ABSTRACT

Title of dissertation: SIMULATION OPTIMIZATION: NEW APPROACHES AND AN APPLICATION

Huashuai Qu, Doctor of Philosophy, 2014

Dissertation directed by: Professor Michael C. Fu
Department of Decision, Operations,
and Information Technologies

Simulation models are commonly used to provide analysis and prediction of the behavior of complex stochastic systems. Simulation optimization integrates optimization techniques into simulation analysis to capture response surface, to choose optimal decision variables and to perform sensitivity analysis. Objective functions usually cannot be computed in closed form and are computationally expensive to evaluate. Many methods are proposed by researchers for problems with continuous and discrete variables, respectively. The dissertation is comprised of both optimization methods and a real-world application. In particular, our goal is to develop new methods based on direct gradient estimates and variational Bayesian techniques.

The first part of the thesis considers the setting where additional direct gradient information is available and introduces different approaches for enhancing regression models and stochastic kriging with this additional gradient information, respectively. For regression models, we propose Direct Gradient Augmented Regression (DiGAR) models to incorporate direct gradient estimators. We characterize

the variance of the estimated parameters in DiGAR and compare them analytically with the standard regression model for some special settings. For stochastic kriging, we propose Gradient Extrapolated Stochastic Kriging (GESK) to incorporate direct gradient estimates by extrapolating additional responses. We show that GESK reduces mean squared error (MSE) compared to stochastic kriging under certain conditions on step sizes. We also propose maximizing penalized likelihood and minimizing integrated mean squared error to determine the step sizes.

The second part of the thesis focuses on the problem of learning unknown correlation structures in ranking and selection (R&S) problems. We proposes a computationally tractable Bayesian statistical model for learning unknown correlation structures in fully sequential simulation selection. We derive a Bayesian procedure that allocates simulations based on the value of information, thus anticipating future changes to our beliefs about the correlations. The proposed approach is able to simultaneously learn unknown mean performance values and unknown correlations, whereas existing approaches in the literature assume independence or known correlations to learn unknown mean performance values only.

Finally we consider an application in business-to-business (B2B) pricing. We propose an approximate Bayesian statistical model for predicting the win/loss probability for a given price and an approach for recommending target prices based on the approximate Bayesian model.

SIMULATION OPTIMIZATION: NEW METHODS AND AN APPLICATION

by

Huashuai Qu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor Michael C. Fu, Chair/Advisor
Professor Ilya O. Ryzhov
Professor Paul Smith
Professor Leonid Koralov
Professor Steve Marcus

Dedication

I dedicate this dissertation to my loving and supportive wife, Xuan Liu.

Acknowledgments

First, I would like to express my most sincere gratitude to my advisors Professor Michael C. Fu and Professor Ilya O. Ryzhov, for their vision, advice and patience to help me proceed through my graduate studies. The completion of this dissertation would not have been possible without their consistent guidance and support.

Special thanks to my other dissertation committee members, Professor Paul Smith, Professor Steve Marcus and Professor Leonid Koralov, for taking the time to read the thesis and attend my defense. I also would like to thank Professor Grace Yang for teaching me the course on stochastic process and helping in my job searching process.

I would like to thank everyone within the department of Mathematics in general. I would particularly like to thank: Marie Chau, Zhixin Lu, Ran Ji, Changhui Tan and Zi Ding for their friendship and feedback.

My family has provided their love and untiring support during the process. Most importantly, doing the research and writing the thesis would be impossible without the support and understanding form my wife, Xuan. I thank my parents, Wenbin Qu and Ping Zhao for their encouragement and their belief in the value of learning. Nothing in a simple paragraph can express the love I have for all of you.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Simulation Optimization

We consider optimization problems where the objective is to minimize an expected value that cannot be computed in closed form. Instead, the expectation must be estimated via simulation. Therefore, deterministic optimization algorithms are not applicable and simulation optimization algorithms are needed.

The general formulation of the simulation optimization problem is as follows:

$$\min_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\left[L(\boldsymbol{\theta}, \omega)\right], \tag{1.1}$$

where $\boldsymbol{\theta} \in \Theta$ is a p-dimensional vector of the decision variables and Θ is the feasible region. It assumes that little knowledge (linearity or convexity) of the objective function $J(\boldsymbol{\theta})$ is known, and moreover $J(\boldsymbol{\theta})$ is the expectation of another quantity $L(\boldsymbol{\theta}, \omega)$, so it cannot be obtained directly. $L(\boldsymbol{\theta}, \omega)$ is the performance measure of interest and ω represents a simulation replication, which comprises the uncertainty of the system. The optimal decision variable is defined as

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}). \tag{1.2}$$

Optimization problems are generally classified into continuous and discrete problems depending on the types of values the decision variables $\boldsymbol{\theta}$ can take. An alternative classification involves the size of the feasible region: finite versus infinite for discrete problems, bounded versus unbounded for continuous problems. Extensive review on the topic of simulation optimization can be found in [1–5]. Three different approaches are briefly discussed since they are closely related to the new methods proposed in the thesis.

Stochastic approximation is a stochastic optimization technique analogous to gradient methods in deterministic optimization problems. Stochastic approximation uses the following recursion to update the solution:

$$x_{k+1} = x_k + a_k \nabla \tilde{f}(x_k),$$

where $\nabla \tilde{f}(x_k)$ is the gradient estimate and a_k is the step size. Two classical methods, Robbins-Monro (RM) and Kiefer-Wolfowitz (KW), estimate the true gradient $\nabla f(x_n)$ using direct gradient estimates and finite difference gradient estimates, respectively. In the stochastic simulation context, direct gradient estimation methods include perturbation analysis (PA) [6–8] and likelihood ratio/score function methods (LR/SF) [9, 10]; see [11] for more details.

Metamodels, also known as Response Surface Methodology (RSM), provide a functional relationship between the performance measurements and parameters of interest. Metamodel-based methods decouple optimization from simulation, as metamodels approximate stochastic responses through an algebraic function and deterministic optimization procedures are applied to the metamodel. Polynomial models are one of the commonly used metamodels since they usually have compact forms and are easy to construct and evaluate. However, due to their lack of flexibility, kriging, splines, neural networks and radial basis functions are more adequate to capture global characteristics of a response surface.

Both stochastic approximation and metamodeling are generally designed for solving stochastic optimization problems with continuous variables. Statistical ranking and selection (R&S) addresses stochastic optimization problem with discrete variables. Three R&S procedures are indifference-zone (IZ) methods, value of information procedures and optimal computing budget allocation (OCBA). IZ methods guarantee asymptotic lower bounds for the probability of correct selection (PCS), as long as the true underlying performance values are sufficiently far apart. Bayesian models for R&S consider the tradeoff between estimates of the performance values and uncertainty about those estimates, using the concept of "value of information." OCBA is designed with the flexibility to adapt to both frequentist and Bayesian models.

1.2 Direct Gradient Enhanced Metamodels

A metamodel is commonly used in simulation optimization to provide an auxiliary functional relationship between the input and output of a simulation model. Conducting simulations to collect experimental data is necessary to build metamodels, where the simulated data collected are usually performance measurements for parameters of interest. However, direct derivative information may also be available in stochastic simulation settings, where the output responses include not only the performance measurement, but also values of the gradient of performance measurement with respect to the parameters. Perturbation analysis (PA) [6–8] and likelihood ratio/score function methods (LR/SF) [9,10] are techniques that aim at estimating the gradient the performance measure. Applications of direct gradient estimates have been studied extensively, including queueing, inventory and finance applications [12,13].

In general, there are two types of metamodeling strategies: iterated local metamodels and global metamodels. An overview of local and global metamodel-based optimization is given in [14] and [15].

Iterated local metamodels, also known as sequential response surface methodology, rely on low-order polynomial regression. A first-order polynomial is usually used to fit local response surface in a small region to determine the search direction. Following a line search, new regions for the parameters of interest are exploited repeatedly until the region of most interest is determined. At the final step, a quadratic approximation is chosen and deterministic optimization methods are applied to locate the optimum. Regression techniques and experiment design are critical in this procedure; see [16] for details.

In global metamodels, high-order polynomial regression or nonlinear regression techniques based on existing knowledge about the response surface are appropriate; see [17] for an example. To capture global characteristics of a response surface, more flexibility in the models is required. Therefore, kriging, splines, neural networks and radial basis functions are more appropriate for fitting global metamodels. Among all these, kriging has received a lot of attention in the stochastic simulation community over the past decade [18–20]. Recently, [21] proposed stochastic kriging as an extension of kriging, which explicitly takes the uncertainties in simulation noise into consideration. Stochastic kriging is considered to be flexible and promising in fitting global response surfaces, especially in stochastic simulation settings.

In the stochastic simulation setting, direct derivative information may be available, i.e., the simulation output may include not only the performance measurements, but also estimates of the gradients of performance measurement with respect to the parameters. Techniques for estimating gradients, including perturbation analysis (PA) and likelihood ratio/score function methods (LR/SF), are discussed in [7], [10] and [13]; see also references therein.

The availability of additional gradient information suggests the potential for improving the quality of metamodels. Combining gradient information has been investigated for building metamodels under deterministic computer simulation settings; see [22] and [23] for approaches to approximate response surface with artificial neutral networks and kriging. In stochastic simulation settings, researchers have also made attempts to incorporate gradient estimates into metamodeling approaches. [24] proposed a gradient surface method (GSM) that uses the gradient estimates only to iteratively fit lower-order polynomial models. [25] introduced stochastic kriging with gradient estimators (SKG) to exploit gradient estimates in stochastic kriging, showing that the new approach provides better prediction with smaller mean squared error (MSE). This approach is similar to cokriging proposed in deterministic simulations [26], and requires differentiability of the correlation functions, since derivatives of random processes or random fields are used to model gradient estimates.

1.3 Variational Bayesian Inference in Simulation Optimization

Bayesian statistical models can be used to represent the beliefs of a decision-maker about an uncertain environment. For example, in revenue management, a seller formulates beliefs about customers' willingness to pay; in energy, we may have a belief about the suitability of a candidate location for a new wind farm.

In R&S, Bayesian models consider the tradeoff between estimates of the performance values and uncertainty about those estimates. This is known as the "value of information" (VIP) approach, going back to [27] and extended in later work. Theoretical properties of the policy were studied in [28]. VIP-based policies considering unknown measurement noise were developed in [29] and [30]. See [31] and [5] for an extensive up-to-date survey of Bayesian learning techniques. [32] compares several sequential procedures and concludes that Bayesian procedures are more efficient when the number of alternatives increases.

In the context of dynamic pricing, Bayesian statistics have been used to model environmental uncertainty [33, 34], and different pricing strategies have been proposed to optimize the balance between revenue and information. For example, [35] proposes a one-step look-ahead strategy for problems with logistic revenue curves, while [36] presents an approach based on multi-armed bandit theory. A recent stream of work, represented by [37], [38], [39], and [40], has focused on establishing long-run convergence rates for policies that are mostly myopic, with occasional periods of exploration spaced increasingly further apart. However, in the specific context of B2B pricing, individual transactions typically have high volume (for example, the seller may be negotiating the price of a year's supply of raw materials) and incur high costs (e.g. the time and money spent during negotiations), making it important to obtain good performance quickly.

Most Bayesian procedures rely on conjugate prior distributions on the unknown model parameters in order to maintain computational tractability. Conjugate priors model the evolution of these beliefs over time as new information is collected, either from stochastic simulation or field experiments. However, there are relatively few of these conjugate models, and they simply do not exist in many problems of interest. Variational Bayesian inference can be used to create computationally tractable, "nearly conjugate" models that optimally approximate the actual belief distributions and enable the use of anticipatory information collection and optimization policies.

1.4 Outline of the Thesis

The thesis centers around simulation optimization, including several new methods based on direct gradient estimates and optimal learning approaches. We now outline the thesis and summarize the contents of each chapter.

Chapter 2 investigates potential modeling improvements that can be achieved by exploiting additional gradient information in the regression setting. Using least squares and maximum likelihood estimation, we propose various Direct Gradient Augmented Regression (DiGAR) models that incorporate direct gradient estimators, starting with a one-dimensional independent variable and then extending to multidimensional input. For some special settings, we are able to characterize the variance of the estimated parameters in DiGAR and compare them analytically with the standard regression model. For a more typical stochastic simulation setting, we investigate the potential effectiveness of the augmented model by comparing it with standard regression in fitting a functional relationship for a simple queueing model, including both a one-dimensional and a four-dimensional example. The preliminary empirical results are quite encouraging, as they indicate how DiGAR can capture trends that the standard model would miss. Even in queueing examples where there is high correlation between the output and the gradient estimators, the basic DiGAR model that does not explicitly account for these correlations performs significantly better than the standard regression model.

Chapter 3 introduces an approach for enhancing stochastic kriging in the setting where additional direct gradient information is available, e.g., provided by techniques such as perturbation analysis or the likelihood ratio method. The new approach, called Gradient Extrapolated Stochastic Kriging (GESK), incorporates direct gradient estimates by extrapolating additional responses. For two simplified settings, we show that GESK reduces mean squared error (MSE) compared to stochastic kriging under certain conditions on step sizes. Since extrapolation step sizes are crucial to the performance of the GESK model, we propose two different approaches to determine the step sizes: maximizing penalized likelihood and minimizing integrated mean squared error. Numerical experiments are conducted to illustrate the performance of the GESK model and to compare it with alternative approaches.

Chapter 4 proposes the first computationally tractable Bayesian statistical model for learning unknown correlation structures in fully sequential simulation selection. Correlations represent similarities or differences between various design alternatives, and can be exploited to extract much more information from each individual simulation. However, in most applications, the correlation structure is unknown, thus creating the additional challenge of simultaneously learning unknown mean performance values and unknown correlations. Based on our new statistical model, we derive a Bayesian procedure that allocates simulations based on the value of information, thus anticipating future changes to our beliefs about the correlations. Our approach outperforms existing methods for known correlation structures in numerical experiments, including one motivated by the problem of optimal wind farm placement, where real data are used to calibrate the simulation model.

Chapter 5 proposes an approximate Bayesian statistical model for predicting the win/loss probability for a given price in business-to-business (B2B) pricing. This model allows us to learn parameters in logistic regression based on binary (win/loss) data and can be quickly updated after each new win/loss observation. We also consider an approach for recommending target prices based on the approximate Bayesian model, thus integrating uncertainty into decision-making. We test the statistical model and the target price recommendation strategy with synthetic data, and observe encouraging empirical results.

Chapter 2: Direct Gradient Augmented Regression

2.1 Introduction

Because regression analysis arose from physically observed processes, it assumes that the only data points generated are measurements of the value of the dependent variable for each combination of values for the independent variable. However, in the stochastic simulation setting that is our primary focus, research over the previous four decades has led to the availability of *direct* derivative information – meaning not from finite-difference approximations, e.g., for each design point or value of the independent variable, the output responses generated include not only a value of the dependent variable, but also a value of the *gradient* of the dependent variable with respect to the independent variable(s). Settings where such direct gradient estimators are available include queueing, inventory, and finance [6–8, 10, 12, 13].

Clearly, in this enhanced data setting, the availability of gradient information should lead to an improvement in the estimated functional relationship between the dependent and independent variables. In this chapter, we investigate such improvements in the regression setting. Specifically, we consider a simple modification of the standard linear regression model to incorporate the additional measurements. We call the new method Direct Gradient Augmented Regression (DiGAR, pronounced "digger"). Using the least squares approach and maximum likelihood estimation, we derive the resulting parameter estimates for the proposed DiGAR models and provide a theoretical analysis of the improvements that can be achieved over a standard model. We also conduct some preliminary numerical experiments to empirically investigate the improvements that can be achieved. In particular, we consider a simple queueing system for which there are analytical results available, which can be used to judge the effectiveness of both standard regression and DiGAR using both a linear and quadratic regression function.

The main motivation for our work is two-fold:

- characterizing a global response surface, a traditional application of regression;
- approximating a local response surface with a view towards local improvements in carrying out simulation-based optimization.

Broadly speaking, our primary objective is to provide improved means for estimating functional relationships using direct higher-order information. As a first step, we consider regression because of its wide application in many fields and because of the central role it plays in the sequential response surface methodology (RSM) approach to optimization, which "is a metamodel-based optimization method that builds linear or quadratic local approximations" to the response function [14]. Polynomial regression is generally used to fit the response surface, as a series of experiments is used along the steepest descent direction to obtain additional improvement [41]; see also [16] for a detailed discussion on the relevant statistical techniques. In Figure 18.1 of [14], simulation optimization strategies are classified according to the nature of the controllable variables and the response function. For the case where the response function is assumed differentiable, there is a dichotomy between direct gradient methods and metamodel methods, of which RSM is the primary technique. DiGAR serves as an example of merging these two categories. By integrating direct gradient estimates into the regression model, the hope is that the fitted curve better approximates the true model, which should also lead to faster convergence when used in sequential procedures.

DiGAR provides a paradigm shift, though it is clearly only applicable in a special setting in which direct gradient estimates are available, which is often the case in stochastic simulation using techniques such as perturbation analysis (PA) and the likelihood ratio/score function (LR/SF) method [6, 7]. These procedures provide estimates in which no resimulation is required, i.e., whenever an estimate of an output performance measure is obtained, an estimate of the derivative(s) with respect to parameters of interest are also obtained at that particular setting of the parameters. This is referred as direct gradient estimates in the thesis, to contrast with indirect estimates obtained by actually changing the value of the parameters and running additional simulations; see [13, 42] for a recent survey and tutorial with references.

In the setting where direct gradient estimates are available, another approach for incorporating gradient estimates into regression models proposed by [24] for a sequential simulation optimization procedure fits the gradient response surface directly using the gradient estimates only; the function estimates themselves are not used in the procedure. In contrast, our approach uses all sets of measurements that are available in augmenting existing functional estimation procedures such as regression. Thus, DiGAR *augments* standard regression rather than replacing it.

To illustrate the approach in a straightforward way, DiGAR uses least squares and maximum likelihood estimation (for the normally distributed setting), but other criteria such as minimum sum of absolute errors (MSAE) and minimization of the maximum absolute errors (MMAE) could also have been used. Least squares is among the most popular criterion in regression [43, 44], due to its simplicity, but it has the drawback of being sensitive to outliers, as is also evident in our numerical examples reported in Section 2.3. There has been much work in robust regression to overcome the drawback of least square regression [45], but our preliminary numerical results on a queueing example indicate that combining the gradient estimations with the least squares approach can lead to noticeable qualitative correction that mitigates sensitivity to outliers, i.e., when some of the observed dependent variable output values exhibit large fluctuations from their means, DiGAR is able to correct the shape of the fitted curve – the slope for a linear fit and the curvature for a quadratic fit. Not surprisingly, the observed variance of the parameter estimates in the DiGAR is also considerably lower than that of the standard regression model.

2.2 Models

In this section, we consider the simplest setting, beginning with a review of the most basic standard linear regression model before introducing the DiGAR models where direct gradient estimators are assumed available.

The following assumptions are used in the section for various theoretical results and calculations.

Assumption 2.1. $E[\epsilon_i] = 0 \ \forall i$.

Assumption 2.2. $E[\epsilon'_i] = 0 \ \forall i$.

Assumption 2.3. *(i)* $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, i \neq j$, *and (ii)* $\mathrm{Var}(\epsilon_i) = \sigma^2 \ \forall i$ *with known* σ^2.

Assumption 2.4. *(i)* $\mathrm{Cov}(\epsilon'_i, \epsilon'_j) = 0 \ \forall \ i \neq j$, $\mathrm{Cov}(\epsilon_i, \epsilon'_j) = 0 \ \forall \ i, j$, *and (ii)* $\mathrm{Var}(\epsilon'_i) = \sigma_g^2 \ \forall \ i$ *with known* σ_g^2.

Assumption 2.5. $\epsilon_i \sim N(0, \sigma^2), \ \epsilon'_i \sim N(0, \sigma_g^2) \ \forall i$ *with known* σ^2 *and* σ_g^2.

2.2.1 Standard Linear Regression Model

Consider the usual regression setting with independent variable x and dependent variable y, where $n > 1$ data points $(x_1, y_1), ..., (x_n, y_n)$ are given. Both independent and dependent variables take values from a continuous domain. The standard linear regression model is the following:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots, n, \tag{2.1}$$

where assumptions on the "noise" terms $\{\epsilon_i\}$ determine properties of resulting estimators.

The least-squares approach minimizes the sum of squared residuals given by $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$, leading to the following estimators for β_0 and β_1:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{2.2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2.3}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of $\{x_i\}$ and $\{y_i\}$, respectively. These estimators are unbiased assuming the noise terms have zero mean.

In the traditional regression model, it is well known that the maximum likelihood estimators (MLEs) for normally distributed residuals coincide with the OLS estimators given by (2.2) and (2.3).

2.2.2 Direct Gradient Augmented Regression

Now consider the enhanced setting where the n data points are $(x_1, y_1, g_1), ..., (x_n, y_n, g_n)$, with g_i representing a direct estimate of the gradient of y_i at x_i. The basic Direct Gradient Augmented Regression (DiGAR) model is the following:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{2.4}$$

$$g_i = \beta_1 + \epsilon'_i, \tag{2.5}$$

where $g_i, i = 1, 2, \cdots, n$ are the gradient estimates with residuals $\{\epsilon'_i\}$.

Again using the OLS approach, the function to be minimized is the sum of the squared deviations in both y_i and g_i,

$$L = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + \sum_{i=1}^{n}(g_i - \beta_1)^2. \tag{2.6}$$

The resulting estimators that minimize (2.6) are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} + n\bar{g}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + n} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \bar{g}}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + 1}, \tag{2.7}$$

where $\bar{x}$, $\bar{y}$ and $\bar{g}$ are the corresponding sample means of x_i, y_i and g_i. Note that in the basic DiGAR model, the form of $\hat{\beta}_0$ remains unchanged, whereas $\hat{\beta}_1$ has the additional terms $\bar{g}$ and 1 in the numerator and denominator, respectively, reflecting the added gradient information.

An alternative approach is to derive estimators using y_i and g_i separately and combine them. However, since we would like to combine two sources of data as opposed to two estimators, we choose to use the objective function in (2.6). Later in the chapter, we will build the connections between these two approaches.

Under the Assumptions 2.3 - 2.5, the parameter estimators given by (2.7) are unbiased, since

$$E[\hat{\beta}_1] = \frac{\sum_{i=1}^{n} x_i E[y_i] - n\bar{x}E[\bar{y}] + nE[\bar{g}]}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + n} = \beta_1,$$

$$E[\hat{\beta}_0] = E[\bar{y}] - E[\hat{\beta}_1]\bar{x} = \beta_0,$$

and the variances of $\hat{\beta}_1$ is calculated as follows:

$$
\begin{aligned}
\operatorname{Var}(\hat{\beta}_1) &= \frac{\sum_{i=1}^{n} x_i^2 \operatorname{Var}(y_i) - n^2\bar{x}^2 \operatorname{Var}(\bar{y}) + n^2 \operatorname{Var}(\bar{g})}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + n\right)^2} \\
&= \frac{\sum_{i=1}^{n} x_i^2\sigma^2 - n^2\bar{x}^2\sigma^2/n + n^2\sigma_g^2/n}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + n\right)^2} \\
&= \frac{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\sigma^2 + n\sigma_g^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + n\right)^2}
\end{aligned}
$$

The DiGAR estimator $\hat{\beta}_1$ can also be viewed as a linear combination of the standard linear regression estimator in (2.3) and another unbiased estimator $\bar{g}$ with weights proportional n and $\sum_{i=1}^{n}(x_i - \bar{x})^2$, respectively. We are particularly interested in investigating the variance of the estimators, and a known results suggest that the optimal weights for combining two unbiased estimators should be inversely proportional to their variances.

A more general form of the least-squares function to be minimized allows relative weighting (convex combination) of the two contributions rather than the equal weighting used in (2.6), i.e.,

$$
L = \alpha \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 + (1-\alpha) \sum_{i=1}^{n} (g_i - \beta_1)^2, \tag{2.8}
$$

where $\alpha \in [0, 1]$. $\alpha = 1$ corresponds to standard regression, $\alpha = 0.5$ corresponds to the OLS DiGAR model introduced earlier, and $\alpha = 0$ uses only the gradient information. Differentiating with respect to β_0 and β_1,

$$\begin{aligned}\frac{\partial L}{\partial \beta_0} &= -2\alpha \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i),\\ \frac{\partial L}{\partial \beta_1} &= -2\alpha \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i - 2(1-\alpha)\sum_{i=1}^{n}(g_i - \beta_1).\end{aligned} \tag{2.9}$$

Setting both equal to 0 and solving yields the following estimators:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \frac{1-\alpha}{\alpha}\bar{g}}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1-\alpha}{\alpha}}. \tag{2.10}$$

If the estimators for the gradients are also unbiased, i.e., $E[\epsilon_i'] = 0$, then the $\alpha-$DiGAR estimators are also unbiased.

Remark 2.1. *The weight α in (2.8) can be viewed as cost to the sum of squared errors. We will look into the weighted objective function in (2.8) and try to build connections between the DiGAR estimators and the estimator obtained by combining two unbiased estimators. In the following sections, we will show that the optimal α provides the same estimator as the optimal estimator obtained by combining two unbiased estimators.*

Proposition 2.2. *Under Assumptions 2.1 & 2.2, the $\alpha-$DiGAR estimators given by (2.10) are unbiased.*

2.2.3 Choice of Weights in α-DiGAR

Instead of thinking α as a fixed cost, we consider the weight α can be chosen in practice. So how should one choose the weight parameter α? One intuitive choice is to choose the relative weights inversely proportional to the corresponding variances, but the variances are unknown and possibly nonhomogeneous across the independent variable range.

Another option is to minimize the prediction variance. The predicted value at x_i can be written as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} - \hat{\beta}_1(\bar{x} - x_i),$$

so that minimizing the prediction variance is equivalent to minimizing the variance of $\hat{\beta}_1$. However, again an analytical expression for this variance is unavailable in most settings.

In the next section we revisit the question of weight selection in some special settings of homogeneous variances where "optimal" weights can be determined explicitly.

2.2.4 Theoretical Comparisons Between Estimators for Special Cases

Both traditional regression and least-squares DiGAR models can be applied in a very general setting, but it is difficult to obtain any analytical results without further assumptions.

For the traditional regression model, if the noise terms have zero mean and in addition are uncorrelated with common variance, the theoretical variances can also be computed explicitly.

Proposition 2.3. *Under Assumptions 2.1 and 2.3, the variances of the estimators*

(2.2) and (2.3) are given by

$$\begin{aligned}\operatorname{Var}(\hat{\beta}_0) &= \sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\right], &(2.11)\\ \operatorname{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}. &(2.12)\end{aligned}$$

Similarly, in the DiGAR model, if all of the residuals are uncorrelated with common variance, it is not difficult to compute the theoretical variances of the α-DiGAR estimators.

Proposition 2.4. *Under Assumptions 2.1-2.4, the variances of the estimators (2.10) are given by*

$$\begin{aligned}\operatorname{Var}(\hat{\beta}_0) &= \frac{\sigma^2}{n}\left[1+\frac{\frac{\bar{x}^2}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}{\left(\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2+\frac{1-\alpha}{\alpha}\right)^2}\right]+\frac{\sigma_g^2}{n}\left[\frac{\bar{x}^2(\frac{1-\alpha}{\alpha})^2}{\left(\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2+\frac{1-\alpha}{\alpha}\right)^2}\right], &(2.13)\\ \operatorname{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{n}\left[\frac{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2+(\frac{1-\alpha}{\alpha})^2\frac{\sigma_g^2}{\sigma^2}}{\left(\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2+\frac{1-\alpha}{\alpha}\right)^2}\right]. &(2.14)\end{aligned}$$

Revisiting Weight Selection in α-DiGAR

Under the assumption of homogenous variances for both the output and its gradient, per Assumptions 2.3(ii) and 2.4(ii), selection of weights proportional to variance implies

$$\frac{\alpha}{1-\alpha}=\frac{\sigma_g^2}{\sigma^2},$$

which leads to

$$\alpha=\frac{\sigma_g^2}{\sigma_g^2+\sigma^2}. \tag{2.15}$$

For the "optimal" choice of weights, which in the previous subsection was said to be equivalent to minimizing the variance of $\hat{\beta}_1$, differentiating (2.14) with respect to α and setting equal to 0 gives the same proportional weights as (2.15).

For the uncorrelated setting, Proposition 2.4 can be used to find necessary and sufficient conditions for the variance of the slope estimator to be lower for the α-DiGAR estimators than for the standard slope estimator given by (2.3).

Proposition 2.5. *Under Assumptions 2.1-2.4,*

$$\frac{\sigma_g^2}{\sigma^2} \leq \frac{2\alpha}{1-\alpha} + \frac{1}{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2} \Longleftrightarrow \mathrm{Var}(\hat{\beta}_1^{DiGAR}) \leq \mathrm{Var}(\hat{\beta}_1^{standard}),$$

where β_1^{DiGAR} and $\beta_1^{standard}$ denote the α-DiGAR slope estimator given by (2.10) and the standard slope estimator given by (2.3), respectively.

Proof. Comparing the variances given by (2.12) and (2.14), the inequality holds if and only if

$$\begin{aligned}
\frac{\sigma^2}{n}\frac{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2 + (\frac{1-\alpha}{\alpha})^2\frac{\sigma_g^2}{\sigma^2}}{\left(\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2+\frac{1-\alpha}{\alpha}\right)^2} &\leq \frac{\sigma^2}{n}\frac{1}{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2} \\
\Longleftrightarrow \quad \left(\frac{\sigma_g}{\sigma}\right)^2 \leq \left(\frac{\alpha}{1-\alpha}\right)^2 & \left(\frac{\left(\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2+\frac{1-\alpha}{\alpha}\right)^2}{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2} - \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right) \\
&= 2\left(\frac{\alpha}{1-\alpha}\right) + \frac{1}{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}
\end{aligned}$$

□

Intuitively, Proposition 2.5 indicates that if σ_g is large relative to σ, then it makes sense to decrease the relative weight for the gradient; also when the simulation budget is limited and thus only a relatively small number of design points are possible and experiment conditions make them close to each other (more likely in sequential RSM applications), the benefit of DiGAR should be more noticeable. The following corollary of Proposition 2.5 provides more explicit bounds for the weighting α. Two special cases include the equal weighting and proportional weighting scenarios, again in the uncorrelated setting.

Corollary 2.6. *Under Assumptions 2.1-2.4,* $\mathrm{Var}(\hat{\beta}_1)$ *for the* α*-DiGAR OLS slope estimator given by (2.10) is strictly lower than that for the standard slope estimator given by (2.3) if*

$$\alpha \geq \frac{\sigma_g^2}{\sigma_g^2 + 2\sigma^2},$$

which includes the following special cases:

(i) $\alpha = \frac{\sigma_g^2}{\sigma_g^2+\sigma^2}$.

(ii) $\alpha = 0.5$ *if* $\sigma_g^2 \leq 2\sigma^2$.

Proof. Main result follows directly from $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$, with the two cases satisfying the requisite condition, or directly from Proposition 2.5, noting that (i) $\frac{\alpha}{1-\alpha} = \frac{\sigma_g^2}{\sigma^2}$, (ii) $\frac{\alpha}{1-\alpha} = 1$. □

Maximum Likelihood Estimation DiGAR

If we further assume that the residuals are normally distributed and uncorrelated (independent) with *known* variances, we can derive the MLEs for β_0 and β_1, which will **not** coincide with the α-DiGAR estimators in general.

Note that Assumption 2.5 subsumes Assumptions 2.1 and 2.2, and includes the known variance condition.

Proposition 2.7. *Under Assumptions 2.3, 2.4, and 2.5, the MLEs for the DiGAR model specified by (2.4) and (2.5) are given by*

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}, \quad \tilde{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \frac{\sigma^2}{\sigma_g^2}\bar{g}}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\sigma^2}{\sigma_g^2}}. \tag{2.16}$$

Proof. Under Assumptions 2.3 and 2.4, y_i and g_i are independent due to the residuals being uncorrelated. Under Assumption 2.5, both σ^2 and σ_g^2 are known. The likelihood function is given by

$$L(\beta_0, \beta_1) = (2\pi)^{-n}(\sigma\sigma_g)^{-n} \exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 - \frac{1}{2\sigma_g^2}\sum_{i=1}^{n}(g_i - \beta_1)^2 \right\}.$$

Differentiating the log-likelihood function with respect to the parameters,

$$\begin{aligned} \frac{\partial \log(L)}{\partial \beta_1} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i + \frac{1}{\sigma_g^2}\sum_{i=1}^{n}(g_i - \beta_1), \\ \frac{\partial \log(L)}{\partial \beta_0} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i), \end{aligned}$$

which after setting equal to 0, and some algebra, leads to (2.16). □

Remark 2.8. *In the actual implementation, σ^2 and σ_g^2 are usually unknown. The MLE's of β_0, β_1, σ^2 and σ_g^2 are the solutions of a nonlinear system of equations and are not available in closed form. They would need to be solved using Newton-Raphson or some other iterative numerical methods. Approximate MLE's for β_0 and β_1 can be obtained by replacing σ^2 and σ_g^2 by the unbiased and consistent estimators given as follows:*

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad \hat{\sigma}_g^2 = \frac{1}{n-1}\sum_{i=1}^{n}(g_i - \bar{g})^2,$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ *with the estimators* $\hat{\beta}_0$ *and* $\hat{\beta}_1$ *provided in (2.7).*

Proposition 2.9. *Under Assumptions 2.3, 2.4, and 2.5, the variance of the MLE for* β_1 *given in (2.16) is*

$$\mathrm{Var}(\tilde{\beta}_1) = \frac{\sigma^2/n}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\sigma^2}{\sigma_g^2}}. \tag{2.17}$$

As expected, as $\sigma_g \to \infty$, where the additional gradient estimators provide no additional useful information, the MLE DiGAR slope estimator variance given by (2.17) approaches the traditional slope estimator variance given by (2.12), whereas as $\sigma_g \to 0$, $\mathrm{Var}(\tilde{\beta}_1) \to 0$ for MLE DiGAR.

Comparing the variances given by (2.17) and (2.12), we have the following result.

Proposition 2.10. *Under Assumptions 2.3, 2.4 and 2.5, and* $\sigma/\sigma_g > 0$*,* $\mathrm{Var}(\tilde{\beta}_1)$ *for MLE DiGAR given in (2.16) is strictly lower than that for the standard estimator given by (2.3).*

Thus, in the uncorrelated common variance setting, the MLE DiGAR slope estimator is guaranteed to have lower variance than the standard regression model slope estimator. However, it should also be noted that the MLEs given by (2.16) contain the variances σ^2 and σ_g^2, which are unknown in practice and thus need to be estimated from the data, whereas either of the two sets of DiGAR least-squares estimators given by (2.7) or (2.10) do not contain these terms. However, choosing the “optimal” weights in the $\alpha-$DiGAR estimator (2.10) as specified by (2.15) leads to the same requirement, since α depends on both σ and σ_g. Furthermore, this choice actually leads back to the MLEs, as summarized in the following.

Proposition 2.11. *The α-DiGAR estimators given by (2.10) using the weight α given by (2.15) coincide with the DiGAR MLEs given by (2.16).*

Remark 2.12. *It is mentioned earlier in the chapter that combing the two unbiased estimators, namely $\hat{\beta}_1$ in (2.3) and $\bar{g}$, is an alternative approach. The optimal weighted estimator obtained in this fashion is*

$$\frac{1}{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2}+\frac{n}{\sigma_g^2}}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+\frac{n}{\sigma_g^2}\bar{g}\right).$$

Simplifying this leads to the same estimator α-DiGAR estimators given by (2.10) using the weight α given by (2.15) and the DiGAR MLEs given by (2.16). This suggests that when we combing two sources of data using the loss function in (2.8), we are able to obtain the same estimator as combining two unbiased estimators, namely $\hat{\beta}_1$ in (2.3) and $\bar{g}$.

2.2.5 Multi-dimensional Linear Models

Now we consider the d-dimensional multiple regression linear model setting. For standard regression,

$$y_i = \beta_0 + \sum_{j=1}^{d}\beta_j x_{ij} + \epsilon_i, \quad i = 1, 2, \cdots, n. \tag{2.18}$$

The least-squares approach minimizes the sum of squared residuals given by $\sum\limits_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{d}\beta_j x_{ij})^2$, leading to the following estimators for β_0 and β_j:

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \sum_{j=1}^{d}\hat{\beta}_j\bar{x}_j, &(2.19)\\ \hat{\beta}_j &= \frac{\sum\limits_{i=1}^{n}(x_{ij}-\bar{x}_j)(y_i-\bar{y}) - \sum_{k\neq j}\hat{\beta}_k\sum\limits_{i=1}^{n}(x_{ik}-\bar{x}_k)(x_{ij}-\bar{x}_j)}{\sum\limits_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}, &(2.20)\end{aligned}$$

where $\bar{x}_j$ is the sample mean of $\{x_{ij}\}$.

The analogous multi-dimensional DiGAR model is given by

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{d}\beta_j x_{ij} + \epsilon_i, &(2.21)\\ g_{ij} &= \beta_j + \epsilon'_{ij}, &(2.22)\end{aligned}$$

where $g_{ij},\ j = 1, 2, \cdots, k,\ i = 1, 2, \cdots, n$ are the gradient estimates with residuals $\{\epsilon'_{ij}\}$. The corresponding least-squares function to be minimized given by

$$\alpha_0\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{d}\beta_j x_{ij})^2 + \sum_{j=1}^{d}\alpha_j\sum_{i=1}^{n}(g_{ij}-\beta_j)^2,$$

where $\sum_{i=0}^{d}\alpha_i = 1,\ \alpha_i \geq 0$. Taking the partial derivatives and setting equal to zero yields again (2.19) and

$$\hat{\beta}_j = \frac{\sum\limits_{i=1}^{n}(x_{ij}-\bar{x_j})(y_i-\bar{y}) - \sum_{k\neq j}\beta_k\sum\limits_{i=1}^{n}(x_{ik}-\bar{x}_k)(x_{ij}-\bar{x}_j) + n\frac{\alpha_j}{\alpha_0}\bar{g}_j}{\sum\limits_{i=1}^{n}(x_{ij}-\bar{x_j})^2 + n\frac{\alpha_j}{\alpha_0}},\quad j > 0, \qquad (2.23)$$

which reduces to the previous expression with $\alpha_0 = \alpha, \alpha_1 = 1-\alpha$, when there is just a single input. We will also refer to these as the α-DiGAR estimators.

2.2.6 DiGAR Model for Gradient Estimates Correlated with Performance Outputs

The least-squares estimators for the basic α-DiGAR model, as given by (2.10) and (2.23), were derived without consideration of the correlation structure, although explicit computation of the theoretical variance as given by Proposition 2.4 required further assumptions. Here we consider the generalized least squares (GLS) setting where the outputs y_i and gradient estimates g_i are correlated as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.24}$$

where for the univariate case

$$\mathbf{y} = \begin{pmatrix} y_1 \\ g_1 \\ \vdots \\ y_n \\ g_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 0 & 1 \\ \vdots & \vdots \\ 1 & x_n \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \tag{2.25}$$

and $\boldsymbol{\epsilon}$ would contain the noise terms for the corresponding output and gradient estimates. The OLS estimator is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}, \tag{2.26}$$

which matches (2.23) in the uncorrelated case with equal weights ($\alpha_i = 1/n$).

The weighted least-squares (WLS) solution is given by

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{W}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{W}\mathbf{y}, \tag{2.27}$$

where $\mathbf{W}$ is a diagonal matrix such that $\operatorname{diag}(\mathbf{W}) = [\alpha_1, \alpha_2, \ldots, \alpha_{2n}]$.

Now we explicitly consider the setting where the covariance matrix of $\boldsymbol{\epsilon}$ is given by $\mathbf{V}$, which is non-diagonal due to the correlations between y_i and g_i. This can also include the case of common random numbers, which would induce correlations within the $\{y_i\}$ and $\{g_i\}$. Again assume the residuals have zero mean, i.e., $E[\boldsymbol{\epsilon}] = \mathbf{0}$ (Assumptions 2.1 and 2.2), so $E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$. In the general correlated, non-homogeneous setting of (2.24)/(2.25), (2.26) is not the best (i.e., variance minimizing) linear unbiased estimator (BLUE) for $\boldsymbol{\beta}$, but the Gauss-Markov Theorem for the uncorrelated homogeneous variance setting can be used to prove that the BLUE for the general setting is given by the following GLS estimator:

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-1}y}, \tag{2.28}$$

and the covariance matrix for $\hat{\boldsymbol{\beta}}_G$ is $\operatorname{Cov}(\hat{\boldsymbol{\beta}}_G) = (\mathbf{X'V^{-1}X})^{-1}$.

Proposition 2.13. *Under Assumptions 2.1 and 2.2, the BLUE for the model given by (2.24) and (2.25) is the GLS estimator given by (2.28). Furthermore, if the residuals are assumed to be normally distributed, the MLE of $\boldsymbol{\beta}$ is the same as the GLS estimator.*

Proof. The proof uses the Gauss-Markov Theorem.

Theorem 2.14 (Gauss-Markov Theorem)**.** *For the regression model (2.24) with $E[\boldsymbol{\epsilon}] = 0$ and $\operatorname{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, the OLS estimators have minimum variance among all linear unbiased estimators.*

Since $\mathbf{V}$ is positive definite, there exists an $2n \times 2n$ nonsingular matrix $\mathbf{P}$ such that $\mathbf{V} = \mathbf{P}\mathbf{P}'$. Multiplying $\mathbf{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ by $\mathbf{P}^{-1}$, we obtain

$$\mathbf{P}^{-1}\mathbf{y} = \mathbf{P}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}^{-1}\boldsymbol{\epsilon},$$

where $E[\mathbf{P}^{-1}\boldsymbol{\epsilon}] = \mathbf{P^{-1}}E[\boldsymbol{\epsilon}] = \mathbf{0}$ and

$$\operatorname{Cov}(\mathbf{P}^{-1}\boldsymbol{\epsilon}) = \mathbf{P}^{-1}\operatorname{Cov}(\boldsymbol{\epsilon}(\mathbf{P}^{-1})' = \mathbf{P}^{-1}\sigma^2\mathbf{V}(\mathbf{P}^{-1})' = \sigma^2\mathbf{P}^{-1}\mathbf{P}\mathbf{P}'(\mathbf{P}')^{-1} = \sigma^2\mathbf{I}.$$

Therefore, the assumptions in the Gauss-Markov Theorem are satisfied, and the least-squares estimator

$$\hat{\boldsymbol{\beta}} = \left[(\mathbf{P^{-1}X})'(\mathbf{P^{-1}X})\right]^{-1}(\mathbf{P^{-1}X})'\mathbf{P^{-1}y},$$

is the best linear unbiased estimator (BLUE). The estimator $\hat{\boldsymbol{\beta}}$ can be written as

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left[(\mathbf{P^{-1}X})'(\mathbf{P^{-1}X})\right]^{-1}(\mathbf{P^{-1}X})'\mathbf{P^{-1}y} \\
&= [\mathbf{X}'(\mathbf{P}')^{\mathbf{-1}}\mathbf{P^{-1}X}]^{-1}\mathbf{X}'(\mathbf{P}')^{\mathbf{-1}}\mathbf{P^{-1}y} \\
&= [\mathbf{X}'(\mathbf{PP}')^{\mathbf{-1}}\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{PP}')^{\mathbf{-1}}\mathbf{y} \\
&= (\mathbf{X}'\mathbf{V^{-1}X})^{\mathbf{-1}}\mathbf{X}'\mathbf{V^{-1}y}.
\end{aligned}$$

$\square$

We now analyze the variance of the slope portion of the GLS estimator by considering a very special case.

Assumption 2.6. $\mathbf{V}$ *is a positive definite matrix such that* y_i *is correlated with* g_j *only when* $i = j$*, with correlation* ρ*.*

Proposition 2.15. *Under Assumptions 2.1, 2.2, 2.3 (ii), 2.4 (ii), and 2.6, for the model given by (2.24)/(2.25), the variance of* $\hat{\beta}_1$ *in (2.28) is given by*

$$\operatorname{Var}(\hat{\beta}_1) = \frac{\sigma^2/n}{\frac{1}{1-\rho^2}\frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\sigma^2}{\sigma_g^2}}, \tag{2.29}$$

which vanishes if either σ or σ_g goes to zero.

Proof. The lower right element of the covariance matrix $(\mathbf{X}'\mathbf{V}^{\mathbf{-1}}\mathbf{X})^{-1}$ gives $\text{Var}(\hat{\beta}_1)$, so writing the covariance matrix of the residuals as

$$\mathbf{V} = \begin{pmatrix} \sigma^2 & \rho\sigma\sigma_g & & & \\ \rho\sigma\sigma_g & \sigma_g^2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \sigma^2 & \rho\sigma\sigma_g \\ & & & \rho\sigma\sigma_g & \sigma_g^2 \end{pmatrix},$$

If the residuals are normally distributed, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{V})$, then $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, and the likelihood function is

$$L(\boldsymbol{\beta}) = \frac{1}{(2\pi)^n|\mathbf{V}|^{1/2}} \exp\left\{-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{V})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/2\right\},$$

so the log-likelihood function is

$$\ln L(\boldsymbol{\beta}) = -n\ln(2\pi) - \frac{1}{2}\ln(|\mathbf{V}|) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Differentiating with respect to $\boldsymbol{\beta}$,

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = -\left(\mathbf{X}'\mathbf{V}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{V}^{\mathbf{-1}}\mathbf{y}\right),$$

setting equal to zero and solving for $\boldsymbol{\beta}$ gives the estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{\mathbf{-1}}\mathbf{y})^{-1}\mathbf{X}'\mathbf{V}^{\mathbf{-1}}\mathbf{y},$$

which is the same as the best linear unbiased estimator (BLUE).

□

In the special correlated setting of Proposition 2.15, it is interesting to note that the variance of the GLS DiGAR slope estimator will go to zero if *either* of the variances of the underlying noises vanish, whereas for the basic OLS DiGAR estimator, both variances need to be zero in the uncorrelated setting (see Proposition 2.4). Also, by comparing (2.29) with the variance of the standard slope estimator given by (2.12), the following is easily established.

Proposition 2.16. *Under Assumptions 2.1, 2.2, 2.3 (ii), 2.4 (ii), and 2.6, for the model given by (2.24)/(2.25), if* $0 < \sigma^2 < \infty$ *and* $0 < \sigma_g^2 < \infty$*, then for* $-1 < \rho < 1$*, the DiGAR slope estimator given by (2.28) has lower variance than the standard slope estimator.*

Note that the DiGAR estimators given by (2.28) for the uncorrelated case will differ from the DiGAR estimators derived earlier using OLS. Generally, ρ is unknown and must be estimated based on the data. As noted in [16], estimating the correlation matrix changes the linear estimator (2.28) into a nonlinear estimator. Thus, although the theoretical analysis indicates potential for variance reduction from estimating the correlation, the extra computational budget spent on estimating ρ (and the resulting nonlinearities) must be traded off with any potential performance gains, an issue that is also investigated empirically in the numerical examples in the next section.

2.3 Numerical Examples

Queueing systems are one of the main application areas for stochastic simulation, and the earliest application of direct gradient estimation in simulation was queueing; thus, we chose a simple queueing model as the setting on which to empirically investigate the performance of the DiGAR estimators, comparing them with standard regression in a setting where direct gradient estimates are available but where one or more of the assumptions of the theoretical results are generally not satisfied. In particular, although the least-squares α-DiGAR models are applicable in the general correlated setting, it is difficult to compute the theoretical variances explicitly for such queueing settings, so simulation experiments are used to empirically compare their performance with standard regression.

Perhaps the simplest queueing system is the first-come, first-served (FCFS), single-server queue considered here. Specifically, we consider the mean total time in system for a customer (queue or delay time plus service time) as a function of the parameters of the (common) service time or interarrival time distribution of customers. Two settings are considered: a univariate setting (single input) and a multivariate setting. In both settings, four outputs, i.e., four different y's, are considered. In queueing theory, the interest is frequently steady-state performance, but since the focus here is on improving regression models rather than on queueing theory, per se, we consider the 2nd, 3rd, 4th, and 5th customers, for which we can easily obtain analytical results that can be used to compare the quality of the standard linear regression and DiGAR models without having to worry about whether the simulation has reached steady state. Letting T_k denote the system time of the kth customer, the outputs of interest are

$$y^{(k)} = E[T_k], k = 2, 3, 4, 5.$$

The gradient estimators for all of the examples are provided in the Electronic Companion Section A.2. The system time performance and its gradient estimate are clearly highly correlated, and the variance of both the performance and its gradient are not homogeneous across the range of independent values considered (although since the queue will be far from the heavy traffic intensity regime, the violation may not be particularly severe). As a result, Assumptions 2.3 and 2.4 are violated and the conclusions of Propositions 2.3 through 2.10 cannot be applied, although Proposition 2.2 holds, as the gradient estimators are unbiased. However, it should be noted that the α-DiGAR estimators are not derived under Assumptions 2.3 and 2.4, which are only sufficient conditions to establish the theoretical results.

2.3.1 Example: $M/M/1$ Queue

The univariate example takes the arrival process to be Poisson with fixed rate and service times to be i.i.d. exponentially distributed, i.e., an $M/M/1$ queue, where the independent variable x is the mean service time. It is straightforward to compute the true theoretical dependence of the expected system time y on the mean service time x (cf. [46]), which can then be used to judge the quality of the fitted curves obtained by the various regression models.

For all experiments in the univariate setting of the $M/M/1$ queue, the arrival rate is fixed at 0.2, and the number of design points is $n = 10$, where output y (mean system time) is obtained at the following values of the independent variable x (mean service time):

$$x_1 = 3.6, x_2 = 3.7, x_3 = 3.8, x_4 = 3.9, x_5 = 4.0, x_6 = 4.1, x_7 = 4.2, x_8 = 4.3, x_9 = 4.4, x_{10} = 4.5.$$

At each of these independent variable values, 10 independent simulation replications (runs) were generated to obtain the system times (and gradients for DiGAR) of the customers; nine additional macroreplications were carried out only for the purposes of estimating the slope variances as described below. The relatively small number of replications highlights the challenges that the standard regression approach faces, since it leads to output estimates with relatively high variances.

Two metrics were used to evaluate the quality of fit for each data set:

(i) the sample mean-squared error from the true model over the independent variable range of interest defined by

$$L_2 = \int_{x_{\min}}^{x_{\max}} (\widehat{y}(x) - y(x))^2 dx,$$

where $\widehat{y}(x)$ and $y(x)$ denotes the fitted and true models, respectively; and

(ii) the theoretical (from the uncorrelated model) variance of the slope estimator denoted by $\widehat{\mathrm{Var}}(\hat{\beta}_1)$ and estimated using (2.12) and (2.14) for the standard and DiGAR models, respectively, with the pooled (over all input values) sample variance estimators s^2 and s_g^2 used in place of the respective output variances.

For the $M/M/1$ queue example, $x_{\min} = 3.6$ and $x_{\max} = 4.5$. To provide a reference for the fitted linear models, a "true" linear model was also computed based on fitting a linear model that minimizes the L_2 error from the true $M/M/1$ queue model given in Electronic Companion Section A.1 (since the true function $y(x)$ is not linear). The $\widehat{\text{Var}}(\hat{\beta}_1)$ metric estimated is based on a formula that assumes conditions not satisfied in these queueing examples. It is calculated to see if the relative behavior resembles the actual variances of the slope estimators, estimated using $N = 10$ macroreplications via

$$s^2(\hat{\beta}_1) = \frac{1}{N-1}\sum_{j=1}^{N}(\hat{\beta}_i^j - \bar{\hat{\beta}}_i)^2, \text{ where } \bar{\hat{\beta}}_1 = \frac{1}{N}\sum_{j=1}^{N}\hat{\beta}_1^j.$$

The first set of experiments performed used the sample mean for the output variable y_i (and g_i for DiGAR) at each of the 10 values of the independent variable x_i to carry out the fitting using standard regression and the various DIGAR models: OLS DiGAR with parameters given by (2.7) – denoted simply by DiGAR throughout; α-DiGAR models with parameters given by (2.10); MLE DiGAR with parameters given by (2.16) – denoted by DiGARn throughout; and the correlated DIGAR models with parameters given by (2.28) – denoted by DiGAR* and DiGAR** for the cases where the correlation matrix is estimated by using the given number of replications (initially 10) and 100,000 (off-line) replications, respectively.

The results are given in Table 2.1, which provide the values of the estimated parameters, along with the calculation of the various metrics. Overall, the DiGAR models all outperform the standard model on all metrics, especially for the variance of the slope estimators, with a superiority of one to two orders of magnitude. The numbers in the last two columns are also encouraging, in that they indicate that the estimated theoretical variances for all of the estimators are in the same ballpark as the sample variances, and more importantly the ratio of improvement is reasonably consistent, so that this metric appears to provide a pretty accurate estimate of relative performance between the various models. The performance of the DiGAR models is also fairly insensitive to the choice of the weight parameter, with the OLS DiGAR model performing adequately, although the “optimal” choice of weights (approximately 0.072, 0.087, 0.101, 0.114 for $y^{(2)}, y^{(3)}, y^{(4)}, y^{(5)}$, respectively) does show improvement and is usually the best performance overall, indistinguishable from DIGAR**, to be discussed shortly.

A clearer visual comparison is provided by the graphs in Figure 2.1, which plot the simulation data, true model, and three fitted models, where the circles are the data points (sample mean of 10 simulated values); the solid line is the true model; the dashed line is the fitted model from standard regression; the line with dots is the fitted OLS DiGAR model; and the dotted line is correlated DiGAR model. The graphs indicate that both the standard and DiGAR models fit the data reasonably well for $y^{(2)}$ and $y^{(3)}$, but there are dramatic differences for $y^{(4)}$, and $y^{(5)}$, in which the DiGAR models correctly capture the orientation of the curve, whereas the slope of the standard linear regression model has the incorrect sign. For all four cases, the normal estimators are indistinguishable from the basic OLS DiGAR model in the graphs, and hence were omitted.

Surprisingly, the correlated model clearly performs worse than the OLS model, which suggests that the misspecification caused by estimating the correlation based on 10 points outweighs the potential gains of a better model. The purpose of DiGAR** – which used 100,000 separate replications to estimate the covariance matrix

Table 2.1: Parameter estimates and performance metrics for $M/M/1$ queue (boxed entries indicate incorrect sign of slope)

i	$y^{(i)}$ model	$\hat{\beta}_1$	$\hat{\beta}_0$	L_2	$\widehat{\text{Var}}(\hat{\beta}_1)$	$s^2(\hat{\beta}_1)$
2	standard	1.68	-1.66	0.48	3.75	5.68
	DiGAR	1.56	-1.19	0.48	0.057	0.047
	DiGAR ($\alpha = 0.25$)	1.58	-1.26	0.48	0.046	0.029
	DiGAR ($\alpha = 0.75$)	1.56	-1.17	0.48	0.173	0.217
	DiGAR ($\alpha \propto 1/\sigma^2$)	1.55	-1.14	0.45	0.045	0.030
	DiGARn	1.55	-1.14	0.48	0.057	0.029
	DiGAR*	1.15	-0.11	1.62	1.99	0.036
	DiGAR**	1.53	-1.10	0.52	0.046	0.030
	"true" linear	1.69	-1.00	4E-6		
3	standard	0.68	4.04	0.27	8.82	7.17
	DiGAR	2.04	-1.50	0.11	0.134	0.066
	DiGAR ($\alpha = 0.25$)	1.86	-0.77	0.12	0.092	0.032
	DiGAR ($\alpha = 0.75$)	2.12	-1.80	0.11	0.369	0.300
	DiGARn	2.14	-1.91	0.11	0.123	0.029
	DiGAR ($\alpha \propto 1/\sigma^2$)	2.14	-1.81	0.051	0.090	0.029
	DiGAR*	1.53	-0.55	2.00	4.35	0.098
	DiGAR**	2.14	-1.85	0.088	0.091	0.028
	"true" linear	2.30	-2.18	2E-5		
4	standard	$\boxed{-2.97}$	20.2	2.06	3.10	5.35
	DiGAR	2.33	-1.26	0.019	0.237	0.074
	DiGAR ($\alpha = 0.25$)	1.63	1.58	0.093	0.046	0.041
	DiGAR ($\alpha = 0.75$)	2.61	-2.41	0.006	0.177	0.265
	DiGAR ($\alpha \propto 1/\sigma^2$)	2.71	-2.53	0.10	0.046	0.035
	DiGARn	2.70	-2.77	0.004	0.081	0.038
	DiGAR*	1.98	-1.56	2.52	0.122	0.110
	DiGAR**	2.73	-2.78	0.027	0.046	0.036
	"true" linear	2.85	-3.43	7E-5		
5	standard	$\boxed{-1.76}$	15.2	2.29	1.30	9.18
	DiGAR	2.89	-3.63	0.71	0.191	0.114
	DiGAR ($\alpha = 0.25$)	2.27	-1.14	0.77	0.053	0.068
	DiGAR ($\alpha = 0.75$)	3.14	-4.64	0.70	0.094	0.419
	DiGAR ($\alpha \propto 1/\sigma^2$)	3.22	-4.73	0.35	0.054	0.064
	DiGARn	3.19	-4.88	0.70	0.094	0.072
	DiGAR*	2.45	-2.98	3.66	0.228	0.096
	DiGAR**	3.24	-4.98	0.55	0.055	0.064
	"true" linear	3.37	-4.70	2E-4		

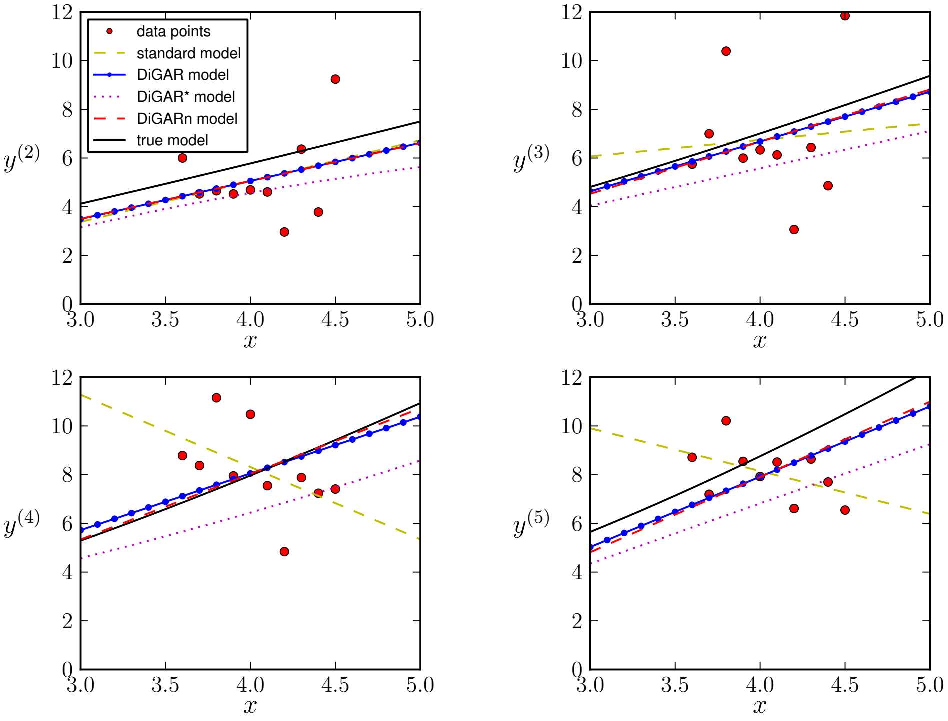


Figure 2.1: $M/M/1$ queue: true model, simulation data and several fitted models, where each data point is the sample mean of 10 independent replications.

to serve as a proxy for the true covariance matrix – was to investigate the conjecture that the error is in fact due to the poor estimation of the correlation matrix. Tables 2.1 indicate that DiGAR** does in fact outperform OLS DiGAR, though the difference is not substantial. Interestingly, as noted earlier, the performance of DIGAR** is nearly identical to that of α-DiGAR with the optimal choice of α. However, it is clear that at least for this example, it is preferable to use OLS DiGAR rather than DiGAR* if only a small number of replications are available to estimate the covariance matrix, which is typically the situation in most of the simulation settings.

Since the variances are much smaller in the DiGAR models, it would seem that a reasonable fit could be obtained using DiGAR with fewer replications at each design point. Thus, a followup experiment used only a *single* output point to measure y_i (and g_i for DiGAR) at each design point. Since there were 10 simulation runs, this experiment can be performed 10 independent times each. As expected, the variance of the slope estimator is substantially lower for DiGAR, and the ratios of the estimated variances range from 11 to 111, with an overall mean over 60, which represents a substantial improvement.

We also have an example showing the incorrect slope sign estimated from the standard regression model happens frequently if only a small number are conducted at each design point, specifically considering the fitted models for just $y^{(2)}$ in each of the 10 individual runs of data set 1. In half (5 out of 10) of the sample, the standard model gives the incorrect sign for the slope. The better match in the slope of the curves is critical in simulation-based optimization procedures such as sequential RSM. Similar results not reported here were also observed for $y^{(3)}$, $y^{(4)}$,

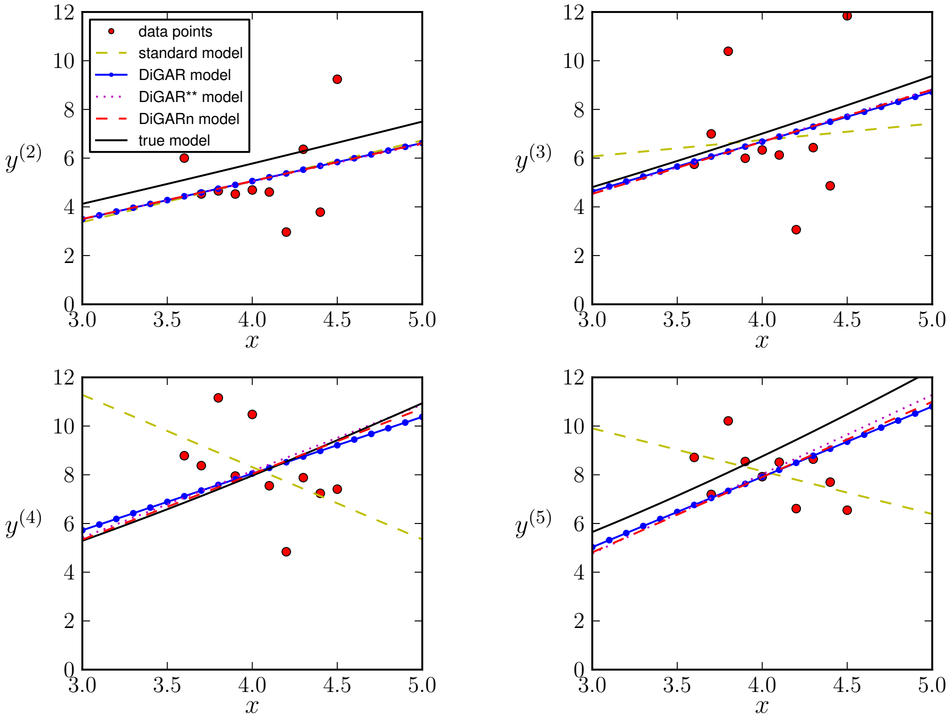


Figure 2.2: $M/M/1$ queue: using correlation matrix estimated based on 100,000 "offline" simulation replications.

and $y^{(5)}$.

2.3.2 Regression using a Quadratic Function

Sequential RSM generally employs a quadratic function to fit the data when the algorithm approaches the optimal value, i.e, the linear fit has slope close to 0. Using the same single-server queue as in Example 1, we consider the following objective function (used previously in many simulation optimization settings, e.g., [47, 48]):

$$y^{(k)}(x) = E[T_k] + c/x, \;\; k = 2, 3, 4, 5, \tag{2.30}$$

where c is a given constant. The additional term c/x in (2.30), which can be viewed as a cost on server speed, makes the function convex with a unique minimizer x^* that can be found analytically for comparison purposes.

The standard regression model using a quadratic function (in x) is given by $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, and the DiGAR model is given by adding $g_i = \beta_1 + 2\beta_2 x_i + \epsilon_i'$. The resulting explicit parameter estimators for the OLS DiGAR model are provided below.

Consider the loss function

$$L = \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_2 x_i^2 - \beta_1 x_i - \beta_0)^2 + \frac{1}{2}\sum_{i=1}^{n}(g_i - 2\beta_2 x_i - \beta_1)^2$$

Differentiating with respect to $\beta_0, \beta_1, \beta_2$,

$$\begin{aligned}
\frac{\partial L}{\partial \beta_0} &= -\sum_{i=1}^{n}(y_i - \beta_2 x_i^2 - \beta_1 x_i - \beta_0) \\
\frac{\partial L}{\partial \beta_1} &= -\sum_{i=1}^{n}(y_i - \beta_2 x_i^2 - \beta_1 x_i - \beta_0)x_i - \sum_{i=1}^{n}(g_i - 2\beta_2 x_i - \beta_1) \qquad (2.31) \\
\frac{\partial L}{\partial \beta_2} &= -\sum_{i=1}^{n}(y_i - \beta_2 x_i^2 - \beta_1 x_i - \beta_0)x_i^2 - \sum_{i=1}^{n}(g_i - 2\beta_2 x_i - \beta_1)2x_i
\end{aligned}$$

Setting them equal to 0 and solving yields the following estimators, $\boldsymbol{\beta} = \mathbf{A}\mathbf{y}$, where

$$\boldsymbol{\beta} \equiv [\beta_0, \beta_1, \beta_2]^T, \ \mathbf{y} = \left[\sum_{i=1}^{n} x_i^2 y_i + 2\sum_{i=1}^{n} x_i g_i, \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} g_i, \sum_{i=1}^{n} y_i\right]^T,$$

and

$$\mathbf{A} = \frac{1}{abc - af^2 - be^2 - cd^2 + 2def}\begin{bmatrix} bc - f^2 & ef - cd & df - be \\ ef - cd & ac - e^2 & de - af \\ df - be & de - af & ab - d^2 \end{bmatrix},$$

with $a = \sum_{i=1}^{n} x_i^4 + 4\sum_{i=1}^{n} x_i^2$, $b = \sum_{i=1}^{n} x_i^2$, $c = n$, $d = \sum_{i=1}^{n} x_i^3 + 2\sum_{i=1}^{n} x_i$, $e = \sum_{i=1}^{n} x_i^2$, $f = \sum_{i=1}^{n} x_i$.

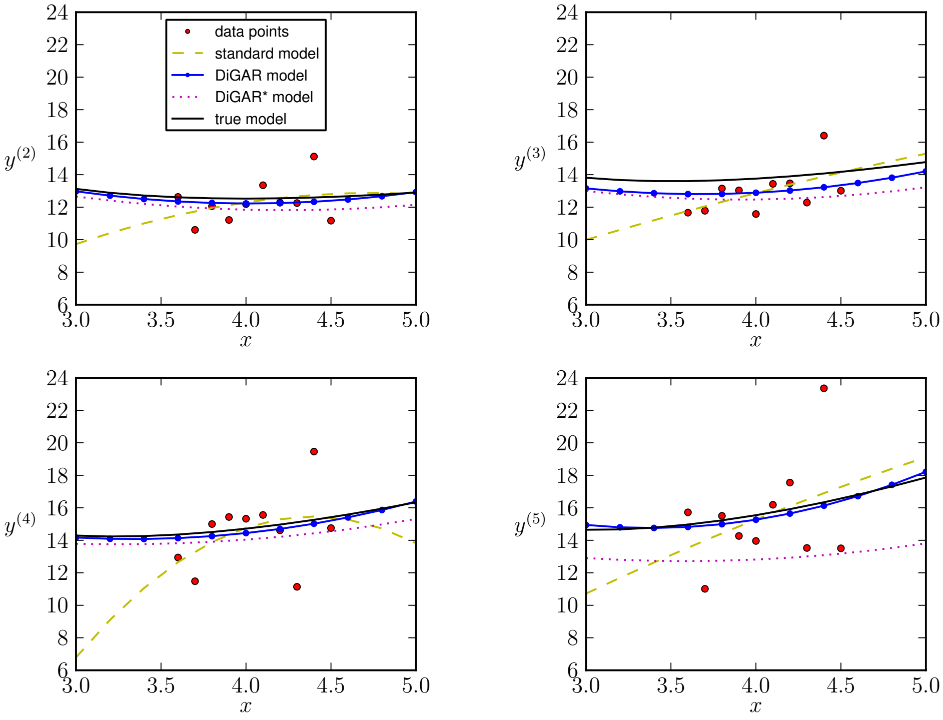


Figure 2.3: $M/M/1$ queue quadratic fit ($c \approx 27$, 10 replications).

Again using 10 replications per design point, the true and fitted models for $c \approx 27$ (actual value of c was chosen to make $x^* = 4$ for $y^{(2)}$), are plotted in Figure 2.3, including the correlated model indicated by DiGAR*, where the correlation matrix is also estimated using the small number of data points. The differences between both DiGAR models and standard regression are substantial, as the convexity/concavity curvature is often incorrect in the standard regression fitted model, e.g., for $y^{(4)}$ and $y^{(5)}$ in several cases. Once again the OLS DiGAR model outperforms the correlated version, again indicating the inadequacy of 10 replications for estimating the correlation matrix.

Two additional cases, $c = 1$ (where x^* lies to the left of the interval) and $c = 100$ (where x^* lies to the right of the interval), showing the fit of each of the models over a much wider range outside the set of design points, where the incorrect curvature of standard regression for several cases is more evident. Visually, it is clear that DiGAR fits the curve much better than standard regression in all three cases.

We also investigated the performance gains when more replications are carried out at each design point: 100, 1000, and 10000. indicate that the differences between both DiGAR models and standard regression can be substantial, even up to 1000 replications at each design point, where the standard model still has the incorrect curvature. Furthermore, even with 100 replications at each design point, the DiGAR* models still appear to be inferior to the OLS DiGAR models. Only when 1000 or 10000 replications are used at each design point do the DiGAR* models show better performance than the OLS DiGAR model, but the superiority is relatively slight (and not visually obvious in the figures), again indicating that the

Table 2.2: $M/M/1$ queue quadratic function optimal value x^* obtained as a function of # replications ($c \approx 27$),
where boxed entries indicate a maximum rather than a minimum

		10 reps			100 reps		
	true	standard	DiGAR	DiGAR*	standard	DiGAR	DiGAR*
$y^{(2)}$	4.0	$\boxed{4.8}$	4.0	4.2	$\boxed{4.0}$	4.1	4.1
$y^{(3)}$	3.5	$\boxed{9.9}$	3.7	3.9	$\boxed{4.0}$	3.1	3.5
$y^{(4)}$	3.4	$\boxed{4.4}$	3.3	3.3	$\boxed{4.0}$	1.8	3.1
$y^{(5)}$	3.1	$\boxed{9.8}$	3.4	3.6	$\boxed{4.1}$	1.0	2.8
		1000 reps			10000 reps		
	true	standard	DiGAR	DiGAR*	standard	DiGAR	DiGAR*
$y^{(2)}$	4.0	$\boxed{3.9}$	4.0	4.0	4.3	4.0	4.0
$y^{(3)}$	3.5	$\boxed{4.2}$	3.4	3.4	0.3	3.4	3.4
$y^{(4)}$	3.4	$\boxed{4.4}$	2.9	3.1	-25.2	3.0	3.0
$y^{(5)}$	3.1	$\boxed{4.4}$	2.5	2.8	-6.4	2.6	2.7

OLS DiGAR model may be sufficiently robust for applications such as sequential RSM.

Finally, Table 2.2 provides the estimates of x^* implied by the fitted functions as a function of the number of replications, where it is evident that the quadratic fit provided by standard regression is problematic for optimization purposes even with 10000 replications at each point; with less than 10000 replications per point, the curvature is always in the incorrect direction for the $c \approx 27$ case.

2.3.3 Multi-dimensional Example: $U/U/1$ Queue

Now we consider a multi-dimensional case using the same single-server queueing example. Specifically, we take both the interarrival times and service times to be i.i.d. following uniform distributions, i.e., a $U/U/1$ queue with respective distribution parameterizations $U(\theta_1 - \delta_1, \theta_1 + \delta_1)$ and $U(\theta_2 - \delta_2, \theta_2 + \delta_2)$, so that the input variable is the four-dimensional vector $x = (\theta_1, \theta_2, \delta_1, \delta_2)$. The output functions considered are the same four as in the previous example, i.e., the mean system time of the 2nd, 3rd, 4th, and 5th customers. Again using the Lindley equation, the true theoretical dependence of the expected system time on the input distributional parameters can be calculated analytically and are included in the Electronic Companion Section A.1.

The standard regression model is $y_i = \beta_0 + \beta_1\theta_1 + \beta_2\theta_2 + \beta_3\delta_1 + \beta_4\delta_2 + \epsilon_i$, and the DiGAR model adds $g_i^j = \beta_j + \epsilon_i^j$, where $g_i^j, j = 1, 2, 3, 4$ represents the derivative of y_i with respect to $\theta_1, \theta_2, \delta_1, \delta_2$, respectively. We simulated a two-level centered full-factorial design (thus, 17 design points), with center point $(\theta_1, \theta_2, \delta_1, \delta_2) = (10, 8, 8, 7)$, corresponding to $U(2, 18)$ and $U(1, 15)$ interarrival and service time distributions, respectively, with a spacing of ±0.1 (for all four dimensions) for the design points. The metrics used are analogous to the ones used previously: L_2 error and mean squared error (MSE), where the latter also takes into account the bias, i.e., MSE $= \mathrm{Var}(\hat{\beta}_i) + (\hat{\beta}_i - \beta_i^*)^2$, where β_i^* is the true value obtained using the analytical formula, and the sample mean and sample variance are calculated for each $\hat{\beta}_i^j,\ i = 1, 2, 3, 4,\ j = 1, 2, \cdots, N$, based on N macroreplications with an equal number of replications at each design point:

$$\bar{\hat{\beta}}_i = \frac{1}{N}\sum_{j=1}^{N}\hat{\beta}_i^j, \quad s_i^2 = \mathrm{Var}(\hat{\beta}_i) = \frac{1}{N-1}\sum_{j=1}^{N}(\hat{\beta}_i^j - \bar{\hat{\beta}}_i)^2.$$

For simplicity, only two DiGAR models are used: the basic one where no weights need to be determined, i.e., (2.23) with $\alpha_j = \alpha_0\ \forall j$; and α^*-DiGAR, for which the "optimal" (proportional to variance) weights are estimated offline using 10,000 replications. As before, the output at each design point is based on the mean of 10 replications, and the number of macroreplications is also $N = 10$. The results shown in Table 2.3 indicate that both DiGAR estimators have smaller MSE than the slope estimators from standard linear regression models by about two orders of magnitude, where α^*-DiGAR models reduce MSEs as well as variances further compared to the basic DiGAR models. In terms of L_2 errors, both DiGAR models are substantially better than standard linear regression model, but here the improvement achieved by using optimal weights is small. Note that as in the 1-D $M/M/1$ queue example, there are cases where standard linear regression model gives the incorrect sign for an estimated slope.

To visualize the results in terms of the slopes, boxplots for estimators $\hat{\beta}_i,\ i = 1, 2, 3, 4$, for the 2nd, 3rd, 4th and 5th customers are shown in Figure 2.4. Each boxplot is labelled as "`parameter:model`". For instance, "θ_1`:Linear Reg`" represents the boxplot for $\hat{\beta}_1$ estimators using the standard linear regression model. The true gradient values calculated from the analytical formulas are indicated in the boxplots by stars. The boxplots further illustrate that the variances of the estimators obtained from standard linear regression are significantly larger than variances of the estimators using the DiGAR models, differing by about two orders of magnitude. The differences between DiGAR and α^*-DiGAR models are not clear from the boxplots, but the sample standard deviations suggest that α^*-DiGAR models are able to reduce variances of the estimators by choosing optimal weights.

DiGAR and α^*-DiGAR models also provide better estimators in terms of absolute errors. The median values are indicated by red segments in all boxplots. Visually, the median values from the DiGAR models are indistinguishable from the true gradient values indicated by red stars, while median values from standard linear regression models are far away from the true gradient values in most cases. Compared to standard linear regression model, the sample means $\bar{\hat{\beta}}_i$ from DiGAR and α^*-DiGAR models are also much closer to the true gradient values.

The effect of increasing the number of replications at each design point is shown in Table 2.4, which provides experiment results for all slope estimators for $y^{(2)}$. As expected, all of the models show improvement, but the relative advantages of the DiGAR models – well over two orders of magnitude improvement in MSE and between one to two orders of magnitude improvement in L_2 – is retained.

2.3.4 Multi-dimensional Example: Sphere Function

Lastly, we consider a more stylized example to test the robustness of DiGAR models when the gradient estimates have much larger variances and for different levels of correlations between the response and gradient estimates. We consider the sphere function defined by $f(\mathbf{x}) = \sum_{i=1}^{n}(x_i)^2$, with partial derivative $\partial f(\mathbf{x})/\partial x_i =$

Table 2.3: Parameter estimates and performance metrics for $U/U/1$ queue (10 replications per design point),

based on 10 macroreplications (boxed entries indicate incorrect sign of slope)

			Linear Regression		DiGAR		α^*-DiGAR	
$y^{(2)}$		true value	value	MSE	value	MSE	value	MSE
	$\hat{\beta}_1$	-0.375	-2.498	26.1	-0.397	0.0024	-0.378	0.0009
	$\hat{\beta}_2$	1.375	$\boxed{-1.031}$	26.2	1.355	0.0030	1.377	0.0009
	$\hat{\beta}_3$	0.171	4.486	25.9	0.215	0.0036	0.175	0.0008
	$\hat{\beta}_4$	0.146	0.835	19.9	0.155	0.0038	0.149	0.0021
	L_2		8.31		0.141		0.139	
$y^{(3)}$	$\hat{\beta}_1$	-0.720	-2.140	22.7	-0.742	0.0048	-0.729	0.0068
	$\hat{\beta}_2$	1.720	$\boxed{-1.292}$	22.5	1.701	0.0078	1.728	0.0069
	$\hat{\beta}_3$	0.279	3.052	18.3	0.337	0.0060	0.312	0.0032
	$\hat{\beta}_4$	0.255	$\boxed{-0.164}$	18.7	0.233	0.0036	0.237	0.0014
	L_2		6.95		0.088		0.086	
$y^{(4)}$	$\hat{\beta}_1$	-1.065	-3.243	22.0	-1.029	0.0169	-1.009	0.0141
	$\hat{\beta}_2$	2.065	0.538	31.2	1.995	0.0203	2.008	0.0142
	$\hat{\beta}_3$	0.362	5.199	50.1	0.430	0.0108	0.386	0.0029
	$\hat{\beta}_4$	0.360	$\boxed{-0.319}$	26.7	0.305	0.0114	0.311	0.0069
	L_2		11.0		0.220		0.216	
$y^{(5)}$	$\hat{\beta}_1$	-1.427	-1.544	13.1	-1.309	0.0367	-1.307	0.0327
	$\hat{\beta}_2$	2.427	1.111	20.4	2.296	0.0396	2.306	0.0328
	$\hat{\beta}_3$	0.431	3.542	45.0	0.483	0.0059	0.455	0.0024
	$\hat{\beta}_4$	0.467	1.469	39.4	0.413	0.0099	0.403	0.0064
	L_2		10.0		0.174		0.171	

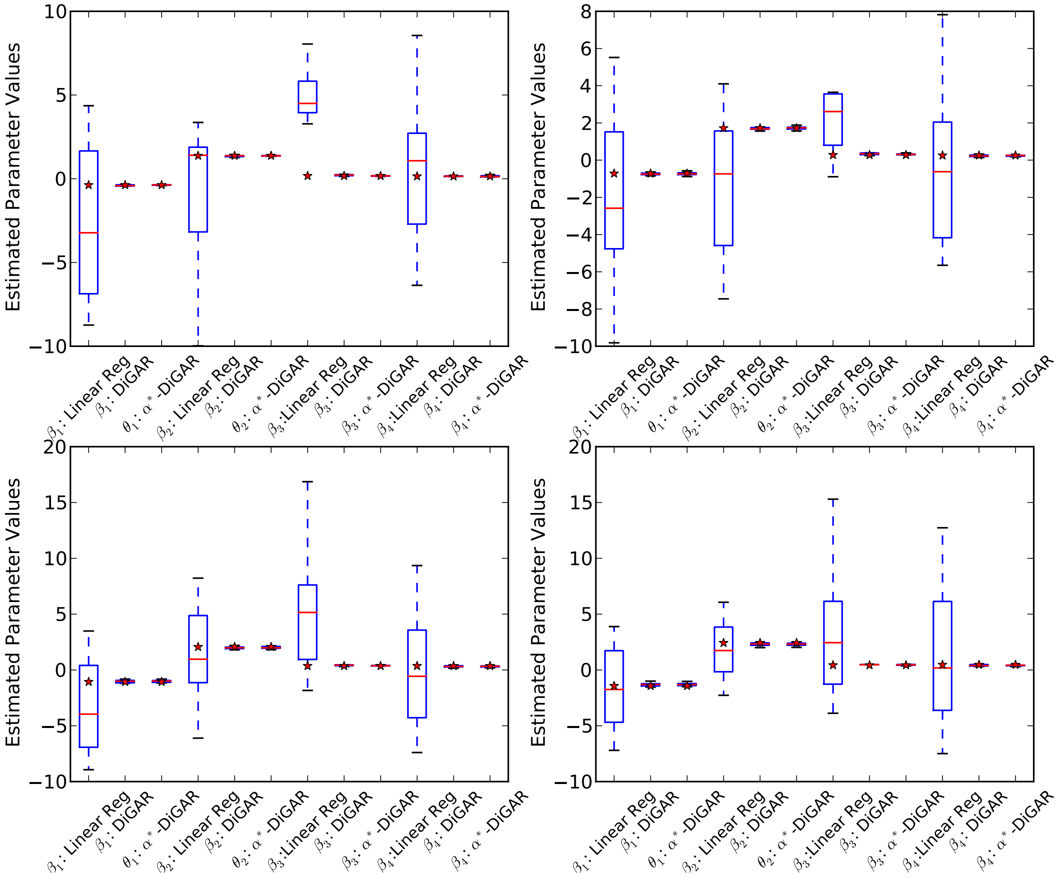


Figure 2.4: $U/U/1$ queue box plots of four estimated slopes for $y^{(k)}$ based on 10 macroreplications

Table 2.4: Parameter estimates and performance metrics for $U/U/1$ queue $y^{(2)}$ w.r.t. # replications/design point

			Linear Regression		DiGAR		α^*-DiGAR	
# reps		true value	value	MSE	value	MSE	value	MSE
1	$\hat{\beta}_1$	-0.375	-3.168	155	-0.397	0.025	-0.371	0.009
	$\hat{\beta}_2$	1.375	$\boxed{-9.977}$	270	1.265	0.035	1.370	0.009
	$\hat{\beta}_3$	0.171	6.012	141	0.236	0.019	0.182	0.007
	$\hat{\beta}_4$	0.146	$\boxed{-6.365}$	67.2	0.057	0.036	0.116	0.026
	L_2		54.0		1.227		1.213	
10	$\hat{\beta}_1$	-0.375	-2.498	26.1	-0.397	0.0024	-0.378	0.0009
	$\hat{\beta}_2$	1.375	$\boxed{-1.031}$	26.2	1.355	0.0030	1.377	0.0009
	$\hat{\beta}_3$	0.171	4.486	25.9	0.215	0.0036	0.175	0.0008
	$\hat{\beta}_4$	0.146	0.835	19.9	0.155	0.0038	0.149	0.0021
	L_2		8.31		0.141		0.139	
100	$\hat{\beta}_1$	-1.065	-0.914	3.08	-1.004	0.00434	-1.005	0.00422
	$\hat{\beta}_2$	2.065	2.074	2.39	2.006	0.00482	2.005	0.00424
	$\hat{\beta}_3$	0.362	0.996	4.84	0.379	0.00084	0.373	0.00026
	$\hat{\beta}_4$	0.360	$\boxed{-0.484}$	1.61	0.309	0.00278	0.316	0.00227
	L_2		1.03		0.036		0.037	
1000	$\hat{\beta}_1$	-0.375	-0.338	0.230	-0.377	0.00005	-0.378	0.00003
	$\hat{\beta}_2$	1.375	1.673	0.162	1.380	0.00005	1.378	0.00003
	$\hat{\beta}_3$	0.171	0.294	0.105	0.174	0.00002	0.173	0.00001
	$\hat{\beta}_4$	0.146	0.181	0.124	0.146	0.00002	0.145	0.00001
	L_2		0.0525		0.0009		0.0008	

$2x_i$. We used $n = 4$ in our numerical experiments, so the standard and DiGAR models are the same as in the previous example. To control the variances and correlations, the noises ϵ_i and $(\epsilon_i^1, \epsilon_i^2, \ldots, \epsilon_i^4)$ form a random vector in $\mathbb{R}^5$ following a multivariate normal distribution with mean 0 and covariance matrix $\mathbf{\Sigma}$. Four different levels of correlations are considered: the independent case ($\rho = 0$), and relatively low, medium, and high levels $\rho = 0.2, 0.5, 0.8$. The noise variances were set at $10, 20, 30, 40, 50$ for $\epsilon_i, \epsilon_i^1, \epsilon_i^2, \epsilon_i^3, \epsilon_i^4$, respectively. A two-level centered full factorial design with center point $(x^1, x^2, x^3, x^4) = (1.0, -0.6, 0.8, -0.5)$ was considered. Two sets of experiments were performed at two different grid sizes for the two-level design, namely 0.5 and 0.05, which correspond to cases where design points are spread out or fairly close. The mean at each design point was based on 10 replications, and 1000 macroreplications were used to estimate the MSE, with the results summarized in Tables 2.5 and 2.6, where DiGAR refers to the OLS DiGAR and α^*-DiGAR uses "optimal" (proportional to variance) weights.

In Table 2.5, results across different levels of correlations are consistent. The α^*-DiGAR model is the best among three models. Since the variances increase from ϵ_i^1 to ϵ_i^4, this leads to increases in MSE from $\hat{\beta}_1$ to $\hat{\beta}_4$ in both DiGAR models, but both DiGAR models still outperform the standard linear regression models and the α^*-DiGAR model reduces the MSE further from the OLS DiGAR model.

The results in Table 2.6 show higher MSEs, due to the more tightly spaced design points, which leads to larger variance in the estimators. For the standard model, the MSEs are about 100 times higher, whereas the DiGAR models are only two to three times higher, i.e., the relative advantage of the DiGAR models increases significantly in this setting, consistent with Proposition 2.5 and the remarks following it.

As expected, increasing the variances of gradient estimators leads to inflation of MSE in the DiGAR models; however, the relative advantage of the DiGAR models is retained. The effects of changing the level of correlation does not affect the performances of both DiGAR models under these two experiment designs as shown in Table 2.5 and 2.6. We also conducted experiments at three different levels of negative correlations. In this case, the MSE increases noticeably for all models with increasing magnitude of the level of correlation, but the relative performance of the models is similar to those for positive correlation, with the advantage of the DiGAR models even greater for the larger magnitudes.

2.4 Conclusion

In this chapter we proposed an augmented regression method that exploits the availability of direct gradient estimators in certain simulation settings. In some basic settings, we analytically characterized the improvement obtained over the standard model by calculating the variance of the estimated parameters and showing under which conditions guaranteed performance improvement can be expected. A simple queue was then used to numerically investigate the improvements, and the numerical results indicated great promise for the approach, with the general observation that the DiGARs models are qualitatively able to capture trends that the standard model might miss, e.g., there were several cases where the standard model gave the incorrect sign of the slope or was oriented in the opposite direction for the quadratic fit. Not surprisingly, the DiGAR slope estimators had much smaller variance than the standard estimator in all of the experiments conducted, retaining the relative advantage even in the higher-dimensional numerical examples. Although only a small portion of results are provided here, these observations hold for many other experiments for this simple queueing model.

Of particular note is the observation that in the queueing example where the gradient estimators are highly correlated with the dependent output data, the α-DiGAR models clearly outperform the standard regression models and for all practical purposes do as well as the correlated DiGAR models, and significantly better when the number of simulation replications at each design point is relatively small. Since the basic DiGAR estimators are quite easy to implement, this has practical implications for immediate use in those settings in which gradient estimators are available. Thus, we recommend using an α-DiGAR estimator in settings where the number of replications is relatively low and the gradient estimate is reasonably accurate, which is generally the case if IPA is applicable. However, more investigation into the effects of misspecification of the correlation structure is warranted before more conclusive statements can be made for the DiGAR* GLS estimators, and the effects are likely to be highly dependent on the application setting. Furthermore, using design of experiments, e.g., along the lines of [49], in choosing the design points could possibly ameliorate the misspecification problem.

Table 2.5: Parameter estimates and performance metrics for the sphere function for two-level centered full factorial design around $(1, -0.6, 0.8, -0.5)$; 10 replication per design point, gridsize 0.5, 1000 macroreplications.

			Linear Regression		DiGAR		α^*-DiGAR	
correlation		true value	value	MSE	value	MSE	value	MSE
zero $\rho = 0$	$\hat{\beta}_1$	2.0	2.001	0.249	2.007	0.093	2.006	0.088
	$\hat{\beta}_2$	-1.2	-1.238	0.259	-1.206	0.124	-1.215	0.109
	$\hat{\beta}_3$	1.6	1.589	0.265	1.615	0.174	1.605	0.127
	$\hat{\beta}_4$	-1.0	-1.016	0.265	-0.976	0.202	-0.993	0.136
low $\rho = 0.2$	$\hat{\beta}_1$	2.0	2.009	0.241	1.999	0.091	2.000	0.083
	$\hat{\beta}_2$	-1.2	-1.235	0.250	-1.200	0.132	-1.209	0.106
	$\hat{\beta}_3$	1.6	1.571	0.239	1.602	0.159	1.591	0.113
	$\hat{\beta}_4$	-1.0	-1.013	0.258	-1.002	0.207	-1.007	0.140
medium $\rho = 0.5$	$\hat{\beta}_1$	2.0	2.002	0.252	2.019	0.084	2.017	0.076
	$\hat{\beta}_2$	-1.2	-1.203	0.266	-1.192	0.127	-1.195	0.108
	$\hat{\beta}_3$	1.6	1.597	0.252	1.605	0.165	1.602	0.125
	$\hat{\beta}_4$	-1.0	-1.002	0.265	-1.000	0.202	-1.001	0.137
high $\rho = 0.8$	$\hat{\beta}_1$	2.0	2.016	0.246	1.980	0.093	1.985	0.087
	$\hat{\beta}_2$	-1.2	-1.191	0.249	-1.220	0.123	-1.212	0.104
	$\hat{\beta}_3$	1.6	1.612	0.230	1.584	0.163	1.594	0.120
	$\hat{\beta}_4$	-1.0	-1.017	0.262	-1.033	0.200	-1.026	0.128

Table 2.6: Parameter estimates and performance metrics for the sphere function for two-level centered full factorial design around $(1, -0.6, 0.8, -0.5)$; 10 replication per design point, gridsize 0.05, 1000 macroreplications.

			Linear Regression		DiGAR		α^*-DiGAR	
correlation		true value	value	MSE	value	MSE	value	MSE
zero $\rho = 0$	$\hat{\beta}_1$	2.0	2.070	23.660	2.007	0.116	2.007	0.115
	$\hat{\beta}_2$	-1.2	-1.395	24.683	-1.194	0.169	-1.195	0.169
	$\hat{\beta}_3$	1.6	1.584	24.645	1.607	0.248	1.607	0.247
	$\hat{\beta}_4$	-1.0	-1.099	24.780	-1.002	0.301	-1.003	0.300
low ρ = 0.2	$\hat{\beta}_1$	2.0	2.087	24.469	1.991	0.124	1.992	0.124
	$\hat{\beta}_2$	-1.2	-1.481	26.743	-1.201	0.182	-1.202	0.182
	$\hat{\beta}_3$	1.6	1.503	25.016	1.608	0.228	1.608	0.226
	$\hat{\beta}_4$	-1.0	-0.919	26.534	-0.991	0.311	-0.990	0.310
medium ρ = 0.5	$\hat{\beta}_1$	2.0	2.003	25.409	1.980	0.124	1.980	0.125
	$\hat{\beta}_2$	-1.2	-1.206	25.095	-1.232	0.177	-1.232	0.177
	$\hat{\beta}_3$	1.6	1.672	26.327	1.590	0.222	1.590	0.221
	$\hat{\beta}_4$	-1.0	-1.144	24.180	-1.011	0.290	-1.012	0.288
large ρ = 0.8	$\hat{\beta}_1$	2.0	1.839	24.427	1.979	0.116	1.979	0.116
	$\hat{\beta}_2$	-1.2	-1.367	25.998	-1.221	0.171	-1.222	0.170
	$\hat{\beta}_3$	1.6	1.494	25.830	1.563	0.229	1.563	0.227
	$\hat{\beta}_4$	-1.0	-1.052	24.366	-1.042	0.281	-1.043	0.278

Chapter 3: Gradient Extrapolated Stochastic Kriging

3.1 Introduction

Consider the same stochastic simulation setting as in Chapter 2, where direct gradient information are available. DiGAR model proposed in Chapter 2 incorporate gradient estimates into regression models by building a separate model for the gradient estimates. Aside from regression, kriging has also been studied extensively in the deterministic simulation community (see, for example, [23] and [50]). Stochastic kriging was introduced by [21] as an extension of kriging in the stochastic simulation setting. Stochastic kriging provides flexible metamodels of simulation output performance measurements while taking simulation noise into consideration.

[25] introduced stochastic kriging with gradient estimators (SKG) to exploit gradient estimates in stochastic kriging, showing that the new approach provides better prediction with smaller mean squared error (MSE). This approach is similar to cokriging proposed in deterministic simulations [26], and requires differentiability of the correlation functions, since derivatives of random processes or random fields are used to model gradient estimates.

We take a different approach to incorporate gradient estimates into stochastic kriging and investigate the potential improvements. A new approach called Gradient

Extrapolated Stochastic Kriging (GESK) is proposed, which extrapolates additional responses in the neighborhood of each design point using the original responses and gradient estimates. These additional responses, which might be biased, lead to better predictions than stochastic kriging if step sizes for extrapolations are chosen carefully. The main idea is to further explore the response surface with simulation responses and gradient estimates, so that a metamodel with better overall accuracy can be constructed. This suggests that GESK models are superior when there are limited number of design points or a response surface contains multiple extreme values.

To investigate the performance of GESK, we analyze the possible reduction in MSE of the GESK model over the standard stochastic kriging model, under two simplified and tractable settings. Conditions that guarantee reduction in MSE are provided as well. We also conduct numerical experiments to illustrate the effectiveness of the GESK model. Numerical results show that GESK performs comparably well or outperforms competing approaches such as stochastic kriging and SKG. Moreover, in certain cases, GESK captures fluctuations of the response surface that are usually missed by the other two approaches.

An important part of implementing the GESK model is the choice of step size. Large step sizes usually lead to large approximation errors and deteriorate prediction accuracy; small step sizes gain little information from extrapolations and might lead to numerical stability issues. We formalize two different strategies, penalized maximum likelihood estimation (PMLE) and minimizing integrated mean squared error (IMSE), to determine optimal step sizes. A cross validation method is presented to determine the regularization parameters required by each of the PMLE and IMSE approaches. We discuss pros and cons for each approach and compare them empirically with numerical examples.

In this section, we review stochastic kriging introduced in [21] and stochastic kriging with gradient estimators (SKG) introduced in [25], and then present the GESK approach.

3.1.1 Stochastic Kriging

Stochastic kriging was introduced in [21], focusing on modeling unknown response surfaces in stochastic simulation settings. Given an experiment design $\{(\mathbf{x}_i, n_i)\}$, $i = 1, 2, \cdots, k$, n_i simulation replications are run at each design point $\mathbf{x}_i$. Let $\mathcal{Y}_j(\mathbf{x}_i)$ be the simulation output from replication j at design point $\mathbf{x}_i$, $j = 1, \ldots, n_i$, and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^\intercal \in \mathbb{R}^d$. Stochastic kriging models the output as

$$\mathcal{Y}_j(\mathbf{x}_i) = \mathbf{f}(\mathbf{x}_i)^\intercal \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_i) + \epsilon_j(\mathbf{x}_i), \tag{3.1}$$

where $\mathbf{f}(\mathbf{x}_i) \in \mathbb{R}^p$ is a vector with known functions of $\mathbf{x}_i$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector with unknown parameters to be estimated. Components in $\mathbf{f}(\mathbf{x}_i)$ can be viewed as basis functions and a polynomial basis is usually adopted in the literature. The term $\mathbf{f}(\mathbf{x}_i)^\intercal \boldsymbol{\beta}$ represents the trend of the overall response surface. It is assumed that M is a realization of a zero-mean stationary random process (or random field) of the second order. This assumption is inherited from the deterministic kriging literature, where the stochastic nature of M is imposed on the problem so that statistical inference can be applied. For this reason, M is sometimes referred to as *extrinsic uncertainty*. This is contrasted with the term $\epsilon_j(\mathbf{x}_i)$, which is the simulation noise for replication j taken at $\mathbf{x}_i$. The uncertainty in $\epsilon_j(\mathbf{x}_i)$ comes from the nature of stochastic simulation, and it is sometimes referred to as *intrinsic uncertainty*.

Given the simulation responses $\{\mathcal{Y}_j(\mathbf{x}_i)\}_{j=1}^{n_i}$, $i = 1, 2, \ldots, k$, the sample mean of response output and simulation noise values at $\mathbf{x}_i$ are denoted by

$$\bar{\mathcal{Y}}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathcal{Y}_j(\mathbf{x}_i), \quad \bar{\epsilon}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i}\epsilon_j(\mathbf{x}_i). \tag{3.2}$$

The averaged responses $\bar{\mathcal{Y}}(\mathbf{x}_i)$ at $\mathbf{x}_i$ is modeled as

$$\bar{\mathcal{Y}}(\mathbf{x}_i) = \mathbf{f}(\mathbf{x}_i)^{\intercal}\boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_i) + \bar{\epsilon}(\mathbf{x}_i).$$

Suppose that we would like to predict the response $\mathsf{Y}(\mathbf{x}_0)$ at any point $\mathbf{x}_0$. Let $\bar{\mathcal{Y}} = \big(\bar{\mathcal{Y}}(\mathbf{x}_1), \bar{\mathcal{Y}}(\mathbf{x}_2), \ldots, \bar{\mathcal{Y}}(\mathbf{x}_k)\big)^{\intercal}$. Let $\boldsymbol{\Sigma}_{\mathsf{M}}$ be the $k \times k$ covariance matrix implied by the random field M and $\boldsymbol{\Sigma}_{\epsilon}$ be the $k \times k$ covariance matrix implied by the simulation noise across all design points $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}$. Let $\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ be the $k \times 1$ vector $(\mathrm{Cov}(\mathsf{M}(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_1)), \ldots, \mathrm{Cov}(\mathsf{M}(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_k)))^{\intercal}$, which represents spatial covariances between a prediction point $\mathbf{x}_0$ and all design points. Also, define the $k \times p$ design matrix $\mathbf{F}$ as $\mathbf{F} = (\mathbf{f}(\mathbf{x}_1), \mathbf{f}(\mathbf{x}_2), \ldots, \mathbf{f}(\mathbf{x}_k))^{\intercal}$. Suppose that $\boldsymbol{\Sigma}_{\mathsf{M}}$, $\boldsymbol{\Sigma}_{\epsilon}$ and $\boldsymbol{\beta}$ are known. Then the MSE-optimal predictor at $\mathbf{x}_0$ is of the form

$$\widehat{\mathsf{Y}}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)^{\intercal}\boldsymbol{\beta} + \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)^{\intercal}[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\epsilon}]^{-1}(\bar{\mathcal{Y}} - \mathbf{F}\boldsymbol{\beta}), \tag{3.3}$$

with corresponding MSE

$$\mathrm{MSE}\left(\widehat{\mathsf{Y}}(\mathbf{x}_0)\right) = \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_0) - \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)^{\intercal}\left[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\epsilon}\right]^{-1}\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot), \tag{3.4}$$

where $\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_0)$ is the spatial variance of the random field at $\mathbf{x}_0$. To build a stochastic kriging metamodel in practice requires imposing some structure on the spatial covariance matrix $\mathbf{\Sigma}_{\mathsf{M}}(\cdot, \cdot)$. It is usually assumed that the spatial covariance between $\mathsf{M}(\mathbf{x}_i)$ and $\mathsf{M}(\mathbf{x}_j)$ is

$$\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{Cov}\left[\mathsf{M}(\mathbf{x}_i), \mathsf{M}(\mathbf{x}_j)\right] = \tau^2 \mathcal{R}_{\mathsf{M}}(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}), \tag{3.5}$$

where τ^2 is the spatial variance of the random field and $\mathcal{R}_{\mathsf{M}}$ is a correlation function with parameter $\boldsymbol{\theta}$. The assumption that M is second-order stationary allows us to write $\mathcal{R}_{\mathsf{M}}(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = \mathcal{R}_{\mathsf{M}}(|\mathbf{x}_i - \mathbf{x}_j|; \boldsymbol{\theta})$, i.e., the correlation depends only on the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Common candidates for the correlation function include the triangular correlation function, the Gaussian correlation function and the Matérn correlation function, etc. See [51] for a detailed discussion on effects of using different correlation functions in stochastic kriging.

3.1.2 Stochastic Kriging With Gradient Estimators

We review the framework of stochastic kriging with gradient estimators (SKG) introduced by [25]. SKG builds stochastic kriging models for gradient estimators upon the stochastic kriging model for simulation responses. These two types of models are estimated together and applied to approximate response surfaces.

Suppose that we observe not only the simulation responses $\mathcal{Y}_j(\mathbf{x}_i)$, but also unbiased gradient estimates $\mathcal{G}_j(\mathbf{x}_i) \in \mathbb{R}^d$ for the jth simulation replication at design point $\mathbf{x}_i$. Given an experimental design $\{(\mathbf{x}_i, n_i)\}_{i=1}^k$, let the gradient estimate from replication j at design point $\mathbf{x}_i$ be $\mathcal{G}_j(\mathbf{x}_i) = \big(\mathcal{G}_j^1(\mathbf{x}_i), \ldots, \mathcal{G}_j^d(\mathbf{x}_i)\big)^{\mathsf{T}}$. In the SKG framework, each response $\mathcal{Y}_j(\mathbf{x}_i)$ is modeled the same as in stochastic kriging and each gradient estimator $\mathcal{G}_j^r(\mathbf{x}_i)$, $r = 1, \ldots, d$, is modeled as

$$\mathcal{G}_j^r(\mathbf{x}_i) = \left(\frac{\partial \mathbf{f}(\mathbf{x}_i)}{\partial x_{ir}}\right)^{\intercal} \boldsymbol{\beta} + \frac{\partial \mathsf{M}(\mathbf{x}_i)}{\partial x_{ir}} + \delta_j^r(\mathbf{x}_i). \tag{3.6}$$

This is valid under the following conditions:

- The function $\mathbf{f}(\mathbf{x}_i)$ is differentiable with respect to $\mathbf{x}_i$.
- The second-order mixed derivative of the correlation function $\mathcal{R}_{\mathsf{M}}$ in (3.5) exists and is continuous.

Let $\bar{\mathcal{G}}^r(\mathbf{x}_i)$ and $\bar{\delta}^r(\mathbf{x}_i)$, $r = 1, \ldots, d$, be the sample average of the gradient estimates and simulation noise, respectively, associated with $\mathbf{x}_i$:

$$\bar{\mathcal{G}}^r(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathcal{G}_j^r(\mathbf{x}_i), \quad \bar{\delta}^r(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} \delta_j^r(\mathbf{x}_i).$$

The SKG framework models the averaged simulation responses and gradient estimates as follows:

$$\begin{aligned}
\bar{\mathcal{Y}}(\mathbf{x}_i) &= \mathbf{f}(\mathbf{x}_i)^{\intercal}\boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_i) + \bar{\epsilon}(\mathbf{x}_i),\\
\bar{\mathcal{G}}^r(\mathbf{x}_i) &= \left(\frac{\partial \mathbf{f}(\mathbf{x}_i)}{\partial x_{ir}}\right)^{\intercal} \boldsymbol{\beta} + \frac{\partial \mathsf{M}(\mathbf{x}_i)}{\partial x_{ir}} + \bar{\delta}^r(\mathbf{x}_i).
\end{aligned}$$

To satisfy the conditions required for (3.6) to hold, a common choice for the correlation function is the Gaussian correlation function. Let $\boldsymbol{\Sigma}_{\mathsf{M}_+}$ be the variance-covariance matrix including spatial covariances induced by M, spatial covariances induced by derivatives of M and those between M and its partial derivatives. Let $\boldsymbol{\Sigma}_{\mathsf{M}_+}(\mathbf{x}_0, \cdot)$ be the vector analogous to $\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ in stochastic kriging. We assume replications across design points are independent. In addition, simulation noise ϵ_j

and δ_j are assumed to be independent from M. The covariance matrix $\mathbf{\Sigma}_{\epsilon_+}$ induced by simulation noise can be estimated by the sample covariances in practice.

Let $\bar{\mathcal{Y}}_+$ be the vector containing sample averages of response estimates and gradient estimates at all design points:

$$\bar{\mathcal{Y}}_+ = \left(\bar{\mathcal{Y}}(\mathbf{x}_1), \ldots, \bar{\mathcal{Y}}(\mathbf{x}_k), \bar{\mathcal{G}}^1(\mathbf{x}_1), \ldots, \bar{\mathcal{G}}^1(\mathbf{x}_k), \ldots, \bar{\mathcal{G}}^d(\mathbf{x}_1), \ldots, \bar{\mathcal{G}}^d(\mathbf{x}_k)\right)^{\mathsf{T}}.$$

The design matrix $\mathbf{F}$ in Section 2.1 now becomes $\mathbf{F}_+$, which can be written as

$$\mathbf{F}_+ = \left(\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_k), \left(\frac{\partial \mathbf{f}(\mathbf{x}_1)}{\partial x_{11}}\right), \ldots, \left(\frac{\partial \mathbf{f}(\mathbf{x}_k)}{\partial x_{k1}}\right), \ldots, \left(\frac{\partial \mathbf{f}(\mathbf{x}_1)}{\partial x_{1k}}\right), \ldots, \left(\frac{\partial \mathbf{f}(\mathbf{x}_k)}{\partial x_{kk}}\right)\right)^{\mathsf{T}}.$$

When $\boldsymbol{\beta}$ is known, the SKG predictor and the corresponding MSE can be obtained by substituting $\bar{\mathcal{Y}}_+, \mathbf{F}_+, \mathbf{\Sigma}_{\mathsf{M}_+}, \mathbf{\Sigma}_{\mathsf{M}_+}(\mathbf{x}_0, \cdot)$ and $\mathbf{\Sigma}_{\epsilon_+}$ for $\bar{\mathcal{Y}}, \mathbf{F}, \mathbf{\Sigma}_{\mathsf{M}}, \mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ and $\mathbf{\Sigma}_{\epsilon}$ in (3.3) and (3.4), respectively. Under some simplified settings, [25] shows that SKG can reduce MSE by incorporating gradient estimates. Numerical experiments also demonstrate the advantage of SKG in improving prediction performance over stochastic kriging.

3.2 Gradient Extrapolated Stochastic Kriging

We propose a different approach for incorporating the gradient estimates called Gradient Extrapolated Stochastic Kriging (GESK). Again, let $\mathcal{G}_j(\mathbf{x}_i) \in \mathbb{R}^d$ be the gradient estimator at $\mathbf{x}_i$ from replication j. Instead of modeling gradient estimates $\mathcal{G}_j(\mathbf{x}_i)$ as partial derivatives of the response surface, the gradient estimates are simply viewed as noisy measurements of the true gradient $\mathcal{G}(\mathbf{x}_i) \in \mathbb{R}^d$, i.e., $\mathcal{G}_j(\mathbf{x}_i) = \mathcal{G}(\mathbf{x}_i) + \boldsymbol{\delta}_j(\mathbf{x}_i)$, where $\{\boldsymbol{\delta}_j(\mathbf{x}_i)\}_{j=1}^{n_i}$ represent the zero-mean independent identically distributed noise across different replications at the design point $\mathbf{x}_i$. Denote the sample mean of gradient estimates at $\mathbf{x}_i$ by

$$\bar{\mathcal{G}}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathcal{G}_j(\mathbf{x}_i).$$

Notice that the response estimate $\mathcal{Y}_j(\mathbf{x}_i)$ and the gradient estimate $\mathcal{G}_j(\mathbf{x}_i)$ within the same replication j are generally correlated.

To incorporate gradient estimates into stochastic kriging, we *extrapolate* in the neighborhood of the original design points $\mathbf{x}_i$, $i = 1, 2, \cdots, k$. Specifically, linear extrapolation is used to obtain additional responses as follows:

$$\mathbf{x}_i^+ = \mathbf{x}_i + \Delta\mathbf{x}_i, \quad \mathcal{Y}_j(\mathbf{x}_i^+) = \mathcal{Y}_j(\mathbf{x}_i) + \mathcal{G}_j(\mathbf{x}_i)^\intercal \Delta\mathbf{x}_i, \tag{3.7}$$

where $\Delta\mathbf{x}_i = (\Delta x_{i1}, \Delta x_{i2}, \ldots, \Delta x_{id})^\intercal$, and the step size $\Delta\mathbf{x}_i$ needs to be small relative to the spacing of $\mathbf{x}_i$. For simplicity, we assume that only one additional point is added in the neighborhood of $\mathbf{x}_i$ and that the same step size is used for all design points, i.e., $\Delta\mathbf{x}_i = \Delta\mathbf{x}$, $i = 1, 2, \ldots, k$. Extensions include using more sophisticated extrapolation techniques and extrapolating multiple additional responses in the neighborhood of $\mathbf{x}_i$.

Let $\bar{\mathcal{Y}}(\mathbf{x}_i^+)$ be the sample average of these extrapolated response outputs, which is defined similarly as $\bar{\mathcal{Y}}(\mathbf{x}_i)$ in (3.2). For ease of notation, let $\bar{\mathcal{Y}}_i = \bar{\mathcal{Y}}(\mathbf{x}_i)$ and $\bar{\mathcal{Y}}_i^+ = \bar{\mathcal{Y}}(\mathbf{x}_i^+)$. Let $\bar{\mathcal{Y}}^+$ be the $2k \times 1$ vector containing both the original responses and the additional responses:

$$\bar{\mathcal{Y}}^+ = \left(\bar{\mathcal{Y}}_1, \bar{\mathcal{Y}}_2, \cdots, \bar{\mathcal{Y}}_k, \bar{\mathcal{Y}}_1^+, \bar{\mathcal{Y}}_2^+, \cdots, \bar{\mathcal{Y}}_k^+\right)^\intercal.$$

Similarly, $\mathbf{x}^+$ is defined as

$$\mathbf{x}^+ = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k, \mathbf{x}_1^+, \mathbf{x}_2^+, \cdots, \mathbf{x}_k^+)^\intercal.$$

The sample average of the additional responses $\bar{\mathcal{Y}}_i^+$ are modeled similarly to the original responses $\bar{\mathcal{Y}}_i$, i.e.,

$$\bar{\mathcal{Y}}_i^+ = \bar{\mathcal{Y}}(\mathbf{x}_i^+) = \mathbf{f}(\mathbf{x}_i^+)^\intercal \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_i^+) + \bar{\epsilon}(\mathbf{x}_i^+).$$

It is worth mentioning that this approach of incorporating gradient information is not restricted to stochastic kriging, but it is a general approach that can be applied to other metamodel approaches. The following assumptions are made:

Assumption 3.1.

1. *Simulations across design points are conducted independently, i.e., the use of common random numbers (CRN) is not considered.*
2. *For any design point $\mathbf{x}_i$, the noise $\epsilon_j(\mathbf{x}_i)$ are independent across replications.*
3. *The random field M is independent of all noise $\epsilon_j(\mathbf{x}_i)$ and $\epsilon_j(\mathbf{x}_i^+)$, for each design point $\mathbf{x}_i$ and replication j.*
4. *The simulation noise $\bar{\epsilon}(\mathbf{x}_l)$ is independent of $\bar{\epsilon}(\mathbf{x}_h^+)$ for $h \neq l$.*

[52] finds that using CRN in stochastic kriging inflates mean squared errors generally. Assuming independence across replications and independence between M and simulation noise is inherited from the stochastic kriging literature. The last assumption says that the original simulation response is correlated with its corresponding extrapolated response, but not other extrapolated responses.

Let $\boldsymbol{\Sigma}^+_{\mathsf{M}}$ be the $2k \times 2k$ variance-covariance matrix implied by the extrinsic spatial correlation model with $2k$ design points, including extrapolated design points:

$$\boldsymbol{\Sigma}^+_{\mathsf{M}} = \begin{pmatrix} \mathrm{Cov}[\mathsf{M}(\mathbf{x}_1), \mathsf{M}(\mathbf{x}_1)] & \cdots & \mathrm{Cov}[\mathsf{M}(\mathbf{x}_1), \mathsf{M}(\mathbf{x}_k)] & \mathrm{Cov}[\mathsf{M}(\mathbf{x}_1), \mathsf{M}(\mathbf{x}_1^+)] & \cdots & \mathrm{Cov}[\mathsf{M}(\mathbf{x}_1), \mathsf{M}(\mathbf{x}_k^+)] \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}[\mathsf{M}(\mathbf{x}_k), \mathsf{M}(\mathbf{x}_1)] & \cdots & \mathrm{Cov}[\mathsf{M}(\mathbf{x}_k), \mathsf{M}(\mathbf{x}_k)] & \mathrm{Cov}[\mathsf{M}(\mathbf{x}_k), \mathsf{M}(\mathbf{x}_1^+)] & \cdots & \mathrm{Cov}[\mathsf{M}(\mathbf{x}_k), \mathsf{M}(\mathbf{x}_k^+)] \\ & & & & & \\ \mathrm{Cov}[\mathsf{M}(\mathbf{x}_1^+), \mathsf{M}(\mathbf{x}_1)] & \cdots & \mathrm{Cov}[\mathsf{M}(\mathbf{x}_1^+), \mathsf{M}(\mathbf{x}_k)] & \mathrm{Cov}[\mathsf{M}(\mathbf{x}_1^+), \mathsf{M}(\mathbf{x}_1^+)] & \cdots & \mathrm{Cov}[\mathsf{M}(\mathbf{x}_1^+), \mathsf{M}(\mathbf{x}_k^+)] \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}[\mathsf{M}(\mathbf{x}_k^+), \mathsf{M}(\mathbf{x}_1)] & \cdots & \mathrm{Cov}[\mathsf{M}(\mathbf{x}_k^+), \mathsf{M}(\mathbf{x}_k)] & \mathrm{Cov}[\mathsf{M}(\mathbf{x}_k^+), \mathsf{M}(\mathbf{x}_1^+)] & \cdots & \mathrm{Cov}[\mathsf{M}(\mathbf{x}_k^+), \mathsf{M}(\mathbf{x}_k^+)] \end{pmatrix}$$

where each entry in $\boldsymbol{\Sigma}^+_{\mathsf{M}}$ can be computed by (3.5) with a given correlation function $\mathcal{R}_{\mathsf{M}}$ and spatial variance τ^2.

Let $\bar{\boldsymbol{\epsilon}}^+ \in \mathbb{R}^{2k}$ be the augmented vector of mean simulation noise:

$$\bar{\boldsymbol{\epsilon}}^+ = \left(\bar{\epsilon}(\mathbf{x}_1), \ldots, \bar{\epsilon}(\mathbf{x}_k), \bar{\epsilon}(\mathbf{x}_1^+), \ldots \bar{\epsilon}(\mathbf{x}_k^+)\right)^{\intercal}.$$

Under Assumption 3.1, let $\boldsymbol{\Sigma}^+_{\epsilon}$ be the $2k \times 2k$ variance-covariance matrix induced by $\bar{\boldsymbol{\epsilon}}^+$, which can be expressed as

$$\boldsymbol{\Sigma}^+_{\epsilon} = \begin{pmatrix} \mathrm{Var}[\bar{\epsilon}(\mathbf{x}_1)] & 0 & \ldots & 0 & \mathrm{Cov}[\bar{\epsilon}(\mathbf{x}_1), \bar{\epsilon}(\mathbf{x}_1^+)] & 0 & \ldots & 0 \\ 0 & \mathrm{Var}[\bar{\epsilon}(\mathbf{x}_2)] & \ldots & 0 & 0 & \ddots & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathrm{Var}[\bar{\epsilon}(\mathbf{x}_k)] & 0 & 0 & \ldots & \mathrm{Cov}[\bar{\epsilon}(\mathbf{x}_k), \bar{\epsilon}(\mathbf{x}_k^+)] \\ & & & & & & & \\ \mathrm{Cov}[\bar{\epsilon}(\mathbf{x}_1^+), \bar{\epsilon}(\mathbf{x}_1)] & 0 & \ldots & 0 & \mathrm{Var}[\bar{\epsilon}(\mathbf{x}_1^+)] & 0 & \ldots & 0 \\ 0 & \ddots & \ldots & 0 & 0 & \mathrm{Var}[\bar{\epsilon}(\mathbf{x}_1^+)] & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathrm{Cov}[\bar{\epsilon}(\mathbf{x}_k^+), \bar{\epsilon}(\mathbf{x}_k)] & 0 & 0 & \ldots & \mathrm{Var}[\bar{\epsilon}(\mathbf{x}_k^+)] \end{pmatrix}.$$

Let $\mathbf{x}_0$ be a prediction point and $\boldsymbol{\Sigma}^+_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ be a $2k \times 1$ vector

$$\boldsymbol{\Sigma}^+_{\mathsf{M}}(\mathbf{x}_0, \cdot) = \left(\mathrm{Cov}[\mathsf{M}(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_1)], \ldots, \mathrm{Cov}[\mathsf{M}(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_k^+)]\right)^{\intercal},$$

which represents spatial covariances between $\mathbf{x}_0$ and design points, including those extrapolated design points. The augmented design matrix $\mathbf{F}^+$ can be expressed as

$$\mathbf{F}^+ = \big(\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_k), \mathbf{f}(\mathbf{x}_1^+), \ldots, \mathbf{f}(\mathbf{x}_k^+)\big)^\intercal .$$

When $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\mathsf{M}}^+$ and $\boldsymbol{\Sigma}_{\epsilon}^+$ are known, the MSE-optimal predictor from the GESK model and its corresponding MSE can be constructed by substituting $\bar{\mathcal{Y}}^+$, $\mathbf{F}^+$, $\boldsymbol{\Sigma}_{\mathsf{M}}^+(\mathbf{x}_0, \cdot)$, $\boldsymbol{\Sigma}_{\mathsf{M}}^+$ and $\boldsymbol{\Sigma}_{\epsilon}^+$ for $\bar{\mathcal{Y}}$, $\mathbf{F}$, $\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)$, $\boldsymbol{\Sigma}_{\mathsf{M}}$ and $\boldsymbol{\Sigma}_{\epsilon}$ in (3.3) and (3.4), respectively.

In practice, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\mathsf{M}}^+$ and $\boldsymbol{\Sigma}_{\epsilon}^+$ are unknown and need to be estimated. The augmented matrix $\boldsymbol{\Sigma}_{\mathsf{M}}^+$ is characterized by the spatial variance τ^2 and correlation function with parameters $\boldsymbol{\theta}$. We assume that the simulation noise vectors $\boldsymbol{\epsilon}_j^+ = \big(\epsilon_j(\mathbf{x}_1), \ldots, \epsilon_j(\mathbf{x}_k), \epsilon_j(\mathbf{x}_1^+), \ldots \epsilon_j(\mathbf{x}_k^+)\big)^\intercal$ are multivariate normally distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}_{\epsilon}^+$. Given the assumption, we first estimate $\boldsymbol{\Sigma}_{\epsilon}^+$. Our approach to estimate $\mathrm{Var}[\bar{\epsilon}(\mathbf{x}_i)]$, $\mathrm{Var}[\bar{\epsilon}(\mathbf{x}_i^+)]$ and $\mathrm{Cov}[\bar{\epsilon}(\mathbf{x}_i), \bar{\epsilon}(\mathbf{x}_i^+)]$, $i = 1, 2, \ldots, k$ will be described in the following. Estimation of $\mathrm{Var}[\bar{\epsilon}(\mathbf{x}_i)]$ is

$$\widehat{\mathrm{Var}}[\bar{\epsilon}(\mathbf{x}_i)] = \frac{1}{n_i}\left[\frac{1}{n_i - 1}\sum_{j=1}^{n_i}\big(\mathcal{Y}_j(\mathbf{x}_i) - \bar{\mathcal{Y}}(\mathbf{x}_i)\big)^2\right].$$

Estimation for $\mathrm{Var}[\bar{\epsilon}(\mathbf{x}_i^+)]$ can be done in a similar fashion by replacing $\mathcal{Y}_j(\mathbf{x}_i)$ by $\mathcal{Y}_j(\mathbf{x}_i^+)$. The covariance $\mathrm{Cov}[\bar{\epsilon}(\mathbf{x}_i), \bar{\epsilon}(\mathbf{x}_i^+)]$ is estimated by the sample covariance as

$$\widehat{\mathrm{Cov}}[\bar{\epsilon}(\mathbf{x}_i), \bar{\epsilon}(\mathbf{x}_i^+)] = \frac{1}{n_i}\left[\frac{1}{n_i - 1}\sum_{j=1}^{n_i}\big(\mathcal{Y}_j(\mathbf{x}_i) - \bar{\mathcal{Y}}(\mathbf{x}_i)\big)\big(\mathcal{Y}_j(\mathbf{x}_i^+) - \bar{\mathcal{Y}}(\mathbf{x}_i^+)\big)\right].$$

This provides us an estimate $\widehat{\boldsymbol{\Sigma}}_{\epsilon}^+$ for $\boldsymbol{\Sigma}_{\epsilon}^+$. Combining this with normality assumptions, we can estimate the set of parameters $(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta})$ together using maximum likelihood estimators (MLEs) as described in [21].

Key to implementing the GESK model is the choice of step sizes for the extrapolated points, which depends on analyzing the potential improvements in performance from the GESK model as well as the approximation errors introduced by extrapolation. A good GESK model should take this bias-variance type tradeoff into consideration. We consider two tractable models: a two-point problem and a k-point problem with known model parameters. Under these two settings, we analyze the potential improvement in MSE by the GESK model over the stochastic kriging model, and provide conditions under which such improvement can be guaranteed.

In addition, we also analyze the effects of step sizes on MSE following the discussions of the two-point problem and the k-point problem in Sections 3.2.1 and 3.2.2. Understanding the effects of step sizes will provide insights for determining step sizes, which will be discussed later in detail in Section 3.3. In the following discussion, we continue to assume that the same step size is used for extrapolation at each design point and only one additional response is extrapolated in the neighborhood of each original design point.

The step size $\Delta\mathbf{x}$ determines the MSE of the GESK predictor through several factors: the biases ζ_i in the extrapolated responses; the correlation ρ_i between the simulation noise of original responses and extrapolated responses; and the variances $\sigma^2_{i_+}$ of the simulation noise in extrapolated responses. Since linear extrapolation is employed in the GESK model, the bias ζ_i in the extrapolated response $\bar{\mathcal{Y}}_i^+$ is bounded by $\mathcal{K}||\Delta\mathbf{x}||^2$ for some $\mathcal{K} > 0$. This suggests that the upper bound of biases can be controlled by choosing different step sizes. The correlation ρ_i depends on both the step size and the covariance between the simulation noise of the responses and those of the gradient estimators. A larger step size or smaller covariance leads to smaller correlation factor ρ_i, whereas $\sigma^2_{i_+}$ changes as the step size changes, but also depends on the sign of the correlation ρ_i. Effects of these factors will be discussed in detail in the following using the two-point problem and k-point problem as in Sections 3.2.1 and 3.2.2.

3.2.1 A Two-Point Problem with Single Extrapolated Point

Consider a one-dimensional problem ($d = 1$) of two design points x_1 and x_2 with numbers of replications n_1 and n_2, respectively. Without loss of generality, let $x_1 < x_2$ and the prediction point be $x_0 \in [x_1, x_2]$. The simulation outputs include responses $\{\mathcal{Y}_j(x_i)\}_{j=1}^{n_i}$ for $i = 1, 2$ at both design points and gradient estimators $\{\mathcal{G}_j(x_1)\}_{j=1}^{n_1}$ at x_1 only. A constant trend is used to represent the overall surface mean, i.e., $\mathbf{f}(x_i)^\intercal \boldsymbol{\beta} = \beta_0$. All parameters $(\beta_0, \tau^2, \theta)$ are assumed to be known.

Let the spatial variance $\tau^2 > 0$ and r_{il} be the correlation between $\mathsf{M}(x_i)$ and $\mathsf{M}(x_l)$, $i, l = 0, 1, \ldots, k$. The correlation r_{il} can be calculated from the correlation function $\mathcal{R}_{\mathsf{M}}(x_i, x_l; \theta)$, but no specific correlation function is specified in this discussion. Let the variance of the simulation noise at x_i from replication j be $\mathrm{Var}[\epsilon_j(x_i)] = \sigma_i^2$.

Let $\bar{\mathcal{Y}} = (\bar{\mathcal{Y}}_1, \bar{\mathcal{Y}}_2)^\intercal$ be the vector containing the sample means of responses at x_1 and x_2. The stochastic kriging predictor at x_0 is given as

$$\hat{\mathsf{Y}}(x_0) = \beta_0 + \tau^2 \frac{(r_1(\tau^2 + \frac{\sigma_2^2}{n_2}) - r_2\tau^2 r_{12})(\bar{\mathcal{Y}}_1 - \beta_0) + (r_2(\tau^2 + \frac{\sigma_1^2}{n_1}) - r_1\tau^2 r_{12})(\bar{\mathcal{Y}}_2 - \beta_0)}{(\tau^2 + \frac{\sigma_1^2}{n_1})(\tau^2 + \frac{\sigma_2^2}{n_2}) - \tau^4 r_{12}^2}, \tag{3.8}$$

with corresponding MSE

$$\text{MSE}\left(\hat{\mathsf{Y}}(x_0)\right) = \tau^2\left(1-\tau^2\frac{(r_{01}^2+r_{02}^2)\tau^2+\frac{r_{01}^2\sigma_2^2}{n_2}+\frac{r_{02}^2\sigma_1^2}{n_1}-2r_{01}r_{02}r_{12}\tau^2}{(\tau^2+\frac{\sigma_1^2}{n_1})(\tau^2+\frac{\sigma_2^2}{n_2})-\tau^4 r_{12}^2}\right). \tag{3.9}$$

With a pre-specified step size Δx, a new design point $x_1^+ = x_1 + \Delta x$ in the interval $[x_1, x_2]$ is added and GESK extrapolates its response as $\mathcal{Y}_j(x_1^+) = \mathcal{Y}_j(x_1) + \Delta x \mathcal{G}_j(x_1)$. This additional response output is modeled as $\mathcal{Y}_j(x_1^+) = \beta_0 + \mathsf{M}(x_1^+) + \epsilon_j(x_1^+)$. To address approximation error introduced by extrapolation, we assume that $\epsilon_j(x_1^+)$ is normally distributed with mean $\zeta_i = \zeta(x_i)$ and variance $\sigma^2_{1_+}$; thus, the extrapolated responses $\mathcal{Y}_j(x_1^+)$ at x_1^+ are biased unless $\zeta_i = 0$.

Let $\bar{\mathcal{Y}}_1^+$ be the sample mean of responses at x_1^+ and the vector $\bar{\mathcal{Y}}^+ = (\bar{\mathcal{Y}}_1, \bar{\mathcal{Y}}_2, \bar{\mathcal{Y}}_1^+)^\intercal$. Let ρ_1 be the correlation between $\bar{\epsilon}(x_1)$ and $\bar{\epsilon}(x_1^+)$ and r_{i1_+} be the correlation between $\mathsf{M}(x_i)$ and $\mathsf{M}(x_1^+)$ for $i = 0, 1, 2$. The variance-covariance matrix $\mathbf{\Sigma}^+ = \mathbf{\Sigma}^+_{\mathsf{M}} + \mathbf{\Sigma}^+_{\epsilon}$ takes the form

$$\mathbf{\Sigma}^+ = \tau^2\begin{pmatrix} 1 & r_{12} & r_{11_+} \\ r_{12} & 1 & r_{21_+} \\ r_{11_+} & r_{21_+} & 1 \end{pmatrix} + \begin{pmatrix} \frac{\sigma_1^2}{n_1} & 0 & \rho_1\frac{\sigma_1\sigma_{1_+}}{n_1} \\ 0 & \frac{\sigma_2^2}{n_2} & 0 \\ \rho_1\frac{\sigma_1\sigma_{1_+}}{n_1} & 0 & \frac{\sigma_{1+}^2}{n_1} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{b} \\ \mathbf{b}^\intercal & c \end{pmatrix},$$

where $\mathbf{\Sigma}$ is the 2×2 covariance matrix of the vector $(\bar{\mathcal{Y}}_1, \bar{\mathcal{Y}}_2)^\intercal$, $\mathbf{b}$ is a 2×1 vector and $c = \tau^2 + \sigma^2_{1+}/n_1$. The vector containing covariances between $\mathsf{M}(x_0)$ and $\left(\mathsf{M}(x_1), \mathsf{M}(x_2), \mathsf{M}(x_1^+)\right)^\intercal$ is

$$\mathbf{\Sigma}^+_{\mathsf{M}}(x_0, \cdot) = \tau^2\begin{pmatrix} r_{01} \\ r_{02} \\ r_{01_+} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}_{\mathsf{M}}(x_0, \cdot) \\ \tau^2 r_{01_+} \end{pmatrix}.$$

The new predictor at x_0 from the GESK model is

$$\widehat{\mathsf{Y}}^+(x_0) = \widehat{\mathsf{Y}}(x_0) + \frac{1}{v}\left[\mathbf{b}^\intercal \boldsymbol{\Sigma}^{-1}(\bar{\mathcal{Y}} - \beta_0 \mathbf{1}_2) - (\bar{\mathcal{Y}}_1^+ - \beta_0)\right]\left[\boldsymbol{\Sigma}_{\mathsf{M}}(x_0,\cdot)^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{b} - \tau^2 r_{01_+}\right], \tag{3.10}$$

where $\widehat{\mathsf{Y}}(x_0)$ is defined in (3.8) and $v = c - \mathbf{b}^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{b}$.

The following theorem provides an expression for MSE($\widehat{\mathsf{Y}}^+(x_0)$) and conditions under which the GESK predictor in (3.10) has smaller MSE than that in (3.8).

Theorem 3.1. *The MSE of the predictor in* (3.10) *can be expressed as*

$$MSE\left(\widehat{\mathsf{Y}}^+(x_0)\right) = MSE\left(\widehat{\mathsf{Y}}(x_0)\right) + \left(\frac{\zeta_1^2}{v^2} - \frac{1}{v}\right)\left[\boldsymbol{\Sigma}_{\mathsf{M}}(x_0,\cdot)^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{b} - \tau^2 r_{01_+}\right]^2, \tag{3.11}$$

and the GESK predictor has a smaller MSE if $\zeta_1^2 < v$.

Proof. The MSE of GESK predictor MSE $\left(\hat{\mathsf{Y}}^+(x_0)\right)$ follows from straightforward calculation and details are provided in the Appendix. Both $\boldsymbol{\Sigma}^+$ and $\boldsymbol{\Sigma}$ are variance-covariance matrices, and

$$\det\left(\boldsymbol{\Sigma}^+\right) = \det\left(\boldsymbol{\Sigma}\right)\det\left(c - \mathbf{b}^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{b}\right) = v\det\left(\boldsymbol{\Sigma}\right),$$

and it follows that $v > 0$ since both det($\boldsymbol{\Sigma}^+$) and det($\boldsymbol{\Sigma}$) are positive. Since $v > 0$, the condition $\zeta_1^2 < v$ is well defined. Under this condition, the GESK predictor has a smaller MSE than the stochastic kriging predictor. □

Theorem 3.1 provided the change in MSE at a prediction point x_0:

$$\Delta_{\text{MSE}} = \left(\frac{\zeta_1^2}{v^2} - \frac{1}{v}\right)\left[\boldsymbol{\Sigma}_{\mathsf{M}}(x_0,\cdot)^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{b} - \tau^2 r_{01_+}\right]^2.$$

We summarize our findings in this setting as follows:

1. The bias ζ_1 must satisfy $\zeta_1^2 < v$ as shown in Theorem 3.1 to guarantee reduction in MSE. Since ζ_1 is proportional to $(\Delta x)^2$, intuitively the step size should be relatively small.

2. The greater the correlation ρ_1, the greater the reduction in MSE. The parameter ρ_1 also depends on the correlation between $\epsilon_j(x_1)$ and $\delta_j(x_1)$, namely, the simulation noise of output responses and gradient estimators. The parameter ρ_1 increases as the correlation between $\epsilon_j(x_1)$ and $\delta_j(x_1)$ increases.

3. The parameter $\sigma_{1_+}^2$ represents the noise in an extrapolated response $\mathcal{Y}_j(x_1^+)$. The reduction in MSE is greater if $\sigma_{1_+}^2$ is smaller.

All conditions seem to prefer using smaller step sizes. However, other difficulties arise if the step sizes are too small: first, because the quantity v becomes smaller as Δx becomes smaller, the condition $\zeta_1^2 < v$ may not hold; second, as Δx approaches zero, the correlation ρ_1 approaches 1 and this may make the matrix $\mathbf{\Sigma}^+$ ill-conditioned, which leads to numerical issues.

3.2.2 A k-Point Problem

In this section, we consider a tractable problem with k design points, where $\mathbf{x}_i \in \mathbb{R}^d$, under the following assumptions:

1. Along with the response outputs $\mathcal{Y}_j(\mathbf{x}_i)$, gradient estimators $\mathcal{G}_j(\mathbf{x}_i)$ are also collected from simulations at design points $\{\mathbf{x}_i\}_{i=1}^k$.

2. Only one additional response is extrapolated in the neighborhood of each

design point.

3. The trend $\mathbf{f}(\mathbf{x}_i)^\intercal\boldsymbol{\beta} = \beta_0$ and all parameters $(\beta_0, \tau^2, \boldsymbol{\theta})$ are known.

Within the region of interest, an additional response $\mathcal{Y}_j^+(\mathbf{x}_i)$ is extrapolated using each pair of observations $(\mathcal{Y}_j(\mathbf{x}_i), \mathcal{G}_j(\mathbf{x}_i))$. All extrapolated design points should be in the interior of the design region; therefore, extrapolations from design points at the boundary should be done cautiously. Let $\bar{\mathcal{Y}}^+$ be the $2k \times 1$ vector that consists of sample means of all response outputs

$$\bar{\mathcal{Y}}^+ = (\bar{\mathcal{Y}}_1, \bar{\mathcal{Y}}_2, \ldots, \bar{\mathcal{Y}}_k, \bar{\mathcal{Y}}_1^+, \bar{\mathcal{Y}}_2^+, \ldots, \bar{\mathcal{Y}}_k^+)^\intercal,$$

where $\bar{\mathcal{Y}}_i = \bar{\mathcal{Y}}(\mathbf{x}_i)$ and $\bar{\mathcal{Y}}_i^+ = \bar{\mathcal{Y}}(\mathbf{x}_i^+)$. The sample mean of original responses at $\mathbf{x}_i$ are modeled as in Section 3.2. The sample mean of extrapolated responses are modeled similarly, i.e., $\bar{\mathcal{Y}}(\mathbf{x}_i^+) = \beta_0 + \mathsf{M}(\mathbf{x}_i^+) + \bar{\epsilon}(\mathbf{x}_i^+)$.

Let ρ_i denote the correlation between $\bar{\epsilon}(\mathbf{x}_i)$ and $\bar{\epsilon}(\mathbf{x}_i^+)$. The spatial correlations between original design points and extrapolated design points are denoted as $r_{il} = \mathrm{Corr}[\mathsf{M}(\mathbf{x}_i), \mathsf{M}(\mathbf{x}_l)]$, $r_{il_+} = \mathrm{Corr}[\mathsf{M}(\mathbf{x}_i), \mathsf{M}(\mathbf{x}_l^+)]$ and $r_{i_+l_+} = \mathrm{Corr}[\mathsf{M}(\mathbf{x}_i^+), \mathsf{M}(\mathbf{x}_l^+)]$ for $i, l = 1, 2, \ldots, k$. The $2k \times 2k$ variance-covariance matrix $\boldsymbol{\Sigma}^+ = \boldsymbol{\Sigma}_{\mathsf{M}}^+ + \boldsymbol{\Sigma}_{\epsilon}^+$ can be expressed in a block form

$$\boldsymbol{\Sigma}^+ = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{B} \\ \mathbf{B}^\intercal & \mathbf{C} \end{pmatrix}, \tag{3.12}$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \tau^2 + \sigma_1^2/n_1 & r_{12} & \cdots & r_{1k} \\ r_{12} & \tau^2 + \sigma_2^2/n_2 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1k} & r_{2k} & \cdots & \tau^2 + \sigma_k^2/n_k \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} \tau^2 r_{11_+} + \rho_1 \frac{\sigma_1 \sigma_{1_+}}{n_1} & \tau^2 r_{12_+} & \cdots & \tau^2 r_{1k_+} \\ \tau^2 r_{12_+} & \tau^2 r_{22_+} + \rho_2 \frac{\sigma_2 \sigma_{2_+}}{n_2} & \cdots & \tau^2 r_{2k_+} \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2 r_{1k_+} & \tau^2 r_{2k_+} & \cdots & \tau^2 r_{kk_+} + \rho_k \frac{\sigma_k \sigma_{k_+}}{n_k} \end{pmatrix},$$

$$\mathbf{C} = \begin{pmatrix} \tau^2 + \sigma_{1_+}^2/n_1 & \tau^2 r_{1_+2_+} & \cdots & \tau^2 r_{1_+k_+} \\ \tau^2 r_{1_+2_+} & \tau^2 + \sigma_{2_+}^2/n_2 & \cdots & \tau^2 r_{2_+k_+} \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2 r_{1_+k_+} & \tau^2 r_{2_+k_+} & \cdots & \tau^2 + \sigma_{k_+}^2/n_k \end{pmatrix}.$$

Given a prediction point $\mathbf{x}_0$, let $\mathbf{\Sigma}_{\mathsf{M}}^{+}(\mathbf{x}_0, \cdot)$ be a $2k \times 1$ vector that consists of the spatial covariances between $\mathbf{x}_0$ and all design points,

$$\begin{aligned} \mathbf{\Sigma}_{\mathsf{M}}^{+}(\mathbf{x}_0, \cdot) &= \big(\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_1), \ldots, \mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_k), \mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_1^{+}), \ldots, \mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_k^{+})\big)^{\intercal} \\ &= \big(\mathbf{\Sigma}_{\mathsf{M}}^{\intercal}(\mathbf{x}_0, \cdot) \quad \mathbf{\Sigma}_{\mathsf{M}^{+}}^{\intercal}(\mathbf{x}_0, \cdot)\big)^{\intercal}, \end{aligned}$$

where the both $\mathbf{\Sigma}_{\mathsf{M}^{+}}(\mathbf{x}_0, \cdot)$ and $\mathbf{\Sigma}_{\mathsf{M}^{+}}(\mathbf{x}_0, \cdot)$ are $k \times 1$ vectors.

As in the analysis of the two-point problem, an important issue to address is the approximation error introduced by extrapolation. Let the noise terms $\epsilon(\mathbf{x}_i^{+})$

at $\mathbf{x}_i^+$ follow normal distributions with means $\zeta_i = \zeta(\mathbf{x}_i)$, which implies that the additional response outputs $\mathcal{Y}_j(\mathbf{x}_i^+)$ are biased if $\zeta_i \neq 0$. We will analyze the effects of incorporating them in the following. Let the vector $\boldsymbol{\zeta} \in \mathbb{R}^{2k}$ be

$$\boldsymbol{\zeta} = (0, 0, \ldots, 0, \zeta_1, \zeta_2, \ldots, \zeta_k)^\intercal = (\mathbf{0}_k^\intercal \;\; \boldsymbol{\zeta}_k^\intercal)^\intercal\,,$$

which represents the expectation of the $2k \times 1$ noise vector $\bar{\boldsymbol{\epsilon}}^+$.

Let $\widehat{\mathsf{Y}}^+(\mathbf{x}_0)$ be the GESK predictor at $\mathbf{x}_0$. The MSE of the GESK predictor for this k-point problem is

$$\begin{aligned}\text{MSE}\left(\widehat{\mathsf{Y}}^+(\mathbf{x}_0)\right) =& \boldsymbol{\Sigma}_{\mathsf{M}}^+(\mathbf{x}_0,\mathbf{x}_0) - \boldsymbol{\Sigma}_{\mathsf{M}}^+(\mathbf{x}_0,\cdot)^\intercal \left[\boldsymbol{\Sigma}_{\mathsf{M}}^+ + \boldsymbol{\Sigma}_{\epsilon}^+\right]^{-1} \boldsymbol{\Sigma}_{\mathsf{M}}^+(\mathbf{x}_0,\cdot) \\ &+ \left(\boldsymbol{\Sigma}_{\mathsf{M}}^+(\mathbf{x}_0,\cdot)^\intercal \left[\boldsymbol{\Sigma}_{\mathsf{M}}^+ + \boldsymbol{\Sigma}_{\epsilon}^+\right]^{-1} \boldsymbol{\zeta}\right)^2 \\ =& \text{MSE}\left(\widehat{\mathsf{Y}}_{2k}(\mathbf{x}_0)\right) + \left(\boldsymbol{\Sigma}_{\mathsf{M}}^+(\mathbf{x}_0,\cdot)^\intercal \left[\boldsymbol{\Sigma}_{\mathsf{M}}^+ + \boldsymbol{\Sigma}_{\epsilon}^+\right]^{-1} \boldsymbol{\zeta}\right)^2.\end{aligned} \tag{3.13}$$

The first term $\text{MSE}\left(\widehat{\mathsf{Y}}_{2k}(\mathbf{x}_0)\right)$ is the MSE of prediction that one would obtain if unbiased responses are collected at $2k$ design points, namely, running simulations at $\mathbf{x}_i^+$ to collect response estimates rather than extrapolating additional response estimates. The second term is the inflation of MSE caused by approximation errors $\boldsymbol{\zeta}$ in the additional extrapolated responses.

Let $\widehat{\mathsf{Y}}(\mathbf{x}_0)$ be the stochastic kriging predictor with k design points. Our interest is to compare the MSE of the GESK predictor $\widehat{\mathsf{Y}}^+(\mathbf{x}_0)$ with that of $\widehat{\mathsf{Y}}(\mathbf{x}_0)$. To achieve this, we begin by looking into the MSE of $\widehat{\mathsf{Y}}_{2k}(\mathbf{x}_0)$.

Using the Woodbury matrix identity and block inverse formula in linear algebra, the MSE of $\widehat{\mathsf{Y}}_{2k}(\mathbf{x}_0)$ can be expressed as

$$\begin{aligned}\text{MSE}\left(\widehat{\mathsf{Y}}_{2k}(\mathbf{x}_0)\right) &= \boldsymbol{\Sigma}_{\mathsf{M}}^+(\mathbf{x}_0,\mathbf{x}_0) - \boldsymbol{\Sigma}_{\mathsf{M}}^+(\mathbf{x}_0,\cdot)^\intercal \left(\boldsymbol{\Sigma}^+\right)^{-1} \boldsymbol{\Sigma}_{\mathsf{M}}^+(\mathbf{x}_0,\cdot) \\ &= \text{MSE}\left(\widehat{\mathsf{Y}}(x_0)\right) - \boldsymbol{\omega}^\intercal \mathbf{V} \boldsymbol{\omega},\end{aligned} \tag{3.14}$$

where $\boldsymbol{\omega} = \mathbf{B}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot) - \boldsymbol{\Sigma}_{\mathsf{M}^+}(\mathbf{x}_0, \cdot)$ and $\mathbf{V} = \left(\mathbf{C} - \mathbf{B}^{\intercal}\boldsymbol{\Sigma}^{-1}\mathbf{B}\right)^{-1}$.

Lemma 3.2. *The matrix* $\mathbf{V} = \left(\mathbf{C} - \mathbf{B}^{\intercal}\boldsymbol{\Sigma}^{-1}\mathbf{B}\right)^{-1}$ *is positive definite.*

Proof. Consider the $2k \times 2k$ covariance matrix $\boldsymbol{\Sigma}^+$,

$$\boldsymbol{\Sigma}^+ = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{B} \\ \mathbf{B}^{\intercal} & \mathbf{C} \end{bmatrix}.$$

First it is easy to see that $\boldsymbol{\Sigma}^+$ is positive definite, so

$$\begin{bmatrix} \mathbf{u} & \mathbf{v} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{B} \\ \mathbf{B}^{\intercal} & \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} > 0,$$

for any $k \times 1$ vector $\mathbf{u}, \mathbf{v} \in \mathbb{R}^k$. This leads to

$$\mathbf{u}^{\intercal}\boldsymbol{\Sigma}\mathbf{u} + 2\mathbf{v}^{\intercal}\mathbf{B}^{\intercal}\mathbf{u} + \mathbf{v}^{\intercal}\mathbf{C}\mathbf{v} > 0.$$

For a fixed vector $\mathbf{v}$, consider $f(\mathbf{u}) = \mathbf{u}^{\intercal}\boldsymbol{\Sigma}\mathbf{u} + 2\mathbf{v}^{\intercal}\mathbf{B}^{\intercal}\mathbf{u} + \mathbf{v}^{\intercal}\mathbf{C}\mathbf{v}$ as a function of $\mathbf{u}$. The first-order condition shows that the minimum of $f(\mathbf{u})$ is

$$\min_{\mathbf{u}} f(\mathbf{u}) = \mathbf{v}^{\intercal}(\mathbf{C} - \mathbf{B}^{\intercal}\boldsymbol{\Sigma}^{-1}\mathbf{B})\mathbf{v},$$

which has to be positive for any $\mathbf{v} \in \mathbb{R}^k$. Therefore the matrix $\mathbf{C} - \mathbf{B}^{\intercal}\boldsymbol{\Sigma}^{-1}\mathbf{B}$ is positive definite and its inverse $\mathbf{V} = (\mathbf{C} - \mathbf{B}^{\intercal}\boldsymbol{\Sigma}^{-1}\mathbf{B})^{-1}$ is also positive definite. □

Since the matrix $\mathbf{V}$ is positive definite, it follows immediately that

$$\mathrm{MSE}\left(\widehat{\mathsf{Y}}_{2k}(\mathbf{x}_0)\right) = \mathrm{MSE}\left(\widehat{\mathsf{Y}}(x_0)\right) - \boldsymbol{\omega}^{\intercal}\mathbf{V}\boldsymbol{\omega} \leq \mathrm{MSE}\left(\widehat{\mathsf{Y}}(x_0)\right),$$

where equality only holds if and only if $\boldsymbol{\omega} = \mathbf{0}$. Thus, not surprisingly, the MSE is reduced if the k additional response outputs are unbiased.

Next we investigate the effect of the extrapolated bias on the overall MSE. Combining (3.13) and (3.14) gives

$$\begin{aligned}
\mathrm{MSE}\left(\widehat{\mathsf{Y}}^{+}(\mathbf{x}_0)\right) &= \mathrm{MSE}\left(\widehat{\mathsf{Y}}_{2k}(\mathbf{x}_0)\right) + \left(\boldsymbol{\Sigma}_{\mathsf{M}}^{+}(\mathbf{x}_0,\cdot)^{\intercal}\left[\boldsymbol{\Sigma}_{\mathsf{M}}^{+} + \boldsymbol{\Sigma}_{\epsilon}^{+}\right]^{-1}\boldsymbol{\zeta}\right)^2 \\
&= \mathrm{MSE}\left(\widehat{\mathsf{Y}}(\mathbf{x}_0)\right) - \boldsymbol{\omega}^{\intercal}\mathbf{V}\boldsymbol{\omega} + \left(\boldsymbol{\Sigma}_{\mathsf{M}}^{+}(\mathbf{x}_0,\cdot)^{\intercal}\left[\boldsymbol{\Sigma}_{\mathsf{M}}^{+} + \boldsymbol{\Sigma}_{\epsilon}^{+}\right]^{-1}\boldsymbol{\zeta}\right)^2 \\
&= \mathrm{MSE}\left(\widehat{\mathsf{Y}}(\mathbf{x}_0)\right) - \boldsymbol{\omega}^{\intercal}\mathbf{V}\boldsymbol{\omega} + \left(\left(\boldsymbol{\Sigma}_{\mathsf{M}^{+}}^{\intercal}(\mathbf{x}_0,\cdot)\mathbf{V} - \boldsymbol{\Sigma}_{\mathsf{M}}^{\intercal}(\mathbf{x}_0,\cdot)\boldsymbol{\Sigma}^{-1}\mathbf{B}\mathbf{V}\right)\boldsymbol{\zeta}_k\right)^2 \\
&= \mathrm{MSE}\left(\widehat{\mathsf{Y}}(\mathbf{x}_0)\right) - \boldsymbol{\omega}^{\intercal}\mathbf{V}\boldsymbol{\omega} + \left(\boldsymbol{\omega}^{\intercal}\mathbf{V}\boldsymbol{\zeta}_k\right)^2 \\
&= \mathrm{MSE}\left(\widehat{\mathsf{Y}}(\mathbf{x}_0)\right) + \boldsymbol{\omega}^{\intercal}\mathbf{V}\boldsymbol{\zeta}_k\boldsymbol{\omega}^{\intercal}\mathbf{V}\boldsymbol{\zeta}_k - \boldsymbol{\omega}^{\intercal}\mathbf{V}\boldsymbol{\omega} \\
&= \mathrm{MSE}\left(\widehat{\mathsf{Y}}(\mathbf{x}_0)\right) + \boldsymbol{\omega}^{\intercal}\mathbf{V}\boldsymbol{\zeta}_k\boldsymbol{\zeta}_k^{\intercal}\mathbf{V}\boldsymbol{\omega} - \boldsymbol{\omega}^{\intercal}\mathbf{V}\boldsymbol{\omega} \\
&= \mathrm{MSE}\left(\widehat{\mathsf{Y}}(\mathbf{x}_0)\right) + \boldsymbol{\omega}^{\intercal}\left(\mathbf{V}\boldsymbol{\zeta}_k\boldsymbol{\zeta}_k^{\intercal}\mathbf{V} - \mathbf{V}\right)\boldsymbol{\omega}.
\end{aligned} \tag{3.15}$$

The next theorem provides a sufficient condition under which $\mathrm{MSE}\left(\widehat{\mathsf{Y}}^{+}(\mathbf{x}_0)\right)$ is smaller than $\mathrm{MSE}\left(\widehat{\mathsf{Y}}(\mathbf{x}_0)\right)$.

Theorem 3.3. *Let $\lambda_i(\mathbf{A})$ denote the ith largest eigenvalue of matrix $\mathbf{A}$, $i = 1, 2, \ldots, k$. The symmetric matrix $\mathbf{W} = \mathbf{V}\boldsymbol{\zeta}_k\boldsymbol{\zeta}_k^{\intercal}\mathbf{V} - \mathbf{V}$ is negative definite if*

$$\boldsymbol{\zeta}_k^{\intercal}\boldsymbol{\zeta}_k \le \frac{\lambda_k(\mathbf{V})}{\left[\lambda_1(\mathbf{V})\right]^2}. \tag{3.16}$$

Proof. Using *Weyl's inequality* in matrix theory and Corollary 11 in [53], the largest eigenvalue $\lambda_1(\mathbf{W})$ of $\mathbf{W}$ satisfies

$$\begin{aligned}
\lambda_1(\mathbf{W}) &= \lambda_1\left(\mathbf{V}\boldsymbol{\zeta}_k\boldsymbol{\zeta}_k^{\intercal}\mathbf{V} - \mathbf{V}\right) \\
&\le \lambda_1\left(\mathbf{V}\boldsymbol{\zeta}_k\boldsymbol{\zeta}_k^{\intercal}\mathbf{V}\right) + \lambda_1(-\mathbf{V}) \\
&= \lambda_1\left(\mathbf{V}\boldsymbol{\zeta}_k\boldsymbol{\zeta}_k^{\intercal}\mathbf{V}\right) - \lambda_k(\mathbf{V}) \\
&\le \left[\lambda_1(\mathbf{V})\right]^2\lambda_1(\boldsymbol{\zeta}_k\boldsymbol{\zeta}_k^{\intercal}) - \lambda_k(\mathbf{V}).
\end{aligned}$$

The $k \times k$ matrix $\boldsymbol{\zeta}_k \boldsymbol{\zeta}_k^\intercal$ is known as a dyad, which has one positive eigenvalue $\boldsymbol{\zeta}_k^\intercal \boldsymbol{\zeta}_k$ and $k-1$ zero eigenvalues provided that $\boldsymbol{\zeta}_k \neq \mathbf{0}$. It follows that the largest eigenvalue of $\boldsymbol{\zeta}_k \boldsymbol{\zeta}_k^\intercal$ is $\lambda_1(\boldsymbol{\zeta}_k \boldsymbol{\zeta}_k^\intercal) = \boldsymbol{\zeta}_k^\intercal \boldsymbol{\zeta}_k$. Applying condition (3.16), we have $\lambda_1(\mathbf{W}) < 0$, namely, the largest eigenvalue of $\mathbf{W}$ is negative, so all eigenvalues of $\mathbf{W}$ are negative, and therefore the matrix $\mathbf{W}$ is negative definite. □

When the matrix $\mathbf{W}$ is negative definite, the quantity $\boldsymbol{\omega}^\intercal \mathbf{W} \boldsymbol{\omega}$ is always negative unless $\boldsymbol{\omega} = \mathbf{0}$, so the GESK model reduces MSE for the k-point problem if (3.16) holds.

Remark 3.4. *The condition in (3.16) is well defined, as the matrix* $\mathbf{V}$ *is shown to be positive definite in Lemma 3.2. Theorem 3.3 shows that when the biases are relatively small, the reduction in MSE from including the additional extrapolated points still exceeds the inflation in MSE introduced from the bias of the extrapolated points.*

In addition to assumptions in Section 3.2.2, we also assume that the k design points are widely spread such that the spatial correlation between design points is approximately 0. A similar assumption is used in [25] to isolate the impacts of incorporating gradient estimators from spatial covariances. This implies that the matrix $\boldsymbol{\Sigma}$ in (3.12) is a diagonal matrix. As the step size $\Delta\mathbf{x}$ is usually small, we assume the same property holds for $\mathbf{B}$ and $\mathbf{C}$ in (3.12) also. The change in MSE of the GESK predictor is

$$\Delta_{\text{MSE}} = \boldsymbol{\omega}^\intercal \left(\mathbf{V} \boldsymbol{\zeta}_k \boldsymbol{\zeta}_k^\intercal \mathbf{V} - \mathbf{V} \right) \boldsymbol{\omega},$$

where $\boldsymbol{\omega} = \mathbf{B}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot) - \boldsymbol{\Sigma}_{\mathsf{M}^{+}}(\mathbf{x}_0, \cdot)$ and $\mathbf{V} = \left(\mathbf{C} - \mathbf{B}^{\intercal}\boldsymbol{\Sigma}^{-1}\mathbf{B}\right)^{-1}$. The effects of $\Delta\mathbf{x}$, ρ_i and $\sigma^2_{i_+}$ are summarized as follows:

1. Theorem 3.3 suggests that the quantity $\boldsymbol{\zeta}_k^{\intercal}\boldsymbol{\zeta}_k$ needs to be small enough to guarantee that GESK can reduce MSE. This condition requires the step size Δx_j in each dimension to be small.

2. Regarding the correlation ρ_i, the preferable sign of the correlation actually depends on the location of the prediction point $\mathbf{x}_0$. A condition between $\Delta\mathbf{x}$ and $\mathbf{x}_i - \mathbf{x}_0$ determines the preferable sign of ρ_i. A specific analytical form for the condition depends on the type of correlation function, larger $|\rho_i|$ is better in each favorable case.

3. If the correlation $\rho_i \approx 0$, smaller $\sigma^2_{i_+}$ is preferable, since it suggests that there is less noise in the extrapolated responses. The same conclusion holds when the correlation ρ_i is positive. However, if the correlation $\rho_i < 0$, smaller $\sigma^2_{i_+}$ is not necessarily better, as there exists an optimal $\sigma^2_{i_+}$ that reduces MSE the most.

Analyzing the effects of step size on MSE in a general setting is more difficult, especially in multidimensional problems. For example, step size used in a multidimensional problem may be different along different directions. Choosing good step sizes is crucial for building the GESK models. In the next section, we propose two different approaches to determine the optimal step size.

3.3 Implementations of GESK

In this section, we focus on two important questions in the implementation of the GESK model: choosing step sizes and choosing gradient estimators. We provide two different techniques for determining steps sizes and discuss their pros and cons. We also make recommendations between the infinitesimal perturbation analysis (IPA) and the likelihood ratio/score function (LR/SF) techniques for gradient estimation.

As discussed in the previous section, central to building a GESK model is the determination of appropriate step sizes. A good choice of step size is crucial to the performance of the GESK model. Different step sizes, even with the same data set, may lead to dramatical performance differences of GESK models. The linear extrapolation used in GESK is only appropriate in a small neighborhood of the design points, so the step size cannot be too large. If the step size is too small, the additional points obtained from linear extrapolations provide little information and it may cause numerical stability issues.

Two natural choices for determining step sizes are maximum likelihood estimation (MLE) and minimizing integrated mean squared error (IMSE). However, the MLE approach is not suitable for determining step sizes in this context. The MLE approach leads to step sizes as small as possible, which results in numerical stability issues when building the GESK model. One unique characteristic of the GESK model is that biases are introduced during extrapolations. Although the biases are unknown, they should be taken into consideration during parameter estimations. To accomplish this, penalty terms are introduced in MLE and IMSE. We formalize two approaches for determining step sizes: penalized maximum likelihood estimation and minimizing integrated mean squared error.

3.3.1 Penalized Maximum Likelihood Estimation

One natural choice for determining the step size $\Delta\mathbf{x}$ is to treat it as a new parameter in addition to the other parameters $(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta})$. Under Assumption 1 in [21], we can write down the likelihood function. However, as mentioned earlier, naive MLE is not suitable for choosing step sizes in this case. Assuming the correlation function in (3.5) is used, we propose a penalized maximum likelihood method where the penalized likelihood function takes the following general form:

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}, \Delta\mathbf{x}) = &-\ln\left[(2\pi)^k\right] - \frac{1}{2}\ln\left[\left|\boldsymbol{\Sigma}_{\mathsf{M}}^{+} + \boldsymbol{\Sigma}_{\epsilon}^{+}\right|\right] \\ &- \frac{1}{2}(\bar{\mathcal{Y}}^{+} - \mathbf{F}^{+}\boldsymbol{\beta})^{\intercal}\left[\boldsymbol{\Sigma}_{\mathsf{M}}^{+} + \boldsymbol{\Sigma}_{\epsilon}^{+}\right]^{-1}(\bar{\mathcal{Y}}^{+} - \mathbf{F}^{+}\boldsymbol{\beta}) - p_{\lambda}(\Delta\mathbf{x}),\end{aligned}$$

where $p_\lambda(\cdot)$ is a given nonnegative penalty function with a regularization parameter λ. Common choices of penalty functions include L_1 penalty, L_2 penalty and smoothly clipped absolute deviation (SCAD).

The proposed penalty function is

$$p_{\lambda}(\Delta\mathbf{x}) = \lambda||\Delta\mathbf{x}||^{-2},$$

where $\Delta\mathbf{x} = (\Delta x_1, \Delta x_2, \ldots, \Delta x_d)$ and $||\Delta\mathbf{x}||^{-2} := \sum_{i=1}^{d}(\Delta x_i)^{-2}$. Therefore the proposed penalized likelihood function is

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}, \Delta\mathbf{x}) = &-\ln\left[(2\pi)^k\right] - \frac{1}{2}\ln\left[\left|\boldsymbol{\Sigma}_{\mathsf{M}}^{+} + \boldsymbol{\Sigma}_{\epsilon}^{+}\right|\right] \\ &- \frac{1}{2}(\bar{\mathcal{Y}}^{+} - \mathbf{F}^{+}\boldsymbol{\beta})^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\mathsf{M}}^{+} + \boldsymbol{\Sigma}_{\epsilon}^{+}\right]^{-1}(\bar{\mathcal{Y}}^{+} - \mathbf{F}^{+}\boldsymbol{\beta}) - \lambda||\Delta\mathbf{x}||^{-2}. \quad (3.17)\end{aligned}$$

Penalized maximum likelihood estimation (PMLE) has been used to do variable selection [54] and overcome flat likelihood function issues [55] for kriging. One key difference is that previous PMLE approaches try to improve the quality of estimates for $(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta})$, while we propose to use PMLE for choosing step sizes.

3.3.2 Minimizing Integrated MSE

Another view is to connect the problem of finding step sizes with design of experiments (DOE). Choosing step sizes is similar to adding new design points in DOE. In deterministic and stochastic kriging literature, many criteria have been proposed to find the "best" experiment design, most of which are based on MSE. Using integrated mean squared error (IMSE) as the objective function, the problem can be formulated as

$$\underset{\Delta\mathbf{x}}{\text{Minimize}} \quad \text{IMSE} = \int_{\mathbf{x}_0 \in \Omega} \text{MSE}^{+}\left(\widehat{\mathsf{Y}}(\mathbf{x}_0; \Delta\mathbf{x})\right) d\mathbf{x}_0, \quad (3.18)$$

where Ω is the region of interest.

Lower IMSE suggests smaller deviation associated with the approximation over the region of interest. In practice, a penalty term involving step sizes is added to MSE$^+$,

$$\text{MSE}^+(\widehat{\mathsf{Y}}(\mathbf{x}_0;\Delta\mathbf{x})) = \boldsymbol{\Sigma}^+_{\mathsf{M}}(\mathbf{x}_0,\mathbf{x}_0) - \boldsymbol{\Sigma}^+_{\mathsf{M}}(\mathbf{x}_0,\cdot)^\intercal[\boldsymbol{\Sigma}^+_{\mathsf{M}} + \boldsymbol{\Sigma}^+_{\epsilon}]^{-1}\boldsymbol{\Sigma}^+_{\mathsf{M}}(\mathbf{x}_0,\cdot) + \lambda||\Delta\mathbf{x}||^2, \tag{3.19}$$

where the Euclidean norm of $\Delta\mathbf{x}$ is used. Adding a penalty term that is proportional to $||\Delta\mathbf{x}||^2$ in MSE$^+$ follows the discussion earlier.

The PMLE approach estimates all the parameters $(\boldsymbol{\beta},\tau^2,\boldsymbol{\theta})$ with $\Delta\mathbf{x}$ simultaneously. However, the IMSE approach requires τ^2 and $\boldsymbol{\theta}$ to be known in advance. In practice, a two-stage strategy is proposed to address this issue:

1. In Stage 1, use the original dataset $\left\{\mathbf{x}_i, \bar{\mathcal{Y}}(\mathbf{x}_i)\right\}_{i=1}^k$ to obtain MLEs for $(\hat{\boldsymbol{\beta}},\hat{\tau}^2,\hat{\boldsymbol{\theta}})$.
2. Calculate MSE$^+\left(\widehat{\mathcal{Y}}(\mathbf{x}_0;\Delta x)\right)$ in (3.18) with the estimated $(\hat{\boldsymbol{\beta}},\hat{\tau}^2,\hat{\boldsymbol{\theta}})$ and a predetermined penalty constant λ.
3. In Stage 2, minimize the IMSE in (3.19) to find the optimal step size.

In our implementation, we use the optimization routine `fmincon` in Matlab.

3.3.3 Choosing Regularization Parameter

The question of selecting the regularization parameter λ in both approaches remains to be answered. We propose to use cross validation (CV), which is widely used in statistics and machine learning community, to choose the regularization parameters. Cross validation allows us to assess the performance with different regularization parameters without running additional simulations. When a J-fold cross validation is applied for a given regularization parameter λ, a corresponding score CV(λ) can be calculated. Given the design points $\{\mathbf{x}_i\}_{i=1}^k$ with the averaged simulation output $\{\bar{\mathcal{Y}}(\mathbf{x}_i), \bar{\mathcal{G}}(\mathbf{x}_i)\}_{i=1}^k$, the CV score is calculated as follows:

1. Split the dataset $\mathcal{D} = \{\mathbf{x}_i, \bar{\mathcal{Y}}(\mathbf{x}_i), \bar{\mathcal{G}}(\mathbf{x}_i)\}_{i=1}^k$ into J subsets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_J$. All the design points located on the boundary are handled separately so that extrapolated design points are still in the space of interest.

2. For $j = 1, 2, \ldots, J$, choose all the design points in $\mathcal{D}_j$ as prediction points and build a GESK model using $\mathcal{D}\backslash\mathcal{D}_j$. Predict the response $\widehat{\mathsf{Y}}(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{D}_j$.

3. Compute the CV score for a given parameter λ as a sum of squared errors between the prediction $\widehat{\mathsf{Y}}(\mathbf{x}_i)$ and the averaged output $\bar{\mathcal{Y}}(\mathbf{x}_i)$ on $\mathcal{D}_j$,

$$\mathrm{CV}(\lambda) = \sum_{j=1}^{J} \sum_{(\mathbf{x}_i, \bar{\mathcal{Y}}_i) \in \mathcal{D}_j} \left(\bar{\mathcal{Y}}_i - \widehat{\mathsf{Y}}(\mathbf{x}_i) \right)^2 .$$

To start the cross validation, we need to choose a set of regularization parameters $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_L\}$. We compute the CV score for each $\lambda_l \in \Lambda$ and choose the best regularization parameter λ^* as

$$\lambda^* = \underset{\lambda_l \in \Lambda}{\arg\min}\, \mathrm{CV}(\lambda_l).$$

Regarding the set of parameters Λ, if Λ contains a sufficiently large number of points and computational time is not an issue, cross validation will choose the best regularization parameter for building a GESK model. In practice, one way to choose the candidate parameter linearly on a logarithmic scale, for example. $10^{-1}, 10^0, \ldots, 10^3$. This is essentially an optimization problem. When the size of the set Λ is small, an exhaustive search algorithm can find λ^* easily. If the size of

Λ get large, randomized search algorithm can be applied as well. In our numerical experiments, CV score is computed for each parameter in Λ to find λ^*.

The performance of the two proposed approaches, PMLE and IMSE, will be investigated via numerical examples in Section 3.4 using different test problems. We summarize their main features and differences as follows:

- Both methods need a regularization parameter to be calibrated, as unknown biases are taken into consideration in both methods. We propose using cross validation methods to determine regularization parameters.
- PMLE uses a penalty function to overcome small step size issues in using naive MLE. The PMLE approach takes biases into consideration, but does not guarantee good performance in MSE or IMSE. However, in high-dimensional problems, maximizing a penalized likelihood is usually computationally faster than integrating MSE.
- IMSE minimizes the IMSE over the design region, but, in high-dimensional problems, numerical integrations become computationally expensive, requiring Monte Carlo methods with long computation time.

3.3.4 Choosing Gradient Estimators

In this chapter, we only consider direct gradient estimators. Specifically, we focus on the infinitesimal perturbation analysis (IPA) and likelihood ratio/score function (LR/SF) methods. Under mild conditions, both techniques are able to provide unbiased gradient estimators, but we would like to know which technique is preferable in building a GESK model, provided that both methods are applicable.

Observations made in [56] and [25] suggest the following: (i) at a given point, correlations between the responses and the corresponding gradient estimates are higher when IPA is applied, (ii) IPA gradient estimators usually have smaller variances than LR/SF estimators, (iii) IPA gradient estimators have better performance when applied in stochastic kriging with gradient estimators (SKG).

Discussions in Sections 3.2.1 and 3.2.2 suggests that the GESK model prefers gradient estimators that are highly correlated with response estimates and have smaller variances. Therefore, under most settings, IPA gradient estimators are preferable to build a GESK model. IPA gradient estimators are employed in the M/M/1 example conducted in Section 3.4.

3.4 NUMERICAL EXAMPLES

In this section, several numerical experiments are conducted to illustrate the proposed GESK model. Our goal in this section is three-fold: To demonstrate the effects of different step sizes on the performance of the GESK model; to empirically compare the effectiveness of the PMLE and IMSE approaches in determining step sizes; to examine the performance of the GESK model in different settings and compare it with stochastic kriging [21] and stochastic kriging with gradient estimators (SKG) [25]. Implementation of SKG and GESK are built upon software for stochastic kriging downloaded from `http://www.stochastickriging.net`.

Across all experiments, we assume little information is known about the response surface and choose constant trends for all models, i.e., $\mathbf{f}(\mathbf{x})^\intercal\boldsymbol{\beta} = \beta_0$. A Gaussian correlation function $R_{\mathsf{M}}(\mathbf{x}, \mathbf{x}') = \exp\{-\theta\|\mathbf{x} - \mathbf{x}'\|^2\}$ is used for all the experiments, since it satisfies the conditions required by SKG.

We implemented both PMLE and IMSE approaches discussed in Section **??** to determine step sizes, together with the cross validation method to choose regularization parameters. The corresponding GESK models are named `GESK-PMLE` and `GESK-IMSE`. The measure of performance we chose is the Empirical IMSE (EIMSE), as used in [57] and other kriging literature:

$$\text{EIMSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\mathsf{Y}}(\mathbf{x}_i) - \mathsf{Y}(\mathbf{x}_i)\right)^2, \tag{3.20}$$

where N is the number of predictions, $\widehat{\mathsf{Y}}(\mathbf{x}_i)$ is the predicted response at $\mathbf{x}_i$ and $\mathsf{Y}(\mathbf{x}_i)$ is the true value at $\mathbf{x}_i$.

3.4.1 Experiment on Step Sizes in GESK

We investigate the effects of using different step sizes in the GESK models using an M/M/1 queue example [58]. The M/M/1 queue has arrival rate 1 and service rate $x \in [1.1, 2]$. We are interested in the steady-state expected waiting time $y(x)$, which has an analytical solution $y(x) = 1/(x(x-1))$. In our simulation, each sample path was initialized in steady state and simulated for 5000 customers. The outputs collected were the average waiting time and its derivative with respect to the service rate x.

Six different experiment designs, $(6, 50)$, $(6, 200)$, $(6, 1000)$, $(8, 200)$, $(10, 200)$, $(20, 200)$, were used in the experiment, where the first element in each pair gives the

Design	GESK-1	GESK-2	GESK-3	GESK-PMLE	GESK-IMSE
$(6, 50)$	0.085 (0.0036)	0.167 (0.0035)	0.618 (0.0149)	0.042 (0.0020)	0.027 (0.0017)
$(6, 200)$	0.094 (0.0023)	0.181 (0.0019)	0.731 (0.0064)	0.034 (0.0011)	0.024 (0.0010)
$(6, 1000)$	0.092 (0.0011)	0.180 (0.0010)	0.747 (0.0037)	0.038 (0.0006)	0.021 (0.0004)
$(8, 200)$	0.006 (0.0004)	0.016 (0.0006)	0.194 (0.0023)	0.005 (0.0003)	0.002 (0.0002)
$(10, 200)$	0.006 (0.0005)	0.007 (0.0011)	0.017 (0.0012)	0.007 (0.0007)	0.004 (0.0003)
$(20, 200)$	0.002 (0.0008)	0.006 (0.0003)	0.048 (0.0020)	0.003 (0.0003)	0.001 (0.0001)

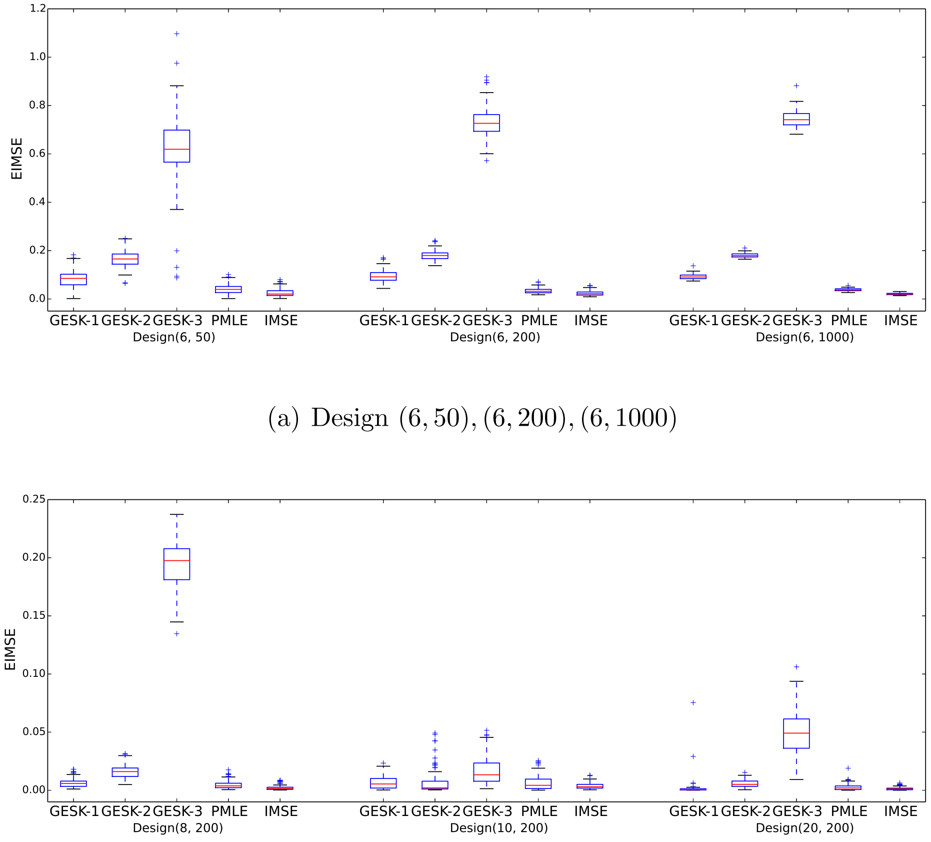
Table 3.1: Averaged EIMSE from 100 macroreplications for GESK models under six designs (# of design points, # of reps) to predict expected waiting time in the M/M/1 queue example. Three fixed step sizes with those determined by PMLE and IMSE are compared. Standard errors are shown in parenthesis.

number of design points and the second element represents the number of replications at each design point.

With equally spaced design points, three predetermined step sizes were chosen for each design, which correspond to 1/10, 1/5 and 1/2 of the length of the subinterval. GESK models built with these step sizes are labelled as `GESK-1`, `GESK-2` and `GESK-3`, respectively. We ran the experiments for 100 macroreplications. Within each macroreplication, we chose $N = 1000$ to estimate the EIMSE in (3.20). Table 3.1 shows the sample mean and standard errors of EIMSE, and Figure 3.1 contains boxplots for the EIMSE.

Our findings are summarized as follows:

- **Predetermined step sizes vs. Optimal step sizes**. Performances of



(a) Design $(6, 50), (6, 200), (6, 1000)$

(b) Design $(8, 200), (10, 200), (20, 200)$

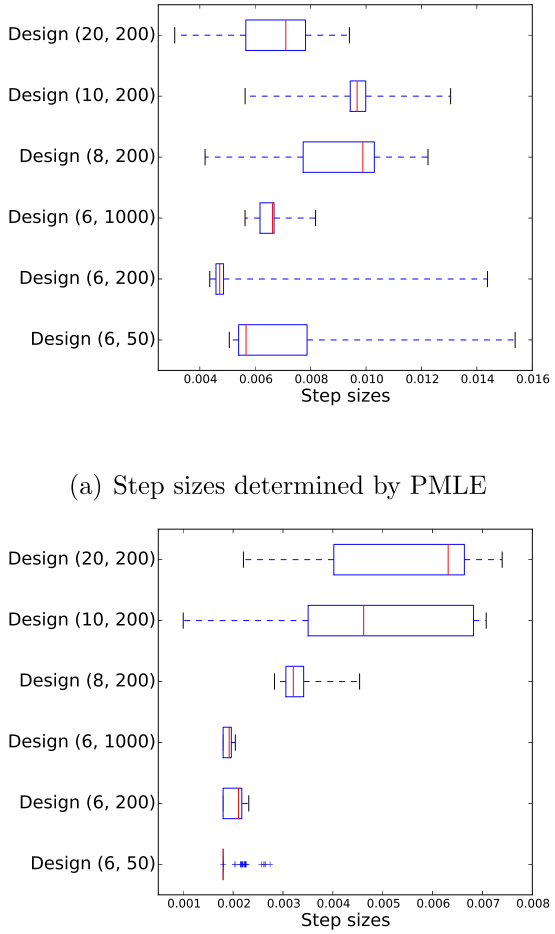
Figure 3.1: Boxplots of EIMSE from 100 macroreplications for the GESK models under six designs (# of design points, # of reps) to predict expected steady-state waiting time in the M/M/1 queue example.

the two optimal step sizes are better than those of predetermined step sizes, especially when the number of design points is small. This is expected, as the choice of step sizes should adapt to the experiment design and simulation output.

- **PMLE vs. IMSE**. The performance of IMSE is better than that of PMLE under most experiment designs, in terms of having smaller averaged EIMSE, smaller variances of EIMSE and smaller number of outliers. Figure 3.2 shows boxplots for step sizes determined by PMLE and IMSE under all six designs.

- **Effect of number of design points**. When the number of design points is small, for example $k = 6$, improvements in EIMSE are more significant. However, when there are already enough design points, improvements are hardly noticeable. In addition, for both PMLE and IMSE, the relative step size (ratio to the size of the subinterval) generally increases as the number of design points increases.

- **Effect of number of replications**. As the number of replications increases, the variances of EIMSE become smaller as shown in Table 3.1 and Figure 3.1. However, changes in the averaged IMSE are not significant. Variances of the chosen step sizes seem to decrease as well, as shown in Figure 3.2.



(a) Step sizes determined by PMLE

(b) Step sizes determined by IMSE

Figure 3.2: Boxplots for step sizes determined by PMLE and IMSE based on 100 macroreplications under six designs (# of design points, # of reps) to predict the expected steady-state waiting time in the M/M/1 queue example.

3.4.2 Comparisons among SK, SKG and GESK

In this section, we will compare the performances of three different metamodels: stochastic kriging, stochastic kriging with gradient estimators (SKG) and gradient extrapolated stochastic kriging (GESK) in three different experiments. The theoretical analysis of GESK in Section **??** assumes that all parameters are known, whereas the comparison here is empirical when all parameters must be estimated.

3.4.2.1 A stochastic simulation example

We used the same M/M/1 queue example as in Section 3.4.1. Six different experiment designs were adopted as well. We ran the experiments for 100 macroreplications. Within each macroreplication, we chose $N = 1000$ to estimate the EIMSE in (3.20). Results are shown in Table 3.2 and Figure 3.3. These 100 macroreplications used the same random numbers as those used in Section 3.4.1, so numbers for the two GESK models in Table 3.3 and corresponding boxplots in Figure 3.4 are the same as those in Section 3.4.1.

Our findings are summarized as follows:

- **SK vs. SKG vs. GESK**. Not surprisingly, SKG and GESK perform better than SK, as incorporating gradient estimators provides more information about the response surface. GESK-IMSE and SKG perform better than GESK-PMLE in most cases, since objective adopted by PMLE is not directly related to MSE. Both GESK models perform comparably well or better than the SKG model. It is not easy to distinguish GESK-IMSE and SKG, since

Design	SK	SKG	GESK-PMLE	GESK-IMSE
$(6, 50)$	0.313 (0.0134)	0.031 (0.0036)	0.042 (0.0020)	0.027 (0.0017)
$(6, 200)$	0.324 (0.0062)	0.016 (0.0007)	0.034 (0.0011)	0.024 (0.0010)
$(6, 1000)$	0.328 (0.0027)	0.016 (0.0003)	0.038 (0.0006)	0.021 (0.0004)
$(8, 200)$	0.054 (0.0019)	0.002 (0.0003)	0.005 (0.0003)	0.002 (0.0002)
$(10, 200)$	0.009 (0.0004)	0.004 (0.0014)	0.007 (0.0007)	0.004 (0.0003)
$(20, 200)$	0.004 (0.0002)	0.004 (0.0004)	0.003 (0.0003)	0.001 (0.0001)

Table 3.2: Averaged EIMSE from 100 macroreplications for SK, SKG and GESK with six different designs on estimating the expected steady-state waiting time in an M/M/1 queue problem. The design $(6, 50)$ means 6 design points with 50 replications at each design point. Standard errors are shown in parentheses.

their performances are really close.

- **Number of design points**. Incorporating gradient estimators improves performance considerably when the design points are sparse. For example, both SKG and GESK have more significant improvement over stochastic kriging when $k = 6$. As the number of design points increases, performance of most models improves.

- **Number of replications**. As the number of replications increases with a fixed number of design points, the variance of EIMSE decreases for all three methods, as shown in Figures 3.3(a). However, the averaged EIMSE does not improve significantly.

3.4.2.2 A stylized example with added noise

We consider a one-dimensional example from [23], where the true response surface is $\mathsf{Y}(x) = \exp(-1.4x)\cos(7\pi x/2)$ with $x \in [-2, 0]$. The presence of multiple local extreme values on the response surface makes building a good metamodel difficult. The simulation response output at x from replication j is $\mathcal{Y}_j(x) = \exp(-1.4x)\cos(7\pi x/2) + \epsilon_j(x)$, with $\epsilon_j(x) \sim \mathcal{N}(0, 1)$. Direct gradient estimates are assumed of the form $\mathcal{G}_j(x) = \mathsf{Y}'(x) + \delta_j(x)$ as the gradient estimate at x from simulation replication j, with $\delta_j(x) \sim \mathcal{N}(0, 25)$. We let $\delta_j(x)$ have a larger variance in order to empirically investigate the performances of SKG and GESK when gradient estimates are noisier.

We ran the experiments for 100 macroreplication. Within each macroreplica-

Design	`SK`	`SKG`	`GESK-PMLE`	`GESK-IMSE`
$(6, 50)$	39.616 (0.0374)	2.044 (0.0106)	1.909 (0.0181)	1.828 (0.0111)
$(6, 200)$	39.586 (0.0192)	2.023 (0.0049)	1.830 (0.0091)	1.757 (0.0050)
$(6, 1000)$	39.581 (0.0084)	7.652 (0.8823)	1.829 (0.0033)	1.758 (0.0023)
$(8, 200)$	2.793 (0.0039)	0.069 (0.0009)	0.063 (0.0026)	0.068 (0.0025)
$(10, 200)$	0.949 (0.0026)	1.243 (0.4204)	0.178 (0.0008)	0.012 (0.0007)
$(20, 200)$	0.008 (0.0002)	0.001 (0.0001)	0.046 (0.0004)	0.004 (0.0004)

Table 3.3: Averaged EIMSE from 100 macroreplications for SK, SKG and GESK with five different designs on $y(x) = \exp(-1.4x)\cos(7\pi x/2) + \epsilon$. Standard errors are shown in parentheses.

tion, we chose $N = 1000$ to estimate the EIMSE in (3.20). Six different experiment designs, $(6, 50)$, $(6, 200)$, $(6, 1000)$, $(8, 200)$, $(10, 200)$ and $(20, 200)$ were adopted, with results shown in Table 3.3 and Figure 3.4. Notice that ln(EIMSE) values are shown in Figure 3.4, as EIMSE results from the three models differ substantially.

- **SK vs. SKG vs. GESK**. As shown in Figure 3.4, both the SKG and the GESK models are better than SK when there is a limited number of design points. The GESK models perform better than SKG when $k = 6$. The explanation is that the response surface has several fluctuation and extrapolation allows GESK models to explore and approximate the response surface better than the others. SKG performs better when there are enough design point, for example, $k = 20$. SKG experiences numerical issues under designs $(6, 1000)$ and $(10, 200)$.

- **Number of replications**. When the number of replications increases, the variance of EIMSE decreases in general, as shown in Table 3.3 and Figure 3.4(a), except for SKG in design $(6, 1000)$. However, the averaged EIMSE does not change much as the number of replications increases, similar to the M/M/1 queue example.

- **Number of design points**. We fixed the number of replications at 200 and increases the number of design points up to 20. Boxplots are shown in Figures 3.4(b). EIMSE results for all models improve as the number of design points increases, with the exception of SKG and GESK-PMLE with design $(10, 200)$.

- **Step sizes**. Step sizes determined by the PMLE and IMSE approaches are shown in Figure 3.5. The plots suggest relationships between experiment designs and step sizes: (i) relative step size (ratio to the size of the subinterval) increases generally when the number of design points increases, (ii) the variability of step sizes decreases as the number of replications increases.

- **Remark**. Performance of SKG with design $(10, 200)$ shown in Table 3.3 and Figure 3.4(b) doesn't seem to match each other. The reason is that several outliers outside of the range shown in Figure 3.4(b) are omitted.

3.4.2.3 A multidimensional example

Lastly, we consider a stylized multidimensional example to test the performance of GESK models, especially when gradient estimates have much larger variances. We consider the sphere function defined by $\mathsf{Y}(\mathbf{x}) = \sum_{i=1}^{2} x_i^2 + \sum_{i=3}^{4} 10x_i^2$. We chose the experimental design space as $[-1, 1]^4$. The simulated response at $\mathbf{x} = (x_1, x_2, x_3, x_4)$ from replication j is $\mathcal{Y}_j(\mathbf{x}) = \mathsf{Y}(\mathbf{x}) + \epsilon_j(\mathbf{x})$ with the noise $\epsilon_j(\mathbf{x}) \sim \mathcal{N}(0, 1)$. The gradient estimate with respect to x_r at $\mathbf{x}$ from replication j is given by $\mathcal{G}_j^r(\mathbf{x}) = \frac{\partial \mathsf{Y}(\mathbf{x})}{\partial x_r} + \delta_j^r(\mathbf{x})$ with $\delta_j^r(\mathbf{x}) \sim \mathcal{N}(0, 25)$. The added noise terms are mutually independent.

We chose two different experiment designs: $(20, 500)$ and $(40, 500)$, which correspond to 20-point and 40-point Latin-hypercube designs with 500 independent replications at each design point, respectively. We collected simulation responses $\mathcal{Y}_j(\mathbf{x})$ and gradient estimates $\mathcal{G}_j^r(\mathbf{x})$ for $r = 1, 2, 3, 4$, $j = 1, 2, \ldots, 500$ to build metamodels.

We ran the experiments for 100 macroreplications. Within each replication, we chose $N = 1000$ to estimate the EIMSE in (3.20). Figure 3.6 contains boxplots for the EIMSE calculated from the 100 macroreplications. Our findings are summarized as follows:

- SKG and both GESK models perform better than SK. As the number of design points increases, the performances of all models improve. Under design (20500), GESK-IMSE seems to be the best; under design $(40, 500)$, SKG is preferred due to its low average and low variance in EIMSE.

- Between the two GESK models, PMLE scales better than IMSE for high-dimensional problems. The IMSE approach requires multidimensional integrations to determine step sizes, which is expensive and depends on the accuracy of integration approximation in high-dimensional problems.

- Step sizes determined by PMLE are generally much larger than those determined by IMSE. Along each dimension, step sizes determined by PMLE and IMSE have similar behavior. If the step size chosen by PMLE is relatively smaller on one dimension, so is the step size chosen by IMSE. Step sizes chosen for a dimension with higher gradient values are not necessarily smaller than others.

3.5 CONCLUSIONS AND FUTURE RESEARCH

In this paper we investigated gradient extrapolated stochastic kriging (GESK), which exploits the availability of direct gradient estimates in stochastic simulation settings. The performance of the GESK models was analyzed theoretically and numerically, with a focus on analyzing the approximation errors introduced by extrapolation. Since step sizes are crucial to GESK models, two methods for determining step sizes were proposed and tested in numerical examples, which indicated substantial gains in performance over SK in all of the experiments. Between the proposed PMLE and IMSE approaches for determining step sizes, IMSE demonstrated better performance in numerical experiments, but it becomes computationally expensive for high-dimensional problems. The numerical experiments showed comparable performance for GESK and SKG, except when the number of design points is very small, where GESK shows some advantage.

From our analysis and numerical experiments, we offer the following overall conclusions:

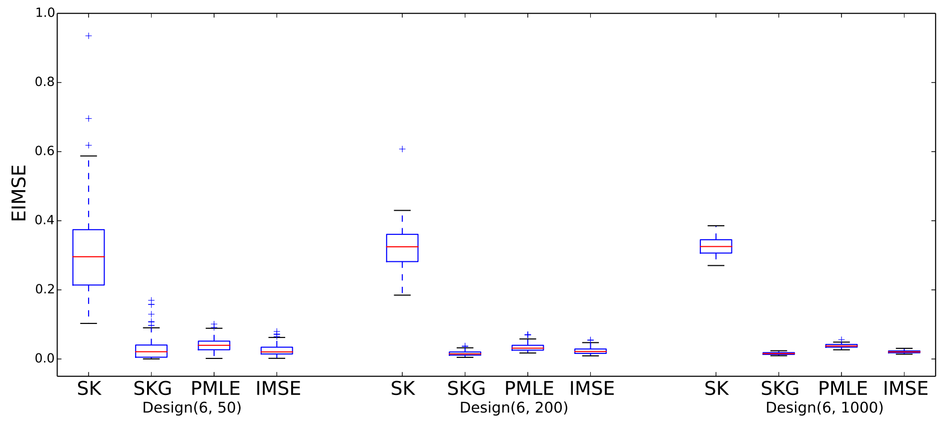
- GESK can be especially effective when the number of design points is relatively small, e.g., in the setting where simulation is expensive.

- GESK offers additional flexibility in choosing the correlation function; in particular, differentiability is not a constraint.

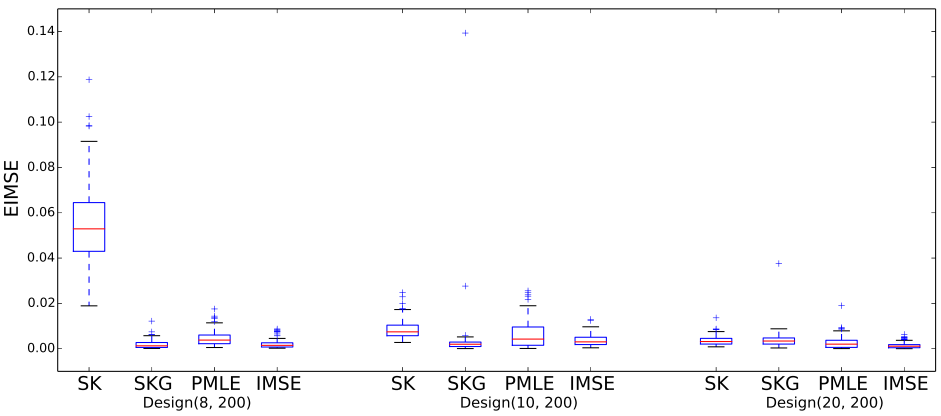
- For high-dimensional problems, GESK using the PMLE is recommended, since its computation does not increase with dimension, whereas the computational burden increases exponentially for IMSE minimization and at least quadratically for SKG.

Our work points to several other directions for future research. The first direction is to focus on the extrapolation strategy in GESK. For this paper, we use linear extrapolation with the same step size and assume that only one additional point is extrapolated from each design point. More sophisticated techniques could use the local response surface information and adaptively determine the extrapolation strategy. This is especially important in higher-dimensional problems with multiple extreme values.

Another direction is to further investigate a comparison of the SKG and GESK models. Improvements from incorporating gradient estimates can be expected from both models. However, it would be valuable to be able to characterize when one model is likely to be more effective. A theoretical analysis of various properties comparing the two models can lead to useful guidelines for practitioners.

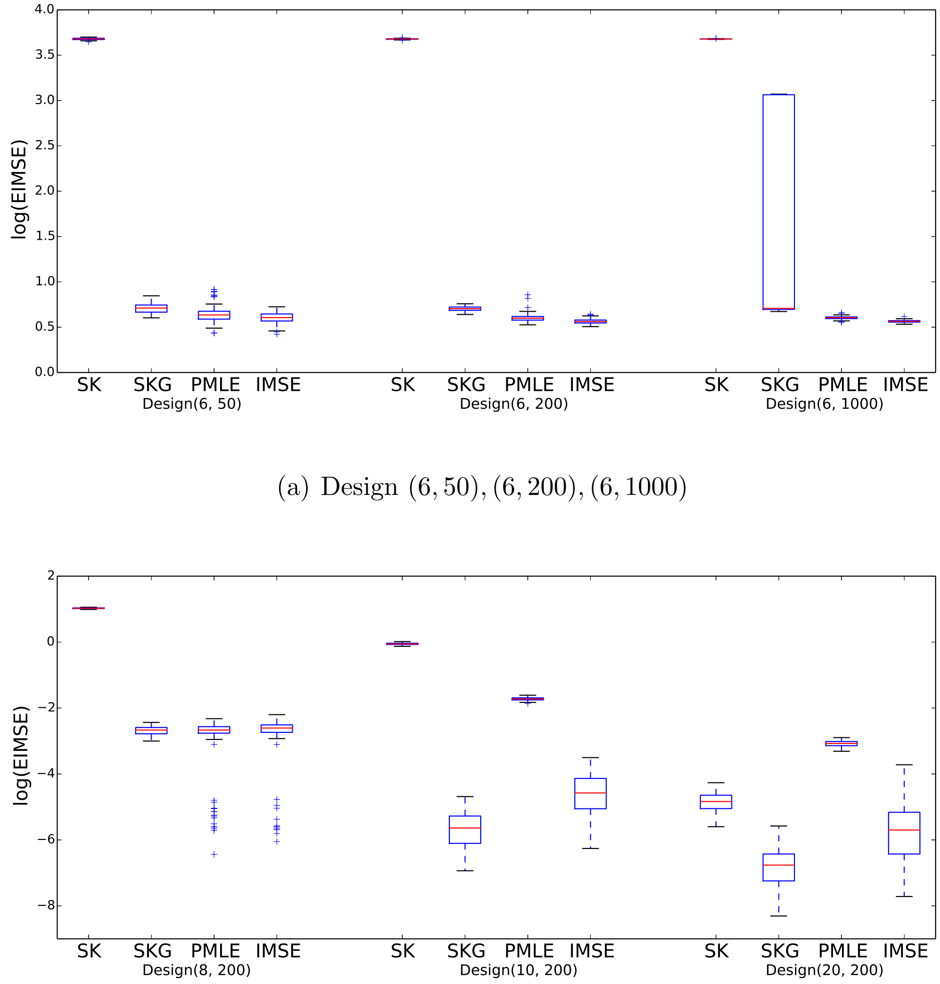


(a) Design $(6, 50), (6, 200), (6, 1000)$



(b) Design $(8, 200), (10, 200), (20, 200)$

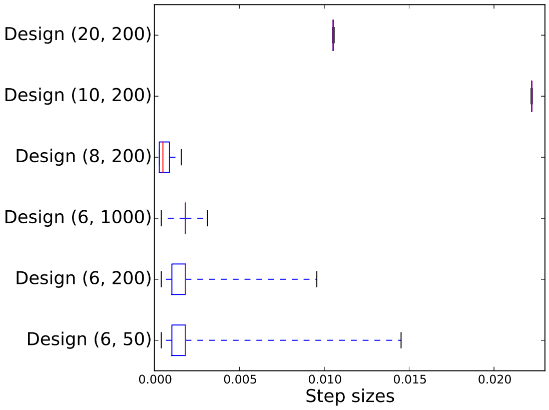
Figure 3.3: Boxplots of EIMSE from 100 macroreplications for SK, SKG and GESK with six different designs on estimating the expected steady-state waiting time in an M/M/1 queue problem, corresponding to results in Table 3.2.



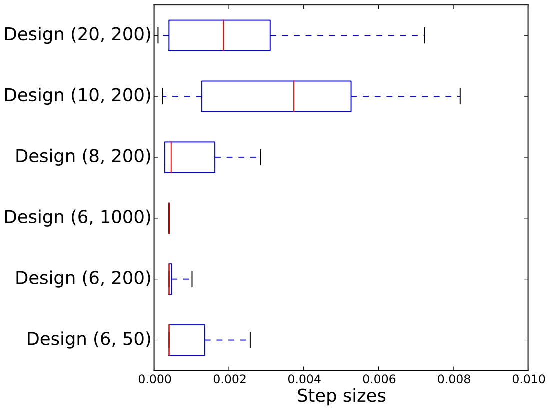
(a) Design $(6, 50), (6, 200), (6, 1000)$

(b) Design $(8, 200), (10, 200), (20, 200)$

Figure 3.4: Boxplots of EIMSE from 100 macroreplications for SK, SKG and GESK with five different designs on $y(x) = \exp(-1.4x)\cos(7\pi x/2) + \epsilon$, corresponding to results in Table 3.3.



(a) Step sizes determined by PMLE



(b) Step sizes determined by IMSE

Figure 3.5: Boxplots for step sizes determined by PMLE and IMSE based on100 macroreplication in the stylized example with added noise.

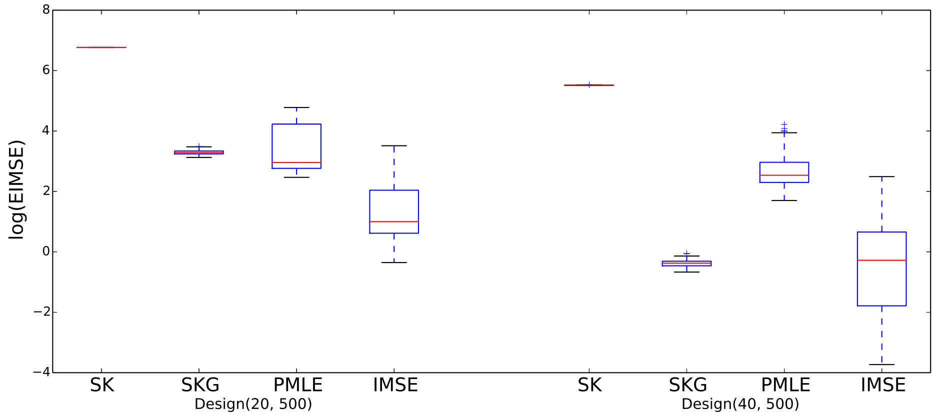


Figure 3.6: Boxplots of EIMSE from 100 macroreplications for SK and GESK with two different Latin-hypercube designs on a four-dimensional function with added noise.

Chapter 4: Simulation Selection with Unknown Correlation Structure

4.1 Introduction

Consider a decision-maker who must identify the best among a finite set of design alternatives with unknown performance values. Stochastic simulation is used to estimate the performance of an alternative. More simulation experiments will produce better estimates; however, these experiments are expensive and time-consuming, limiting the simulation budget. We must use this budget efficiently to maximize the quality of the final selection decision.

In many applications, the simulation budget is comparable to, or smaller than, the number of design alternatives. However, there may be correlations between the underlying mean performance values. Correlations can potentially allow us to handle much larger problems: a single piece of information about one alternative can now be used to learn about other alternatives with "similar" values. However, these similarities can be difficult to quantify or guess heuristically. Consider the following examples:

1. *Wind farm placement.* Given a set of candidate locations for a new wind farm

installation, we wish to select the one with the highest average power output. However, power output depends on volatile wind speeds and other physical factors like pressure gradient, frictional forces, wind currents, and topographical features. These factors are difficult to quantify, but simulation can be used to estimate the net result [59]. Physical and topographical similarities induce complex correlations between locations.

2. *Logistics management.* In a vehicle routing problem with service choice [60], customers can request privilege for early delivery through bidding. The service provider will accept a set of requests if the total bid price exceeds the additional cost incurred by deviating from the optimal route. The number of acceptable sets of requests grows combinatorially. To solve this problem, we have to use the routing cost computed for one set of requests to infer the costs of other sets that contain one or more of the same customers.

3. *Call center control.* A call center administrator assigns agents in shifts to minimize average call waiting time. The administrator is uncertain about employee efficiency [61], making it difficult to determine the best assignment. Simulation can be used to test performances from different assignments. The performance of two different assignments will be correlated if the assignments involve the same agents.

Simulation selection procedures consist of a statistical model of the decision-maker's estimates of the performance values, and an optimization algorithm for choosing an alternative to simulate based on the current statistical estimates. In the classical literature on ranking and selection (R&S), estimates are constructed using frequentist statistics, and decisions are made using the indifference-zone (IZ) approach pioneered by [62]; see also [63] for an overview of classical results. IZ methods guarantee asymptotic lower bounds for the probability of correct selection (PCS), as long as the true underlying performance values are sufficiently far apart. The best-performing IZ methods include those by [64,65] and [66]. Numerous reviews and surveys are available, including [67], [68], [69] and [70].

Bayesian models for R&S consider the tradeoff between our estimates of the performance values and our uncertainty about those estimates. This is known as the "value of information" approach, going back to [27] and extended in later work. See [31] or [5] for an overview of value of information procedures (VIP). The optimal computing budget allocation (OCBA) methodology (see e.g. [71–73]), designed to maximize a Bayesian version of PCS, can also be included in the Bayesian category.

Both frequentist [66] and Bayesian [29,30] methods are able to handle problems where the simulation output has unknown variance. However, most work on R&S typically makes independence assumptions on the estimates of performance values: under this assumption, a single experiment only provides information about a single alternative, making it difficult to handle large problems with a small simulation budget. Correlations have largely been studied in the context of common random numbers inside simulators; see [74] and [75] for IZ methods in this setting. [76] considers this problem from the perspective of OCBA.

The present chapter, however, uses the term "correlation" in a broader sense. In the Bayesian setting, correlations can be used inside a distribution of belief as a measure of the inherent similarities or differences between alternatives (e.g., the geographical similarities between two wind farm locations). As we show in this chapter, correlated beliefs can significantly improve the performances of simulation selection, even when the simulation output is completely independent. Recently, [77] studied Bayesian R&S with correlated beliefs ([78] extends this analysis to include correlated simulation output), but under the restrictive assumption that the correlation structure was correctly specified by the decision-maker. By contrast, we develop a model where the correlation structure is unknown, and has to be learned together with the performance values. Our model has the ability to correct inaccurate prior beliefs as new information arrives. If the simulation output is correlated, we can also learn that correlation structures provided that some prior information about it is available. To our knowledge, [79] is the only work on Bayesian R&S to consider unknown correlation structures.

Bayesian R&S procedures rely on conjugate prior distributions on the unknown model parameters in order to maintain computational tractability. The Wishart distribution is a well-known conjugate prior for an unknown covariance matrix, assuming that we can simultaneously observe the performance of every alternative. See e.g. [80] or [81] for applications of the Wishart distribution in simulation metamodeling and input uncertainty. However, in fully sequential R&S, we only sample from one alternative at a time. There is no standard conjugate prior for this problem, although the statistics community has made several attempts to create one; see [82], [83] or [84] for examples. Unfortunately, these models either present computational difficulties in an R&S setting or cannot extract information about multiple alternatives from a single scalar observation. [79] resolves this problem by imposing restrictions on the sampling procedure: the simulation budget is allocated equally among a certain subset of alternatives.

We propose a different approach, where we have the flexibility to simulate any one alternative at any time. Although our prior is not exactly conjugate, we create an optimal approximation of conjugacy by minimizing the Kullback-Leibler divergence between the true posterior and the normal-Wishart distribution, leading to a computationally efficient learning model. The approximate model enables us to derive a new VIP that generalizes previous procedures on R&S with known correlations. We establish intuitive analogies between the new model and classical statistical results on unknown sampling variance. We also show that, all else being equal, information has greater value when the correlation structure is unknown, making it important to consider this uncertainty when allocating the next simulation experiment.

4.2 Learning Unknown Correlation Structures

Let $\{1, 2, \cdots, K\}$ be a set of alternatives. Let $\widehat{\mathbf{Y}}$ be a multivariate normal random vector in $\mathbb{R}^K$ with mean $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K)$ and covariance matrix $\mathbf{\Sigma}$. Our goal is to discover the alternative x with the largest underlying mean μ_x. Assuming that $\mathbf{\Sigma}$ is invertible, we define $\mathbf{R} = \mathbf{\Sigma}^{-1}$ to be the *precision matrix* of $\widehat{\mathbf{Y}}$. For ease of computation and presentation, we will work with the precision matrix instead of the covariance matrix throughout this chapter.

The vector $\widehat{\mathbf{Y}}$ describes the behavior of K alternatives, all observed concurrently. We suppose that both $\boldsymbol{\mu}$ and $\mathbf{R}$ are unknown. Let $\hat{y}_x$ represents the simulation output of the behavior of the xth alternative. The sampling distribution of $\hat{y}_x$ given $\boldsymbol{\mu}$ and $\mathbf{R}$ is univariate normal with the following probability density function

$$p(\hat{y}_x|\boldsymbol{\mu}, \mathbf{R}) \propto \frac{1}{|\mathbf{e}_x'\mathbf{R}^{-1}\mathbf{e}_x|^{\frac{1}{2}}} \exp\left\{-\frac{(\hat{y}_x - \mu_x)^2}{2\mathbf{e}_x'\mathbf{R}^{-1}\mathbf{e}_x}\right\}, \tag{4.1}$$

where $\mathbf{e}_x = (0, \ldots, 1, \ldots, 0)$ is a $K \times 1$ vector, with 1 at the xth component, and 0 at others. The prime denotes transpose.

We allow the precision matrix $\mathbf{R}$ to be non-diagonal, implying correlations between components of $\widehat{\mathbf{Y}}$. As the rest of this section will show, when $\boldsymbol{\mu}$ and $\mathbf{R}$ are both unknown, a set of beliefs about $\mathbf{R}$ will induce correlations between our beliefs about different components of $\boldsymbol{\mu}$, implying similarities and differences between alternatives. We can expect that a single observation $\hat{y}_x$ should also provide some information about other alternatives that are correlated with x. However, the nature of this information is not clear as the correlation structure is unknown.

4.2.1 Learning from complete observations

Taking the Bayesian viewpoint, we view the unknown mean vector $\boldsymbol{\mu}$ and the precision matrix $\mathbf{R}$ as a random vector and a random matrix, respectively. In accordance with the Bayesian approach, we assume that our prior knowledge about these unknown quantities is reflected by a prior distribution, which we write as

$$\boldsymbol{\mu}|\mathbf{R} \sim \mathcal{N}_K(\boldsymbol{\theta}^0, q^0\mathbf{R}), \quad \mathbf{R} \sim \mathcal{W}_K(b^0, \mathbf{B}^0).$$

The precision matrix $\mathbf{R}$ is assumed to follow a Wishart distribution parametrized by a scalar b^0 and a $K \times K$ matrix $\mathbf{B}^0$. The conditional distribution of $\boldsymbol{\mu}$ given $\mathbf{R}$ is multivariate normal with mean vector $\boldsymbol{\theta}^0$ and precision matrix $q^0\mathbf{R}$, where $\boldsymbol{\theta}^0$ is a K vector and q^0 is a scalar. The probability density function of the Wishart distribution, see e.g. [85], is given as

$$p(\mathbf{R}) = \frac{1}{Z(b^0, \mathbf{B}^0)} |\mathbf{R}|^{\frac{b^0 - K - 1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{B}^0\mathbf{R})\right\},$$

with a normalizing constant

$$Z(b^0, \mathbf{B}^0) = \pi^{\frac{K(K-1)}{4}} \left|\frac{\mathbf{B}^0}{2}\right|^{-\frac{b^0}{2}} \prod_{i=1}^{K} \Gamma\left(\frac{b^0 + 1 - i}{2}\right).$$

Therefore the joint prior distribution of $\boldsymbol{\mu}$ and $\mathbf{R}$ is

$$p^0(\boldsymbol{\mu}, \mathbf{R}) = \frac{1}{Z(b^0, \mathbf{B}^0)} |\mathbf{R}|^{\frac{b^0 - K - 1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{B}^0\mathbf{R})\right\} \left(\frac{q^0}{2\pi}\right)^{\frac{K}{2}} |\mathbf{R}|^{\frac{1}{2}} \exp\left\{-\frac{q^0}{2}(\boldsymbol{\mu} - \boldsymbol{\theta}^0)'\mathbf{R}(\boldsymbol{\mu} - \boldsymbol{\theta}^0)\right\}.$$

The Wishart distribution has the property that $\mathbb{E}(\mathbf{R}) = b^0(\mathbf{B}^0)^{-1}$, whence $\mathbb{E}(\boldsymbol{\Sigma}) = \frac{\mathbf{B^0}}{b^0 - K + 1}$. The matrix $\mathbf{B}^0$ can be viewed as a generalized "sum of squares". If the prior parameters are constructed from historical data (known as a "first-stage sample" in [79]), the diagonal entries of $\mathbf{B}^0$ will be the sums of squared deviations of the first-stage observations from their means. The scalar b^0 is analogous to the size of the first-stage sample, so that $\frac{\mathbf{B}^0}{b^0 - K + 1}$ is precisely the empirical covariance matrix constructed from the first-stage data. The parameter q^0 is also analogous to a sample size; if first-stage sampling is used, $\frac{\mathbf{R}^{-1}}{q^0}$ will be the covariance matrix of the sample mean $\boldsymbol{\mu}$.

If our distribution of belief at stage n is normal-Wishart, and our next observation is the entire vector $\widehat{\mathbf{Y}}^{n+1} \sim \mathcal{N}_K(\boldsymbol{\mu}, \mathbf{R})$, standard results from Bayesian analysis [86] tell us that the posterior density

$$p^{n+1}(\boldsymbol{\mu}, \mathbf{R}|\widehat{\mathbf{Y}}^{n+1}) = \frac{p(\widehat{\mathbf{Y}}^{n+1}|\boldsymbol{\mu}, \mathbf{R})p^n(\boldsymbol{\mu}, \mathbf{R})}{\iint p(\widehat{\mathbf{Y}}^{n+1}|\boldsymbol{\mu}, \mathbf{R})p^n(\boldsymbol{\mu}, \mathbf{R})d\boldsymbol{\mu}d\mathbf{R}}$$

is another normal-Wishart distribution with parameters

$$q^{n+1} = q^n + 1, \tag{4.2}$$

$$b^{n+1} = b^n + 1, \tag{4.3}$$

$$\boldsymbol{\theta}^{n+1} = \frac{q^n\boldsymbol{\theta}^n + 1}{q^n + 1}, \tag{4.4}$$

$$\mathbf{B}^{n+1} = \mathbf{B}^n + \frac{q^n}{q^n + 1}(\boldsymbol{\theta}^n - \widehat{\mathbf{Y}}^{n+1})(\boldsymbol{\theta}^n - \widehat{\mathbf{Y}}^{n+1})'. \tag{4.5}$$

This is known as the conjugacy property of the normal-Wishart distribution. Conjugacy allows us to represent a distribution of belief with a finite, small number of parameters, which can be easily updated after each new observation. It is convenient to denote these parameters by the notational shorthand $\mathbf{S}^n = (q^n, b^n, \boldsymbol{\theta}^n, \mathbf{B}^n)$, representing the *state* of our beliefs at time n.

If we are able to observe the entire vector $\widehat{\mathbf{Y}}^{n+1}$ (also known as a "complete observation"), the decision-maker's objective can be easily formulated as follows. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an appropriate probability space, and define a filtration $\mathcal{F}^n$, where $\mathcal{F}^n$ is the σ-algebra generated by the first n observations $\mathbf{Y}^1, ..., \mathbf{Y}^n$. Then, $\boldsymbol{\theta}^n = \mathbb{E}(\boldsymbol{\mu}|\mathcal{F}^n)$ intuitively represents our time-n beliefs about $\boldsymbol{\mu}$. We wish to find

$$\mathbb{E}(\max_x \mathbb{E}(\mu_x|\mathcal{F}^N)), \tag{4.6}$$

where N is the simulation budget. The maximum inside the outer expectation in (4.6) represents the decision-maker's implementation decision: at time N, we will select the alternative that appears to be the best. If complete observations are available, (4.6) is simply a statistical estimation problem with no elements of simulation optimization. However, we will now move to a setting where only a single component of $\widehat{\mathbf{Y}}^{n+1}$ can be observed at a time, giving rise to the problem of choosing the right component.

4.2.2 Learning model for scalar observations

As is common in fully sequential ranking and selection, we now suppose that at stage n we sample from alternative x only, with an observation $\hat{y}_x^{n+1} \sim \mathcal{N}(\mu_x, (\mathbf{e}_x'\mathbf{R}^{-1}\mathbf{e}_x)^{-1})$. The distribution of $\hat{y}_x^{n+1}$ is simply the marginal distribution of the xth component of $\widehat{\mathbf{Y}}^{n+1}$. Using Bayes' rule, the joint posterior distribution of $\boldsymbol{\mu}$ and $\mathbf{R}$, given the observation $\hat{y}_x^{n+1}$, can be written as

$$\begin{aligned} p^{n+1}(\boldsymbol{\mu}, \mathbf{R} | \hat{y}_x^{n+1}) \propto & \frac{1}{Z(b^n, \mathbf{B}^n)} |\mathbf{R}|^{\frac{b^n-K-1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{B}^n\mathbf{R})\right\} \left(\frac{q^n}{2\pi}\right)^{\frac{K}{2}} |\mathbf{R}|^{\frac{1}{2}} \\ & \cdot \exp\left\{-\frac{q^n}{2}(\boldsymbol{\mu}-\boldsymbol{\theta}^n)'\mathbf{R}(\boldsymbol{\mu}-\boldsymbol{\theta}^n)\right\} \frac{1}{(2\pi\mathbf{R}^{-1})_{xx}^{\frac{1}{2}}} \exp\left\{-\frac{(\hat{y}_x^{n+1}-\mu_x)^2}{2(\mathbf{R}^{-1})_{xx}}\right\}. \end{aligned} \tag{4.7}$$

After decomposing the posterior distribution in (4.7) into the conditional posterior distribution of $\boldsymbol{\mu}$ given $\mathbf{R}$ and the marginal posterior distribution of $\mathbf{R}$, we observe that the conditional distribution of $\boldsymbol{\mu}$ given $\mathbf{R}$ is multivariate normal, but the marginal distribution of $\mathbf{R}$ is no longer a Wishart distribution.

Computational difficulties arise from here. Equation (4.7) suggests that the conjugacy property of the normal-Wishart distribution is lost if we can no longer observe the entire vector $\widehat{\mathbf{Y}}^{n+1}$. Conjugacy of the prior distribution is necessary in order to build a computationally tractable learning model and develop efficient sequential decision procedures that make sampling choices based on a small set of belief parameters. In what follows, we force conjugacy using the density projection technique. To be precise, by minimizing the Kullback-Leibler (KL) divergence, we find the best approximation for the posterior distribution in (4.7) from the normal-Wishart family. The posterior distribution is then replaced by its normal-Wishart approximation, and the decision-maker's beliefs are assumed to be normal-Wishart after each successive observation.

Let $\xi(\boldsymbol{\mu}, \mathbf{R})$ be a distribution from the normal-Wishart family with parameters $(q, b, \boldsymbol{\theta}, \mathbf{B})$ such that

$$\boldsymbol{\mu}|\mathbf{R} \sim \mathcal{N}_K(\boldsymbol{\theta}, q\mathbf{R}), \quad \mathbf{R} \sim \mathcal{W}_K(b, \mathbf{B}).$$

Define $\mathcal{D}^n_{KL}(\xi \| p^{n+1})$ to be the Kullback-Leibler (KL) divergence between $\xi(\boldsymbol{\mu}, \mathbf{R})$ and $p^{n+1}(\boldsymbol{\mu}, \mathbf{R}|\hat{y}^{n+1}_x)$, which is given by

$$\mathcal{D}^n_{KL}(\xi \| p^{n+1}) = \mathbb{E}_\xi \left(\log \frac{\xi(\boldsymbol{\mu}, \mathbf{R})}{p^{n+1}(\boldsymbol{\mu}, \mathbf{R}|\hat{y}^{n+1}_x)} \right), \tag{4.8}$$

where $\mathbb{E}_\xi[\cdot]$ is the expectation with respect to $\xi(\boldsymbol{\mu}, \mathbf{R})$. This quantity, bounded below by zero, is used to measure the "distance" between the distributions ξ and p^{n+1}. Lower KL divergence suggests that there is more similarity between the two distributions. For simplicity of notation, we write $\mathcal{D}^n_{KL}(\xi \| p)$ instead of $\mathcal{D}^n_{KL}(\xi \| p^{n+1})$. We wish to find

$$(q^{n+1}, b^{n+1}, \boldsymbol{\theta}^{n+1}, \mathbf{B}^{n+1}) = \arg\min_{(q,b,\boldsymbol{\theta},\mathbf{B})} \mathcal{D}^n_{KL}(\xi \| p), \tag{4.9}$$

the set of parameters that projects (according to KL divergence) the normal-Wishart distribution onto the true posterior in (4.7).

We first give a closed-form expression for (4.8), and then solve (4.9). We briefly sketch the proof of the solution, but readers are referred to the Appendix for the complete details.

Proposition 4.1. *The KL divergence $\mathcal{D}^n_{KL}(\xi\|p)$ defined in* (4.8) *can be expressed as*

$$\begin{aligned}\mathcal{D}^n_{KL}(\xi\|p) =& \frac{b^{n+1}-b^n}{2}\left(-\log\left|\frac{\mathbf{B^{n+1}}}{2}\right| + \sum_{i=1}^{K}\psi\left(\frac{b^{n+1}-i+1}{2}\right)\right) - \frac{b^{n+1}K}{2} \\ &+ \frac{b^{n+1}}{2}\operatorname{tr}\left(\mathbf{B}^n(\mathbf{B}^{n+1})^{-1}\right) + \log\frac{Z(b^n,\mathbf{B}^n)}{Z(b^{n+1},\mathbf{B}^{n+1})} + \frac{1}{2}\log B^{n+1}_{xx} \\ &+ \frac{1}{2}\left[K\log\frac{q^{n+1}}{q^n} + K\frac{q^n}{q^{n+1}} - K + q^n(\boldsymbol{\theta}^n-\boldsymbol{\theta}^{n+1})' b^{n+1}(\mathbf{B}^{n+1})^{-1}(\boldsymbol{\theta}^n-\boldsymbol{\theta}^{n+1})\right] \\ &- \frac{1}{2}\psi\left(\frac{b^{n+1}-K+1}{2}\right) + \frac{1}{2q^{n+1}} + \frac{1}{2}(\hat{y}^{n+1}_x-\theta^{n+1}_x)^2\frac{b^{n+1}-K+1}{B^{n+1}_{xx}} + C,\end{aligned}$$

where $\psi(x) = d\ln\Gamma(x)/dx$ is the digamma function and C is some constant that does not depend on the parameters of ξ.

Proof. Proof: First notice that the posterior distribution in (4.7) can be written as

$$p^{n+1}(\boldsymbol{\mu},\mathbf{R}|\hat{y}^{n+1}_x) = \frac{p^{n+1}(\boldsymbol{\mu}|\mathbf{R})p^{n+1}(\mathbf{R})p(\hat{y}^{n+1}_x|\boldsymbol{\mu},\mathbf{R})}{p(\hat{y}^{n+1}_x)}.$$

Then the KL divergence is given as

$$\begin{align}
\mathcal{D}_{KL}^{n}(\xi|p) &= \mathbb{E}_{\xi}\left(\log\frac{\xi(\boldsymbol{\mu},\mathbf{R})}{p^{n+1}(\boldsymbol{\mu},\mathbf{R}|\hat{y}_x^{n+1})}\right) \nonumber\\
&= \iint \xi(\boldsymbol{\mu},\mathbf{R})\log\frac{\xi(\mathbf{R})\xi(\boldsymbol{\mu}|\mathbf{R})p(\hat{y}_x^{n+1})}{p^{n+1}(\boldsymbol{\mu}|\mathbf{R})p^{n+1}(\mathbf{R})p(\hat{y}_x^{n+1}|\boldsymbol{\mu},\mathbf{R})}d\boldsymbol{\mu}d\mathbf{R} \nonumber\\
&= \iint \xi(\boldsymbol{\mu},\mathbf{R})\log\frac{\xi(\mathbf{R})}{p^{n+1}(\mathbf{R})}d\boldsymbol{\mu}d\mathbf{R} \tag{4.10}\\
&\quad + \iint \xi(\boldsymbol{\mu},\mathbf{R})\log\frac{\xi(\boldsymbol{\mu}|\mathbf{R})}{p^{n+1}(\boldsymbol{\mu}|\mathbf{R})}d\boldsymbol{\mu}d\mathbf{R} \tag{4.11}\\
&\quad - \iint \xi(\boldsymbol{\mu},\mathbf{R})\log p(\hat{y}_x^{n+1}|\boldsymbol{\mu},\mathbf{R})d\boldsymbol{\mu}d\mathbf{R} \tag{4.12}\\
&\quad + \iint \xi(\boldsymbol{\mu},\mathbf{R})\log p(\hat{y}_x^{n+1})d\boldsymbol{\mu}d\mathbf{R}, \tag{4.13}
\end{align}$$

where (4.10) can be computed as

$$\begin{align}
&\iint \xi(\boldsymbol{\mu},\mathbf{R})\log\frac{\xi(\mathbf{R})}{p^{n+1}(\mathbf{R})}d\boldsymbol{\mu}d\mathbf{R} \nonumber\\
&= \int \xi(\mathbf{R})\log\frac{\xi(\mathbf{R})}{p^{n+1}(\mathbf{R})}d\mathbf{R} \nonumber\\
&= \int \xi(\mathbf{R})\log\left\{\frac{Z(b^{n},\mathbf{B}^{n})}{Z(b^{n+1},\mathbf{B}^{n+1})}|\mathbf{R}|^{\frac{b^{n+1}-b^{n}}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{B}^{n+1}\mathbf{R})-\frac{1}{2}\operatorname{tr}(\mathbf{B}\mathbf{R})\right\}\right\}d\mathbf{R} \nonumber\\
&= \frac{b^{n+1}-b^{n}}{2}\mathbb{E}_{\xi}\left(\log|\mathbf{R}|\right)+\log\frac{Z(b^{n},\mathbf{B}^{n})}{Z(b^{n+1},\mathbf{B}^{n+1})}-\frac{1}{2}\mathbb{E}_{\xi}\left[\operatorname{tr}((\mathbf{B}^{n+1}+\mathbf{B}^{n})\mathbf{R})\right] \nonumber\\
&= \frac{b^{n+1}-b^{n}}{2}\left(-\log\left|\frac{\mathbf{B^{n+1}}}{2}\right|+\sum_{i=1}^{K}\psi\left(\frac{b^{n+1}-i+1}{2}\right)\right)+\log\frac{Z(b^{n},\mathbf{B}^{n})}{Z(b^{n+1},\mathbf{B}^{n+1})} \nonumber\\
&\qquad + \frac{b^{n+1}}{2}\operatorname{tr}\left(\mathbf{B}^{n}(\mathbf{B}^{n+1})^{-1}\right)-\frac{b^{n+1}}{2}K, \tag{4.14}
\end{align}$$

the term in (4.11) can be computed as

$$\begin{aligned}
&\iint \xi(\boldsymbol{\mu},\mathbf{R}) \log \frac{\xi(\boldsymbol{\mu}|\mathbf{R})}{p^{n+1}(\boldsymbol{\mu}|\mathbf{R})} d\boldsymbol{\mu} d\mathbf{R} \\
=& \iint \xi(\boldsymbol{\mu},\mathbf{R}) \log \left\{ \frac{|q^{n+1}\mathbf{R}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\theta}^{n+1})' q^{n+1}\mathbf{R}(\boldsymbol{\mu}-\boldsymbol{\theta}^{n+1})\right\}}{|q^{n}\mathbf{R}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\theta}^{n})' q^{n}\mathbf{R}(\boldsymbol{\mu}-\boldsymbol{\theta}^{n})\right\}} \right\} \\
=& \iint \xi(\boldsymbol{\mu},\mathbf{R}) \left[\frac{K}{2} \log \frac{q^{n+1}}{q^n} - \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\theta}^{n+1}) q^{n+1}\mathbf{R}(\boldsymbol{\mu}-\boldsymbol{\theta}^{n+1}) + \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\theta})' q^{n}\mathbf{R}(\boldsymbol{\mu}-\boldsymbol{\theta}) \right] d\boldsymbol{\mu} d\mathbf{R} \\
=& \frac{K}{2} \log \frac{q^{n+1}}{q^n} + \frac{K}{2}\frac{q^n}{q^{n+1}} - \frac{K}{2} + \frac{q^n}{2}(\boldsymbol{\theta}^{n+1}-\boldsymbol{\theta}^{n})' b^{n+1}(\mathbf{B}^{n+1})^{-1}(\boldsymbol{\theta}^{n+1}-\boldsymbol{\theta}^{n}), \qquad (4.15)
\end{aligned}$$

and the term in (4.12) can be computed as

$$\begin{aligned}
&\iint \xi(\boldsymbol{\mu},\mathbf{R}) \log p(\hat{y}^{n+1}|\boldsymbol{\mu},\mathbf{R}) d\boldsymbol{\mu} d\mathbf{R} \\
=& \iint \xi(\boldsymbol{\mu},\mathbf{R}) \log \left[(2\pi\mathbf{R}^{-1})_{xx}^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{R}^{-1})_{xx}^{-1}(\hat{y}_x^{n+1}-\mu_x)^2 \right] \right] d\boldsymbol{\mu} d\mathbf{R} \\
=& -\frac{1}{2}\mathbb{E}_{\xi}\left(\log(R^{-1})_{xx}\right) - \frac{1}{2}\log(2\pi) - \frac{1}{2}(\hat{y}_x^{n+1}-\theta_x^{n+1})^2 \mathbb{E}\left((\mathbf{R}^{-1})_{xx}^{-1}\right) - \frac{1}{2q^{n+1}} \\
=& -\frac{1}{2}\left(\log B_{xx}^{n+1} - \log 2 - \psi\left(\frac{b^{n+1}-K+1}{2}\right)\right) - \frac{1}{2}\log(2\pi) - \frac{1}{2q^{n+1}} - \frac{1}{2}(\hat{y}_x^{n+1}-\theta_x^{n+1})^2 \frac{b^{n+1}-K}{B_{xx}^{n+1}}
\end{aligned}$$
$$(4.16)$$

Notice that $p(\hat{y}_x^{n+1})$ is the marginal distribution of $\hat{y}_x^{n+1}$, which is not a function of $\boldsymbol{\mu}$ and $\mathbf{R}$, so (4.13) doesn't depend on the parameters of ξ. The proof then follows from equations (4.14), (4.15) and (4.16). □

Theorem 4.2. *There exists a finite value* Δb^n *s.t. the solution to* (4.9) *can be* *expressed as*

$$q^{n+1} = q^n + \frac{1}{K}, \tag{4.17}$$

$$b^{n+1} = b^n + \Delta b^n, \tag{4.18}$$

$$\begin{aligned}\boldsymbol{\theta}^{n+1} &= \left(q^n b^{n+1} (\mathbf{B}^{n+1})^{-1} + \frac{b^{n+1} - K + 1}{\mathbf{e}_x' \mathbf{B}^{n+1} \mathbf{e}_x} \mathbf{e}_x \mathbf{e}_x' \right)^{-1} \\ &\quad \cdot \left(q^n b^{n+1} (\mathbf{B}^{n+1})^{-1} \boldsymbol{\theta}^n + \frac{b^{n+1} - K + 1}{\mathbf{e}_x' \mathbf{B}^{n+1} \mathbf{e}_x} \hat{y}_x^{n+1} \mathbf{e}_x \right),\end{aligned} \tag{4.19}$$

$$\mathbf{B}^{n+1} = \frac{b^{n+1}}{b^n} \mathbf{B}^n + \frac{b^{n+1}}{b^n + 1} \left(\frac{q^n (\hat{y}_x^{n+1} - \theta_x^n)^2}{\frac{q^n b^{n+1}}{b^{n+1} - K + 1} + 1} - \frac{B_{xx}^n}{b^n} \right) \frac{\mathbf{B}^n \mathbf{e}_x \mathbf{e}_x' \mathbf{B}^n}{(B_{xx}^n)^2}. \tag{4.20}$$

Proof. Proof: Taking derivatives of (4.8) with respect to the parameters and applying matrix calculus, we obtain

$$\frac{\partial \mathcal{D}_{KL}^n(\xi \| p)}{\partial q^{n+1}} = \frac{1}{2} \left\{ \frac{K}{q^{n+1}} - \frac{q^n K}{(q^{n+1})^2} - \frac{1}{(q^{n+1})^2} \right\}, \tag{4.21}$$

$$\begin{aligned}\frac{\partial \mathcal{D}_{KL}^n(\xi \| p)}{\partial b^{n+1}} &= \frac{1}{2} \left\{ \frac{(\hat{y}_x^{n+1} - \theta_x^n)^2}{B_{xx}^{n+1}} \frac{(q^n)^2 b^{n+1} (K-1)}{(q^n b^{n+1} + b^{n+1} - K + 1)^2} + \frac{b^n K + 1}{b^{n+1}} - K \right. \\ &\quad \left. + \frac{b^{n+1} - b^n}{2} \sum_{i=1}^{K} \psi' \left(\frac{b^{n+1} - i + 1}{2} \right) - \frac{1}{2} \psi' \left(\frac{b^{n+1} - K + 1}{2} \right) \right\},\end{aligned} \tag{4.22}$$

$$\frac{\partial \mathcal{D}_{KL}^n(\xi \| p)}{\partial \boldsymbol{\theta}^{n+1}} = q^n b^{n+1} (\mathbf{B}^{n+1})^{-1} (\boldsymbol{\theta}^n - \boldsymbol{\theta}^{n+1}) + \frac{b^{n+1} - K + 1}{B_{xx}^{n+1}} (\mathbf{e}_x \mathbf{e}_x') (\hat{y}_x^{n+1} \mathbf{e}_x - \boldsymbol{\theta}^{n+1}), \tag{4.23}$$

$$\begin{aligned}\frac{\partial \mathcal{D}_{KL}^n(\xi \| p)}{\partial \mathbf{B}^{n+1}} &= -\frac{1}{2} q^n b^{n+1} \left(\mathbf{B}^{n+1} \right)^{-1} (\boldsymbol{\theta}^n - \boldsymbol{\theta}^{n+1})(\boldsymbol{\theta}^n - \boldsymbol{\theta}^{n+1})' (\mathbf{B}^{n+1})^{-1} + \frac{b^n}{2} (\mathbf{B}^{n+1})^{-1} \\ &\quad - \frac{b^{n+1}}{2} (\mathbf{B}^{n+1})^{-1} \mathbf{B}^n (\mathbf{B}^{n+1})^{-1} + \left(\frac{1}{2 B_{xx}^{n+1}} - \frac{b^{n+1} - K + 1}{2} \frac{(\hat{y}_x^{n+1} - \theta_x^{n+1})^2}{(B_{xx}^{n+1})^2} \right) \mathbf{e}_x \mathbf{e}_x'.\end{aligned} \tag{4.24}$$

Setting (4.21) - (4.24) to zero and solving, we notice that (4.17) follows immediately from (4.21). However, (4.22)-(4.24) are more difficult to solve because each equation depends on multiple parameters. We denote the change in degrees of freedom by

$\Delta b^n \equiv b^{n+1} - b^n$. Then we can derive (4.19) and (4.20) as functions of Δb^n. Finally, we compute Δb^n itself (see Remark 4.4 for a discussion).

Setting (4.23) to zero, we obtain

$$\left(q^n b^{n+1} (\mathbf{B}^{n+1})^{-1} + \frac{b^{n+1} - K + 1}{B^{n+1}_{xx}} \right) \boldsymbol{\theta}^{n+1} = q^n b^{n+1} (\mathbf{B}^{n+1})^{-1} \boldsymbol{\theta}^n + \frac{b^{n+1} - K + 1}{B^{n+1}_{xx}} \mathbf{e}_x \mathbf{e}_x' \hat{y}^{n+1}_x \mathbf{e}_x,$$

solving for $\boldsymbol{\theta}^{n+1}$ and it yields (4.19).

Setting (4.24) to zero and multiplying $\mathbf{B}^{n+1}$ from left and right, we obtain,

$$\begin{aligned} \mathbf{B}^{n+1} = & \frac{b^{n+1}}{b^n} \mathbf{B}^n + \frac{q^n b^{n+1}}{b^n} (\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^n)(\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^n)' \\ & - \frac{1}{b^n} \left(\frac{1}{B^{n+1}_{xx}} - \frac{(b^{n+1} - K + 1)(\hat{y}^{n+1}_x - \theta^{n+1}_x)^2}{(B^{n+1}_{xx})^2} \right) \mathbf{B}^{n+1} \mathbf{e}_x \mathbf{e}_x' \mathbf{B}^{n+1}. \end{aligned}$$

Since

$$(\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^n)(\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^n)' = \frac{(\hat{y}^{n+1}_x - \theta^n_x)^2}{(\gamma^n)^2} \frac{\mathbf{B}^{n+1} \mathbf{e}_x \mathbf{e}_x' \mathbf{B}^{n+1}}{(B^{n+1}_{xx})^2},$$

and

$$(\hat{y}^{n+1}_x - \theta^{n+1}_x)^2 = \frac{(\hat{y}^{n+1}_x - \theta^n_x)^2 (\gamma^n - 1)^2}{(\gamma^n)^2},$$

where

$$\gamma^n = 1 + \frac{q^n b^{n+1}}{b^{n+1} - K + 1},$$

it follows that

$$\begin{aligned} \mathbf{B}^{n+1} &= \frac{b^{n+1}}{b^n} \mathbf{B}^n + \left(\frac{q^n b^{n+1} (\hat{y}^{n+1}_x - \theta^n_x)^2}{b^n \gamma^n} - \frac{B^{n+1}_{xx}}{b^n} \right) \frac{\mathbf{B}^{n+1} \mathbf{e}_x \mathbf{e}_x' \mathbf{B}^{n+1}}{(B^{n+1}_{xx})^2} \\ &= \frac{b^{n+1}}{b^n} \mathbf{B}^n + \frac{b^{n+1}}{b^n} \left(\frac{q^n (\hat{y}^{n+1}_x - \theta^n_x)^2}{\frac{q^n b^{n+1}}{b^{n+1} - K + 1} + 1} - \frac{B^{n+1}_{xx}}{b^{n+1}} \right) \frac{\mathbf{B}^{n+1} \mathbf{e}_x \mathbf{e}_x' \mathbf{B}^{n+1}}{(B^{n+1}_{xx})^2}. \qquad (4.25) \end{aligned}$$

The matrix $\mathbf{B}^{n+1}$ shows up in both sides of (4.25). We will show how to derive updating equations for all entries in the matrix B^{n+1}.

Consider B_{xx}^{n+1}, it follows from (4.25) that

$$b^n B_{xx}^{n+1} = b^{n+1} B_{xx}^n + \frac{q^n b^{n+1} (\hat{y}_x^{n+1} - \theta_x^n)^2}{\gamma^n} - B_{xx}^{n+1},$$

then

$$B_{xx}^{n+1} = \frac{b^{n+1}}{b^n + 1} B_{xx}^n + \frac{1}{b^n + 1} \frac{q^n b^{n+1} (\hat{y}_x^{n+1} - \theta_x^n)^2}{\gamma^n}.$$

It follows from symmetry of the matrix $\mathbf{B}^n$ that $B_{ix}^{n+1} = B_{xi}^{n+1}$.

Consider B_{xi}^{n+1} and B_{ix}^{n+1} for $i \neq x$,

$$b^n B_{xi}^{n+1} = b^{n+1} B_{xi}^n + \left(\frac{q^n b^{n+1} (\hat{y}_x^{n+1} - \theta_x^n)^2}{\gamma^n} - B_{xx}^{n+1} \right) \frac{B_{xx}^{n+1} B_{xi}^{n+1}}{(B_{xx}^{n+1})^2},$$

then it follows that

$$B_{xi}^{n+1} = \frac{b^{n+1}}{b^n + 1} B_{xi}^n + \frac{1}{b^n + 1} \frac{q^n b^{n+1} (\hat{y}_x^{n+1} - \theta_x^n)^2}{\gamma^n} \frac{B_{xi}^n}{B_{xx}^n}. \tag{4.26}$$

The following result from (4.26)is worth mentioning,

$$\frac{B_{xi}^{n+1}}{B_{xx}^{n+1}} = \frac{B_{xi}^n}{B_{xx}^n}.$$

Consider B_{ij}^{n+1} for $i, j \neq x$,

$$b^n B_{ij}^{n+1} = b^{n+1} B_{ij}^n + \left(\frac{q^n b^{n+1} (\hat{y}_x^{n+1} - \theta_x^n)^2}{\gamma^n} - B_{xx}^{n+1} \right) \frac{B_{xi}^{n+1} B_{xj}^{n+1}}{(B_{xx}^{n+1})^2,}$$

and it follows that

$$B_{ij}^{n+1} = \frac{b^{n+1}}{b^n} B_{ij}^n + \frac{1}{b^n + 1} \left(\frac{q^n b^{n+1} (\hat{y}_x^{n+1} - \theta_x^n)^2}{\gamma^n} - \frac{b^{n+1}}{b^n} B_{xx}^n \right) \frac{B_{xi}^n B_{xj}^n}{(B_{xx}^n)^2}.$$

□

Using the Sherman-Morrison-Woodbury formula [**?**, see e.g.]]GoLo96, we can rewrite (4.19) without using inverse matrices as

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \frac{\hat{y}_x^{n+1} - \theta_x^n}{\frac{q^n b^{n+1}}{b^{n+1} - K + 1} B_{xx}^n + B_{xx}^n} \mathbf{B}^n \mathbf{e}_x. \tag{4.27}$$

The most crucial aspect of (4.27) is that a single scalar observation is now used to update the entire posterior mean vector through the matrix $\mathbf{B}^n$. Similar behavior occurs in the Kalman filter-like update used by [77] in the case of known correlation structures. In that setting, the updating equation incorporates both the variance of the current belief and the known variance of the observations. However, when the correlation structure is unknown, the matrix $\mathbf{B}^n$ is used to estimate both types of variances.

Equations (4.17)-(4.20) allow us to conveniently represent and update a joint distribution of belief about $\boldsymbol{\mu}$ and $\mathbf{R}$ using a finite number of parameters, which can be compactly encoded in the belief state $\mathbf{S}^n$. We can now connect the mechanism of approximate Bayesian inference back to a formal objective function. Recall from Section 4.2.1 that the sampling model is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathbb{P}$ is the law of the process $\mathbf{S}^n$. Essentially, our approximate Bayesian learning model replaces $\mathbb{P}$ by an alternate probability measure $\bar{\mathbb{P}}$ under which $(\boldsymbol{\mu}, \mathbf{R})$ is normal-Wishart, given $\mathcal{F}^n$ with parameters obtained through KL minimization.

We use the notation $\mathbb{E}_{\bar{\mathbb{P}}}(\cdot)$ for expectations under the probability measure $\bar{\mathbb{P}}$. Given a measurement budget of N, the experimenter chooses a measurement policy, which is a function x^π mapping a belief state $\mathbf{S}^n$ to an alternative $x^\pi(\mathbf{S}^n) \in \{1, \ldots, K\}$. Under the probability space $(\Omega, \mathcal{F}, \bar{\mathbb{P}})$, the policy makes measurement decisions sequentially. As before, $\mathcal{F}^n$ is the σ-algebra generated by all the decisions made in the first n stages, as well as the observations collected. Let π be a measurement policy. The notation $\mathbb{E}^\pi$ indicates that the expectation is taken when the measurement policy π is applied. We also define Π as the set of measurement policies. The challenge is to choose a measurement policy π for allocating the simulation budget one measurement at a time, and our objective can be written as

$$\sup_{\pi \in \Pi} \mathbb{E}^{\pi}_{\mathbb{P}} \left(\max_{x} \mathbb{E}^{\pi}_{\mathbb{P}} (\mu_x | \mathcal{F}^N) \right). \tag{4.28}$$

As in (4.6), the maximum in (4.28) represents the decision-maker's implementation decision to select the alternative that seems to be the best at time N. However, unlike (4.6), equation (4.28) now contains the optimization problem of choosing a policy π, i.e., a sequence of measurement decisions.

We close our discussion of the learning model with several remarks on the interpretation of the model parameters. The approximate updating equations (4.17)-(4.20) are intuitive generalizations of the conjugate update in (4.2)-(4.5). For example, in (4.5), the squared error matrix $(\boldsymbol{\theta}^n - \hat{\mathbf{Y}}^{n+1})(\boldsymbol{\theta}^n - \hat{\mathbf{Y}}^{n+1})'$ is used to update $\mathbf{B}^n$. In (4.20), the full matrix is not available, so the update uses the scalar squared error $(\theta^n_x - \hat{y}^{n+1}_x)^2$ to update all covariances between x and other alternatives.

Remark 4.3. *The parameter q^n in the prior distribution is intended to be a reflection of prior precision relative to the sample size that is tunable by the researcher or practitioner to reflect their prior confidence. Recall from (4.2) that, when we have complete observations, this parameter is always increased by* 1. *By analogy, if we only collect information about one out of K alternatives, q^n is increased by $1/K$.*

Remark 4.4. *Although one might expect b^n to behave in the same way as q^n, this is not exactly the case. The parameter b^n increases by 1 when we have complete observations. However, when we sample from only one alternative, the increment Δb^n actually depends on $(q^n, b^n, \hat{y}^{n+1}_x, \theta^n_x, B^n_{xx})$. The quantity Δb^n does not have a closed-*

form expression, but can easily be obtained numerically via a bisection procedure or Newton's method on the interval $[0, 1]$. *We have also observed in our numerical experiments that the optimal values of* Δb^n *appear to be smaller than* $1/K$ *and approach* $1/K$ *asymptotically over time.*

Remark 4.5. *Note that the computational complexity of the updates* (4.17)-(4.20) *is* $O(K^2)$, *identical to that of the conjugate updates for R&S with known correlations; see equations (2.22) and (2.23) in [5]. The number of iterations of the bisection method needed to compute* Δb^n *within a fixed, pre-specified tolerance level does not depend on* K. *However, the effort needed for a single iteration of the bisection method is* $O(K)$, *since the terms in (4.22) have to be recomputed when different values of* Δb^n *are considered.*

4.2.3 Predictive distribution of the next observation

Given the prior distribution on the unknown parameters, the distribution of $\hat{y}_x^{n+1}$ represents the decision-maker's beliefs about the next observation (assuming that alternative x will be measured). For this reason, it is known as the *predictive distribution*. In Section 4.3, we introduce a policy that uses the predictive distribution to look ahead to the outcome of a simulation decision. In preparation for this discussion, we now present the predictive distribution for the normal-Wishart model under the approximate Bayesian learning scheme of Section 4.2.1. That is, we assume that the decision-maker's beliefs at time n are represented by a normal-Wishart distribution whose parameters are contained in the state $\mathbf{S}^n$, and use this assumption to characterize $\hat{y}_x^{n+1}$.

For completeness, we provide the definition of the multivariate t-distribution [87].

Definition 4.1. *A p-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_p)$ is said to have the p-variate t-distribution with ν degrees of freedom, mean vector $\boldsymbol{\mu}$, and correlation matrix $\mathbf{V}$ if its joint pdf is given by*

$$f(\mathbf{x}) = \frac{\Gamma(\frac{\nu+p}{2})}{(\pi\nu)^{p/2}|\mathbf{V}|^{1/2}\Gamma(\frac{\nu}{2})}\left[1 + \frac{1}{\nu}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{(\nu+p)/2}.$$

The predictive distribution of $\hat{y}_x^{n+1}$ follows from the following two lemmas.

Lemma 4.6. *Suppose that $(\boldsymbol{\mu}, \mathbf{R})$ follows a normal-Wishart distribution with parameters $(q^n, b^n, \boldsymbol{\theta}^n, \mathbf{B}^n)$. Then the predictive distribution of a complete observation $\widehat{\mathbf{Y}}^{n+1}$ is a multivariate t-distribution with $b^n - K + 1$ degrees of freedom, mean vector $\boldsymbol{\theta}^n$ and correlation matrix $(q^n + 1)\mathbf{B}^n/q^n(b^n - K + 1)$.*

Proof. Proof: The predictive distribution of the vector $\widehat{\mathbf{Y}}^{n+1}$,

$$p(\widehat{\mathbf{Y}}^{n+1}) = \iint p(\widehat{\mathbf{Y}}^{n+1}, \boldsymbol{\mu}, \mathbf{R})d\boldsymbol{\mu}d\mathbf{R},$$

and

$$\begin{aligned} p(\widehat{\mathbf{Y}}^{n+1}, \boldsymbol{\mu}, \mathbf{R}) \propto |\mathbf{R}|^{\frac{b^n-K+1}{2}} &\exp\left\{-\frac{1}{2}(\widehat{\mathbf{Y}}^{n+1} - \boldsymbol{\mu})'\mathbf{R}(\widehat{\mathbf{Y}}^{n+1} - \boldsymbol{\mu})\right\} \\ &\cdot \exp\left\{-\frac{q^n}{2}(\boldsymbol{\mu} - \boldsymbol{\theta}^n)'\mathbf{R}(\boldsymbol{\mu} - \boldsymbol{\theta}^n)\right\} \exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{B}^n\mathbf{R})\right\}. \end{aligned}$$

It can be verified that

$$\begin{aligned} &(\widehat{\mathbf{Y}}^{n+1} - \boldsymbol{\mu})'\mathbf{R}(\widehat{\mathbf{Y}}^{n+1} - \boldsymbol{\mu}) + q^n(\boldsymbol{\mu} - \boldsymbol{\theta}^n)'\mathbf{R}(\boldsymbol{\mu} - \boldsymbol{\theta}^n) \\ &= (q^n + 1)(\boldsymbol{\mu} - \bar{\boldsymbol{\theta}}^n)'\mathbf{R}(\boldsymbol{\mu} - \bar{\boldsymbol{\theta}}^n) + \frac{q^n}{q^n + 1}(\boldsymbol{\theta} - \widehat{\mathbf{Y}}^{n+1})'\mathbf{R}(\boldsymbol{\theta} - \widehat{\mathbf{Y}}^{n+1}), \end{aligned}$$

with

$$\bar{\boldsymbol{\theta}}^n = \frac{q^n\boldsymbol{\theta}^n + \widehat{\mathbf{Y}}^{n+1}}{q^n+1}.$$

It follows that

$$p(\widehat{\mathbf{Y}}^{n+1}, \boldsymbol{\mu}, \mathbf{R}) \propto |\mathbf{R}|^{\frac{1}{2}} \exp\left\{-\frac{q^n+1}{2}(\boldsymbol{\mu}-\bar{\boldsymbol{\theta}}^n)'\mathbf{R}(\boldsymbol{\mu}-\bar{\boldsymbol{\theta}}^n)\right\} |\mathbf{R}|^{\frac{b^n-K}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(\bar{\mathbf{B}}^n\mathbf{R})\right\},$$

where

$$\bar{\mathbf{B}}^n = \mathbf{B}^n + \frac{q^n}{q^n+1}(\boldsymbol{\theta}^n - \widehat{\mathbf{Y}}^{n+1})(\boldsymbol{\theta}^n - \widehat{\mathbf{Y}}^{n+1})'.$$

Integration with respect to $\boldsymbol{\mu}$ and $\mathbf{R}$ yields

$$p(\widehat{\mathbf{Y}}^{n+1}) \propto \left(1 + \frac{q^n}{q^n+1}(\widehat{\mathbf{Y}}^{n+1}-\boldsymbol{\theta}^n)'(\mathbf{B}^n)^{-1}(\widehat{\mathbf{Y}}^{n+1}-\boldsymbol{\theta}^n)\right)^{-\frac{b^n+1}{2}}.$$

This shows that $\widehat{\mathbf{Y}}^{n+1}$ has a multivariate t-distribution with degrees of freedom $b^n - K + 1$, mean vector $\boldsymbol{\theta}^n$ and correlation matrix $\frac{q^n+1}{q^n(b^n-K+1)}\mathbf{B}^n$. □

Lemma 4.7. *The predictive distribution of $\hat{y}_x^{n+1}$ is*

$$\hat{y}_x^{n+1} \sim t\left(b^n - K + 1, \theta_x^n, \frac{q^n(b^n-K+1)}{(q^n+1)B_{xx}^n}\right), \tag{4.29}$$

which is a univariate Student's t-distribution with $b^n - K + 1$ degrees of freedom, mean θ_x^n and variance $\frac{q^n+1}{q^n(b^n-K-1)}B_{xx}^n$.

Proof. Proof: This result follows by combining Lemma 4.6 together with results in Section 1.10 from [87]. □

Using the predictive distribution found in (4.29), we can derive another expression for the updating equation of $\boldsymbol{\theta}^n$ in (4.27). Define

$$T^n = \frac{\hat{y}_x^{n+1} - \theta_x^n}{\left(\frac{q^n+1}{q^n(b^n-K+1)} B_{xx}^n\right)^{1/2}},$$

$$\tilde{\mathbf{s}}^n(q^n, b^n, \mathbf{B}^n, x) = \frac{\left(\frac{q^n+1}{q^n(b^n-K+1)}\right)^{1/2}}{\left(\frac{q^n b^{n+1}}{b^{n+1}-K+1} + 1\right)(B_{xx}^n)^{1/2}} \mathbf{B}^n \mathbf{e}_x.$$

Then, (4.27) can be rewritten as

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \tilde{\mathbf{s}}^n(q^n, b^n, \mathbf{B}^n, x) T^n, \tag{4.30}$$

where T^n has a Student's t-distribution with $b^n - K + 1$ degrees of freedom, mean 0 and scale parameter 1.

At time n, the vector $\boldsymbol{\theta}^{n+1}$ of future beliefs is unknown. However, we see from (4.30) that our uncertainty about this vector originates from a single scalar random variable. This is in line with previous work on ranking and selection with known correlation structures [77], where the scalar random variable is normally distributed. When the correlations are unknown, we use Student's t-distribution, forming a precise analogy to classical frequentist statistics.

4.2.4 Information loss due to approximate Bayesian inference

The KL divergence $\mathcal{D}_{KL}^n(\xi \parallel p)$ can be thought of as the incremental information loss incurred by forcing conjugacy after the $(n+1)$st observation, under the assumption that our beliefs at time n are accurately represented by a normal-Wishart distribution. This section shows that, under the probability measure $\bar{\mathbb{P}}$, the incremental information loss converges to zero in probability as $n \to \infty$. That is, if conjugacy is maintained up to time n, the error due to a single application of approximate Bayesian inference at time $n+1$ will become vanishingly small for large enough n.

This result is intended to provide the intuition that, over time, the learning model with scalar observations bears greater resemblance to a conjugate learning model. As in Section 4.2.3, we assume that the decision-maker's beliefs at time n are represented by the normal-Wishart distribution. We begin by showing in Proposition 4.8 that the degrees of freedom parameter b^n goes to infinity, eventually leading to the result that the incremental loss from one additional observation vanishes to zero.

Proposition 4.8. *If b^0 is sufficiently large, then $\Delta b^n \in (0,1)$ $\bar{\mathbb{P}}$-a.s. and $b^n \to \infty$ as $n \to \infty$.*

Proof. Proof: Let $\Delta b^n = 1$ in (4.22) and from which we have

$$\begin{aligned}
\frac{\partial \mathcal{D}^n(\xi|p)}{\partial b^{n+1}}\Big|_{b^{n+1}=b^n+1} &\geq \frac{1}{2}\left(\frac{b^n K+1}{b^n+1} - K\right) + \frac{1}{4}\sum_{i=1}^{K}\psi'\left(\frac{b^n-i+2}{2}\right) - \frac{1}{4}\psi'\left(\frac{b^n-K+2}{2}\right)\\
&= -\frac{K-1}{2(b^n+1)} + \frac{1}{4}\sum_{i=1}^{K-1}\psi'\left(\frac{b^n-i+2}{2}\right)\\
&\geq -\frac{K-1}{2(b^n+1)} + \frac{1}{4}\sum_{i=1}^{K-1}\frac{2}{b^n-i+2}\\
&> -\frac{K-1}{2(b^n+1)} + \frac{K-1}{2(b^n+1)} = 0.
\end{aligned}$$

Let $\Delta b^n = 0$ in (4.22), we can show that for any $\epsilon > 0$, there exists sufficiently large b^n such that the first term is less than ϵ. We also observe that

$$\frac{1}{b^n} - \frac{1}{2}\psi'\left(\frac{b^n-K+1}{2}\right) < \frac{1}{b^n} - \frac{1}{b^n-K+1} < 0.$$

Since (4.22) is a continuous function of Δb^n on $[0,1]$, we know that $\Delta b^n \in (0,1)$, whence b^n has a limit by the monotone convergence theorem. In the following, we will prove that b^n goes to infinity by contradiction. Assume that there exists $M < \infty$ such that b^n converges to M. This suggests that Δb^n converges to zero. Taking the limit of (4.22) as n goes to infinity yields,

$$\lim_{n\to\infty} \frac{(\hat{y}_x^{n+1} - \theta_x^n)^2}{B_{xx}^{n+1}} \frac{(q^n)^2 b^{n+1}(K-1)}{(q^n b^{n+1} + b^{n+1} - K + 1)^2} = \psi'\left(\frac{M-K+1}{2}\right) - \frac{1}{M} \tag{4.31}$$

From Lemma 4.7, the predictive distribution of $\hat{y}_x^{n+1}$ is a Student's t-distribution with $b^n - K + 1$ degrees of freedom, mean θ_x^n and variance $\frac{q^n+1}{q^n(b^n-K+1}B_{xx}^n$. It follows that

$$\frac{(\hat{y}_x^{n+1} - \theta_x^n)}{(B_{xx}^n)^{1/2}} = T^n \left(\frac{q^n+1}{q^n(b^n-K+1)}\right)^{1/2}. \tag{4.32}$$

As $q^n \to \infty$,

$$\lim_{n\to\infty} \frac{(q^n)^2 b^{n+1}(K-1)}{(q^n b^{n+1} + b^{n+1} - K + 1)^2} = \frac{1}{M},$$

whence (4.31) can be rewritten as

$$\lim_{n\to\infty} l(T^n) = \left[\psi'\left(\frac{M-K+1}{2}\right) - \frac{1}{M}\right] M^2, \tag{4.33}$$

where $l(T^n)$ is a function of the random variable T^n. Since $b^n \to M$, the random variable T^n converges weakly to a Student's t-distribution with $M-K+1$ degrees of freedom. That means that (4.33) cannot hold almost surely. Therefore, we conclude that the degrees of freedom b^n goes to infinity as $n \to \infty$. □

The fact that the degrees of freedom parameter b^n goes to infinity is a key to the other results in this section. We next show several preliminary results concerning the updating equation for $\mathbf{B}^n$.

Proposition 4.9. *Let*

$$M_x^n = \left(\frac{b^{n+1}}{b^n+1}\right)\left(\frac{q^n(b^{n+1}-K+1)}{q^n b^{n+1}+b^{n+1}-K+1}\frac{(\hat{y}_x^{n+1}-\theta_x^n)^2}{(B_{xx}^n)^2}-\frac{1}{b^n B_{xx}^n}\right).$$

Then, $M_x^n B_{xx}^n$ converges to zero in $\bar{\mathbb{P}}$-probability.

Proof. Proof: First note that

$$M_x^n B_{xx}^n = \left(\frac{b^{n+1}}{b^n+1}\right)\left(\frac{q^n(b^{n+1}-K+1)}{q^n b^{n+1}+b^{n+1}-K+1}\frac{(\hat{y}_x^{n+1}-\theta_x^n)^2}{B_{xx}^n}-\frac{1}{b^n}\right),$$

therefore showing that $M_x^n B_{xx}^n$ converges to zero in probability is equivalent to showing that $(\hat{y}_x^{n+1}-\theta_x^n)^2/B_{xx}^n$ converges to zero in probability. For any $\epsilon > 0$, using (4.32) and Chebyshev's inequality, we know that

$$\bar{\mathbb{P}}\left(\left|\frac{(\hat{y}_x^{n+1}-\theta_x^n)}{(B_{xx}^n)^{1/2}}\right| > \epsilon\right) \leq \frac{q^n+1}{q^n(b^n-K+1)}\frac{1}{\epsilon^2}.$$

Then we have

$$\lim_{n\to\infty}\bar{\mathbb{P}}\left(\left|\frac{(\hat{y}_x^{n+1}-\theta_x^n)}{(B_{xx}^n)^{1/2}}\right| > \epsilon\right) = 0,$$

as required. □

If we view the updating equation (4.13) from the viewpoint of stochastic approximation, then the quantity $M_x^n B_{xx}^n$ can be considered as the stepsize. Since the stepsize converges to zero, this guarantees that the change in the matrix $\mathbf{B}^n$ will not be too large.

We will provide two propositions related to the determinant and the trace of the matrix $\mathbf{B}^n$. Instead of checking the matrix $\mathbf{B}^n$ componentwise, these two results provide the changes in the determinant and the trace analytically, which are useful for studying the asymptotic behavior of the matrix $\mathbf{B}^n$.

Proposition 4.10. *The determinant of* $\mathbf{B}^n$ *is updated recursively through*

$$\det(\mathbf{B}^{n+1}) = \det(\mathbf{B}^n)\left(\frac{b^{n+1}}{b^n}\right)^K \left(1 + \frac{b^n}{b^{n+1}} M_x^n B_{xx}^n\right). \tag{4.34}$$

Proof. Proof: First rewrite (4.20) as

$$\mathbf{B}^{n+1} = \mathbf{B}^n + \mathbf{B}^n \left(\frac{\Delta b^n}{b^n}\mathbf{I}_K + M_x^n \mathbf{e}_x \mathbf{e}_x' \mathbf{B}^n\right)$$

It follows from the matrix determinant lemma [88] that

$$\begin{aligned}
\det(\mathbf{B}^{n+1}) &= \det(\mathbf{B}^n)\det\left(\mathbf{I}_K + \frac{\Delta b^n}{b^n}\mathbf{I}_K + M_x^n \mathbf{e}_x \mathbf{e}_x' \mathbf{B}^n\right) \\
&= \det(\mathbf{B}^n)\det\left(\frac{b^{n+1}}{b^n}\mathbf{I}_K + M_x^n \mathbf{e}_x \mathbf{e}_x' \mathbf{B}^n\right) \\
&= \det(\mathbf{B}^n)\det\left(\frac{b^{n+1}}{b^n}\mathbf{I}_K\right)\left(1 + \frac{b^n}{b^{n+1}} M_x^n \mathbf{e}_x' \mathbf{B}^n \mathbf{e}_x\right) \\
&= \det(\mathbf{B}^n)\left(\frac{b^{n+1}}{b^n}\right)^K \left(1 + \frac{b^n}{b^{n+1}} M_x^n B_{xx}^n\right),
\end{aligned}$$

as desired. □

Proposition 4.11.

$$\operatorname{tr}\left(\mathbf{B}^n(\mathbf{B}^{n+1})^{-1}\right) = \frac{b^nK+1}{b^{n+1}} - \frac{q^n}{2}(\boldsymbol{\theta}^{n+1}-\boldsymbol{\theta}^n)'(\mathbf{B}^{n+1})^{-1}(\boldsymbol{\theta}^{n+1}-\boldsymbol{\theta}^n) - \frac{1}{2}(\hat{y}_x^{n+1}-\theta_x^{n+1})^2\frac{b^{n+1}-K+1}{B_{xx}^{n+1}}.$$

Proof. Proof: Multiplying (4.24) by $\mathbf{B}^{n+1}$ from the left yields

$$\begin{aligned}
&\frac{b^n}{2}\mathbf{I}_K - \frac{b^{n+1}}{2}\mathbf{B}^n(\mathbf{B}^{n+1})^{-1} - \frac{q^n b^{n+1}}{2}(\boldsymbol{\theta}^{n+1}-\boldsymbol{\theta})(\boldsymbol{\theta}^{n+1}-\boldsymbol{\theta}^n)'(\mathbf{B}^{n+1})^{-1} \\
&\qquad + \left(\frac{1}{2B_{xx}^{n+1}} - \frac{b^{n+1}-K+1}{2(B_{xx}^{n+1})^2}(\hat{y}_x^{n+1}-\theta_x^{n+1})^2\right)\mathbf{B}^{n+1}\mathbf{e}_x\mathbf{e}_x' = 0
\end{aligned}$$

Taking `trace` on both sides, it gives

$$\begin{aligned}
&\frac{b^nK}{2} - \frac{b^{n+1}}{2}\operatorname{tr}\left(\mathbf{B}^n(\mathbf{B}^{n+1})^{-1}\right) - \frac{q^n b^{n+1}}{2}(\boldsymbol{\theta}^{n+1}-\boldsymbol{\theta}^n)'(\mathbf{B}^{n+1})^{-1}(\boldsymbol{\theta}^{n+1}-\boldsymbol{\theta}^n) \\
&\qquad \left(\frac{1}{2B_{xx}^{n+1}} - \frac{b^{n+1}-K+1}{2(B_{xx}^{n+1})^2}(\hat{y}_x^{n+1}-\theta_x^{n+1})^2\right)B_{xx}^{n+1} = 0
\end{aligned} \tag{4.35}$$

Solving for $\operatorname{tr}(\mathbf{B}^n(\mathbf{B}^{n+1})^{-1})$ from (4.35) completes the proof. □

The next lemma finds the limit of a sequence of expressions involving the gamma and digamma functions. The limit will appear repeatedly in the proof of our main results later.

Lemma 4.12. *For any $\alpha, \beta, \gamma \in \mathbb{R}$,*

$$\lim_{x\to\infty} \log \frac{\Gamma(x+\alpha)}{\Gamma(x+\beta)} - (\alpha-\beta)\psi(x+\gamma) = 0. \tag{4.36}$$

Proof. Proof: Notice that

$$\log \frac{\Gamma(x+\alpha)}{\Gamma(x+\beta)} - (\alpha-\beta)\psi(x+\gamma) = \log\left(\frac{\Gamma(x+\alpha)}{\Gamma(x+\beta)} e^{-(\alpha-\beta)\psi(x+\gamma)}\right),$$

then to prove (4.36) is equivalent to prove

$$\lim_{x\to\infty} \frac{\Gamma(x+\alpha)}{\Gamma(x+\beta)} e^{-(\alpha-\beta)\psi(x+\gamma)} = 1.$$

From [89], we have the asymptotic expansion

$$\frac{\Gamma(x+\alpha)}{\Gamma(x+\beta)} = x^{\alpha-\beta}\left[1 + \frac{(\alpha-\beta)(\alpha+\beta-1)}{2x} + O(x^{-2})\right]. \tag{4.37}$$

For $x > 0$, we have [90]

$$\log x - \frac{1}{x} < \psi(x) < \log x - \frac{1}{2x}.$$

Without loss of generality, we assume that $\alpha > \beta$. This gives

$$(x+\gamma)^{-(\alpha-\beta)} e^{\frac{\alpha-\beta}{2(x+\gamma)}} < e^{-(\alpha-\beta)\psi(x+\gamma)} < (x+\gamma)^{-(\alpha-\beta)} e^{\frac{\alpha-\beta}{x+\gamma}}.$$

Therefore, by (4.37), we find that

$$\lim_{x\to\infty} \frac{\Gamma(x+\alpha)}{\Gamma(x+\beta)} e^{-(\alpha-\beta)\psi(x+\gamma)} = 1.$$

□

We now state the key theorem. As the number of measurements goes to infinity, the KL divergence converges to zero in probability. This suggests that the incremental information loss incurred by forcing conjugacy vanishes.

Theorem 4.13. *As $n \to \infty$, the KL divergence $\mathcal{D}^n_{KL}(\xi^{n+1} \parallel p^{n+1})$ converges to zero in $\bar{\mathbb{P}}$-probability.*

Proof. Proof: The constant C omitted in Proposition (4.1) can be given explicitly as

$$C = \frac{1}{2}\log\frac{q^n}{q^n+1} - \frac{1}{2}\log B^n_{xx} + \log\left(\Gamma\left(\frac{b^n-K+2}{2}\right)\right) - \log\left(\Gamma\left(\frac{b^n-K+1}{2}\right)\right)$$
$$- \frac{b^n-K+2}{2}\log\left(1+\frac{q^n}{(q^n+1)B^n_{xx}}(\hat{y}^{n+1}_x-\theta_x)^2\right),$$

whence the KL divergence $\mathcal{D}^n_{KL}(\xi^{n+1} \parallel p^{n+1})$ can be expressed as

$$\mathcal{D}^n_{KL}(\xi^{n+1} \parallel p^{n+1}) = \frac{b^n}{2}\log\frac{\det(\mathbf{B}^{n+1})}{\det(\mathbf{B}^n)} + \frac{b^{n+1}}{2}\operatorname{tr}\left(\mathbf{B}^n(\mathbf{B}^{n+1})^{-1}\right) - \frac{b^{n+1}}{2}K \tag{4.38}$$

$$+ \frac{q^n b^{n+1}}{2}(\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^n)'(\mathbf{B}^{n+1})^{-1}(\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^n) + \frac{1}{2}(\hat{y}^{n+1}_x - \theta^{n+1}_x)^2\frac{b^{n+1} - K + 1}{B^{n+1}_{xx}} \tag{4.39}$$

$$- \frac{b^n - K + 2}{2}\log\left(1 + \frac{q^n}{(q^n+1)B^n_{xx}}(\hat{y}^{n+1}_x - \theta^n_x)^2\right) \tag{4.40}$$

$$+ \frac{b^{n+1} - b^n}{2}\sum_{i=1}^{K}\psi\left(\frac{b^{n+1} - i + 1}{2}\right) + \sum_{i=1}^{K}\log\Gamma\left(\frac{b^n - i + 1}{2}\right) - \sum_{i=1}^{K}\log\Gamma\Bigg(\frac{b^{n+1}}{} \tag{4.41}$$

$$- \frac{1}{2}\psi\left(\frac{b^{n+1} - K + 1}{2}\right) + \log\Gamma\left(\frac{b^n - K + 2}{2}\right) - \log\Gamma\left(\frac{b^n - K + 1}{2}\right) \tag{4.42}$$

$$+ \frac{1}{2}K\log\left(\frac{q^{n+1}}{q^n}\right) + \frac{1}{2}K\frac{q^n}{q^{n+1}} - \frac{1}{2}K + \frac{1}{2q^{n+1}} + \frac{1}{2}\log\frac{q^n}{q^n+1} \tag{4.43}$$

$$+ \frac{1}{2}\log B^{n+1}_{xx} - \frac{1}{2}\log B^n_{xx}. \tag{4.44}$$

Following Propositions 4.10 and 4.11, the terms in (4.38) and (4.39) can be simplified as

$$\frac{b^n K}{2}\log\frac{b^{n+1}}{b^n} + \frac{b^n}{2}\log\left(1 + \frac{b^n}{b^{n+1}}M^n_x B^n_{xx}\right) + \frac{b^n K + 1}{2} - \frac{b^{n+1}K}{2}.$$

As $b^n \to \infty$, it is easy to show that

$$\lim_{n\to\infty}\frac{b^n K}{2}\log\frac{b^{n+1}}{b^n} + \frac{b^n K}{2} - \frac{b^{n+1}K}{2} = 0.$$

Also, we can show that

$$\lim_{n\to\infty}\frac{b^n}{2}\log\left(1 + \frac{b^n}{b^{n+1}}M^n_x B^n_{xx}\right) - \frac{b^n - K + 2}{2}\log\left(1 + \frac{q^n}{(q^n+1)B^n_{xx}}(\hat{y}^{n+1}_x - \theta^n_x)^2\right) = -\frac{1}{2}.$$

This suggests that the sum of (4.38), (4.39) and (4.40) approaches zero as $n \to \infty$. Using Lemma 4.12, we can show that both (4.41) and (4.42) go to zero as $n \to \infty$. It is easy to check that (4.43) approaches zero as $n \to \infty$. It follows from (4.20) that

$$B_{xx}^{n+1} = \frac{b^{n+1}}{b^n + 1} B_{xx}^n + \frac{1}{b^n + 1} \frac{q^n b^{n+1} (\hat{y}_x^{n+1} - \theta_x^n)^2}{1 + \frac{q^n b^{n+1}}{b^{n+1} - K + 1}}.$$

Therefore,

$$\log \frac{B_{xx}^{n+1}}{B_{xx}^n} = \log \left(\frac{b^{n+1}}{b^n + 1} + \frac{1}{b^n + 1} \frac{q^n b^{n+1}}{1 + \frac{q^n b^{n+1}}{b^{n+1} - K + 1}} \frac{(\hat{y}_x^{n+1} - \theta_x^n)^2}{B_{xx}^n} \right),$$

which is easily shown to converge to zero. This completes the proof. □

4.3 The Value Of Information

Value of information procedures allocate the simulation budget by evaluating the potential of new observations to improve the current estimate of the best value (see [31] for a survey). The information potential is defined in terms of the expected difference in the estimated objective value before and after the next observation occurs. We do not know exactly how an observation of alternative x will change our beliefs about the best alternative, but we can compute an expectation over the predictive distribution in (4.30). In this way, we can "look ahead" to the random outcome of the next observation, attempting to anticipate the results before we see them. If we sample from alternative x at time n and collect observation $\hat{y}_x^{n+1}$, the value of information is defined as

$$\mathcal{V}_x(S^n) = \mathbb{E}^n \left[\max_i \theta_i^{n+1} \mid x^n = x \right] - \max_i \theta_i^n,$$

where $\mathbb{E}^n$ is the conditional expectation taken with respect to the decision-maker's distribution of belief at time n, and x^n denotes the alternative measured at time n.

Note that the predictive distribution of $\boldsymbol{\theta}^{n+1}$ depends on q^n, b^n and $\mathbf{B}^n$ only through the vector $\tilde{\mathbf{s}}^n$ from (4.30). As a result, the expected value of information can be rewritten as

$$\mathcal{V}(\boldsymbol{\theta}^n, \tilde{\mathbf{s}}(S^n, x), m) = \mathbb{E}^n \left[\max_i \theta_i^n + \tilde{\mathbf{s}}(S^n, x^n) T_m \mid x^n = x \right] - \max_i \theta_i^n, \tag{4.45}$$

where $\mathcal{V}$ is defined by $\mathcal{V}(\mathbf{a}, \mathbf{b}, m) = \mathbb{E}[\max_i a_i + b_i T_m] - \max_i a_i$, $\mathbf{a}$ and $\mathbf{b}$ are $K \times 1$ vectors. The random variable T_m follows a univariate Student's t-distribution with degrees of freedom m, mean 0 and variance 1.

Once again, (4.45) assumes that the decision-maker's beliefs are represented by a normal-Wishart distribution at each time step. In practice, the normal-Wishart distribution is an approximation of the true posterior beliefs, updated using (4.17)-(4.20). By using this approximation to represent our uncertainty, we can solve (4.45) in closed form, leading to a computationally efficient procedure.

We introduce a fully sequential policy called *Projected Learning of Unknown Correlations with Knowledge Gradients* (PLUCK). The PLUCK policy chooses an alternative by computing

$$x^{PLUCK}(\mathbf{S}^n) = \arg\max_x \mathcal{V}(\boldsymbol{\theta}^n, \tilde{\mathbf{s}}(S^n, x), m). \tag{4.46}$$

We now show how the value of information can be computed exactly under $\bar{\mathbb{P}}$.

4.3.1 Computation of the Value of Information

To compute the expected value of information, we start by defining a function $h : \mathbb{R} \mapsto \{1, 2, \ldots, K\}$ as

$$h(t) := \max(\operatorname*{argmax}_i \; a_i + b_i t).$$

The function h tells us which alternative is the best among $\{1, 2, \cdots, K\}$ in the sense of having largest value of $a_i + b_i t$ given $T_m = t$. The largest index is chosen if multiple alternatives tie. Instead of calculating $\mathcal{V}(\mathbf{a}, \mathbf{b}, m)$ directly, we notice that

$$\max_i a_i + b_i T_m = a_{h(T_m)} + b_{h(T_m)} T_m,$$

and rewrite $a_{h(T_m)} + b_{h(T_m)} T_m$ as a telescoping sum,

$$a_{h(0)} + b_{h(0)} T_m + \left[\sum_{i=h(0)}^{h(T_m)-1} (a_{i+1} - a_i) + (b_{i+1} - b_i) T_m \right] + \left[\sum_{i=h(T_m)}^{h(0)-1} (a_i - a_{i+1}) + (b_i - b_{i+1}) T_m \right].$$

Using standard techniques (see Section 5.3 of [5]), we can find a non-decreasing sequence $\{c_i\}_{i=0}^{K}$ such that $h(z) = i$ if and only if $z \in [c_{i-1}, c_i)$. It follows that $\mathcal{V}(\mathbf{a}, \mathbf{b}, m)$ can be written as

$$\mathcal{V}(\mathbf{a}, \mathbf{b}, m) = \sum_{i=1}^{K-1} (b_{i+1} - b_i) \mathbb{E}[(T_m - |c_i|)^+].$$

To continue the computational procedure, we need an analytical form for the tail expectation of a univariate Student's t-distribution. Denote the pdf and cdf of a standard Student's t-distribution with m degrees of freedom as $g_m(\cdot)$ and $G_m(\cdot)$, respectively. We can easily rewrite $\mathbb{E}\left[(T_m - |c_i|)^+\right]$ as

$$\mathbb{E}\left[(T_m - |c_i|)^+\right] = \mathbb{E}(T_m \cdot \mathbf{1}_{\{T_m > |c_i|\}}) - |c_i|(1 - G_m(|c_i|)).$$

It also can be shown [30] that

$$\mathbb{E}(T_m \cdot \mathbf{1}_{\{T_m > |c_i|\}}) = \frac{m + c_i^2}{m - 1} g_m(|c_i|).$$

With the analytical form for the tail expectation, the value of information can be expressed as:

$$\mathcal{V}(\mathbf{a}, \mathbf{b}, m) = \sum_{i=1}^{K-1} (b_{i+1} - b_i) \left(\frac{m + c_i^2}{m - 1} g_m(|c_i|) - |c_i|(1 - G_m(|c_i|)) \right). \tag{4.47}$$

We note that (4.46) has the same computational complexity as the analogous VIP for R&S with known correlation structures [5, 77]. The breakpoints c_i can be computed in $O(K \log K)$ time. Repeating this for every alternative yields $O(K^2 \log K)$. Just as in the learning model of Section 4.2.1, we can account for unknown correlations for the same computational cost.

4.3.2 Monotonicity of the Value of Information

The value of information calculated in (4.47) depends on the degrees of freedom m of the Student's t-distribution. Lemma 4.7 shows that in the nth stage, the predictive distribution of the new observation $\hat{y}_x$ follows a univariate Student's t-distribution with degrees of freedom $m = b^n - K + 1$. The parameter b^n is updated through (4.18) and increases as the PLUCK policy collects information. The relationship between the value of information and the degrees of freedom is summarized in the next theorem.

Theorem 4.14. *For fixed $K \times 1$ vectors $\mathbf{a}$ and $\mathbf{b}$, the value of information $\mathcal{V}(\mathbf{a}, \mathbf{b}, m)$ is a decreasing function in the degrees of freedom parameter m.*

Proof. Proof: Let $\mathcal{V}(\mathbf{a}, \mathbf{b}, m)$ and $\mathcal{V}(\mathbf{a}, \mathbf{b}, n)$ be the values of information for two different values of the degrees of freedom parameter, with $m \geq n$. Since $\mathbf{a}$ and $\mathbf{b}$ are fixed, it is sufficient to consider $\mathbb{E}(T_m - |c_i|)^+$ and $\mathbb{E}(T_n - |c_i|)^+$.

Let $g_m(t)$ and $g_n(t)$ be the probability density function of Student's t distributions with m and n degrees of freedom, respectively. There exists $c^* > 0$ such that $g_m(c^*) = g_n(c^*)$ with $g_m(t) \leq g_n(t)$ on $[c^*, \infty)$ and $g_m(t) > g_n(t)$ on $[0, c^*)$. We will consider two cases:

(i) If $|c_i| \geq c^*$, then $g_m(t) \leq g_n(t)$ for $t \in [|c_i|, \infty)$.

$$E(T_m - |c_i|)^+ = \int\limits_{|c_i|}^{\infty} (t - |c_i|)\phi_m(t)dt \leq \int\limits_{|c_i|}^{\infty} (t - |c_i|)\phi_n(t)dt = \mathbb{E}(T_n - |c_i|)^+.$$

(ii) If $|c_i| < c^*$, then $g_m(t) > g_n(t)$ for $t \in [0, |c_i|)$.

$$\begin{aligned}
&\mathbb{E}(T_m - |c_i|)^+ - \mathbb{E}(T_n - |c_i|)^+ \\
&= \int\limits_{|c_i|}^{c^*} (t - |c_i|)g_m(t)dt + \int\limits_{c^*}^{\infty} (t - |c_i|)g_m(t)dt - \int\limits_{|c_i|}^{c^*} (t - |c_i|)g_n(t)dt - \int\limits_{c^*}^{\infty} (t - |c_i|)g_n(t)dt \\
&= \int\limits_{|c_i|}^{c^*} (t - |c_i|)g_m(t)dt - \int\limits_{|c_i|}^{c^*} (t - |c_i|)g_n(t)dt + \int\limits_{c^*}^{\infty} (t - |c_i|)g_m(t)dt - \int\limits_{c^*}^{\infty} (t - |c_i|)g_n(t)dt \\
&= \int\limits_{|c_i|}^{c^*} (t - |c_i|)g_m(t)dt - \int\limits_{|c_i|}^{c^*} (t - |c_i|)g_n(t)dt + \int\limits_{0}^{c^*} (t - |c_i|)g_m(t)dt - \int\limits_{0}^{c^*} (t - |c_i|)g_n(t)dt \\
&= \int\limits_{0}^{|c_i|} (t - |c_i|)g_n(t)dt - \int\limits_{0}^{|c_i|} (t - |c_i|)g_m(t)dt < 0
\end{aligned}$$

It follows from (i) and (ii) that for any $|c_i| > 0$,

$$\mathcal{V}(\mathbf{a}, \mathbf{b}, m) = \sum_{i=1}^{K-1} (b_{i+1} - b_i)E(T_m - |c_i|)^+ \leq \sum_{i=1}^{K-1} (b_{i+1} - b_i)E(T_n - |c_i|)^+ = \mathcal{V}(\mathbf{a}, \mathbf{b}, n).$$

Therefore the value of information decreases in the number of degrees of freedom. □

The theorem suggests that the value of information decreases as the degrees of freedom increase, with all else being equal. In other words, the same information, under the same estimated means and covariances, is less valuable when we have already accumulated many other observations.

Theorem 4.14 leads to an interesting comparison with earlier work on R&S with known correlations. It is a well-known result that T_m converges weakly to a standard normal random variable as the degrees of freedom m goes to infinity. Recall that, when the correlation structure is known, we calculate a version of (4.45) using a standard normal random variable; see (5.16) in [5]. Theorem 4.14 suggests that, given the same estimated means and conditional covariances, the value of information is inherently higher when the correlations are unknown. That is, a single measurement provides more information when we are learning both means and covariances.

4.4 Numerical Experiments

We present experimental results demonstrating the value added by learning unknown correlations using PLUCK. Throughout this section, we considered six policies: the PLUCK policy, the correlated KG (CKG) policy in [77], a sequential modified version of proportional to variance (PTV) policy, a sequential OCBA policy designed for opportunity cost in [91], a Greedy policy and the LL policy with linear loss in [79]. We briefly explain the distinctions between the remaining policies below.

Both PLUCK and CKG are designed to sample sequentially, with CKG assuming a known covariance structure and using a conjugate Bayesian learning model. This comparison allows us to see the value added by incorporating unknown correlations into our decision-making. The PTV and greedy policy are also sequential, and make simulation decisions at time n in the following ways: PTV policy chooses the alternative with the highest variance; the greedy policy chooses the alternative $\arg\max_x \theta^n_x$. For both of these methods, we use our approximate Bayesian learning model in (4.17)-(4.20) to update our beliefs about the alternatives. This comparison allows us to see the value added by using the PLUCK policy to make decisions, in addition to the value of learning unknown correlations. Lastly, the LL policy of [79] first screens out a subset of alternatives, then allocates the simulation budget equally among the rest. This structure allows a conjugate normal-Wishart prior to be used, with the drawback that the policy often samples alternatives that do not provide a lot of useful information. The LL policy can also be extended to allow multiple screening stages; however, this approach works best with large sampling budgets. In our experiments, we consider problems where the simulation budget is comparable to the number of alternatives, making it difficult to run LL for more than two stages.

4.4.1 Wind Farm Placement Example

Our study is based on the wind farm placement problem mentioned in the introduction. For the purpose of these experiments, we use the wind speed at a location as a stand-in for power output. In practice, wind speed data are collected at each location simultaneously, rather than sequentially. However, this also allows us to use the data to demonstrate the effectiveness of our policy by comparing its performance to how well we could have done. Practical applications of sequential simulation in wind farm placement use complex physics-based models incorporating factors other than wind speed, thus necessitating the use of sequential simulation. For our purposes, the public availability of wind-speed data allows us to create a realistic test setting for the PLUCK algorithm.

We used hourly wind speed data [92] across the United States with latitude and longitude resolution of 0.125 degrees. The data consist of two components: the zonal component u (in the west-east direction) and the meridional component v (in the north-south direction). Assuming for the purpose of this example that all the wind turbines can be placed in the right direction, we focus on the magnitude of the wind speed, which is defined as $\sqrt{u^2 + v^2}$. The objective is to select the location with the highest wind speed over a set of 64 locations.

We considered data from four regions across the United States: Kansas, Washington, Iowa and Oklahoma. All regions have had a high percentage of wind power generation, or a large amount of wind capacity installed, in recent years. For each of the four regions, we selected 64 different locations sitting on an 8×8 grid (the areas of these grids range from 3500 to 4500 square miles) within the region as alternatives for wind farm placement.

We used 1800 days of data to estimate the mean and covariance matrix of a multivariate normal distribution. These fitted parameters were taken to represent the "true" underlying sampling distribution. In our experiments, individual

		Policies					
Experiment	Performance Measure	PLUCK	CKG	Greedy	OCBA	PTV	LL
Kansas	Opportunity Cost	0.0314	0.1270	0.3073	0.2760	0.0537	0.2848
	Standard Errors	0.0026	0.0068	0.0045	0.0040	0.0036	0.0047
Washington	Opportunity Cost	0.0640	0.0917	0.2651	0.2021	0.1125	0.1647
	Standard Errors	0.0051	0.0050	0.0018	0.0053	0.0060	0.0058
Iowa	Opportunity Cost	0.0512	0.0892	0.1917	0.1526	0.1229	0.1335
	Standard Errors	0.0037	0.0040	0.0046	0.0036	0.0034	0.0040
Oklahoma	Opportunity Cost	0.0509	0.0911	0.2413	0.1401	0.1816	0.2031
	Standard Errors	0.0040	0.0050	0.0013	0.0017	0.0039	0.0042

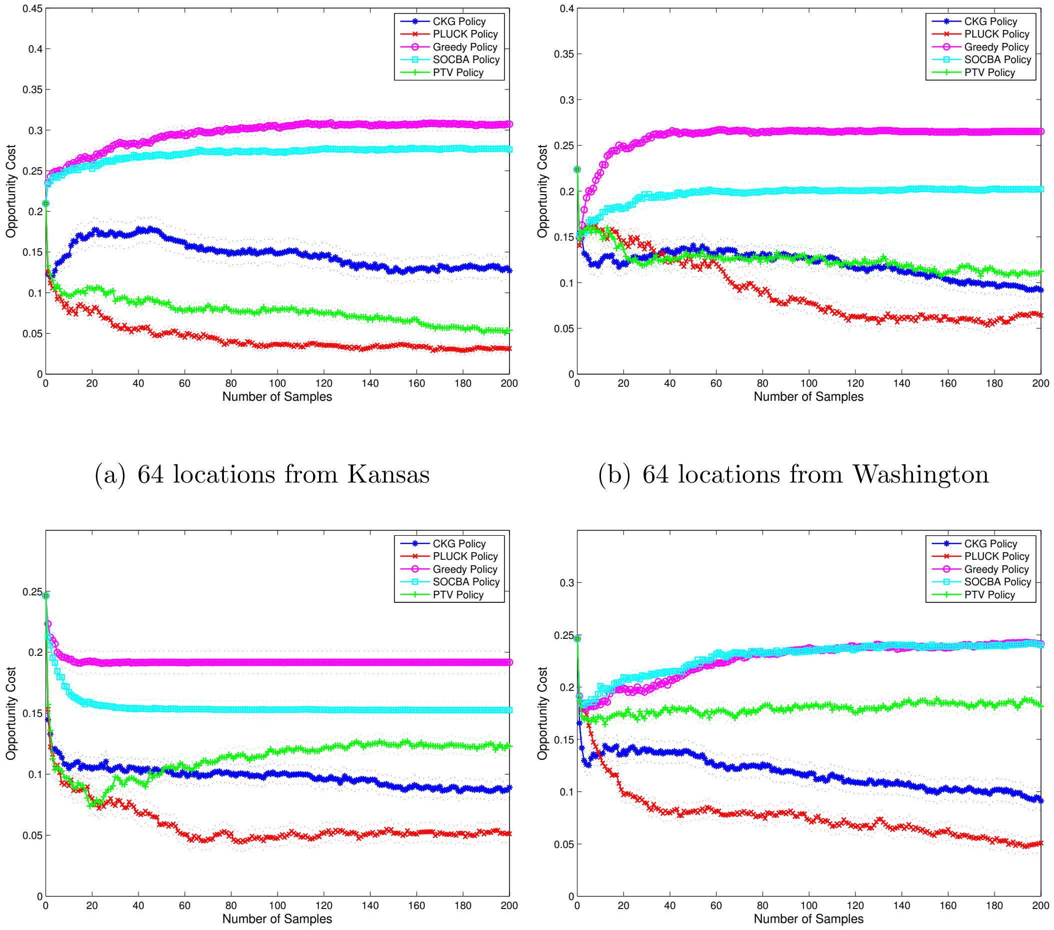
Table 4.1: Final opportunity cost and standard errors for the experiments

observations were generated by simulating from normal distributions with the true parameter values. However, the policies were not allowed to see these true values when making decisions.

It is critical to collect information efficiently when the decision-maker's prior beliefs are inaccurate or misleading. To show that the PLUCK policy is particularly effective in such a situation, we used a small number of data points to create a prior for which the location that appeared to be the best was quite different from the true best location. We discuss this issue further below; for now, we note that each policy was given a budget of $N = 200$ measurements to correct this initial error. Table 4.1 gives the final opportunity cost, which is defined as

$$C^{\pi} = \max_x \mu_x - \mu_{(\arg\max_x \boldsymbol{\theta}_x^N)}$$

after N measurements for each policy π, averaged over 500 sample paths. Lower opportunity cost suggests that a policy selects an alternative closer to the best. The



(a) 64 locations from Kansas

(b) 64 locations from Washington

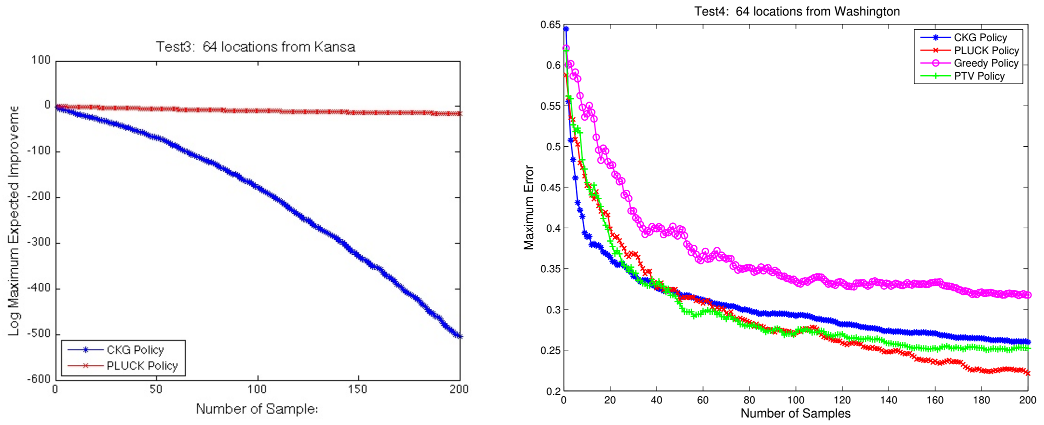
(c) 64 locations from Iowa

(d) 64 locations from Oklahoma

Figure 4.1: Averaged opportunity cost as the number of samples increases, where the dashed lines are the mean plus or minus two standard errors

PLUCK policy outperforms all other competing policies by a statistically significant amount based on Table 4.1, while the CKG policy has smaller final opportunity cost than the other policies in three experiments. The PTV policy performs better than the sequential OCBA policy and the greedy policy. The LL policy performs poorly in all four experiments, possibly because the simulation budget is quite small relative to the number of alternatives.

Figure 4.1 shows how this performance measure changes as the number of sam-



(a) Expected Improvement

(b) Maximum Error

Figure 4.2: Value of information and maximum error as the number of samples increases

ples N varies from 1 to 100. The bands indicate the mean performance measures plus or minus two standard errors. The LL policy is omitted from these figures because it allocates simulations in batch rather than sequentially, by dividing them uniformly across any alternatives that were not screened out. The opportunity cost for PLUCK tends to decrease over time. For CKG, sometimes there is a degradation in performance at the beginning. We conjecture that this behavior arises because CKG assumes a known covariance structure. If the prior beliefs about the correlations are inaccurate, this misdirects the way in which CKG incorporates new information into the posterior. A small amount of information can thus make CKG produce even worse performance than what could be obtained with just the prior. The sequential OCBA policy and the greedy policy tend to work poorly on all cases, and performance of the PTV policy differs dramatically among cases.

We also considered a different set of experiments in which results were averaged across multiple priors constructed from a small sample of wind speed data. Overall, we found that PLUCK still outperformed the competition, with the caveat that all policies were more heavily affected by the initial degradation in performance (the early iterations needed to get a handle on the true correlation structure).

We make two interesting observations from the experimental results. Figure 4.2(a) shows the logarithm of the value of information as computed by both PLUCK and CKG (for a particular experiment), while Figure 4.2(b) shows the maximum absolute difference between the posterior and true means for various policies. Figure 4.2(a) shows that the value of information is much higher when we consider unknown correlations, as suggested by Theorem 4.14. Figure 4.2(b) shows that the

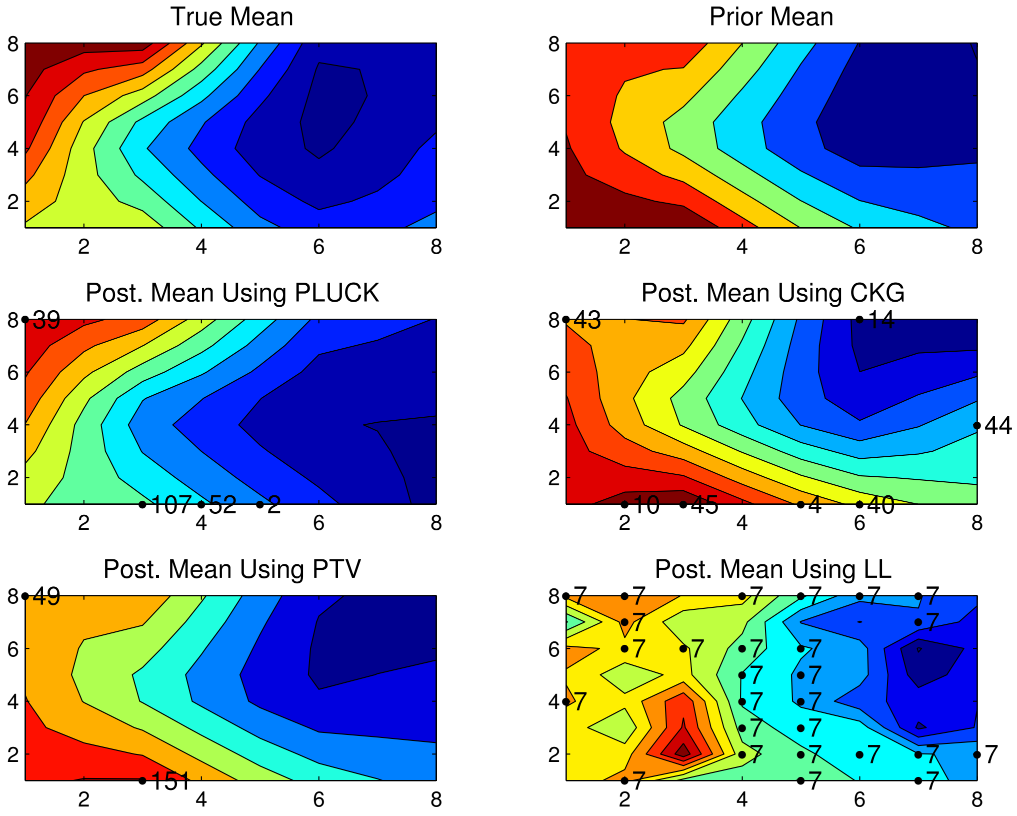


Figure 4.3: Contour map of different policies after 200 measurements

		Policies					
Experiment	Performance Measure	PLUCK	CKG	Greedy	OCBA	PTV	LL
Queue (correlated)	Opportunity Cost	0.3181	0.4677	0.6979	1.0521	0.4862	2.1105
	Standard Errors	0.0380	0.0423	0.0631	0.0645	0.0451	0.034
Network	Opportunity Cost	0.1938	0.2902	0.3339	0.3439	0.3420	0.2374
	Standard Errors	0.0024	0.0063	0.0047	0.0055	0.0066	0.0045

Table 4.2: Final opportunity cost and standard errors for the queue selection and network selection problems

PLUCK policy does a better overall job of estimating the true values.

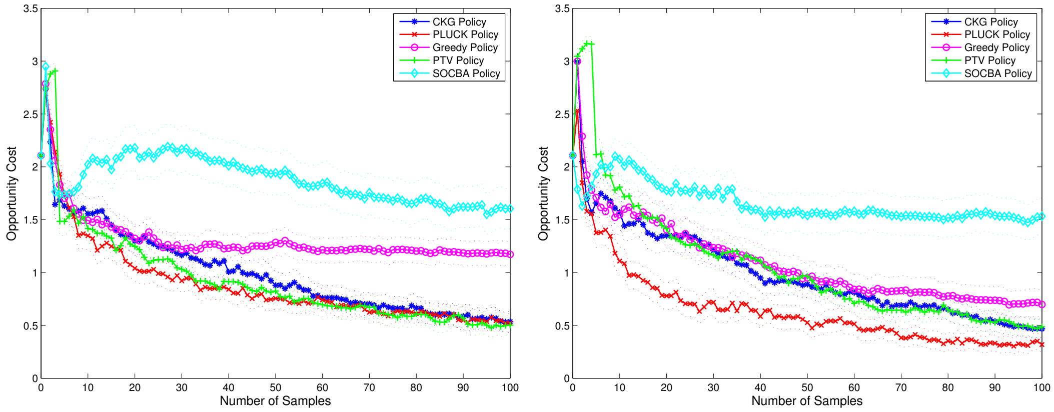
Figure 4.3 gives contour maps of the true means, prior means and posterior means after 100 measurements with four different policies. The sequential OCBA policy and the greedy policy are omitted due to its poor performance. The number of times that each alternative is measured is also shown on the contour maps (zeros are omitted). Red colors indicate higher values. We can see that the true best alternative is in the upper-left corner, whereas the prior misdirects us toward bottom-left. After 200 measurements, the PLUCK policy captures the general trend of the true values, whereas CKG and PTV are still stuck on beliefs that resemble the prior. Observe that both PLUCK and CKG measure the true best alternative in the upper-left corner almost equally often. However, the statistical model used by PLUCK provides more accurate posterior beliefs, leading to a better implementation decision. The LL policy performs poorly and the identified best is far away from the true best. Also, its batch structure allocates many samples to alternatives that do not provide a lot of useful information.

4.4.2 A Single-Server Queue Selection Problem

In simulation, correlations may arise due to common random numbers. However, it is important to keep in mind that correlated beliefs reflect inherent similarities or differences between alternatives, even when the actual simulation output is completely independent. The following example demonstrates that correlated beliefs can enhance performance even when no correlations are present in the simulation output.

Consider 20 first-come, first-served M/G/1 queues. The interarrival times follow an exponential distribution with $\lambda = 0.05$ and the service times follow Pareto distributions with mean service rates $\frac{2}{3}(0.1 + 0.05(i-1))$, $i = 1, 2, ..., 20$. Suppose that the administrator of these queues wishes to reduce costs by closing the worst server, i.e., the one with the largest expected waiting time. System 1 is the worst, having the smallest service rate. However, this is unknown to the administrator.

Observe that, due to the structure of the problem, the performance of queues i and j will exhibit greater similarity if $|i-j|$ is smaller. Thus, even though these queues function independently, our beliefs about their performance can be correlated. Of course, we do not know the problem structure, but we can use an empirical covariance matrix computed from a small sample of observations to initialize our prior distribution, and use PLUCK to improve on this prior. Table 4.2 gives the final opportunity cost. Figure 4.4(a) shows the performance of PLUCK over time when the prior matrix parameter $\mathbf{B}^0$ is diagonal, while 4.4(b) shows performance with $\mathbf{B}^0$ computed using small-sample empirical covariances. We see that, although



(a) With independent belief

(b) With correlated belief

Figure 4.4: Comparing averaged opportunity cost in M/G/1 queue selection problem

the queues function independently, PLUCK can leverage correlated beliefs to learn much more quickly than the other policies.

As before, a small sample of 10 replications was used to create a prior for the covariance matrix.We then compared PLUCK, CKG, PTV, LL, the sequential OCBA policy and the greedy policy by running 1000 macroreplications. The Pollaczek-Khinchin formula can be applied to compute the true expected waiting time. Figure 4.4(b) shows that PLUCK outperforms the competing policies, especially in early stages. The PTV policy and CKG policy are indistinguishable most of the time. The greedy policy and the sequential OCBA policy work poorly in this experiment. The performance of the CKG policy and PTV policy are behind the PLUCK policy initially, but they eventually catch up and lag behind PLUCK slightly. In summary, this experiment suggests that we are learning the similarities-between alternatives, enabling us to discover the optimal solution more quickly even when the actual simulation outputs are independent.

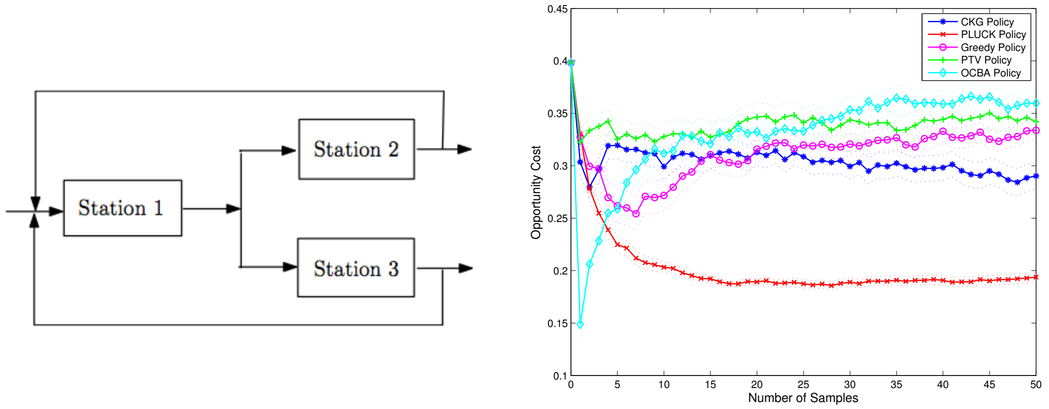
4.4.3 3-Station Jackson Network

Consider a classical 3-station open Jackson network shown in Figure 4.5(a), where the interarrival times and service times follow exponential distributions. Let λ be the total external arrival rate to the system, and let μ_j represent the service rate at station j. Upon completing service at station i, a job leaves the network with probability p_{i0} or is routed to station j with probability p_{ij}.

The goal of the administrator is to minimize the average time spent by customers in the system subject to a constraint on the overall service rate. Suppose that all the available agents can achieve an overall service rate of 3 for stations 2 and 3. Consider 10 different assignments where the service rate at station 2 is $1 + 0.1i$, $i = 1, 2, \cdots, 10$. The performance of different assignments will exhibit correlation due to similarities in the service rates. We chose $\lambda = 0.5$ and the routing probability matrix $\mathbf{P} = [p_{ij}]$ as

$$\mathbf{P} = \begin{array}{c} \begin{array}{cccc} p_{i1} & p_{i2} & p_{i3} & p_{i0} \end{array} \\ \begin{pmatrix} 0 & 0.7 & 0.3 & 0 \\ 0.3 & 0 & 0 & 0.7 \\ 0.2 & 0 & 0 & 0.8 \end{pmatrix} \end{array}$$

Again, a small sample of 10 replications was used to created priors for the mean and covariance matrix. We compared PLUCK, CKG, PTV, LL, the sequential OCBA policy and the greedy policy, where each policy is given a sampling budget of 50. The true expected times in the system for different assignments are computed analytically. The final opportunity costs averaged over 500 sample paths are shown in Figure 4.5(b), and Table 4.2 gives the final opportunity cost. The PLUCK policy again outperforms all the other policies.



(a) A three-station Jackson network (b) Averaged opportunity cost

Figure 4.5: Numerical experiment on a three-station Jackson network

4.5 Conclusion

We have presented the first computationally tractable statistical learning model for fully sequential ranking and selection with unknown correlation structures. The model uses approximate Bayesian inference to represent and update our beliefs about unknown performance means and unknown covariances using the normal-Wishart distribution. We have also derived a value of information procedure that anticipates new information about both the true values and the true correlations when allocating simulations. Previous work in this area has required known correlation structures, an assumption that is likely to be violated in many applications. We relax this assumption, but retain the ability to learn about multiple alternatives from a single observation, for the same computational cost as the known-covariance case.

We believe that our work offers a useful way to tackle large learning problems with difficult correlation structures, and opens up new applications for Bayesian optimal learning.

Chapter 5: Bayesian Learning on Logistic Demand Curves

5.1 Introduction

The problem of business-to-business (B2B) pricing arises in high-volume commercial transactions between businesses. For example, consider the problem faced by a supplier of raw materials negotiating a long-term contract with a large manufacturing concern. After a period of negotiation, the seller quotes a price, which can be accepted or rejected. If the pricing offer is rejected, the seller loses a substantial amount of revenue, but it may not be clear exactly how much lower the offer should have been. If the offer is accepted, the seller makes a profit, but is left wondering whether a somewhat higher offer would still have been accepted. The seller's goal is to maximize total revenue from a sequence of contracts, in the face of uncertainty about buyer behavior.

Dynamic pricing in general is subject to uncertainty. Classic models in revenue management often assume stochastic demand for a product [93, 94], or uncertain customer valuations of it [95]. Recent work, however, has considered the additional dimension that the uncertainty may be *environmental*, that is, the seller does not even know the distribution from which customer valuations are drawn. In practice, this distribution must be estimated, and the estimate must be adjusted over time as new transactions are observed. This gives rise to the problem of "learning and earning," in which the seller does not always prefer the decision that appears to be optimal based on the current demand model (referred to as the "myopic" decision), but rather may engage in more exploratory or experimental behavior. For example, an online retailer may increase or decrease some prices for a period of time, simply to observe the effect on sales. Although this behavior may result in lost revenue, it provides new information that produces a more accurate demand model, enabling better pricing decisions in the future.

The literature has used Bayesian statistics to model environmental uncertainty [33, 34], and different pricing strategies have been proposed to optimize the balance between revenue and information. For example, [35] proposes a one-step look-ahead strategy for problems with logistic revenue curves, while [36] presents an approach based on multi-armed bandit theory. A recent stream of work, represented by [37], [38], [39], and [40], has focused on establishing long-run convergence rates for policies that are mostly myopic, with occasional periods of exploration spaced increasingly further apart. However, in the specific context of B2B pricing, individual transactions typically have high volume (for example, the seller may be negotiating the price of a year's supply of raw materials) and incur high costs (e.g. the time and money spent during negotiations), making it important to obtain good performance quickly.

We consider an application where information arrives in the form of binary win/loss observations, representing customers' yes/no responses to the seller's pricing offers (or "bids"). A common demand model in this setting (used e.g. by [35]) assumes that these binary outcomes follow a logistic distribution, which also allows us to relate the win probability to a set of regression features representing additional information about the product or customer type. Although this is a fairly natural choice of demand model (essentially just an instance of logistic regression), it is quite challenging to connect to the Bayesian way of representing new information and using it to update the seller's beliefs. While, for linear regression, the multivariate normal distribution offers an intuitive and easy-to-use conjugate prior [96], no such model is available for logistic regression, making it difficult to represent a belief over a continuous space of logistic curves.

We approach this problem with approximate Bayesian inference, using the technique of density projection to create a multivariate normal posterior distribution that is "approximately conjugate," in the sense of minimizing the Kullback-Leibler divergence from the actual posterior. See e.g. [97] for an application of this technique to the problem of learning unknown correlation structures in ranking and selection. In the context of logistic regression, our approach is similar to the variational approximation by [98], but involves an additional optimization step using infinitesimal perturbation analysis (see e.g. [11] or [13]) to further improve the quality of the approximation. Using this statistical technique to efficiently update a multivariate normal prior on the parameters of the logistic demand curve, we then apply a policy that optimizes a myopic estimate of the expected revenue curve (see e.g. Ch. 11 of [5]). Our numerical experiments provide evidence in favor of both the approximate Bayesian learning model and the Bayes-greedy pricing policy. Although our Bayesian model has numerous applications outside pricing, in this particular context it enables the seller to compactly model a set of beliefs about win probabilities for a wide range of customer and product segments, and then quickly update this belief in real time.

5.2 Problem Formulation and Learning Model

Section 5.2.1 introduces the demand and revenue curves optimized by a seller in the B2B pricing problem. In Section 5.2.2, we discuss the challenge of developing a Bayesian model for learning the parameters of the demand curve. Then, Sections 5.2.3 and 5.2.4 outline our proposed approach for overcoming this challenge.

5.2.1 Problem Formulation

Consider a seller who must quote prices for a sequence of corporate clients. The $(n+1)$st client will accept a price offer $p^n \geq 0$ with probability ρ, which may also depend on additional properties of the client or product. The function ρ is called the *demand curve*, and is not known exactly to the seller. However, the seller does observe the client's response, modeled as a binary variable Y^{n+1}, where $Y^{n+1} = 1$ with probability ρ, representing a sale (or "win"), and $Y^{n+1} = 0$ represents a "loss." The seller's expected revenue from the client is

$$R(p^n) = p^n \rho, \qquad p^n \geq 0, \tag{5.1}$$

where the demand curve ρ usually depends on the price p^n. In most applications, we need to consider the marginal cost c for the product, and work with the expected profit

$$\Pi(p^n) = (p^n - c)\rho, \qquad p^n \geq c. \tag{5.2}$$

We assume that Y^n follows a logistic distribution, allowing us to write the demand curve as

$$\rho(\mathbf{x}^n) = \mathbb{P}(Y^{n+1} = 1) = \frac{1}{1 + e^{-\boldsymbol{\mu}^\intercal \mathbf{x}^n}}, \tag{5.3}$$

where $\mathbf{x}^n$ is a vector of features, observed by the seller, providing relevant information for the $(n+1)$st pricing decision. In the simplest possible model, the customers are assumed to be homogeneous, $\mathbf{x}^n = [1, p^n]^\intercal$, and the parameter vector $\boldsymbol{\mu}$ consists only of an intercept and a slope term. We use this simple model in our examples throughout this chapter. However, our analysis is readily applicable to the general case, where $\mathbf{x}^n$ may also contain information about the product (type or volume) and the client (region, industry, history with the seller).

In all of these cases, the parameter vector $\boldsymbol{\mu}$ is unknown to the seller and must be inferred using a combination of prior knowledge and incoming win/loss results. The shape of the demand curve is extremely sensitive to the parameter values, making it important to obtain accurate estimates of the parameters as quickly as possible. We now propose a Bayesian framework for learning the demand curve.

5.2.2 Bayesian Model For Dynamic Pricing

We adopt the Bayesian view, and represent our uncertainty about the vector $\boldsymbol{\mu}$ using a multivariate normal prior distribution, that is,

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}). \tag{5.4}$$

The multivariate normal distribution offers a compact and powerful way to model correlations between our beliefs about different components of $\boldsymbol{\mu}$. Because the observation Y^{n+1} provides information about an entire vector $\mathbf{x}^n$, our beliefs about different components of this vector should become correlated due to their dependence on the same observations. A second convenience of the multivariate normal distribution (important for computational purposes) is that the linear combination $\boldsymbol{\mu}^\intercal \mathbf{x}^n$ follows a univariate normal distribution.

In linear regression, where a continuous response variable is related to a linear combination of features, the multivariate normal prior possesses the property of *conjugacy.* That is, if the residual errors are i.i.d. normal, the posterior distribution of the regression parameters, conditional on a sequence of observations, will remain normal [96]. This model makes the learning process highly efficient computationally, as one only needs to recursively update the mean vector and covariance matrix of the belief distribution after each observation. Unfortunately, in logistic regression, there is no known prior distribution that is conjugate with logistic observations. To see this, we first assume that $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\theta}^n, \boldsymbol{\Sigma}^n)$, and write the likelihood function of Y^{n+1} as

$$P(Y^{n+1}) = g(H^{n+1}(\boldsymbol{\mu})), \tag{5.5}$$

where $g(z) = (1 + e^{-z})^{-1}$ and $H^{n+1}(\boldsymbol{\mu}) = (2Y^{n+1} - 1)(\boldsymbol{\mu}^\intercal \mathbf{x}^n)$. Equation (5.5) allows us to represent the win/loss probability in a concise form. Applying Bayes' rule, the posterior distribution, given the bidding price p^n and the observation Y^{n+1}, can be written as

$$p(\boldsymbol{\mu}|p^n, Y^{n+1}) \propto g(H^{n+1}(\boldsymbol{\mu}))|\boldsymbol{\Sigma}^n|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\theta}^n)^{\intercal}(\boldsymbol{\Sigma}^n)^{-1}(\boldsymbol{\mu}-\boldsymbol{\theta}^n)\right\}, \quad (5.6)$$

which is clearly non-normal.

We would like to retain the multivariate normal prior due to its power in modeling correlated beliefs. However, we are now required to use the techniques of approximate Bayesian inference to develop a multivariate normal posterior that is "approximately conjugate." Several such approaches have been proposed, including approximation methods based on Laplace approximation [99] and variational bounds [98]. We take a variational Bayesian approach to approximate the posterior distribution by minimizing the Kullback-Leibler divergence between the true posterior distribution and a multivariate normal distribution.

5.2.3 Variational Bayesian Approximation

Suppose that, after observing n responses, our beliefs about $\boldsymbol{\mu}$ are multivariate normal with parameters $(\boldsymbol{\theta}^n, \boldsymbol{\Sigma}^n)$. Let $P(\boldsymbol{\mu}|p^n, Y^n, \boldsymbol{\theta}^n, \boldsymbol{\Sigma}^n)$ be the likelihood function of this distribution. The variational Bayesian approach approximates the posterior distribution of $\boldsymbol{\mu}$, given Y^{n+1}, with a normal distribution $Q(\boldsymbol{\mu}|\boldsymbol{\theta}^{n+1}, \boldsymbol{\Sigma}^{n+1})$ by minimizing the Kullback-Leibler (KL) divergence. The KL divergence between $P(\boldsymbol{\mu}|p^n, Y^{n+1}, \boldsymbol{\theta}^n, \boldsymbol{\Sigma}^n)$ and $Q(\boldsymbol{\mu}|\boldsymbol{\theta}^{n+1}, \boldsymbol{\Sigma}^{n+1})$ is defined as

$$\mathcal{D}(Q \shortparallel P) := \mathbb{E}_Q\left(\log \frac{Q(\boldsymbol{\mu}|\boldsymbol{\theta}^{n+1}, \boldsymbol{\Sigma}^{n+1})}{P(\boldsymbol{\mu}|p^n, Y^{n+1}, \boldsymbol{\theta}^n, \boldsymbol{\Sigma}^n)}\right), \quad (5.7)$$

where the expectation is taken with respect to Q. This definition can be partially simplified, as stated in the following result.

Proposition 5.1. *The KL divergence can be written as*

$$\mathcal{D}(Q \,\|\, P) = \mathbb{E}_Q\left[\log\left(1 + e^{-H^{n+1}(\boldsymbol{\mu})}\right)\right] + h(\boldsymbol{\theta}^n, \boldsymbol{\Sigma}^n, \boldsymbol{\theta}^{n+1}, \boldsymbol{\Sigma}^{n+1}), \tag{5.8}$$

with the second component specified as

$$h(\boldsymbol{\theta}^n, \boldsymbol{\Sigma}^n, \boldsymbol{\theta}^{n+1}, \boldsymbol{\Sigma}^{n+1}) = \frac{1}{2}\left[tr\left((\boldsymbol{\Sigma}^n)^{-1}\boldsymbol{\Sigma}^{n+1}\right) + (\boldsymbol{\theta}^n - \boldsymbol{\theta}^{n+1})^\intercal(\boldsymbol{\Sigma}^n)^{-1}(\boldsymbol{\theta}^n - \boldsymbol{\theta}^{n+1}) - k - \ln\frac{|\boldsymbol{\Sigma}^{n+1}|}{|\boldsymbol{\Sigma}^n|} + C\right.$$

where C is a constant that does not depend on $\boldsymbol{\theta}^{n+1}$ and $\boldsymbol{\Sigma}^{n+1}$.

To minimize the KL divergence, the first step is to take the gradient of $\mathcal{D}(Q \,\|\, P)$ with respect to its parameter $\boldsymbol{\theta}^{n+1}$ and $\boldsymbol{\Sigma}^{n+1}$. Unfortunately, a closed-form expression for the gradient is not available, because the expectation in equation (5.8) is intractable. However, if our goal is to minimize an expected value, a connection to gradient-based stochastic search [100] comes naturally to mind. The work by [101] uses such an approach, where a likelihood ratio estimate [102] of the gradient is constructed. However, this approach leads to a noisy simulation optimization problem, whose dimensionality is quadratic in the number of features, presenting substantial computational difficulties.

Instead of optimizing with respect to $(\boldsymbol{\theta}^{n+1}, \boldsymbol{\Sigma}^{n+1})$, we utilize a dimension reduction technique and propose the following form for $\boldsymbol{\theta}^{n+1}$ and $\boldsymbol{\Sigma}^{n+1}$:

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\Sigma}^{n+1}\left((\boldsymbol{\Sigma}^n)^{-1}\boldsymbol{\theta}^n + \left(Y^{n+1} - \frac{1}{2}\right)\mathbf{x}^n\right) \tag{5.9}$$

$$\boldsymbol{\Sigma}^{n+1} = \left((\boldsymbol{\Sigma}^n)^{-1} + \lambda\mathbf{x}^n(\mathbf{x}^n)^\intercal\right)^{-1} \tag{5.10}$$

Applying the Sherman-Morrison formula to (5.9) and (5.10), we obtain

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \frac{\frac{Y^{n+1}-1/2}{\lambda} - (\mathbf{x}^n)^\intercal \boldsymbol{\theta}^n}{\frac{1}{\lambda} + (\mathbf{x}^n)^\intercal \boldsymbol{\Sigma}^n \mathbf{x}^n} \boldsymbol{\Sigma}^n \mathbf{x}^n, \tag{5.11}$$

$$\boldsymbol{\Sigma}^{n+1} = \boldsymbol{\Sigma}^n - \frac{\boldsymbol{\Sigma}^n \mathbf{x}^n (\mathbf{x}^n)^\intercal \boldsymbol{\Sigma}^n}{\frac{1}{\lambda} + (\mathbf{x}^n)^\intercal \boldsymbol{\Sigma}^n \mathbf{x}^n}. \tag{5.12}$$

In this form, there is only one parameter λ to be determined. We minimize the KL divergence with respect to λ to find the optimal multivariate normal posterior distribution from the parametrized family in (5.11)-(5.12). Aside from the computational convenience of reducing the size of the problem, we choose precisely this form because it resembles the Kalman-filter-like equations used for Bayesian linear regression; the parameter λ is analogous to the precision of the residuals, while $\frac{Y^{n+1}-1/2}{\lambda}$ stands in for the continuous observation. In this way, our learning model for logistic regression makes an intuitive connection to the well-understood linear setting. Moreover, previous work on logistic regression, including [98] and [99], has derived updating rules with very similar form, based on different approximation techniques for the posterior likelihood function.

For additional convenience, we apply the transformation $v = \frac{1}{\lambda}$ and find

$$v^* = \operatorname*{argmin}_v \mathcal{D}(Q \,||\, P). \tag{5.13}$$

The parameter v is analogous to the variance of the residuals in a linear regression model. Since no such explicit parameter is given in logistic regression, we simply find the value that produces the most accurate approximation.

5.2.4 Minimizing the Kullback-Leibler Divergence

We now propose a stochastic approximation method to solve the minimization problem in (5.13), which requires estimations of the gradient of $\mathcal{D}(Q \shortparallel P)$ with respect to the single parameter v. This results in

$$\nabla_v \mathcal{D}(Q \shortparallel P) = \nabla_v \mathbb{E}_Q \left[\log \left(1 + e^{-H^{n+1}(\boldsymbol{\mu})} \right) \right] + \nabla_v h(\boldsymbol{\theta}^n, \boldsymbol{\Sigma}^n, \boldsymbol{\theta}^{n+1}, \boldsymbol{\Sigma}^{n+1}). \qquad (5.14)$$

Since we do not have a close-form expression for $\nabla_v \mathbb{E}_Q \left[\log \left(1 + e^{-H^{n+1}(\boldsymbol{\mu})} \right) \right]$, we propose to use infinitesimal perturbation analysis (IPA) to obtain noisy samples of the gradient (see e.g. [100] or [13] for an introduction). First, we transform the expectation in (5.14) into an integration with respect to a standard univariate normal distribution, $\mathbb{E}[\bar{f}(Z)]$, where $Z \sim \mathcal{N}(0,1)$ and

$$\bar{f}(z) = \log \left(1 + \exp \left\{ -(2Y^{n+1} - 1) \left[\left(\mathbf{x}^n(p^n(\boldsymbol{\mu}))^\intercal \boldsymbol{\Sigma}^{n+1} \mathbf{x}^n(p^n) \right)^{1/2} z + (\boldsymbol{\theta}^{n+1})^\intercal \mathbf{x}^n(p^n) \right] \right\} \right).$$

The next result shows that the conditions for IPA [103] hold.

Proposition 5.2. $\nabla_v \mathbb{E}\left[\bar{f}(Z)\right] = \mathbb{E}\left[\nabla_v \bar{f}(Z)\right]$.

The IPA estimator itself is given as

$$\mathbb{E}\left[\nabla_v \bar{f}(Z)\right] \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_v \bar{f}(Z^{(i)}),$$

where $Z^{(i)}$ are independent samples from a standard normal distribution. We denote the gradient estimator by $\widehat{\nabla}_v \mathbb{E}_Q \left[\log \left(1 + e^{-H^{n+1}(\boldsymbol{\mu})} \right) \right]$ and plug it into (5.14) for $\nabla_v \mathbb{E}_Q \left[\log \left(1 + e^{-H^{n+1}(\boldsymbol{\mu})} \right) \right]$. This produces an estimator of $\widehat{\nabla}_v \mathcal{D}(Q \shortparallel P)$, and we can apply the Robbins-Monro stochastic approximation algorithm

$$v_{n+1} = v_n - a_n \widehat{\nabla}_v \mathcal{D}(Q \shortparallel P),$$

for some suitably chosen stepsize a_n, to find the optimal v^* and thus the optimal λ^*. Then we can apply the updating rules in (5.11) and (5.12) to determine the approximate posterior distribution after collecting each observation Y^n.

5.3 Dynamic Pricing Policy

We have shown a way in which the seller's beliefs can be updated after observing customer response to a price. It remains to address how that price can be chosen in the first place. In this section, we expand upon the notion of a "Bayes-greedy" pricing policy introduced in Ch. 11 of [5]. Greedy and semi-greedy policies have been widely studied in the literature on dynamic pricing under environmental uncertainty (see e.g. [39]), and our policy may also be viewed as part of that realm. However, in the setting of Bayesian logistic regression, the concept of "greedy" admits important nuances.

Ideally, the seller would like to choose the price that maximizes the true revenue curve,

$$p^* = \arg\max_p \frac{p}{1 + e^{-(\boldsymbol{\mu}^\intercal \mathbf{x}(p))}}, \tag{5.15}$$

where we emphasize that $\mathbf{x}$ depends on p since p is typically one component of the vector of features. A simple "greedy" policy will simply replace $\boldsymbol{\mu}$ in (5.15) by the current posterior mean vector $\boldsymbol{\theta}^n$. This is typically the approach used in frequentist models (e.g. in [40]) where an MLE estimator is used in place of $\boldsymbol{\theta}^n$.

In the Bayesian setting, however, this approach will under-perform, because it does not use all of the available information. In particular, it does not account for the uncertainty in our beliefs, expressed by $\boldsymbol{\Sigma}^n$. The covariance matrix is important because it specifies a whole family of possible revenue curves, parametrized by $\boldsymbol{\mu} \sim \mathcal{N}\left(\boldsymbol{\theta}^n, \boldsymbol{\Sigma}^n\right)$. Thus, a *Bayes-greedy* policy will still myopically optimize the expected single-period revenue, but the expectation will be over the entire space of revenue curves. That is,

$$p^n = \underset{p}{\operatorname{argmax}}\, \mathbb{E}\left[R(p)\right] = \underset{p}{\operatorname{argmax}}\, \mathbb{E}\left[\frac{p}{1+e^{-(\boldsymbol{\mu}^\intercal \mathbf{x}^n(p))}}\right], \tag{5.16}$$

where the expectation is taken with respect to the (approximate) posterior joint distribution of the parameters.

5.3.1 Computation of the Bayes-Greedy Policy

In order to use the Bayes-greedy policy, we require the ability to compute the expectation in (5.16). The approximate Bayesian model suggests that the posterior distribution is multivariate normal, which leads to another convenient dimension reduction. If $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\theta}^n, \boldsymbol{\Sigma}^n)$ after collecting n observations, then

$$\boldsymbol{\mu}^\intercal \mathbf{x} \sim \mathcal{N}\left((\boldsymbol{\theta}^n)^\intercal \mathbf{x}, \mathbf{x}^\intercal \boldsymbol{\Sigma}^n \mathbf{x}\right)$$

for arbitrary $\mathbf{x}$. Therefore, let $W = \boldsymbol{\mu}^\intercal \mathbf{x}$, noticing that W is actually a function of p, and rewrite (5.16) as

$$p^n = \underset{p}{\operatorname{argmax}}\, \mathbb{E}\left[\frac{p}{1+e^{-W}}\right], \tag{5.17}$$

where the expectation is now taken with respect to a univariate normal distribution with appropriately chosen mean and variance.

The expectation in (5.17) is known as the *logistic-normal integral* [104], which plays an important role in statistics. However, this integral is impossible to compute analytically. It may be computed using Monte Carlo simulation, in particular using IPA (it can be shown that the relevant conditions hold). However, [105] offers a tractable approximation

$$\mathbb{E}\left[\frac{1}{1+e^{-W}}\right] \approx \frac{1}{1+e^{-\frac{\mathbb{E}(W)}{\gamma}}},$$

where

$$\gamma = \sqrt{1+\frac{\pi}{8}\mathrm{Var}(W)}.$$

This leads to an approximate Bayes-greedy policy that can be written as

$$p^n = \operatorname*{argmax}_p \frac{p}{1+e^{-\frac{(\boldsymbol{\theta}^n)^\intercal \mathbf{x}^n(p)}{\gamma^n(p)}}}, \tag{5.18}$$

where

$$\gamma^n(p) = \sqrt{1+\frac{\pi}{8}\mathbf{x}^n\,(p)^\intercal\,\boldsymbol{\Sigma}^n\mathbf{x}^n\,(p)}.$$

This approximation gives us a closed-form expression for the expected revenue function, so that making a pricing decision using (5.18) is computationally easier.

5.3.2 Analysis of the Bayes-Greedy Policy

It can be easily shown that the objective function optimized by the point-estimate policy,

$$R_{PE}(p) = \frac{p}{1+e^{-(\boldsymbol{\theta}^n)^\intercal \mathbf{x}^n(p)}},$$

is log-concave (but not concave). As a consequence, this function has a single globally optimal price. We show that the Bayes-greedy objective function in (5.16) possesses the same property, whence it follows that the idea of a "Bayes-greedy price" is well-defined.

Theorem 5.3. *The Bayes-greedy objective function*

$$R_{BG}(p) = \mathbb{E}[R(p)] = \mathbb{E}\left[\frac{p}{1 + e^{-(\boldsymbol{\mu}^{\intercal}\mathbf{x}^{n}(p))}}\right]$$

is quasi-concave in p when $p > 0$.

An important consequence of Theorem 5.3 is that, if we apply IPA to optimize R_{BG}, we are guaranteed to converge to the optimal price. In general, IPA is only guaranteed to find a local optimum. However, in this case, we can apply stochastic approximation to solve the problem directly instead of using the approximation in (5.18). However, we are still interested in understanding the approximate problem, since it is easier to solve. One can observe that (5.18) resembles the point-estimate objective, but with an additional factor $\gamma^{n}(p)$ incorporating our uncertainty about the regression coefficients. The following proposition summarizes structural properties of this factor.

Proposition 5.4. *The factor $\gamma^{n}(p) \geq 1$ is convex in p.*

The variance factor can be viewed as the risk we have to take when choose a price. In most problem instances that we have observed, the factor $\gamma^{n}(p)$ is not only a convex function, but also an increasing function within the domain of bidding prices. This suggests that the risk is higher when we take a higher price, but the possible reward is also higher. Moreover, the probability of success decreases when we choose a higher price. However, since the factor $\gamma^{n}(p)$ is greater than 1, the

Bayes-greedy policy tends to explore higher prices than the point-estimate policy, leading to possible higher profit.

5.4 Numerical Experiments

In this section, we present numerical experiments using the approximate Bayesian learning model with stochastic approximation proposed in Section 5.2 and the Bayes-greedy policy proposed in Section 5.3. We compare this with several alternative approaches, described as follows:

1. The standard frequentist logistic regression approach with the point-estimate policy. In this approach, logistic regression is reapplied after collecting each observation to estimate parameters in the demand function. Then, a pricing decision is made using the point-estimate policy mentioned in Section 5.3.
2. The variational lower bound approach in [98], using the point-estimate policy to make pricing decisions.
3. The variational lower bound approach in [98], with the proposed Bayes-greedy policy. This is used to show the advantage of choosing the parameter λ optimally using IPA.

Suppose that we are running a computer company, and one of our standard desktop computer models has production cost c. When we set a price for selling, we restrict the prices to be within the range $[p_l, p_u]$, where the lower bound $p_l \geq c$. This means that we never set a price that is lower than the cost, and it is very unlikely that we will make a sale if the selling price is above p_u. Our objective is to maximize the profit function as in (5.2). In our experiment, we choose $c = 300$, $p_l = 300$ and $p_u = 500$. We consider a finite number of possible bidding prices from 300 to 500 in increments of 10.

For the purposes of this example, we use a two-parameter model, that is, $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and $\mathbf{x}(p) = [1, p]^\intercal$. We begin with a prior mean $\boldsymbol{\theta}^0 = [500, -1]^\intercal$. The corresponding demand and profit curves, based on this prior, are shown as the solid green lines in Figure 5.1. The figure also shows three different realizations of $\boldsymbol{\mu}$ in which the maximum possible profits are "low," "medium," and "high." Minor changes in the regression parameters can significantly alter the shape of the profit curve. From Figure 5.1(a), we see that the prior is essentially telling us that any customer is highly likely to purchase the computer for prices between \$300 and \$500. Considering the upper and lower bounds we have chosen for our price, this type of prior can be understood as uninformative (since our belief suggests that a customer will buy the product for almost any price in the range).

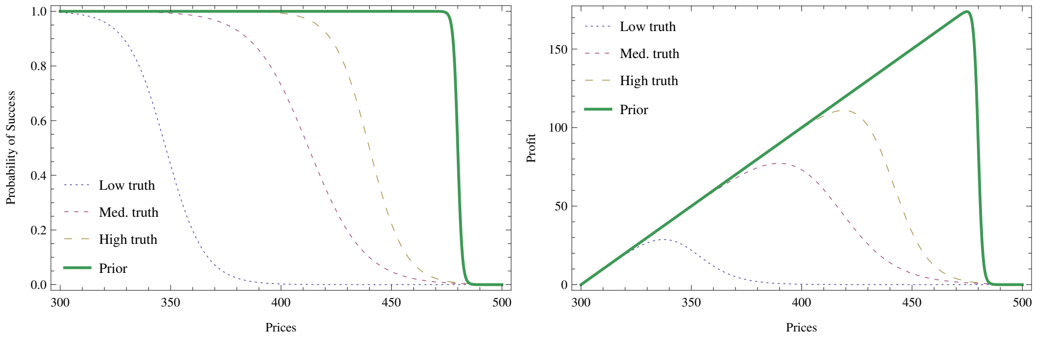
The prior covariance matrix we choose for one specific setting is given by

$$\boldsymbol{\Sigma}^0 = \begin{bmatrix} 100 & 0 \\ 0 & 0.01 \end{bmatrix}.$$

Instead of interpreting this as our uncertainty about the parameters in the prior belief, it may be more meaningful to consider $\boldsymbol{\Sigma}^0$ as the uncertainty about possible prices that buyers will pay. Notice that the magnitudes of the two variances are quite different, again due to the extreme sensitivity of the profit curve to small changes in the regression parameters, particularly the price sensitivity μ_2. If the variance of μ_2 is too large, this essentially means that μ_2 has a high probability of being positive, which is quite unlikely to occur in practice. Furthermore, in a practical application of the Bayes-greedy policy, we may have $\mathbf{x}(p) = [\tilde{\mathbf{x}}, p]^\intercal$, where $\tilde{\mathbf{x}}$ contains product and customer attributes unrelated to the price. For the purpose of myopically optimizing the price, this model is equivalent to a two-parameter model where multiple features are embedded into μ_1, in which case the variance of μ_1 actually represents the variance of a sum of random variables, and should be much larger than the variance of μ_2.

We compare the performances, including pricing decisions, single-period profit and cumulative profit, of the approach proposed in this chapter (referred to as "IPA-Bayes") and three alternative approaches. The results are reported for the first 20 iterations and averaged over 1000 sample paths, with the true value of $\boldsymbol{\mu}$ fixed according to the three scenarios shown in Figure 5.1. In all numerical experiments, the Bayes-greedy policy refers to the approximate policy in equation (5.18). We briefly discuss each of the three scenarios below.

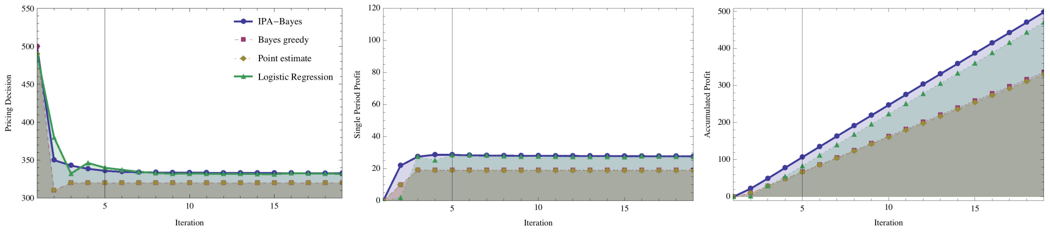
Low-truth scenario. The parameters of the low truth setting are $\mu_1 = 40$ and $\mu_2 = -0.115$. With pre-specified values for μ_1 and μ_2, the optimal bidding price is 340. As shown in Figure 5.2, all four methods start with the same bidding price initially and converge after 6 iterations, but the values they converge to are different. All methods converge to prices below the optimal selling price, but the prices from IPA-Bayes and the logistic regression method are closer to optimal, and produce similar profits. IPA-Bayes adjusts more quickly to new information than the



(a) Probability of success

(b) Profit curves

Figure 5.1: Probability of success and corresponding profit curve as a function of the price under three different scenarios



(a) Bidding prices

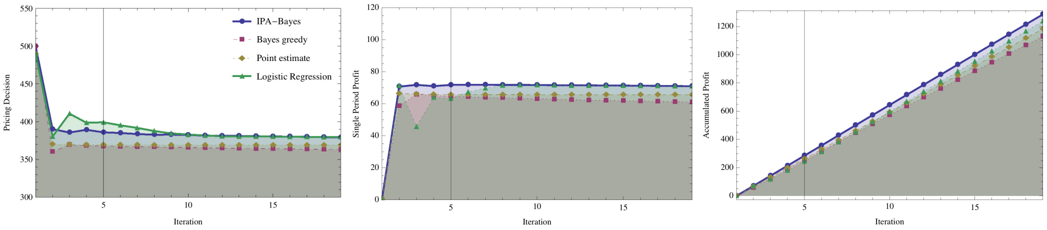
(b) Single-period-profit

(c) Cumulative profit

Figure 5.2: Plots of bidding prices, single-period profit and cumulative profits over time under the low-truth scenario

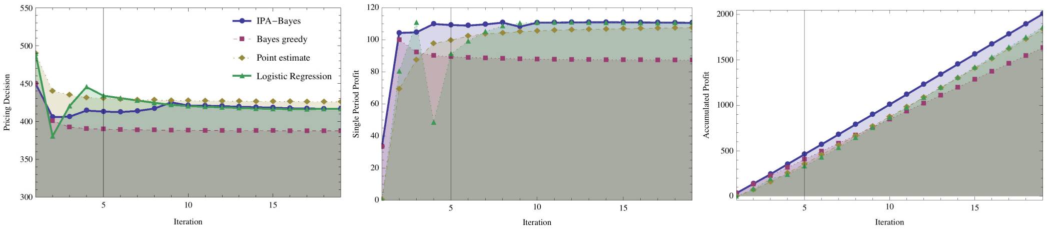
other three methods, without the volatile behavior observed for frequentist logistic regression. Additionally, IPA-Bayes shows advantages in single-period profit during the first 5 iterations, resulting in higher cumulative profit.

Medium-truth scenario. The true parameters are $\mu_1 = 32.5$ and $\mu_2 = -0.08$, and the optimal bidding price is 380. Figure 5.3 shows that all four methods converge to a price close to optimal. Both the single-period profit and the cumulative profit from the IPA-Bayes method dominate those from the other three methods. Note that, while the frequentist method explores higher prices than IPA-Bayes for some



(a) Bidding prices (b) Single-period-profit (c) Cumulative profit

Figure 5.3: Plots of bidding prices, single-period profit and cumulative profits over time under the medium-truth scenario



(a) Bidding prices (b) Single-period-profit (c) Cumulative profit

Figure 5.4: Plots of bidding prices, single-period profit and cumulative profits over time under the high-truth scenario

time, this behavior is actually too aggressive, and produces lower profits.

High-truth scenario. The parameters of the high-truth setting are $\mu_1 = 55$ and $\mu_2 = -0.125$, with the optimal price being \$420. As shown in Figure 5.4, both IPA-Bayes and Bayes-greedy start from a lower bidding price than the other methods, due to the effect of the uncertainty factor $\gamma^n(p)$. However, IPA-Bayes quickly adjusts and increases the bidding price, eventually getting close to optimal, and dominating the other methods in terms of single-period profit. After 10 iterations, frequentist logistic regression catches up and produces similar single-period profit, at the expense of volatile behavior and smaller profits in the early iterations.

Discussion. Frequentist logistic regression generally performs well after a few iterations. However, in the sequential setting, we have to refit a new logistic regression model after every observation, which becomes more time-consuming as the number of observations increases. The Bayesian learning model, while less accurate (due to the approximation of conjugacy), provides a quick and efficient way to update parameters, and generally produces a "smoother" sequence of prices; essentially, the uncertainty encoded in the covariance matrix smooths the pricing decision, compared to the volatile prices chosen by the frequentist method in the early iterations. Among the methods using approximate Bayesian inference, IPA-Bayes is consistently the best.

In these examples, some competing policies tend to perform very similarly to IPA-Bayes after about 10 iterations. However, in the specific context of B2B pricing, the early iterations are especially important because each individual contract tends to have much higher value, and the opportunity cost of pricing suboptimally is more severe. To give some perspective, if the value of every contract is on the order of hundreds of thousands, or millions, of dollars, the overall planning horizon will be shorter, and the first $5 - 10$ iterations will become very significant.

5.5 Conclusion

We have presented an approximate Bayesian approach for learning the parameters in a logistic regression model, with specific application to learning revenue curves in B2B pricing problems. We use infinitesimal perturbation analysis and stochastic search to improve the quality of the approximation. We also consider a pricing policy that incorporates uncertainty about the parameters into the estimated expected revenue curve, and chooses a price that optimizes this aggregated function. The proposed model and pricing policy show encouraging results in our empirical experiments. Future work will test the proposed approach on real-world pricing data, where the underlying statistical model can be high-dimensional.

Chapter 6: Conclusion

In this thesis we have proposed two different metamodeling approaches that incorporates direct gradient estimates for solving simulation optimization problems with continuous variables and a knowledge-gradient method that employs variational Bayesian technique for ranking and selection problems. We have also proposed an approximate Bayesian statistical model and price recommendation strategy in business-to-business (B2B) pricing context.

In Chapter 2, we analyzed regression models that explicitly incorporate direct gradient estimators, and derived the corresponding parameter estimators. We provided preliminary evidence for the potential gains from the DiGAR approach by comparing with standard regression both theoretically via analytical calculations under settings with more restrictive assumptions, and empirically via simple queueing examples where the assumptions under which the theoretical results are established do not hold. More generally, we investigated the idea of augmenting statistical models when direct gradient estimators are available, motivated by stochastic simulation settings. We provided an alternative model for the local improvement step in the sequential RSM approach used in experimental design for optimization.

In Chapter 3, we investigated the idea of incorporating gradient estimates into stochastic kriging by extrapolating additional responses using the original responses and gradient estimates. This approach is not restricted to stochastic kriging, but can be applied to other metamodeling approaches as well. We analyzed the proposed GESK model theoretically under simplified settings and showed that it provides predictions with smaller MSE than stochastic kriging. We also conducted numerical experiments and illustrated the performance of the GESK model. We presented two different strategies, namely PMLE and IMSE, to determine extrapolation step sizes used in GESK. Effectiveness of these two strategies were compared using numerical examples.

In Chapter 4, we created a new Bayesian model for simultaneously learning unknown means and unknown correlations in fully sequential R&S. We derived a new VIP for ranking and selection with unknown correlation structures. The new procedure intuitively generalizes VIP for R&S with known correlations, with the additional ability to incorporate the decision-makers uncertainty about the correlation structure into decision-making. We proved that the value of information is greater when the correlation structure becomes unknown. We also argued that the incremental information loss from a single application of approximate Bayesian inference eventually vanishes. We provided numerical results to show the value added by learning unknown correlations. In particular, we studied a version of the wind farm placement problem using real data.

In Chapter 5, we considered the problem of optimally choosing prices to maximize revenue or profit from transactions with heterogeneous customers. We developed a learning model that maintains a set of beliefs about the effects of the significant characteristics, but is able to adjust and improve those beliefs as new data come in. We also developed an optimization algorithm that recommends a price to maximize average revenue, based on the estimates provided by the predictive and learning models. The uncertainty measured by the learning model is used as a factor in the price calculation. For example, if the estimate from the data suggests that we should quote a high price, but there is a large amount of uncertainty suggesting that the estimate is unreliable, the final recommendation made by the procedure will tend to be more conservative. We conducted simulations to check the performance of the optimal prices. The results of the simulations suggest that the optimal bidding algorithm has the potential to substantially increase cumulative revenue over time.

Our work has initiated some new ideas and points to several other more general directions for future research. The proposed DiGAR model introduced in Chapter 2 can be used in the application to simulation-based optimization, , for example, sequential RSM, which was one of the motivations for choosing regression models for incorporating direct stochastic gradient estimators. How much gains will be realized in the optimization efficiency from the improved linear regression model? Although it is reasonable to expect improvements, since the new method does obtain better fitted values than simple linear regression, both theoretical work and numerical experimentation are needed to characterize and quantify the improvements.

The GESK method proposed in Chapter 3 use linear extrapolation with the same step size and assume that only one additional point is extrapolated from each design point. More sophisticated techniques could use the local response surface information and adaptively determine the extrapolation strategy. This is especially important in higher-dimensional problems with multiple extreme values. Another line of research is on the comparison between GESK and SKG. Improvements from incorporating gradient estimates can be expected from both models. However, the question is when does one model performs better than the other? To compare these two models theoretically will provide more insights about this question and lead to guideline for practitioners as to when to choose each of these two models for practitioners.

A.1 Analytical Results for $M/M/1$ and $U/U/1$ Queues

For the $M/M/1$ queue, the true models for an interarrival mean of 0.2 are given by

$$
\begin{aligned}
y^{(2)}(x) &= x + \frac{x^2}{5+x}, \\
y^{(3)}(x) &= x + \frac{5x^2}{(5+x)^2} + \frac{x^3(15+2x)}{(5+x)^3}, \\
y^{(4)}(x) &= x + \frac{25x^2}{(5+x)^3} + \frac{25x^3}{(5+x)^4} + \frac{5x^3(15+2x)}{(5+x)^4} + \frac{x^4(225+50x+3x^2)}{(5+x)^5}, \\
y^{(5)}(x) &= x + \frac{125x^2}{(5+x)^4} + \frac{250x^3}{(5+x)^5} + \frac{25x^3(15+2x)}{(5+x)^5} + \frac{5x^4(225+50x+3x^2)}{(5+x)^6} \\
&\quad + \frac{25x^4(15+2x)}{(5+x)^6} + \frac{250x^4}{(5+x)^6} + \frac{x^5(10+x)(350+65x+4x^2)}{(5+x)^7}.
\end{aligned}
$$

For the $U/U/1$ queue, the true models are given by

$$\begin{aligned}
y^{(2)} &= \frac{\delta_1}{4} - \frac{\theta_1}{2} + \frac{3\theta_2}{2} + \frac{1}{\delta_1}\left(\frac{\delta_2^2}{12} + \frac{\theta_1^2}{4} - \frac{\theta_1\theta_2}{2} + \frac{\theta_2^2}{4}\right) \\
y^{(3)} &= \frac{5\delta_1}{12} - \theta_1 + 2\theta_2 + \frac{1}{12\delta_1}\left(2\delta_2^2 + 9\theta_1^2 - 18\theta_1\theta_2 + 9\theta_2^2\right) \\
&\qquad - \frac{1}{12\delta_1^2}(\theta_1 - \theta_2)(\delta_2^2 + 2\theta_1^2 - 4\theta_1\theta_2 + 2\theta_2^2) \\
y^{(4)} &= \frac{107\delta_1}{192} - \frac{25\theta_1}{16} + \frac{41\theta_2}{16} + \frac{1}{2880\delta_1}(750\delta_2^2 + 4590\theta_1^2 - 9180\theta_1\theta_2 + 4590\theta_2^2) \\
&\qquad - \frac{1}{48\delta_1^2}(\theta_1 - \theta_2)(13\delta_2^2 + 35\theta_1^2 - 70\theta_1\theta_2 + 35\theta_2^2) \\
&\qquad\quad + \frac{1}{2880\delta_1^3}(13\delta_2^4 + 270\delta_2^2\theta_1^2 - 540\delta_2^2\theta_1\theta_2 + 270\delta_2^2\theta_2^2 \\
&\qquad\qquad + 405\theta_1^4 - 1620\theta_1^3\theta_2 + 2430\theta_1^2\theta_2^2 - 1620\theta_1\theta_2^3 + 405\theta_2^4) \\
y^{(5)} &= \frac{221\delta_1}{320} - \frac{107\theta_1}{48} + \frac{155\theta_2}{48} + \frac{1}{2880\delta_1}(1070\delta_2^2 + 8430\theta_1^2 - 16860\theta_1\theta_2 + 8430\theta_2^2) \\
&\qquad - \frac{1}{48\delta_1^2}(\theta_1 - \theta_2)(29\delta_2^2 + 99\theta_1^2 - 198\theta_1\theta_2 + 99\theta_2^2) \\
&\qquad\quad + \frac{1}{2880\delta_1^3}(49\delta_2^4 + 1230\delta_2^2\theta_1^2 - 2460\delta_2^2\theta_1\theta_2 + 1230\delta_2^2\theta_2^2 \\
&\qquad\qquad + 2325\theta_1^4 - 9300\theta_1^3\theta_2 + 13950\theta_1^2\theta_2^2 - 9300\theta_1\theta_2^3 + 2325\theta_2^4) \\
&\qquad\qquad\quad - \frac{1}{720\delta_1^4}(\theta_1 - \theta_2)(9\delta_2^4 + 80\delta_2^2\theta_1^2 - 160\delta_2^2\theta_1\theta_2 + 80\delta_2^2\theta_2^2 + 96\theta_1^4 \\
&\qquad\qquad\qquad - 384\theta_1^3\theta_2 + 576\theta_1^2\theta_2^2 - 384\theta_1\theta_2^3 + 96\theta_2^4)
\end{aligned}$$

A.2 Gradient Estimation for $G/G/1$ Queue

Let A_k be the interarrival time between the $(k-1)$st and kth customer (by convention, taking A_1 to be the time of the 1st arrival), and let X_k be the service time of the kth customer. The system time of the kth customer, denoted by T_k, satisfies the well-known Lindley equation:

$$T_{k+1} = X_{k+1} + (T_k - A_{k+1})^+, \tag{1}$$

where $a^+ = \max(a, 0)$. The infinitesimal perturbation analysis (IPA) estimator is then obtained by simple differentiation, which for a general parameter θ is given by ([46]):

$$\frac{dT_{k+1}}{d\theta} = \frac{dX_{k+1}}{d\theta} + \left(\frac{dT_k}{d\theta} - \frac{dA_{k+1}}{d\theta}\right) 1\{T_k \geq A_{k+1}\},\ k > 1, \quad \text{with } \frac{dT_1}{d\theta} = \frac{dX_1}{d\theta}. \tag{2}$$

For x a parameter of the (common) customer service time distribution, the unbiased IPA estimator is

$$\frac{dT_{k+1}}{dx} = \frac{dX_{k+1}}{dx} + \frac{dT_k}{dx} 1\{T_k \geq A_{k+1}\},$$

where dX/dx can be calculated based on the distribution for the random variable X. For example, if X is exponentially distributed (with mean x), then dX/dx is simply given by X/x, and (A.2) becomes

$$\frac{dT_{k+1}}{dx} = \frac{X_{k+1}}{x} + \frac{dT_k}{dx} 1\{T_k \geq A_{k+1}\},\ k > 1, \quad \text{with } \frac{dT_1}{dx} = \frac{X_1}{x},$$

the latter assuming that the system starts empty. This is what is used for the $M/M/1$ queue example.

Similarly, for the $U/U/1$ example, where the interarrival time and service time distributions are $U(\theta_1 - \delta_1, \theta_1 + \delta_1)$ $U(\theta_2 - \delta_2, \theta_2 + \delta_2)$, respectively, the four unbiased

IPA estimators are

$$\begin{aligned}
\frac{\partial T_{k+1}}{\partial \theta_1} &= \left(\frac{\partial T_k}{\partial \theta_1} - 1\right) 1\{T_k \geq A_{k+1}\},\ k > 1, \quad \text{with } \frac{\partial T_1}{\partial \theta_1} = 0, \\
\frac{\partial T_{k+1}}{\partial \theta_2} &= 1 + \frac{\partial T_k}{\partial \theta_2} 1\{T_k \geq A_{k+1}\},\ k > 1, \quad \text{with } \frac{\partial T_1}{\partial \theta_2} = 1, \\
\frac{\partial T_{k+1}}{\partial \delta_1} &= \left(\frac{\partial T_k}{\partial \delta_1} - \frac{A_{k+1} - \theta_1}{\delta_1}\right) 1\{T_k \geq A_{k+1}\},\ k > 1, \quad \text{with } \frac{\partial T_1}{\partial \delta_1} = 0, \\
\frac{\partial T_{k+1}}{\partial \delta_2} &= \frac{X_{k+1} - \theta_2}{\delta_2} + \frac{\partial T_k}{\partial \delta_2} 1\{T_k \geq A_{k+1}\},\ k > 1, \quad \text{with } \frac{\partial T_1}{\partial \delta_2} = \frac{X_1 - \theta_2}{\delta_2},
\end{aligned}$$

Bibliography

[1] M. C. Fu. Optimization via simulation: A review. *Annals of Operations Research*, 53:199–248, 1994.

[2] M. C. Fu. Optimization for simulation: Theory vs. practice (Feature Article). *INFORMS Journal on Computing*, 14(3):192–215, 2002.

[3] M. C. Fu. Simulation optimization: Evolution or revolution? *INFORMS Journal on Computing*, 14(3):226–227, 2002.

[4] *Handbook of Simulation Optimization*. Springer, 2014.

[5] W. B. Powell and I. O. Ryzhov. *Optimal Learning*. John Wiley and Sons, 2012.

[6] Y. C. Ho and X. R. Cao. *Perturbation Analysis and Discrete Event Dynamic Systems*. Kluwer Academic, 1991.

[7] P. Glasserman. *Gradient Estimation via Perturbation Analysis*. Kluwer Academic Publishers, Boston, Massachusetts, 1991.

[8] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, New York, 2004.

[9] R. Y. Rubinstein. *Monte Carlo Optimization, Simulation and Sensitivity of Queueing Networks*. John Wiley & Sons, 1986.

[10] R. Y. Rubinstein and A. Shapiro. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley & Sons, 1993.

[11] M. C. Fu. Stochastic gradient estimation. In S. G. Henderson and B. L. Nelson, editors, *Handbooks of Operations Research and Management Science, vol. 13: Simulation*, pages 575–616. North-Holland Publishing, Amsterdam, 2006.

[12] S. Asmussen and P. Glynn. *Stochastic Simulation: Algorithms and Analysis.* Springer, New York, 2007.

[13] M. C. Fu. What you should know about simulation and derivatives. *Naval Research Logistics*, 55(8):723–736, 2008.

[14] R. R. Barton and M. Meckesheimer. Metamodel-based simulation optimization. In S. G. Henderson and B. L. Nelson, editors, *Handbooks in Operations Research and Management Science: Simulation*, chapter 18, pages 535–574. Elsevier, 2006.

[15] R. R. Barton. Simulation optimization using metamodels. In M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, editors, *Proceedings of the 2009 Winter Simulation Conference*, pages 230–238, Piscataway, New Jersey, December 2009. Institute of Electrical and Electronics Engineers, Inc.

[16] J.P.C. Kleijnen. *Design and Analysis of Simulation Experiments.* New York: Springer, 2008.

[17] Feng Yang, Bruce Ankenman, and Barry L Nelson. Efficient generation of cycle time-throughput curves through simulation and metamodeling. *Naval Research Logistics*, 54(1):78–93, 2007.

[18] N.A.C. Cressie. *Statistics for spatial data.* Wiley series in probability and mathematical statistics: Applied probability and statistics. John Wiley and Sons, New York, 2, revised edition, 1993.

[19] Michael L. Stein. *Interpolation of spatial data: some theory for kriging.* Springer series in statistics. Springer-Verlag, New York, 1999.

[20] Jack P. C. Kleijnen, Wim C. M. van Beers, and Inneke van Nieuwenhuyse. Constrained optimization in expensive simulation: Novel approach. *European Journal of Operational Research*, 202(1):164–174, April 2010.

[21] B. E. Ankenman, B. L. Nelson, and J. Staum. Stochastic Kriging for Simulation Metamodeling. *Operations research*, 58(2):371–382, March 2010.

[22] W. Liu. *Development of gradient-enhanced kriging approximations for multidisciplinary design optimization.* PhD thesis, University of Notre Dame, 2003.

[23] T. J. Santner, B. Williams, and W. Notz. *The Design and Analysis of Computer Experiments.* Springer-Verlag, 2003.

[24] Y. C. Ho, L. Shi, L. Dai, and W. Gong. Optimizing discrete event dynamic systems via the gradient surface method. *Discrete Event Dynamic Systems: Theory and Applications*, 2:99–120, Jan 1992.

[25] X. Chen, B. E. Ankenman, and B. L. Nelson. Enhancing stochastic kriging metamodels with gradient estimators. *Operations Research*, 61(2):512–528, 2013.

[26] J. J. Alonso and H. S. Chung. Using gradients to construct cokriging approximation models for high-dimensional design optimization problems. In *40th AIAA Aerospace Sciences Meeting and Exhibit, AIAA*, pages 2002–0317, 2002.

[27] S. S. Gupta and K. J. Miescke. Bayesian look ahead one-stage sampling allocations for selection of the best population. *Journal of Statistical Planning and Inference*, 54(2):229–244, 1996.

[28] P. I. Frazier, W. B. Powell, and S. Dayanik. A knowledge gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

[29] S. E. Chick and K. Inoue. New procedures to select the best simulated system using common random numbers. *Management Science*, 47:1133–1149, 2001.

[30] S. E. Chick, J. Branke, and C. Schmidt. Sequential sampling to myopically maximize the expected value of information. *INFORMS Journal on Computing*, 22(1):71–80, 2010.

[31] H. S. Chang. Converging marriage in honey-bees optimization and application to stochastic dynamic programming. *Journal of Global Optimization*, 35(3):423–441, 2006.

[32] J. Branke, S. E. Chick, and C. Schmidt. Selecting a selection procedure. *Management Science*, 53(12):1916–1932, 2007.

[33] E. Cope. Bayesian strategies for dynamic pricing in e-commerce. *Naval Research Logistics*, 54(3):265–281, 2007.

[34] V. F. Farias and Ben. Van Roy. Dynamic pricing with a prior on market response. *Operations Research*, 58(1):16–29, 2010.

[35] A. X. Carvalho and M. L. Puterman. Learning and pricing in an internet environment with binomial demands. *Journal of Revenue and Pricing Management*, 3(4):320–336, 2005.

[36] M. Chhabra and S. Das. Learning the Demand Curve in Posted-Price Digital Goods Auctions. In *Proceedings of the 10th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 63–70, 2011.

[37] Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

[38] A. V. den Boer and B. Zwart. Dynamic pricing and learning with finite inventories. *Submitted for publication*, 2011.

[39] J. M. Harrison, N. B. Keskin, and A. Zeevi. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science*, 58(3):570–586, 2012.

[40] J. Broder and P. Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

[41] George E. P. Box and Norman R. Draper. *Response surfaces, mixtures, and ridge analyses.* Wiley & Sons, New York, Jan 2007.

[42] M. C. Fu. Gradient estimation. In S. G. Henderson and B. L. Nelson, editors, *Handbooks in Operations Research and Management Science: Simulation*, chapter 19, pages 575–616. Elsevier, 2006.

[43] S. Weisberg. *Applied Linear Regression, 3rd Edition.* Wiley & Sons, New York, 2005.

[44] Alvin C. Rencher and G. Bruce. Schaalje. *Linear Models in Statistics.* Wiley & Sons, New York, 2007.

[45] R. J. Beckman and R. D. Cook. Outliers. *Technometrics*, 25(2):119–163, 1983.

[46] R. Suri and M. A. Zazanis. Perturbation analysis gives strongly consistent sensitivity estimates for the $M/G/1$ queue. *Management Science*, 34:39–64, 1988.

[47] M. C. Fu. Convergence of a stochastic approximation algorithm for the $GI/G/1$ queue using infinitesimal perturbation analysis. *Journal of Optimization Theory and Applications*, 65(1):149–160, 1990.

[48] P. L'Ecuyer, N. Giroux, and P. Glynn. Stochastic optimization by simulation: Numerical experiments with the $M/M/1$ queue in steady-state. *Management Science*, 40(10):1245–1261, 1994.

[49] Anatoly Zhigljavsky, Holger Dette, and Andrey Pepelyshev. A new approach to optimal design for linear models with correlated observations. *Journal of the American Statistical Association*, 105(491):1093–1103, 2010.

[50] J. P. C. Kleijnen and W. C. M. van Beers. Robustness of kriging when interpolating in random simulation with heterogeneous variances: some experiments. *European Journal of Operational Research*, 165(3):826 – 834, 2005.

[51] W. Xie, B. L. Nelson, and J. Staum. The influence of correlation functions on stochastic kriging metamodels. In *Proceedings of the 2010 Winter Simulation Conference*, pages 1067–1078, Piscataway, New Jersey, December 2010. Institute of Electrical and Electronics Engineers, Inc, Winter Simulation Conference.

[52] X. Chen, B. E. Ankenman, and B. L. Nelson. The effects of common random numbers on stochastic kriging metamodels. *ACM Transactions on Modeling and Computer Simulation*, 22(2):7:1–7:20, March 2012.

[53] F. Zhang and Q. Zhang. Eigenvalue inequalities for matrix product. *IEEE Transactions on Automatic Control*, 51(9):1506 –1509, Sept. 2006.

[54] T. Chu, J. Zhu, and H. Wang. Penalized maximum likelihood estimation and variable selection in geostatistics. *The Annals of Statistics*, 39(5):2607–2625, 2011.

[55] R. Li and A. Sudjianto. Analysis of computer experiments using penalized likelihood in gaussian kriging models. *Technometrics*, 47(2):111–121, May 2005.

[56] P. L'Ecuyer. A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Science*, 36:1364–1383, 1990.

[57] W. C. M. van Beers and J. P. C. Kleijnen. Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping. *European Journal of Operational Research*, 186(3):1099 – 1113, 2008.

[58] J. Staum. Better simulation metamodeling: The why, what, and how of stochastic kriging. In M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, editors, *Proceedings of the 2009 Winter Simulation Conference*, pages 119–133, Piscataway, New Jersey, December 2009. Institute of Electrical and Electronics Engineers, Inc, Winter Simulation Conference.

[59] G. Marmidis, S. Lazarou, and E. Pyrgioti. Optimal placement of wind turbines in a wind park using Monte Carlo simulation. *Renewable Energy*, 33(7):1455 – 1460, 2008.

[60] P. Francis, K. Smilowitz, and M. Tzur. The period vehicle routing problem with service choice. *Transportation Science*, 40(4):439–454, 2006.

[61] A. Arlotto, N. Gans, and S. Chick. Optimal employee retention when inferring unknown learning curves. In B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, editors, *Proceedings of the 2010 Winter Simulation Conference*, pages 1178–1188, 2010.

[62] R. E. Bechhofer. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, 25(1):pp. 16–39, 1954.

[63] R. E. Bechhofer, T. J. Santner, and D. M. Goldsman. *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons.* Wiley, 1995.

[64] S.-H. Kim and B. L. Nelson. A fully sequential procedure for indifference-zone selection in simulation. *ACM Transactions on Modeling and Computer Simulation*, 11:251–273, 2001.

[65] S.-H. Kim and B. L. Nelson. On the asymptotic validity of fully sequential selection procedures for steady-state simulation. *Operations Research*, 54(3):475–488, 2006.

[66] T. Homem-de-Mello. A study on the cross-entropy method for rare-event probability estimation. *INFORMS Journal on Computing*, 2006. accepted for publication.

[67] S.-H. Kim and B. L. Nelson. Selecting the best system. In S.G. Henderson and B.L. Nelson, editors, *Handbooks of Operations Research and Management Science, vol. 13: Simulation*, pages 501–534. North-Holland Publishing, Amsterdam, 2006.

[68] S.-H. Kim and B. L. Nelson. Recent advances in ranking and selection. In S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, editors, *Proceedings of the 2007 Winter Simulation Conference*, pages 162–172, 2007.

[69] L. J. Hong and B. L. Nelson. A brief introduction to optimization via simulation. In M.D. Rosetti, R.R. Hill, B. Johansson, A. Dunkin, and R.G. Ingalls, editors, *Proceedings of the 2009 Winter Simulation Conference*, pages 75–85, 2009.

[70] S.-H Kim. Statistical ranking and selection. In S. I. Gass and M. C. Fu, editors, *Encyclopedia of Operations Research and Management Science*. Springer, 2013. To appear.

[71] C. H. Chen, D. He, M. C. Fu, and L. H. Lee. Efficient simulation budget allocation for selecting an optimal subset. *INFORMS Journal on Computing*, 2008. accepted for publication.

[72] D. He, L. H. Lee, C. H. Chen, M. C. Fu, and S. Wasserkrug. Simulation optimization using the cross-entropy method with optimal computing budget allocation. *ACM Transactions on Modeling and Computer Simulation*, 20(1):4:1–4:22, February 2010.

[73] C.-H. Chen and L. H. Lee. *Stochastic Simulation Optimization: An Optimal Computing Budget Allocation*. System Engineering and Operations Research. World Scientific, 2011.

[74] W. N. Yang and B. L. Nelson. Using common random numbers and control variates in multiple-comparison procedures. *Operations Research*, 39:583–591, 1991.

[75] B. L. Nelson and F. J. Matejcik. Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Management Science*, 41:1935–1945, 1995.

[76] M. C. Fu, J. Q. Hu, C. H. Chen, and X. Xiong. Simulation allocation for determining the best design in the presence of correlated sampling. *INFORMS Journal on Computing*, 19(1):101–111, 2007.

[77] P. I. Frazier, W. B. Powell, and S. Dayanik. The knowledge-gradient policy for correlated normal rewards. *INFORMS Journal on Computing*, 21(4):599–613, 2009.

[78] P I. Frazier, J. Xie, and S. E. Chick. Value of information methods for pairwise sampling with correlations. In S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, editors, *Proceedings of the 2011 Winter Simulation Conference*, pages 3974–3986, 2011.

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

[80] R. C. H. Cheng and C. S. M. Currie. Optimization by simulation metamodelling methods. In R.G. Ingalls, M.D. Rossetti, J. S. Smith, and B. A. Peters, editors, *Proceedings of the 2004 Winter Simulation Conference*, pages 485–490, 2004.

[81] B. Biller and C. G. Corlu. Accounting for parameter uncertainty in large-scale stochastic simulations with correlated inputs. *Operations Research*, 59(3):661–673, 2011.

[82] J. Kadane and R. Trader. A bayesian treatment of multivariate normal data with observations missing at random. *Statistical Decision Theory and Related Topics*, 4(1):225–234, 1988.

[83] F. Dominici, G. Parmigiani, and M. Clyde. Conjugate analysis of multivariate normal data with incomplete observations. *Canadian Journal of Statistics*, 28(3):533–550, 2000.

[84] K. Triantafyllopoulos. Missing observation analysis for matrix-variate time series data. *Statistics and Probability Letters*, 78(16):2647–2653, 2008.

[85] A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. Chapman & Hall, 2000.

[86] M. H. DeGroot. *Optimal Statistical Decisions*. Wiley-Interscience, New York, wcl edition, 2004.

[87] S. Kotz and S. Nadarajah. *Multivariate t Distributions and Their Applications*. Cambridge University Press, New York, 2004.

[88] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[89] A. Erdélyi and F. G. Tricomi. The asymptotic expansion of a ratio of gamma functions. *Pacific Journal of Mathematics*, 1(1):133–142, 1951.

[90] G. D. Anderson and S.-L. Qiu. A monotoneity property of the gamma function. *Proceedings of The American Mathematical Society*, 125(11), 1997.

[91] D. He, S. E. Chick, and C. H. Chen. The opportunity cost and OCBA selection procedures in ordinal optimization. *IEEE Transactions on Systems, Man, and Cybernetics–Part C*, 37:951–961, 2007.

[92] B. A. Cosgrove, D. Lohmann, K. E. Mitchell, P. R. Houser, E. F. Wood, J. C. Schaake, A. Robock, C. Marshall, J. Sheffield, Q. Duan, L. Luo, R. W. Higgins, R. T. Pinker, J. D. Tarpley, and J. Meng. Real-time and retrospective forcing in the north american land data assimilation system (NLDAS) project. *Journal of Geophysical Research*, 108(D22):8842–8853, 2003.

[93] G. Gallego and G. Van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science*, 40(8):999–1020, 1994.

[94] Y. Feng and G. Gallego. Optimal starting times for end-of-season sales and optimal stopping times for promotional fares. *Management Science*, 41(8):1371–1391, 1995.

[95] E. P. Lazear. Retail pricing and clearance sales. *The American Economic Review*, 76(1):14–32, 1986.

[96] T. P. Minka. Bayesian linear regression. Technical report, Microsoft Research, 2000.

[97] H. Qu, I. O. Ryzhov, and M. C. Fu. Ranking and selection with unknown correlation structures. In *Proceedings of the 2012 Winter Simulation Conference*, 2012. to appear.

[98] T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

[99] David J. Spiegelhalter and Steffen L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20(5):579–605, 1990.

[100] S. Kim. Gradient-based simulation optimization. In L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, editors, *Proceedings of the 2006 Winter Simulation Conference*, pages 159–167, 2006.

[101] D. M. Blei, M. I. Jordan, and J. W. Paisley. Variational bayesian inference with stochastic search. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1367–1374, 2012.

[102] J. C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*. Wiley-Interscience, 2005.

[103] P. L'Ecuyer. On the interchange of derivative and expectation for likelihood ratio derivative estimators. *Management Science*, 41(4):738–747, 1995.

[104] L. Devroye. Random variate generation. In S. G. Henderson and B. L. Nelson, editors, *Handbooks in Operations Research and Management Science: Simulation*, chapter 4. Elsevier, 2005.

[105] G. E. Crooks. Logistic approximation to the logistic-normal integral. Technical report, Lawrence Berkeley National Laboratory, 2009.