition of the value function in terms of the Q-factors from (2.10) and (2.11), we have

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

Combining this bound with (C.11), and moving the $\tilde{\sigma}(\beta_x)$ outside the maximization and expectation operators, we obtain

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

where in the last step we used the definition of V^N in terms of V^{N-1} from (2.10).

Proof of Theorem 2.6.3

The proof is by induction backward on k. The theorem holds for the base case, $k = N-1$, by Remark 2.3.1. Now let $k < N-1$. Let π^* be an optimal policy, with decision function X^{*k} at time k. Let $s = (\mu, \beta) \in \mathbb{S}^k$. Then

$$V^k(s) = \mathbb{E}\left[V^{k+1}(T(s, X^{*k}(s), Z^{k+1}))\right] = \mathbb{E}\left[V^{KG,k+1}(T(s, X^{*k}(s), Z^{k+1}))\right], \qquad \text{(C.12)}$$

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

Consider two cases. In the first case, suppose $X^{*k}(s) = X^{KG,k}(s)$. By (C.12),

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

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

$$\begin{aligned} V^k(s) &= \mathbb{E}\left[V^{KG,k+2}(T(S^{k+1}, X^{KG}(S^{k+1}), Z^{k+2}))\right] \\ &= \mathbb{E}\left[V^{KG,k+2}(T(S^{k+1}, X^{KG}(s), Z^{k+2}))\right], \end{aligned} \qquad \text{(C.13)}$$

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

Note that S^{k+2} is the state to which we arrive when we measure $X^{*k}(s)$ at time k and $X^{KG}(s)$ at time $k+1$. Let $E_x = e_x(e_x)^T$ be a matrix of all zeros except for a single 1 at row x, column x, and let $\stackrel{d}{=}$ denote equality in distribution. Then the definition (2.8) of the transition function T and $X^{KG}(s) \neq X^{*,k}(s)$ imply

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

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

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

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

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

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

Proof of Theorem 2.6.6

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

$$(\beta_i \neq \infty \text{ and } \beta_j \neq \infty \text{ and } \beta_i < \beta_j) \implies \mu_i \geq \mu_j \tag{C.14}$$

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

Let $n \in \{0 \dots N-2\}$, $x \in \{1, \dots, M\}$, $s \in \mathbb{S}^n$, and $S^n = s$ a.s. Consider $S^{n+1} := T(S^n, x, Z^{n+1})$. Let $i, j \in \{1, \dots, M\}$ meet the conditions of the implication (C.14) for S^{n+1}, so $\beta_i^{n+1} \neq \infty$ and $\beta_j^{n+1} \neq \infty$ and $\beta_i^{n+1} < \beta_j^{n+1}$. We will show that $\mu_i^n \geq \mu_j^n$, which will show that S^{n+1} meets condition (C.14) and is in $\mathbb{S}^{n+1}$.

First, $\beta^n \leq \beta^{n+1}$ component-wise implies that $\beta_i^n \neq \infty$ and $\beta_j^n \neq \infty$. Also, since $(\sigma^\varepsilon)^2 = 0$, $\beta_x^{n+1} = \infty$, which implies that $x \neq i, j$, and the measurement between S^n and S^{n+1} altered neither the i component nor the j component. Thus, $\beta_i^n = \beta_i^{n+1} < \beta_j^{n+1} = \beta_j^n$. This shows that i, j meet the conditions of the implication (C.14) for S^n as well as S^{n+1}. Thus, since $S^n \in \mathbb{S}^n$, $\mu_i^n \geq \mu_j^n$. Then, again because $x \neq i, j$ implies that the means of the i and j components did not change from time n to $n+1$, $\mu_i^{n+1} \geq \mu_j^{n+1}$, showing that S^{n+1} meets the condition (C.14), and $S^{n+1} \in \mathbb{S}^{n+1}$. Thus, $\{\mathbb{S}^n\}$ is a covering of the future from 0.

Now we will show that KG is persistent on $\{\mathbb{S}^n\}$. Let $s \in \mathbb{S}^n$ and $S^n = s$ a.s. Condition (C.14) together with Remark 2.3.3 and Remark 2.3.4 imply $X^{KG}(S^n) \in \arg\min_{x'} \beta^n_{x'}$ with ties broken by smallest index. Let $x \neq X^{KG}(S^n)$. We showed that $S^{n+1} := T(S^n, x, Z^{n+1}) \in \mathbb{S}^{n+1}$ almost surely. Thus, again by condition (C.14), Remark 2.3.3 and Remark 2.3.4, $X^{KG}(S^{n+1}) \in \arg\min_{x'} \beta^{n+1}_{x'}$. We use the state transition function for the case with $(\sigma^\varepsilon) = 0$, $\beta^{n+1}_{x'} = \beta^n_{x'} + \infty 1_{\{x'=x\}}$, and we consider two cases.

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

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

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

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

In both cases KG is persistent on $\{\mathbb{S}^n\}$, and Theorem 2.6.3 shows that $V^{KG,0}(s) = V^0(s)$ for all $s \in \mathbb{S}^0$.

Proof of Lemma 4.4.2

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).

Proof of Lemma 4.4.3

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

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

Proof of Lemma 4.4.4

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

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

Proof of Theorem 4.4.5

Lemma 4.4.2 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 4.4.1. 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 4.2.2 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 4.4.4 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,

$$x^n(\omega) \notin A \quad \forall n > K(\omega). \tag{C.15}$$

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 (C.15) 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 4.4.3, $V^N(S^\infty) = U(S^\infty)$ almost surely.

Proof of Proposition 3.3.1

We will show the result by considering the problem (3.8) as a dynamic program. Our state space will consist of the current time n and the posterior distribution at that time, which is parameterized by the vectors a^n, b^n, ρ^n, μ^n. For compactness we will use S^n to denote this tuple of random vectors $(a^n, b^n, \rho^n, \mu^n)$ and s to denote one possible value that this tuple might take.

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

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

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

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

$$\max_x \mu_x^n - C(n) < \mathbb{E}\left[\max_x \mu^{n+1} - C(n+1) \mid S^n = s\right]. \tag{C.16}$$

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

$$\max_x \mu_x^n - C(n) < \mathbb{E}\left[V(S^{n+1}, n+1) \mid S^n = s\right]. \tag{C.17}$$

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

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

where the inequality in the penultimate line is a consequence of the fact that the deterministic time $n+1$ is a stopping time contained in the set over which the supremum is taken, and the final line is due to the tower property of conditional expectation. Finally, this inequality with (C.16) shows (C.17).

Proof of Lemma 5.1.1

The random variable $\beta(\theta)$ is integrable under Q_0 and hence under $\mathbb{P}$ because the measure Q_0 is in the exponential family with natural parameter $\tau(S_0, N_0) = 0 \in \mathrm{dom}(\Lambda)$, and the openness of $\mathrm{dom}(\Lambda)$ implies that the moment generating function of $\beta(\theta)$ under Q_0, which is $\exp(\Psi(\,\cdot\,))$, is finite in an open ball around 0 in $\mathbb{R}^l$.

The integrability of $\beta(\theta)$, together with the fact $K_t = \mathbb{E}_t[\beta(\theta)]$, implies that $(K_t)_{t\in\mathbb{N}}$ is a uniformly integrable martingale. By (Kallenberg 1997) Theorem 6.21, $(K_t)_{t\in\mathbb{N}}$ converges almost surely to an integrable random variable. This random variable is K_∞. It takes values in $\mathrm{cl}(\mathcal{K})$ because each K_t takes values in $\mathcal{K}$.

Proof of Lemma 5.1.2

For each $T \in \mathbb{N}$, the inner expectation on the right-hand side of (5.3) satisfies

$$\mathbb{E}\left[R(\theta;i) \mid P(\,\cdot\,;x,\theta)\ \forall x \in \mathcal{X}\right] = \mathbb{E}_T\left[R(\theta;i) \mid P(\,\cdot\,;x,\theta)\ \forall x \in \mathcal{X}\right]$$

since $(X_t, Y_t)_{t\le T}$ is conditionally independent of $R(\theta;i)$ given $P(\,\cdot\,;x,\theta)$, $x \in \mathcal{X}$. Using this and the tower property of conditional expectation, the right-hand side of (5.3) can be rewritten

$$\mathbb{E}\left[\mathbb{E}_T\left[\min_{i\in\mathcal{I}} \mathbb{E}_T\left[R(\theta;i) \mid P(\,\cdot\,;x,\theta)\ \forall x \in \mathcal{X}\right]\right]\right].$$

Thus, asymptotic optimality holds if and only if

$$\lim_{T\to\infty} \mathbb{E}\left[\min_i \mathbb{E}_T\left[R(\theta;i)\right] - \mathbb{E}_T\left[\min_i \mathbb{E}_T\left[R(\theta;i) \mid P(\,\cdot\,;x,\theta)\ \forall x \in \mathcal{X}\right]\right]\right] = 0,$$

which is true if and only if

$$g(K_T;\mathcal{X}) = \min_i \mathbb{E}_T\left[R(\theta;i)\right] - \mathbb{E}_T\left[\min_i \mathbb{E}_T\left[R(\theta;i) \mid P(\,\cdot\,;x,\theta)\ \forall x \in \mathcal{X}\right]\right]$$

converges almost surely to 0, where we use the fact that applying Jensen's inequality and the tower property to $\mathbb{E}_T\left[\min_i \mathbb{E}_T\left[R(\theta;i) \mid P(\,\cdot\,;x,\theta)\ \forall x \in \mathcal{X}\right]\right]$ shows $g(K_T;\mathcal{X})$ is nonnegative.

Proof of Lemma 5.2.1

Fix $x \in \mathcal{X}$ and $i \in \mathcal{I}$, and let Ω_x denote the event $\{N_{\infty,x} = \infty\}$.

Let $\mathcal{C}$ denote a countable separating class for $\mathcal{Y}$. Then, any measure on $\mathcal{Y}$ is completely determined by the values it takes on the elements of $\mathcal{C}$, and

$$\sigma\left\{P(\,\cdot\,; x, \theta)\right\} = \bigvee_{C \in \mathcal{C}} \sigma\left\{P(C; x, \theta)\right\}.$$

Thus, $\mathbb{E}_\infty\left[R(\theta; i) \mid P(\,\cdot\,; x, \theta)\right]$ may be written as $f((X_t, Y_t)_{t\in\mathbb{N}}, (P(C; x, \theta))_{C\in\mathcal{C}})$ for some measurable function $f : (\mathcal{X} \times \mathcal{Y})^{\mathbb{N}} \times [0,1]^{\mathcal{C}} \mapsto \mathbb{R}$.

For each $C \in \mathcal{C}$ define a random variable

$$\hat{P}_C := \begin{cases} 0, & \text{if } \sum_{t\in\mathbb{N}} \mathbf{1}_{\{X_t = x\}} = 0, \\ \lim_{t\to\infty} \left(\sum_{t' \le t} \mathbf{1}_{\{X_{t'} = x\}} \mathbf{1}_{\{Y_{t'} \in C\}}\right) / N_{tx}, & \text{otherwise.} \end{cases}$$

This random variable is $\mathcal{F}^\infty$-measurable, and the event $\left\{\hat{P}_C = P(C; x, \theta)\ \forall C \in \mathcal{C}\right\}$ is almost sure on Ω_x by the strong law of large numbers and the countability of $\mathcal{C}$. Thus,

$$\mathbb{E}_\infty\left[R(\theta; i) \mid P(\,\cdot\,; x, \theta)\right] = f((X_t, Y_t)_{t\in\mathbb{N}}, (P(C; x, \theta))_{C\in\mathcal{C}}) = f((X_t, Y_t)_{t\in\mathbb{N}}, (\hat{P}_C)_{C\in\mathcal{C}})$$

almost surely on Ω_x, and

$$\mathbb{E}_\infty\left[R(\theta; i)\mathbf{1}_{\Omega_x} \mid P(\,\cdot\,; x, \theta)\right] = f((X_t, Y_t)_{t\in\mathbb{N}}, (\hat{P}_C)_{C\in\mathcal{C}})\mathbf{1}_{\Omega_x}$$

almost surely, where we use the fact that $\mathbf{1}_{\Omega_x} \in \mathcal{F}^\infty$.

Furthermore, by the tower property and the $\mathcal{F}^\infty$-measurability of $\mathbf{1}_{\Omega_x}$,

$$\begin{aligned} \mathbb{E}_\infty\left[R(\theta; i)\mathbf{1}_{\Omega_x}\right] &= \mathbb{E}_\infty\left[\mathbb{E}_\infty\left[R(\theta; i)\mathbf{1}_{\Omega_x} \mid P(\,\cdot\,; x, \theta)\right]\right] \\ &= \mathbb{E}_\infty\left[f((X_t, Y_t)_{t\in\mathbb{N}}, (\hat{P}_C)_{C\in\mathcal{C}})\mathbf{1}_{\Omega_x}\right] \\ &= f((X_t, Y_t)_{t\in\mathbb{N}}, (\hat{P}_C)_{C\in\mathcal{C}})\mathbf{1}_{\Omega_x} = \mathbb{E}_\infty\left[R(\theta; i)\mathbf{1}_{\Omega_x} \mid P(\,\cdot\,; x, \theta)\right]. \end{aligned}$$

This shows the lemma.

Note that although the proof can be simplified in the case $\sigma\left\{P(\,\cdot\,; x, \theta)\mathbf{1}_{\Omega_x}\right\} \subseteq \mathcal{F}^\infty$, this inclusion does not hold in general because there may be a measure-0 event on which $\hat{P}_C(\omega) \neq P(C; x, \theta(\omega))$.

Proof of Lemma 5.2.2

Fix any $x \in \mathcal{X}$, and let Ω_x denote the event $\{N_{\infty,x} = \infty\}$. Now, fix $i \in \mathcal{I}$ and consider the sequence of conditional expectations $(\mathbb{E}_t [R(\theta; i) \mid P(\cdot\,; x, \theta)])_{t\in\mathbb{N}}$. Since $R(\theta; i)$ is integrable, this is a uniformly integrable martingale with respect to the filtration $\left(\mathcal{F}^t \vee \sigma \{P(\cdot\,; x, \theta)\}\right)_{t\in\mathbb{N}}$, and converges almost surely and in L^1 to the integrable random variable $\mathbb{E}_\infty [R(\theta; i) \mid P(\cdot\,; x, \theta)]$ (see, e.g., (Kallenberg 1997) Theorem 6.21).

Similarly, $(\mathbb{E}_t [R(\theta; i)])_{t\in\mathbb{N}}$ is a uniformly integrable martingale with respect to the filtration $\left(\mathcal{F}^t\right)_{t\in\mathbb{N}}$, and converges almost surely and in L^1 to the integrable random variable $\mathbb{E}_\infty [R(\theta; i)]$.

By Lemma 5.2.1, $\mathbb{E}_\infty [R(\theta; i) \mid P(\cdot\,; x, \theta)] \mathbf{1}_{\Omega_x} = \mathbb{E}_\infty [R(\theta; i)] \mathbf{1}_{\Omega_x}$. For each $t \in \mathbb{N} \cup \{\infty\}$ define two random variables,

$$R_t := \min_{i\in\mathcal{I}} \mathbb{E}_t [R(\theta; i)],$$
$$\widetilde{R}_t := \min_{i\in\mathcal{I}} \mathbb{E}_t [R(\theta; i) \mid P(\cdot\,; x, \theta)].$$

Since $\mathcal{I}$ is a finite set, we have $R_t \to R_\infty$ almost surely and in L^1, $\widetilde{R}_t \to \widetilde{R}_\infty$ almost surely and in L^1, and $R_\infty \mathbf{1}_{\Omega_x} = \widetilde{R}_\infty \mathbf{1}_{\Omega_x}$ almost surely.

We now show $\mathbb{E}_t \widetilde{R}_t$ converges in L^1 to $\mathbb{E}_\infty \widetilde{R}_\infty$. We begin by using the triangle inequality to bound $\lim_t \mathbb{E}\left[\left|\mathbb{E}_t \widetilde{R}_t - \mathbb{E}_\infty \widetilde{R}_\infty\right|\right]$ above by

$$\lim_t \mathbb{E}\left[\left|\mathbb{E}_t \widetilde{R}_t - \mathbb{E}_t \widetilde{R}_\infty\right|\right] + \lim_t \mathbb{E}\left[\left|\mathbb{E}_t \mathbb{E}_\infty \widetilde{R}_\infty - \mathbb{E}_\infty \widetilde{R}_\infty\right|\right].$$

We show that this upper bound is zero. We rewrite the first term and bound it above via Jensen's inequality to obtain,

$$\lim_t \mathbb{E}\left[\left|\mathbb{E}_t \widetilde{R}_t - \mathbb{E}_t \widetilde{R}_\infty\right|\right] = \lim_t \mathbb{E}\left[\left|\mathbb{E}_t\left[\widetilde{R}_t - \widetilde{R}_\infty\right]\right|\right] \leq \lim_t \mathbb{E}\left[\left|\widetilde{R}_t - \widetilde{R}_\infty\right|\right] = 0,$$

which is zero since $\widetilde{R}_t$ converges to $\widetilde{R}_\infty$ in L^1. Examining the second term, $\widetilde{R}_\infty$ is an integrable random variable, so $\left(\mathbb{E}_t \widetilde{R}_\infty\right)_{t\in\mathbb{N}}$ is a uniformly integrable martingale converging in L^1 to $\mathbb{E}_\infty \widetilde{R}_\infty$. This shows the second term is also zero, implying that $\lim_t \mathbb{E}\left[\left|\mathbb{E}_t \widetilde{R}_t - \mathbb{E}_\infty \widetilde{R}_\infty\right|\right] = 0$, and $\mathbb{E}_t \widetilde{R}_t$ converges in L^1 to $\mathbb{E}_\infty \widetilde{R}_\infty$.

Conditioning on $\mathcal{F}^t \vee \sigma \{P(\cdot\,; x, \theta)\}$ and using Jensen's inequality together with the tower property shows that $\mathbb{E}_t\left[\mathbb{E}_{t+1} \widetilde{R}_{t+1}\right] \leq \mathbb{E}_t \widetilde{R}_t$, and so $(\mathbb{E}_t \widetilde{R}_t)_{t\in\mathbb{N}}$ is a supermartingale.

Since it is nonnegative, it converges almost surely, and this convergence must be to the same random variable $\mathbb{E}_\infty \widetilde{R}_\infty$ to which L^1 convergence is shown above.

This implies in turn that $(\mathbb{E}_t \widetilde{R}_t)\mathbf{1}_{\Omega_x}$ converges almost surely to

$$\left(\mathbb{E}_\infty \widetilde{R}_\infty\right)\mathbf{1}_{\Omega_x} = \mathbb{E}_\infty\left[\widetilde{R}_\infty \mathbf{1}_{\Omega_x}\right] = \mathbb{E}_\infty\left[R_\infty \mathbf{1}_{\Omega_x}\right] = R_\infty \mathbf{1}_{\Omega_x}.$$

Also converging almost surely to $R_\infty \mathbf{1}_{\Omega_x}$ is $(R_t \mathbf{1}_{\Omega_x})_{t\in\mathbb{N}}$, and since both $(\mathbb{E}_t \widetilde{R}_t)\mathbf{1}_{\Omega_x}$ and $R_t \mathbf{1}_{\Omega_x}$ converge almost surely to the same random variable, we have that $g(K_t;\{x\})\mathbf{1}_{\Omega_x} = \left(-R_t + \mathbb{E}_t \widetilde{R}_t\right)\mathbf{1}_{\Omega_x}$ converges to zero almost surely. This shows that $(\Omega \setminus \Omega_x) \cup \{\lim_t g(K_t;\{x\}) = 0\}$ is almost sure, completing the proof.

Proof of Theorem 5.2.3

Define four events,

$$\begin{aligned}
\Omega_1 &:= \left\{K_t \in \widetilde{\mathcal{K}},\ \forall t \in \mathbb{N} \cup \{\infty\}\right\},\\
\Omega_2 &:= \left\{K_\infty = \lim_t K_t \text{ exists}\right\},\\
\Omega_3 &:= \bigcap_{x\in\mathcal{X}} \left[\{N_{\infty,x} < \infty\} \cup \left\{\lim_t g(K_t;\{x\}) = 0\right\}\right],\\
\Omega_4 &:= \bigcap_{A\subseteq 2^{\mathcal{X}}} \left[\left\{\sum_{t\in\mathbb{N}} \Pi(t,K_t,A) < \infty\right\} \cup \left\{\sum_{t\in\mathbb{N}} \mathbf{1}_{\{X_{t+1}\in A\}} = \infty\right\}\right].
\end{aligned}$$

Each of these events is almost sure: Ω_1 is almost sure by the assumption of the theorem; Ω_2 is almost sure by Lemma 5.1.1; Ω_3 is almost sure by Lemma 5.2.1; and Ω_4 is almost sure by the extended Borel-Cantelli Lemma (see, e.g., (Kallenberg 1997) Corollary 6.20), since $\left\{\sum_{t\in\mathbb{N}} \mathbf{1}_{\{X_{t+1}\in A\}} = \infty\right\}$ is almost sure on $\left\{\sum_{t\in\mathbb{N}} \mathbb{P}_t\left\{X_{t+1}\in A\right\} = \infty\right\}$ and $\mathbb{P}_t\left\{X_{t+1}\in A\right\} = \Pi(t,K_t,A)$ almost surely. We then define their intersection $\Omega_0 := \Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$ and note that it too is almost sure.

Choose $\omega \in \Omega_0$ and suppose for contradiction that $K_\infty(\omega) \notin M_*$. Since $\omega \in \Omega_1$ implies $K_\infty(\omega) \in \widetilde{\mathcal{K}}$, (5.4) implies $K_\infty(\omega)$ has an open neighborhood U such that $\limsup_t c_t < 1$, where

$$c_t := \sup_{k'\in U\cap\widetilde{\mathcal{K}}\cap\mathcal{K}} \Pi(t,k',A_{K_\infty(\omega)}).$$

Since $\lim_t K_t(\omega) = K_\infty(\omega)$, there exists a $t' \in \mathbb{N}$ such that $K_t(\omega) \in U$ for all $t > t'$. Furthermore, for all $t > t'$ the finiteness of t, together with $\omega \in \Omega_1$, implies $K_t(\omega) \in$

$U \cap \widetilde{\mathcal{K}} \cap \mathcal{K}$. Thus

$$\sum_{t\in\mathbb{N}} \Pi(t, K_t(\omega), \mathcal{X} \setminus A_{K_\infty(\omega)}) \geq \sum_{t>t'} \Pi(t, K_t(\omega), \mathcal{X} \setminus A_{K_\infty(\omega)}) \geq \sum_{t>t'} (1 - c_t) = \infty,$$

where the final equality with infinity is due to $\limsup_t c_t < 1$.

Since $\omega \in \Omega_4$, this implies $\sum_{t\in\mathbb{N}} \mathbf{1}_{\{X_{t+1}(\omega)\notin A_{K_\infty(\omega)}\}} = \infty$. In particular, since $\mathcal{X}$ is finite, there must exist some $x \notin A_{K_\infty(\omega)}$ satisfying $\sum_{t\in\mathbb{N}} \mathbf{1}_{\{X_{t+1}(\omega)=x\}} = \infty$. Finally, $\omega \in \Omega_3$ implies $\lim_t g(K_t(\omega); \{x\}) = 0$, implying that $K_\infty(\omega) \in M_x$, and $x \in A_{K_\infty(\omega)}$. This contradicts $x \notin A_{K_\infty(\omega)}$.

This contradiction shows that $K_\infty(\omega) \in M_*$ for all $\omega \in \Omega_0$, implying $\lim_t g(K_t; \{x\}) = 0$ almost surely for all $x \in \mathcal{X}$. This, together with Assumption 5.1.3 and Lemma 5.1.2, implies asymptotic optimality.

Proof of Lemma 5.3.1

Since μ_i is $\mathbb{P}^{(k)}$ integrable for each $k \in \mathcal{K}$, the inequality $R(\theta; i) = \max_j \mu_j - \mu_i \leq \sum_j |\mu_j|$ shows that $R(\theta; i)$ is also $\mathbb{P}^{(k)}$ integrable for each $k \in \mathcal{K}$, and $\mathrm{dom}(g) = \mathcal{K}$. We now move to the statement about continuity of g. For $k \in \mathrm{dom}(g)$ and $C \subseteq \mathcal{X}$, we have

$$g(k; C) = \min_i \mathbb{E}^{(k)} [R(\theta; i)] - \mathbb{E}^{(k)} \left[\min_i \mathbb{E}^{(k)} [R(\theta; i) \mid (\mu_j, \lambda_j), j \in \mathcal{I}_C] \right].$$

Noting the $\mathbb{P}^{(k)}$ integrability of $\max_{j'} \mu_{j'}$, and the two relations

$$\begin{aligned}\min_i \mathbb{E}^{(k)} [R(\theta; i)] = \mathbb{E}^{(k)} \left[\max_{j'} \mu_{j'} \right] - \max_i \hat{\mu}_i(k), \\ \mathbb{E}^{(k)} \left[\min_i \mathbb{E}^{(k)} [R(\theta; i) \mid (\mu_j, \lambda_j), j \in \mathcal{I}_C] \right] \\ = \mathbb{E}^{(k)} \left[\max_{j'} \mu_{j'} \right] - \mathbb{E}^{(k)} \left[\max_i \mathbb{E}^{(k)} [\mu_i \mid (\mu_j, \lambda_j), j \in \mathcal{I}_C] \right],\end{aligned}$$

we have

$$\begin{aligned} g(k; C) &= \mathbb{E}^{(k)} \left[\max_i \mathbb{E}^{(k)} [\mu_i \mid (\mu_j, \lambda_j), j \in \mathcal{I}_C] \right] - \max_i \hat{\mu}_i(k) \\ &= \mathbb{E}^{(k)} \left[\max \left(\max_{i\in\mathcal{I}_C} \mu_i, \max_{i\notin\mathcal{I}_C} \hat{\mu}_i(k) \right) \right] - \max_{i\in\mathcal{I}} \hat{\mu}_i(k). \end{aligned} \tag{C.18}$$

This expression agrees with (5.6) for $k \in \mathrm{dom}(g)$, and so it only remains to show that (5.6) is continuous on $\mathrm{cl}(\mathrm{dom}(g))$.

Let $f(\theta,k) = \max(\max_{i\in\mathcal{I}_C}\mu_i, \max_{i\notin\mathcal{I}_C}\hat{\mu}_i(k))$. Since $k \mapsto \hat{\mu}_i(k)$ is continuous and finite on $\mathrm{cl}(\mathrm{dom}(g))$, to show continuity of (5.6) it is sufficient to show continuity of $k \mapsto \mathbb{E}^{(k)}[f(\theta,k)]$ on $\mathrm{cl}(\mathcal{K})$. To this end, let $(k_t) \subseteq \mathrm{cl}(\mathrm{dom}(g))$ be a sequence converging to $k_* \in \mathrm{cl}(\mathrm{dom}(g))$. We show $\lim_t \mathbb{E}^{(k_t)}[f(\theta,k_t)] = \mathbb{E}^{(k_*)}[f(\theta,k_*)]$ through Lebesgue's dominated convergence theorem together with the identity

$$\mathbb{E}^{(k)}[f(\theta,k)] = \int_{\mathbb{R}_+} \mathbb{P}^{(k)}\{f(\theta,k) > u\} - \mathbb{P}^{(k)}\{f(\theta,k) < -u\}\ du. \tag{C.19}$$

The integrand in (C.19) is dominated by

$$\begin{aligned}
\mathbb{P}^{(k)}\{|f(\theta,k)| > u\} &\le \mathbb{P}^{(k)}\left\{\max\left(\max_{i\in\mathcal{I}_C}|\mu_i|, \max_{i\notin\mathcal{I}_C}|\hat{\mu}_i(k)|\right) > u\right\} \\
&= 1 - \left[\prod_{i\in\mathcal{I}_C}\mathbb{P}^{(k)}\{|\mu_i| \le u\}\right]\left[\prod_{i\notin\mathcal{I}_C}\mathbf{1}_{\{|\hat{\mu}_i(k)|\le u\}}\right] \\
&\le \begin{cases} 1, & \text{if } \max_i|\hat{\mu}_i(k)| \ge u, \\ 1 - \prod_{i\in\mathcal{I}_C}\mathbb{P}^{(k)}\{|\mu_i| \le u\}, & \text{otherwise.} \end{cases}
\end{aligned}$$

Construct $\tilde{k}$ so that $\hat{\mu}_i(\tilde{k}) = \inf_t|\hat{\mu}_i(k_t)|$, $\sigma_i^2(\tilde{k}) = \sup_t\sigma_i^2(k_t)$ and note that $\tilde{k} \in \widetilde{\mathcal{K}}$.

For $u > |\hat{\mu}_i(k)|$, $\mathbb{P}^{(k)}\{|\mu_i| \le u\}$ is decreasing in $|\hat{\mu}_i(k)|$ and increasing in $\sigma_i^2(k)$, so for all $u \ge \bar{u} := \max_{i\in\mathcal{I}_C}\sup_t|\hat{\mu}_i(k_t)|$ we have

$$1 - \prod_{i\in\mathcal{I}_C}\mathbb{P}^{(k_t)}\{|\mu_i| \le u\} \le 1 - \prod_{i\in\mathcal{I}_C}\mathbb{P}^{(\tilde{k})}\{|\mu_i| \le u\} = \mathbb{P}^{(\tilde{k})}\left\{\max_{i\in\mathcal{I}_C}|\mu_i| > u\right\}.$$

Thus, the integrand in (C.19) is dominated by $\mathbf{1}_{\{u<\bar{u}\}} + \mathbf{1}_{\{u\ge\bar{u}\}}\mathbb{P}^{(\tilde{k})}\{\max_{i\in\mathcal{I}_C}|\mu_i| > u\}$, a function whose integral over $\mathbb{R}_+$ is given by

$$\bar{u} + \int_{[\bar{u},\infty)}\mathbb{P}^{(\tilde{k})}\left\{\max_{i\in\mathcal{I}_C}|\mu_i| > u\right\}\ du \le \bar{u} + \mathbb{E}^{(\tilde{k})}\left[\max_{i\in\mathcal{I}_C}|\mu_i|\right] \le \bar{u} + \sum_{i\in\mathcal{I}_C}\mathbb{E}^{(\tilde{k})}[|\mu_i|],$$

which is finite because $\bar{k} \in \widetilde{\mathcal{K}}$.

Thus, by the dominated convergence theorem,

$$\lim_t \mathbb{E}^{(k_t)}[f(\theta,k_t)] = \int_{\mathbb{R}_+}\lim_t \mathbb{P}^{(k_t)}\{f(\theta,k_t) > u\} - \mathbb{P}^{(k_t)}\{f(\theta,k_t) < -u\}\ du, \tag{C.20}$$

provided that the limit of the integrand exists for all but at most countably many u.

Note that there at most countably many $u \in \mathbb{R}$ with $\mathbb{P}^{(k_*)}\{|f(\theta,k_*)| = u\} > 0$, and choose any other u. We show $\lim_t \mathbb{P}^{(k_t)}\{f(\theta,k_t) > u\} = \mathbb{P}^{(k_*)}\{f(\theta,k_*) > u\}$, and a similar

argument may be used to show $\lim_t \mathbb{P}^{(k_t)}\{f(\theta,k_t) < -u\} = \mathbb{P}^{(k_*)}\{f(\theta,k_*) < -u\}$. This argument is used again later in the proof of Lemma 5.4.1, and so we note here that we only use three facts: $\mathbb{P}^{(k_*)}\{|f(\theta,k_*)| = u\} = 0$; f is uniformly continuous; and $\mathbb{P}^{(k_t)}$ converges weakly to $\mathbb{P}^{(k_*)}$.

By the triangle inequality,

$$\begin{aligned}\lim_t \left|\mathbb{P}^{(k_t)}\{f(\theta,k_t) > u\} - \mathbb{P}^{(k_*)}\{f(\theta,k_*) > u\}\right| \\ \leq \lim_t \left|\mathbb{P}^{(k_t)}\{f(\theta,k_*) > u\} - \mathbb{P}^{(k_*)}\{f(\theta,k_*) > u\}\right| \\ + \lim_t \left|\mathbb{P}^{(k_t)}\{f(\theta,k_t) > u\} - \mathbb{P}^{(k_t)}\{f(\theta,k_*) > u\}\right|.\end{aligned}$$

Since $\mathbb{P}^{(k_*)}\{f(\theta,k_*) = u\} = 0$, $\mathbb{P}^{(k_t)}$ converges weakly to $\mathbb{P}^{(k_*)}$, and f is continuous, $\{\theta : f(\theta,k_*) > 0\}$ is an open set whose boundary has measure 0 under $\mathbb{P}^{(k_*)}$. This implies $\lim_t |\mathbb{P}^{(k_t)}\{f(\theta,k_*) > u\} - \mathbb{P}^{(k_*)}\{f(\theta,k_*) > u\}| = 0$ via the Portmanteau theorem (Billingsley 1999).

Now consider $\lim_t \left|\mathbb{P}^{(k_t)}\{f(\theta,k_t) > u\} - \mathbb{P}^{(k_t)}\{f(\theta,k_*) > u\}\right|$. Choose any $\epsilon > 0$ and let $\delta > 0$ be small enough that $\mathbb{P}^{(k_*)}\{|f(\theta,k_*) - u| < \delta\} < \epsilon$ and $\mathbb{P}^{(k_*)}\{|f(\theta,k_*) - u| = \delta\} = 0$.

Since f is uniformly continuous, there exists a $\delta' > 0$ such that, for all $k \in \widetilde{\mathcal{K}}$ satisfying $||k - k_*|| < \delta'$, we have $|f(\theta,k) - f(\theta,k_*)| < \delta$. Here, $||\cdot||$ is the Euclidean norm. Then, since $\lim_t k_t = k_*$, there exists a T such that $||k_t - k_*|| < \delta'$ for all $t > T$. We then have for all $t > T$ that whenever θ satisfies $f(\theta,k_*) > u + \delta$ it also satisfies $f(\theta,k_t) > u$, and whenever θ satisfies $f(\theta,k_*) < u - \delta$ it also satisfies $f(\theta,k_t) < u$. This implies that

$$\lim_t |\mathbb{P}^{(k_t)}\{f(\theta,k_t) > u\} - \mathbb{P}^{(k_t)}\{f(\theta,k_*) > u\}| \leq \lim_t \mathbb{P}^{(k_t)}\{|f(\theta,k_*) - u| < \delta\}.$$

Since $\mathbb{P}^{(k_*)}\{|f(\theta,k_*) - u| = \delta\} = 0$, the limit $\lim_t \mathbb{P}^{(k_t)}\{|f(\theta,k_*) - u| < \delta\}$ is equal to $\mathbb{P}^{(k_*)}\{|f(\theta,k_*) - u| < \delta\}$, again by the Portmanteau theorem, and this is strictly less than ϵ by assumption. Thus we have shown for every $\epsilon > 0$ that $\lim_t |\mathbb{P}^{(k_t)}\{f(\theta,k_t) > u\} - \mathbb{P}^{(k_t)}\{f(\theta,k_*) > u\}| < \epsilon$, and so this limit must equal 0.

We have shown that $\lim_t \mathbb{P}^{(k_t)}\{f(\theta,k_t) > u\} = \mathbb{P}^{(k_*)}\{f(\theta,k_*) > u\}$ for all but countably many u. It can be shown similarly, for all but countably many u, that

$\lim_t \mathbb{P}^{(k_t)}\{f(\theta,k_t) < -u\} = \mathbb{P}^{(k_*)}\{f(\theta,k_*) < -u\}$. Thus, by (C.20),

$$\begin{aligned}\lim_t \mathbb{E}^{(k_t)}[f(\theta,k_t)] &= \int_{\mathbb{R}_+} \mathbb{P}^{(k_*)}\{f(\theta,k_*) > u\} - \mathbb{P}^{(k_*)}\{f(\theta,k_*) < -u\}\ du \\ &= \mathbb{E}^{(k_*)}[f(\theta,k_*)].\end{aligned}$$

This shows the continuity of (5.6).

Proof of Lemma 5.3.2

First suppose $\mathcal{I}_C \subseteq \mathcal{I}(k)$. Then, the fact $\max_{i\in\mathcal{I}_C}\mu_i = \max_{i\in\mathcal{I}_C}\hat{\mu}_i(k)$ almost surely under $\mathbb{P}^{(k)}$, together with (5.6) from Lemma 5.3.1 implies that $g(k;C) = 0$.

Now suppose $\mathcal{I}_C \setminus \mathcal{I}(k)$ is nonempty, and let i' be one of its elements. Then

$$g(k;C) \geq \mathbb{E}^{(k)}\left[\max\left(\mu_{i'}, \max_{i\neq i'}\hat{\mu}_i(k)\right)\right] - \max_{i\in\mathcal{I}}\hat{\mu}_i(k), \tag{C.21}$$

where we use the tower property on (5.6) to introduce an inner expectation conditioned on $\mu_{i'}$, then exchange it with the maximums using Jensen's inequality, and remove it using the tower property. We now consider two cases. If $\hat{\mu}_{i'}(k) \leq \max_{i\neq i'}\hat{\mu}_i(k)$, then $\max_{i\neq i'}\hat{\mu}_i(k) = \max_{i\in\mathcal{I}}\hat{\mu}_i(k)$ and we have through (C.21) that

$$g(k;C) \geq \mathbb{P}^{(k)}\left\{\mu_{i'} > \max_{i\neq i'}\hat{\mu}_i(k) + 1\right\} + \max_{i\neq i'}\hat{\mu}_i(k) - \max_{i\in\mathcal{I}}\hat{\mu}_i(k) > 0.$$

If $\hat{\mu}_{i'}(k) > \max_{i\neq i'}\hat{\mu}_i(k)$, then $\hat{\mu}_{i'}(k) = \max_{i\in\mathcal{I}}\hat{\mu}_i(k)$ and we add and subtract $\mu_{i'}$ to the term in the expectation in (C.21) to obtain

$$\begin{aligned}g(k;C) &\geq \mathbb{E}^{(k)}\left[\mu_{i'} + \max\left(0, -\mu_{i'} + \max_{i\neq i'}\hat{\mu}_i(k)\right)\right] - \max_{i\in\mathcal{I}}\hat{\mu}_i(k), \\ &= \mathbb{E}^{(k)}\left[\max\left(0, -\mu_{i'} + \max_{i\neq i'}\hat{\mu}_i(k)\right)\right] \geq \mathbb{P}^{(k)}\left\{-\mu_{i'} + \max_{i\neq i'}\hat{\mu}_i(k) > 1\right\} > 0.\end{aligned}$$

Proof of Lemma 5.4.1

We begin by proving the expressions for $\mathrm{dom}(g)$. The expression for $\mathrm{cl}(\mathrm{dom}(g))$ follows from taking the closure in $\mathrm{cl}(\mathcal{K})$ of the expression for $\mathrm{dom}(g)$.

For $k \in \mathcal{K}$, the marginal distribution of $(\rho_i(k)\hat{\lambda}_i(k))^{1/2}(\mu_i - \hat{\mu}_i(k))$ under $\mathbb{P}^{(k)}$ is student-t with $2a_i(k)$ degrees of freedom. From the fact that a student-t distributed random variable is integrable iff it has strictly more than 1 degree of freedom, we have that $\mathbb{E}^{(k)}[|\mu_i|] < \infty$

iff $2a_i(k) > 1$. To show $\mathrm{dom}(g) = \{k \in \mathcal{K} : a_i(k) > 1/2 \ \forall i \in \mathcal{I}\}$ it is then sufficient to show $\mathbb{E}^{(k)}[R(\theta; i)] < \infty$ for all i iff $\mathbb{E}^{(k)}[|\mu_i|] < \infty$ for all i.

Treating one direction, if $\mathbb{E}^{(k)}[|\mu_i|] < \infty$ for all i, then $\infty > \sum_i \mathbb{E}^{(k)}[\sum_i |\mu_i|] \geq \mathbb{E}^{(k)}[\max_i \mu_i - \mu_j] = \mathbb{E}^{(k)}[R(\theta; j)]$. For the converse direction, letting $x^+ = \max(0, x)$ denote the positive-part of x, we have

$$\infty > \mathbb{E}^{(k)}[R(\theta; i)] = \mathbb{E}^{(k)}\left[(\max_{j \neq i} \mu_j - \mu_i)^+\right] \geq \mathbb{E}^{(k)}\left[\mu_j^+ \mathbf{1}_{\{\mu_i \leq 0\}}\right],$$

and $\infty > \mathbb{E}^{(k)}\left[\mu_j^+ \mathbf{1}_{\{\mu_i \leq 0\}}\right] = \mathbb{E}^{(k)}\left[\mu_j^+\right] \mathbb{P}^{(k)}\{\mu_i \leq 0\}$ implies that $\mathbb{E}^{(k)}\left[\mu_j^+\right]$ is finite, which implies in turn that $\mathbb{E}^{(k)}[|\mu_j|]$ is finite. This shows the statement about $\mathrm{dom}(g)$.

We now move to the statement about continuity of g. For $k \in \mathrm{dom}(g)$ and $C \subseteq \mathcal{X}$, we have by an argument identical to the one used in the proof of Lemma 5.3.1 that

$$g(k; C) = \mathbb{E}^{(k)}\left[\max\left(\max_{i \in \mathcal{I}_C} \mu_i, \max_{i \notin \mathcal{I}_C} \hat{\mu}_i(k)\right)\right] - \max_{i \in \mathcal{I}} \hat{\mu}_i(k).$$

This expression agrees with (5.7) for $k \in \mathrm{dom}(g)$, and so it only remains to show that (5.7) is continuous on $\mathrm{cl}(\mathrm{dom}(g))$.

Define a set $\widetilde{\mathcal{K}} = \{k \in \mathrm{cl}(\mathrm{dom}(g)) : a_i(k) > 1/2 \ \forall i \in \mathcal{I}_C\}$, and a function $f(\theta, k) = \max\left(\max_{i \in \mathcal{I}_C} \mu_i, \max_{i \notin \mathcal{I}_C} \hat{\mu}_i(k)\right)$. Since $k \mapsto \hat{\mu}_i(k)$ is continuous and finite on $\mathrm{cl}(\mathrm{dom}(g))$, to show continuity of (5.7) it is sufficient to show continuity of a function $h : \mathrm{cl}(\mathrm{dom}(g)) \mapsto \mathbb{R}_+ \cup \{\infty\}$ defined by

$$h(k) = \begin{cases} \mathbb{E}^{(k)}[f(\theta, k)] & \text{if } k \in \widetilde{\mathcal{K}}, \\ +\infty & \text{if } k \notin \widetilde{\mathcal{K}}. \end{cases}$$

To this end, let $(k_t) \subseteq \mathrm{cl}(\mathrm{dom}(g))$ be a sequence converging to $k_* \in \mathrm{cl}(\mathrm{dom}(g))$. We show $\lim_t h(k_t) = h(k_*)$ by considering two cases.

In the first case, suppose $k_* \notin \widetilde{\mathcal{K}}$. Since $h(k_t) = \infty$ when $k_t \notin \widetilde{\mathcal{K}}$, we may assume without loss of generality that $(k_t) \subseteq \widetilde{\mathcal{K}}$ and show $\lim_t h(k_t) = \infty$.

Since $k_* \notin \widetilde{\mathcal{K}}$, there is some $i' \in \mathcal{I}_C$ such that $a_{i'}(k_*) = 1/2$. Then,

$$\mathbb{E}^{(k_t)}[f(\theta, k)] = \mathbb{E}^{(k_t)}\left[\max(\max_{i \in \mathcal{I}_C} \mu_i, \max_{i \notin \mathcal{I}_C} \hat{\mu}_i(k))\right] \geq \mathbb{E}^{(k_t)}\left[\max(\mu_{i'}, \max_{i \neq i'} \hat{\mu}_i(k))\right],$$

where we use Jensen's inequality and the identity $\mathbb{E}^{(k_t)}[\mu_i \mid \mu_{i'}] = \hat{\mu}_i(k_t)$ for $i \neq i'$. Let $u = \sup_t \max(\max_{i \neq i'} \hat{\mu}_i(k_t), 0)$, and we have

$$\begin{aligned}\mathbb{E}^{(k_t)}[f(\theta,k)] &\geq \mathbb{E}^{(k_t)}\left[\mu_{i'}\mathbf{1}_{\{\mu_{i'}>u\}}\right] + \max_{i\neq i'}\hat{\mu}_i(k_t)\mathbb{P}^{(k_t)}\{\mu_{i'} < u\} \\ &\geq \mathbb{E}^{(k_t)}\left[\mu_{i'}\mathbf{1}_{\{\mu_{i'}>u\}}\right] + \min(\max_{i\neq i'}\hat{\mu}_i(k_t), 0).\end{aligned}$$

Finally, we have $\lim_t \left[\mu_{i'}\mathbf{1}_{\{\mu_{i'}>u\}}\right] = \infty$ because the marginal distribution under $\mathbb{P}^{(k_t)}$ of $(\rho_{i'}(k_t)\hat{\lambda}_{i'}(k_t))^{1/2}(\mu_{i'} - \hat{\mu}_{i'}(k_t))$ is student-t with $2a_{i'}(k_t)$ degrees of freedom, which implies that $\mathbb{E}^{(k_t)}\left[\mu_{i'}\mathbf{1}_{\{\mu_{i'}>u\}}\right]$ goes to infinity as $2a_{i'}(k_t)$ goes to 1.

In the second case, suppose $k_* \in \widetilde{\mathcal{K}}$. Then, for $i \in \mathcal{I}_C$, $\liminf_t a_i(k_t) = a_i(k_*) > 1/2$, allowing for ∞ as a possible value. Since we may always take a T large enough that the tail sequence $(k_t)_{t>T}$ is contained entirely within $\widetilde{\mathcal{K}}$, we may assume without loss of generality that $(k_t) \in \widetilde{\mathcal{K}}$ for all t. Thus $h(k_t) = \mathbb{E}^{(k_t)}[f(\theta,k)]$. To show $\lim_t \mathbb{E}^{(k_t)}[f(\theta,k)] = \mathbb{E}^{(k_*)}[f(\theta,k)]$, we use Lebesgue's dominated convergence theorem together with the identity

$$\mathbb{E}^{(k)}[f(\theta,k)\} = \int_{\mathbb{R}_+} \mathbb{P}^{(k)}\{f(\theta,k) > u\} - \mathbb{P}^{(k)}\{f(\theta,k) < -u\}\ du. \tag{C.22}$$

The integrand in (C.22) is dominated by

$$\begin{aligned}\mathbb{P}^{(k)}\{|f(\theta,k)| > u\} &\leq \mathbb{P}^{(k)}\left\{\max\left(\max_{i\in\mathcal{I}_C}|\mu_i|, \max_{i\notin\mathcal{I}_C}|\hat{\mu}_i(k)|\right) > u\right\} \\ &= 1 - \left[\prod_{i\in\mathcal{I}_C}\mathbb{P}^{(k)}\{|\mu_i| \leq u\}\right]\left[\prod_{i\notin\mathcal{I}_C}\mathbf{1}_{\{|\hat{\mu}_i(k)|\leq u\}}\right] \\ &\leq \begin{cases}1, & \text{if } \max_i|\hat{\mu}_i(k)| \geq u, \\ 1 - \prod_{i\in\mathcal{I}_C}\mathbb{P}^{(k)}\{|\mu_i| \leq u\}, & \text{otherwise.}\end{cases}\end{aligned}$$

Construct $\tilde{k}$ so that $\hat{\mu}_i(\tilde{k}) = \inf_t|\hat{\mu}_i(k_t)|$, $\hat{\lambda}_i(\tilde{k}) = \inf_t \hat{\lambda}_i(k_t)$, $a_i(\tilde{k}) = \inf_t a_i(k_t)$. Note that $\tilde{k} \in \widetilde{\mathcal{K}}$, and that $a_i(\tilde{k}) = \inf_t a_i(k_t)$ implies $\rho_i(\tilde{k}) = \inf_t \rho_i(k_t)$. For $u > |\hat{\mu}_i(k)|$, $\mathbb{P}^{(k)}\{|\mu_i| \leq u\}$ is decreasing in each of $|\hat{\mu}_i(k)|$, $\hat{\lambda}_i(k)$, $a_i(k)$, and $\rho_i(k)$. Thus, for all $u \geq \bar{u} := \max_{i\in\mathcal{I}_C}\sup_t|\hat{\mu}_i(k_t)|$, we have

$$1 - \prod_{i\in\mathcal{I}_C}\mathbb{P}^{(k_t)}\{|\mu_i| \leq u\} \leq 1 - \prod_{i\in\mathcal{I}_C}\mathbb{P}^{(\tilde{k})}\{|\mu_i| \leq u\} = \mathbb{P}^{(\tilde{k})}\left\{\max_{i\in\mathcal{I}_C}|\mu_i| > u\right\}.$$

Thus, the integrand in (C.22) is dominated by $\mathbf{1}_{\{u<\bar{u}\}} + \mathbf{1}_{\{u\geq\bar{u}\}}\mathbb{P}^{(\tilde{k})}\{\max_{i\in\mathcal{I}_C}|\mu_i| > u\}$,

a function whose integral over $\mathbb{R}_+$ is given by

$$\bar{u} + \int_{[\bar{u},\infty)} \mathbb{P}^{(\tilde{k})} \left\{ \max_{i \in \mathcal{I}_C} |\mu_i| > u \right\} \, du \leq \bar{u} + \mathbb{E}^{(\tilde{k})} \left[\max_{i \in \mathcal{I}_C} |\mu_i| \right] \leq \bar{u} + \sum_{i \in \mathcal{I}_C} \mathbb{E}^{(\tilde{k})} \left[|\mu_i| \right],$$

which is finite because $\bar{k} \in \widetilde{\mathcal{K}}$. Then, by the dominated convergence theorem,

$$\lim_t \mathbb{E}^{(k_t)} \left[f(\theta, k_t) \right] = \int_{\mathbb{R}_+} \lim_t \mathbb{P}^{(k_t)} \left\{ f(\theta, k_t) > u \right\} - \mathbb{P}^{(k_t)} \left\{ f(\theta, k_t) < -u \right\} \, du,$$

provided that the limit of the integrand exists for all but at most countably many u.

Note that there at most countably many $u \in \mathbb{R}$ with $\mathbb{P}^{(k_*)} \left\{ |f(\theta, k_*)| = u \right\} > 0$, and choose any other u. Since f is uniformly continuous and $\mathbb{P}^{(k_t)}$ converges weakly to $\mathbb{P}^{(k_*)}$, an argument identical to the one used in the proof of Lemma 5.3.1 shows that $\lim_t \mathbb{P}^{(k_t)} \left\{ f(\theta, k_t) > u \right\} = \mathbb{P}^{(k_*)} \left\{ f(\theta, k_*) > u \right\}$ and $\lim_t \mathbb{P}^{(k_t)} \left\{ f(\theta, k_t) < -u \right\} = \mathbb{P}^{(k_*)} \left\{ f(\theta, k_*) < -u \right\}$. This shows the continuity of (5.7).

Proof of Lemma 5.4.2

The proof of this lemma is quite similar to the proof for Lemma 5.3.2, which is for the variance-known version of the R&S problem. First suppose $\mathcal{I}_C \subseteq \mathcal{I}(k)$. Then we have $\max_{i \in \mathcal{I}_C} \mu_i = \max_{i \in \mathcal{I}_C} \hat{\mu}_i(k)$ almost surely under $\mathbb{P}^{(k)}$. This with (5.7) from Lemma 5.4.1 implies that $g(k; C) = 0$.

Now suppose $\mathcal{I}_C \setminus \mathcal{I}(k)$ is nonempty, and let i' be one of its elements. If $a_i(k) = 1/2$ for some $i \in \mathcal{I}_C$, then $g(k; C) = \infty \neq 0$. If not, we proceed using an argument identical to the one used in the proof of Lemma 5.3.2. The details may be found in that proof. If $\hat{\mu}_{i'}(k) \leq \max_{i \neq i'} \hat{\mu}_i(k)$, we have $g(k; C) \geq \mathbb{P}^{(k)} \left\{ \mu_{i'} > \max_{i \neq i'} \hat{\mu}_i(k) + 1 \right\} > 0$, and if $\hat{\mu}_{i'}(k) > \max_{i \neq i'} \hat{\mu}_i(k)$, we have $g(k; C) \geq \mathbb{P}^{(k)} \left\{ -\mu_{i'} + \max_{i \neq i'} \hat{\mu}_i(k) > 1 \right\} > 0$.

Proof of Lemma 5.5.1

Let $\mathbb{P}^{[x]}$ denote the measure under which the sampling policy used by $\mathbb{P}$ is replaced with one that measures x repeatedly. This sampling policy is defined by $\Pi(t, k, A) = \mathbf{1}_{\{1 \in A\}}$ for $t \in \mathbb{N}$,$k \in \mathcal{K}$, and $A \subseteq \mathcal{X}$. Let $\mathbb{E}^{[x]}$ denote the expectation under $\mathbb{P}^{[x]}$. We then write $g(k; \{x\})$ as the expected improvement observed after measuring x an infinite number of

times. We fix any time t and write

$$g(k;\{x\}) = \min_i \mathbb{E}^{(k)}\left[R(\theta;i)\right] - \mathbb{E}^{[x]}\left[\min_i \mathbb{E}_{\infty}^{[x]}\left[R(\theta;i)\right] \mid K_t = k\right].$$

The second term on the right-hand side may be rewritten using the tower property of conditional expectation as $\mathbb{E}^{[x]}\left[\mathbb{E}_{t+1}^{[x]}\left[\min_i \mathbb{E}_{\infty}^{[x]}\left[R(\theta;i)\right]\right] \mid K_t = k\right]$. Using Jensen's inequality to switch the $\mathbb{E}_{t+1}^{[x]}$ and the minimum immediately after it and then applying the tower property a second time, this is bounded above by $\mathbb{E}^{[x]}\left[\min_i \mathbb{E}_{t+1}^{[x]}\left[R(\theta;i)\right] \mid K_t = k\right]$. Thus we have

$$g(k,\{x\}) \geq \min_i \mathbb{E}^{(k)}\left[R(\theta;i)\right] - \mathbb{E}^{[x]}\left[\min_i \mathbb{E}_{t+1}^{[x]}\left[R(\theta;i)\right] \mid K_t = k\right] = h(k,x).$$

To see that $h(k,x)$ is non-negative, we use Jensen's inequality and the tower property to note

$$\begin{aligned}\mathbb{E}^{[x]}\left[\min_i \mathbb{E}_{t+1}^{[x]}\left[R(\theta;i)\right] \mid K_t = k\right] &\leq \min_i \mathbb{E}^{[x]}\left[\mathbb{E}_{t+1}^{[x]}\left[R(\theta;i)\right] \mid K_t = k\right] \\ &= \min_i \mathbb{E}^{[x]}\left[R(\theta;i) \mid K_t = k\right] = \min_i \mathbb{E}^{(k)}\left[R(\theta;i)\right].\end{aligned}$$

References

Anderson Jr, E. & Parker, G. (2002), 'The Effect of Learning on the Make/Buy Decision', *Production and Operations Management* **11**(3), 313–339.

Auer, P., Cesa-Bianchi, N., Freund, Y. & Schapire, R. (1995), Gambling in a rigged casino: The adversarial multi-armed banditproblem, *in* 'Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on', pp. 322–331.

Babb, J., Rogatko, A. & Zacks, S. (1998), 'Cancer phase i clinical trials: efficient dose escalation with overdose control.', *Stat Med* **17**(10), 1103–1120.

Bechhofer, R. (1954), 'A single-sample multiple decision procedure for ranking means of normal populations with known variances', *The Annals of Mathematical Statistics* **25**(1), 16–39.

Bechhofer, R., Kiefer, J. & Sobel, M. (1968), *Sequential Identification and Ranking Procedures*, University of Chicago Press, Chicago.

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

Bellman, R. (1954), 'The theory of dynamic programming', *Bull. Amer. Math. Soc.* **60**, 503–516.

Berger, J. & Deely, J. (1988), 'A Bayesian Approach to Ranking and Selection of Related Means With Alternatives to Analysis-of-Variance Methodology', *Journal of the American Statistical Association* **83**(402), 364–373.

Berry, D. & Fristedt, B. (1985), *Bandit Problems: Sequential Allocation of Experiments*, Routledge.

Bickel, J. & Smith, J. (2006), 'Optimal sequential exploration: A binary learning model', *Decision Analysis* **14**(15), 16.

Bickel, P. & Doksum, K. (2007), *Mathematical statistics: basic ideas and selected topics*, 2nd ed. updated printing edn, Prentice Hall.

Billingsley, P. (1999), *Convergence of probability measures*, Wiley Series in Probability and Statistics: Probability and Statistics, second edn, John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Box, G., Hunter, W. & Hunter, J. (1978), *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, John Wiley & Sons, New York.

Branin, F. (1972), 'Widely convergent method for finding multiple solutions of simultaneous nonlinear equations', *IBM J. Res. Develop* **16**(5), 504–522.

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

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

Brezzi, M. & Lai, T. (2002), 'Optimal learning and experimentation in bandit problems', *Journal of Economic Dynamics and Control* **27**(1), 87–108.

Chang, H., Fu, M., Hu, J. & Marcus, S. (2007), *Simulation-Based Algorithms for Markov Decision Processes*, Springer, Berlin.

Chen, C. (1995), 'An effective approach to smartly allocate computing budget for discrete event simulation', *IEEE Conference on Decision and Control, 34 th, New Orleans, LA* pp. 2598–2603.

Chen, C., Dai, L. & Chen, H. (1996), 'A gradient approach for smartly allocating computing budget for discrete event simulation', *Simulation Conference Proceedings, 1996. Winter* pp. 398–405.

Chen, C., Donohue, K., Yücesan, E. & Lin, J. (2003), 'Optimal computing budget allocation for monte carlo simulation with application to product design', *Simulation Modelling Practice and Theory* **11**(1), 57–74.

Chen, C., He, D. & Fu, M. (2006), 'Efficient Dynamic Simulation Allocation in Ordinal Optimization', *IEEE Transactions Automatic Control* **51**(12), 2005–2009.

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

Chen, H., Chen, C. & Yucesan, E. (2000), 'Computing efforts allocation for ordinal optimization and discrete event simulation', *Automatic Control, IEEE Transactions on* **45**(5), 960–964.

Chen, H., Dai, L., Chen, C. & Yücesan, E. (1997), 'New development of optimal computing budget allocation for discrete event simulation', *Proceedings of the 29th conference on Winter simulation-Volume 00* pp. 334–341.

Chick, S. (2000), 'Bayesian methods: bayesian methods for simulation', *Proceedings of the 32nd conference on Winter simulation* pp. 109–118.

Chick, S., Branke, J. & Schmidt, C. (2007), New greedy myopic and existing asymptotic sequential selection procedures: preliminary empirical results, *in* 'Proceedings of the 2007 Winter Simulation Conference', Winter Simulation Conference, IEEE, Piscataway, NJ.

Chick, S., Branke, J. & Schmidt, C. (2009), Sequential sampling to myopically maximize the expected value of information. Submitted.

Chick, S. & Gans, N. (2008), Economic analysis of simulation selection problems. Submitted, 2008.

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

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

Chudnovsky, G. (1988), *Search theory: some recent developments*, CRC Press.

Clark, C. (1961), 'The greatest of a finite set of random variables', *Operations Research* **9**, 145–163.

Cohn, D. A., Ghahramani, Z. & Jordan, M. I. (1996), 'Active learning with statistical models', *CoRR* .

Cressie, N. (1993), *Statistics for Spatial Data, revised edition*, Wiley Interscience.

Currin, C., Mitchell, T., Morris, M. & Ylvisaker, D. (1991), 'Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments', *Journal of the American Statistical Association* **86**, 953–963.

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

Diaconis, P. & Freedman, D. (1986), 'On the Consistency of Bayes Estimates', *The Annals of Statistics* **14**(1), 1–26.

Doob, J. L. (1949), Application of the theory of martingales, *in* 'Le Calcul des Probabilités et ses Applications.', Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, Centre National de la Recherche Scientifique, Paris, pp. 23–27.

Eichhorn, B. H. & Zacks, S. (1973), 'Sequential search of an optimal dosage. I', *J. Amer. Statist. Assoc.* **68**, 594–598.

Frazier, P. & Powell, W. (2008*a*), The knowledge-gradient stopping rule for ranking and selection, *in* 'Proceedings of the 2008 Winter Simulation Conference'.

Frazier, P. & Powell, W. (2009*a*), Simulation model calibration with correlated knowledge-gradients. in review.

Frazier, P. & Powell, W. B. (2008*b*), Asymptotic optimality of sequential sampling policies for bayesian information collection. submitted.

Frazier, P. & Powell, W. B. (2009*b*), Properties of the value of information. in preparation.

Frazier, P., Powell, W. B. & Dayanik, S. (2008*a*), 'The knowledge-gradient policy for correlated normal rewards', *INFORMS Journal on Computing* . to appear.

Frazier, P., Powell, W. B. & Dayanik, S. (2008*b*), 'A knowledge-gradient policy for sequential information collection', *SIAM Journal on Control and Optimization* **47**(5).

Frazier, P., Powell, W., Dayanik, S. & Kantor, P. (2009), Approximate dynamic programming in knowledge discovery for rapid response, *in* 'Hawaii International Conference on Systems Science'. to appear.

Freedman, D. (1963), 'On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case', *The Annals of Mathematical Statistics* **34**(4), 1386–1403.

Fu, M. (2002), 'Optimization for simulation: Theory vs. practice', *INFORMS Journal on Computing* **14**(3), 192–215.

Fu, M. C., Hu, J.-Q., Chen, C.-H. & Xiong, X. (2007), 'Simulation allocation for determining the best design in the presence of correlated sampling', *INFORMS J. on Computing* **19**(1), 101–111.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. (2004), *Bayesian data analysis*, second edn, CRC Press.

Ghosh, J. & Ramamoorthi, R. (2003), *Bayesian Nonparametrics*, Springer.

Gittins, J. (1989), *Multi-Armed Bandit Allocation Indices*, John Wiley and Sons, New York.

Gittins, J. C. & Jones, D. M. (1974), A dynamic allocation index for the sequential design of experiments, *in* J. Gani, ed., 'Progress in Statistics', pp. 241–266.

Golub, G. & Van Loan, C. (1996), *Matrix Computations*, John Hopkins University Press, Baltimore, MD.

Gordon, R. (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. & Miescke, K. (1994), 'Bayesian look ahead one stage sampling allocations for selecting the largest normal mean', *Statistical Papers* **35**, 169–177.

Gupta, S. & Miescke, K. (1996), 'Bayesian look ahead one-stage sampling allocations for selection of the best population', *Journal of statistical planning and inference* **54**(2), 229–244.

Hannah, L., Powell, W. B. & Stewart, J. (2009), One-stage r&d portfolio optimization with an application to solid oxide fuel cells. in review.

Hartman, J. (1973), 'Some experiments in global optimization', *Naval Reserach Logistics Quarterly* **20**, 569–576.

Hartmann (1991), 'An improvement on paulson's procedure for selecting the population with the largest mean from k normal populations with a common unknown variance', *Sequential Analysis* **10**(1-2), 1–16.

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

Howard, R. (1966), 'Information Value Theory', *Systems Science and Cybernetics, IEEE Transactions on* **2**(1), 22–26.

Huang, D., Allen, T., Notz, W. & Zeng, N. (2006), 'Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models', *Journal of Global Optimization* **34**(3), 441–466.

Inoue, K., Chick, S. & Chen, C. (1999), 'An empirical evaluation of several methods to select the best system', *ACM Transactions on Modeling and Computer Simulation (TOMACS)* **9**(4), 381–407.

Jennison, C. & Turnbull, B. (2000), *Group sequential methods with applications to clinical trials*, CRC Press.

Jones, D. (2001), 'A Taxonomy of Global Optimization Methods Based on Response Surfaces', *Journal of Global Optimization* **21**(4), 345–383.

Jones, D., Schonlau, M. & Welch, W. (1998), 'Efficient Global Optimization of Expensive Black-Box Functions', *Journal of Global Optimization* **13**(4), 455–492.

Kaelbling, L. (1993), *Learning in embedded systems*, MIT Press, Cambridge, MA.

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

Kapoor, A. & Greiner, R. (2005), Learning and Classifying Under Hard Budgets, *in* '16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005', Springer-Verlag New York Inc.

Kennedy, M. & O'Hagan, A. (2001), 'Bayesian calibration of computer models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(3), 425–464.

Kiefer, J. & Wolfowitz, J. (1952), 'Stochastic estimation of the maximum of a regression function', *Annals Mathematical Statistics* **23**, 462–466.

Kim, S.-H. & Nelson, B. L. (2001), 'A fully sequential procedure for indifference-zone selection in simulation', *ACM Trans. Model. Comput. Simul.* **11**(3), 251–273.

Kim, S. & Nelson, B. (2006*a*), *Handbook in Operations Research and Management Science: Simulation*, Elsevier, chapter Selecting the best system.

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

Kleijnen, J. (2009), 'Kriging metamodeling in simulation: A review', *European Journal of Operational Research* **192**(3), 707–716.

Kushner, H. J. (1964), 'A new method of locating the maximum of an arbitrary multi- peak curve in the presence of noise', *Journal of Basic Engineering* **86**, 97–106.

Lai, T. (1987), 'Adaptive treatment allocation and the multi-armed bandit problem', *The Annals of Statistics* **15**(3), 1091–1114.

Lai, T. (2001), 'Sequential analysis: some classical problems and new challenges', *Statistica Sinica* **11**(2), 303–408.

Lariviere, M. & Porteus, E. (1999), 'Stalking Information: Bayesian Inventory Management with Unobserved Lost Sales', *Management Science* **45**(3), 346–363.

Law, A. M. & Kelton, W. D. (2000), *Simulation Modeling and Analysis*, third edn, McGraw-Hill, New York.

Locatelli, M. (1997), 'Bayesian Algorithms for One-Dimensional Global Optimization', *Journal of Global Optimization* **10**(1), 57–76.

Madigan, D., Mittal, S. & Roberts, F. (2007), Sequential decision making algorithms for port of entry inspection: Overcoming computational challenges, *in* 'IEEE International Conference on Intelligence and Security Informatics (ISI-2007)', pp. 1–7.

Marschak, J. & Radner, R. (1972), *Economic Theory of Teams*, Yale University Press, New Haven.

Mockus, J. (1972), 'On bayesian methods for seeking the extremum', *Automatics and Computers* (3), 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', *Journal of Global Optimization* **4**(4), 347–365.

Mockus, J., Tiesis, V. & Zilinskas, A. (1978), The application of Bayesian methods for seeking the extremum, *in* L. Dixon & G. Szego, eds, 'Towards Global Optimisation', Vol. 2, North Holland, Amsterdam, pp. 117–129.

Myers, R. & Montgomery, D. (2002), *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, John Wiley & Sons, New York.

Negoescu, D., Frazier, P. & Powell, W. B. (2008), Optimal learning policies for the newsvendor problem with censored demand and unobservable lost sales. in preparation.

Negoescu, D., Frazier, P. & Powell, W. B. (2009), Optimal learning for drug discovery in ewing's sarcoma. Princeton University Senior Thesis, in preparation.

Nelson, B. & Staum, J. (2006), 'Control variates for screening, selection, and estimation of the best', *ACM Transactions on Modeling and Computer Simulation (TOMACS)* **16**(1), 52–75.

Nelson, B., Swann, J., Goldsman, D. & Song, W. (2001), 'Simple procedures for selecting the best simulated system when the number of alternatives is large', *Operations Research* **49**(6), 950–963.

Nino-Mora, J. (2001), 'Restless bandits, partial conservation laws and indexability', *Advances in Applied Probability* **33**(1), 76–98.

O'Hagan, A. & Kingman, J. (1978), 'Curve fitting and optimal design for prediction', *Journal of the Royal Statistical Society. Series B (Methodological)* **40**, 1–42.

Papageorgiou, E. & Sircar, R. (2009), 'Multiscale intensity models and name grouping for valuation of multi-name credit derivatives', *Applied Mathematical Finance* . to appear.

Paulson, E. (1964), 'A sequential procedure for selecting the population with the largest mean from k normal populations', *The Annals of Mathematical Statistics* **35**(1), 174–180.

Paulson, E. (1994), 'Sequential procedures for selecting the best one of k koopman-darmois populations', *Sequential Analysis* **13**(3).

Powell, W. B. (2007), *Approximate Dynamic Programming: Solving the curses of dimensionality*, John Wiley and Sons, New York.

Preparata, F. & Shamos, M. (1985), *Computational Geometry: An Introduction*, Springer.

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

Renninger, L., Verghese, P. & Coughlan, J. (2007), 'Where to look next? eye movements reduce local uncertainty', *Journal of Vision* **7**, 1–17.

Rimey, R. & Brown, C. (1994), 'Control of selective perception using bayes nets and decision theory', *International Journal of Computer Vision* **12**(2), 173–207.

Rinott, Y. (1978), 'On two-stage selection procedures and related probability-inequalities', *Communications in Statistics-Theory and Methods* **7**(8), 799–811.

Ryzhov, I. & Powell, W. B. (2009), Information collection on a graph. in review.

Ryzhov, I., Powell, W. B. & Frazier, P. (2008*a*), The knowledge-gradient algorithm for a general class of online learning problems. in review.

Ryzhov, I., Powell, W. & Frazier, P. (2008*b*), 'The knowledge gradient algorithm for a general class of online learning problems'. submitted.

Sacks, J., Welch, W., Mitchell, T. & Wynn, H. (1989), 'Design and analysis of computer experiments', *Statistical Sci.* **4**, 409–423.

Sasena, M. (2002), Flexibility and Efficiency Enhancements for Constrained Global Design Optimization with Kriging Approximations, PhD thesis, University of Michigan.

Schmegner, C. & Baron, M. (2004), 'Principles of optimal sequential planning', *Sequential Analysis* **23**(1), 11–32.

Schmitz, N. (1993), *Optimal sequentially planned decision procedures*, Springer-Verlag Berlin.

Schonlau, M. & Welch, W. (1996), Global optimization with nonparametric function fitting, *in* 'Proceedings of the Section on Physical and Engineering Sciences', pp. 183–186.

Siegmund, D. (1985), *Sequential analysis: tests and confidence intervals*, Springer Verlag.

Singh, S., Jaakkola, T., Littman, M. & Szepesvari, C. (2000), 'Convergence results for single-step on-policy reinforcement-learning algorithms', *Machine Learning* **39**, 287–308.

Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*, John Wiley & Sons, Hoboken, NJ.

Stone, L. (1975), *Theory of Optimal Search*, Academic Press.

Stuckman, B. (1988), 'A global search method for optimizing nonlinear systems', *Systems, Man and Cybernetics, IEEE Transactions on* **18**(6), 965–977.

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

Tomberlin, D. (2008), An approach to managing fisheries when weak and strong stocks mix, *in* 'Proceedings of 2008 International Institute of Fisheries Economics and Trade'.

van Beers, W. & Kleijnen, J. (2008), 'Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping', *European Journal of Operational Research* **186**(3), 1099–1113.

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

Walker, S. (2004), 'New approaches to Bayesian consistency', *The Annals of Statistics* **32**, 2028–2043.

Weitzman, M. (1979), 'Optimal search for the best alternative', *Econometrica: Journal of the Econometric Society* pp. 641–654.

Wetherill, G. & Brown, D. (1991), *Statistical process control: theory and practice*, Chapman & Hall.

Wetherill, G. & Glazebrook, K. (1986), *Sequential Methods in Statistics.*, CRC Press.

Whittle, P. (1988), 'Restless bandits: Activity allocation in a changing world', *Journal of applied probability* pp. 287–298.

Williams, B., Santner, T. & Notz, W. (2000), 'Sequential design of computer experiments to minimize integrated response functions', *Statistica Sinica* **10**, 1133–1152.