Optimal Bayesian experimental design in the presence of model error

by

Chi Feng

B.S., Physics, California Institute of Technology (2012)

Submitted to the Center for Computational Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Computation for Design and Optimization

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

Signature redacted

Author .....
Center for Computational Engineering
February 4, 2015

Signature redacted

Certified by.....
Youssef M. Marzouk
Associate Professor of Aeronautics and Astronautics
Thesis Supervisor

Signature redacted

Accepted by .....
Nicolas Hadjiconstantinou
Professor of Mechanical Engineering
Co-Director, Computation for Design and Optimization

Optimal Bayesian experimental design in the presence of model error

by

Chi Feng

Submitted to the Center for Computational Engineering
on February 4, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Computation for Design and Optimization

Abstract

The optimal selection of experimental conditions is essential to maximizing the value of data for inference and prediction. We propose an information theoretic framework and algorithms for robust optimal experimental design with simulation-based models, with the goal of maximizing information gain in *targeted subsets* of model parameters, particularly in situations where experiments are costly.

Our framework employs a Bayesian statistical setting, which naturally incorporates heterogeneous sources of information. An objective function reflects expected information gain from proposed experimental designs. Monte Carlo sampling is used to evaluate the expected information gain, and stochastic approximation algorithms make optimization feasible for computationally intensive and high-dimensional problems.

A key aspect of our framework is the introduction of model calibration discrepancy terms that are used to "relax" the model so that proposed optimal experiments are more robust to model error or inadequacy. We illustrate the approach via several model problems and misspecification scenarios. In particular, we show how optimal designs are modified by allowing for model error, and we evaluate the performance of various designs by simulating "real-world" data from models not considered explicitly in the optimization objective.

Thesis Supervisor: Youssef M. Marzouk
Title: Associate Professor of Aeronautics and Astronautics

Acknowledgments

The author would like to acknowledge support from the Computational Mathematics program of the Air Force Office of Scientific Research (AFOSR).

Contents

List of Figures

List of Tables

1

Introduction

1.1 Optimal experimental design

Experimental data are important in developing and refining models of physical systems. For example, data may be used for parameter inference, i.e., updating knowledge of parameters in a model, which may in turn be used for making predictions or decisions. Acquiring data from experiments—in the laboratory or in the field—can often be expensive. In addition to material and labor costs costs, some experiments may be difficult to repeat, since they may depend on specific circumstances outside of the experimenter's control, e.g., time-sensitive weather and atmospheric measurements. In this context, maximizing the value of experimental data—designing experiments to be "optimal" by some appropriate measure—is important for improving the efficiency of the modeling process. Experimental design encompasses questions of where, when, and how to make measurements, i.e., which variables to interrogate under what experimental conditions.

These questions have received much attention in the statistics community and in many science and engineering applications. When observables depend linearly on parameters of interest, common solution criteria for the optimal experimental design problem are written as functionals of the information matrix [3]. These criteria include the well-known 'alphabetic optimality' conditions, e.g., A-optimality to minimize the average variance of parameter estimates, or G-optimality to minimize the maximum

variance of model predictions. Bayesian analogues of alphabetic optimality, reflecting prior and posterior uncertainty in the model parameters, can be derived from a decision-theoretic point of view [7]. For instance, Bayesian D-optimality can be obtained from a utility function containing Shannon information while Bayesian A-optimality may be derived from a squared error loss. In the case of linear-Gaussian models, the criteria of Bayesian alphabetic optimality reduce to mathematical forms that parallel their non-Bayesian counterparts [7]. This parallelism is demonstrated in Chapter 2.

For nonlinear models, however, exact evaluation of optimal design criteria is much more challenging. More tractable design criteria can be obtained by imposing additional assumptions, effectively changing the form of the objective; these assumptions include linearizations of the forward model, Gaussian approximations of the posterior distribution, and additional assumptions on the marginal distribution of the data [7]. In the Bayesian setting, such assumptions lead to design criteria that may be understood as approximations of an expected utility. Most of these involve prior expectations of the Fisher information matrix [8]. Cruder "locally optimal" approximations require selecting a "best guess" value of the unknown model parameters and maximizing some functional of the Fisher information evaluated at this point [11]. None of these approximations, though, is suitable when the parameter distribution is broad or when it departs significantly from normality [9]. A more general design framework, free of these limiting assumptions, is preferred [38, 50].

More rigorous information theoretic criteria have been proposed throughout the literature. In [29], Lindley suggests using expected gain in Shannon information, from prior to posterior, as a measure of the information provided by an experiment; the same objective can be justified from a decision theoretic perspective [30, 33]. Sebastiani and Wynn [43] propose selecting experiments for which the marginal distribution of the data has maximum Shannon entropy, which can be interpreted as a special case of Lindley's criterion. Reverting to Lindley's criterion, Ryan [42] introduces a Monte Carlo estimator of expected information gain to design experiments for a model of material fatigue. Terejanu et al. [49] use a kernel estimator of mutual information (equivalent to expected information gain) to identify parameters in chemical kinetic

model. The latter two studies evaluate their criteria on every element of a finite set of possible designs (on the order of ten designs in these examples), and thus sidestep the challenge of optimizing the design criterion over general design spaces. And both report significant limitations due to computation expense; [42] concludes that "full blown search" over the design space is infeasible, and that two order-of-magnitude gains in computational efficiency would be required even to discriminate among the enumerated designs.

1.2 Model discrepancy

> *"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."*
>
> – George E. P. Box *(1987)*

An important component of uncertainty quantification is model discrepancy, which is defined as the difference between the output from a computational model and values of the real physical system that the model is trying to predict. Forward uncertainty quantification—propagating input uncertainty to infer output uncertainty—quantifies only a subset of uncertainty about the physical system since it fails to account for model discrepancy (and other sources of uncertainty) [5].

Model discrepancy was formally introduced as a source of uncertainty in computer model predictions by Kennedy and O'Hagan [24], who referred to it as model inadequacy (other names include model error, model form error, model bias, and structural uncertainty). They considered the problem of using observations of the real physical system to learn about uncertain input parameters through model calibration—what we refer to as parameter inference—and showed how to account for model discrepancy in calibration and in subsequent predictions of the physical system. Since then, their modeling framework have been widely adopted and developed further in [22].

The principle findings of Brynjarsdóttir and O'Hagan paper [5] are:

1. If model discrepancy is ignored, then predictions and parameter inference are

biased, and this bias persists with increasing numbers of observations, i.e,. the posterior mean will not converge to the value of the physical parameter, even with an unlimited number of measurements.

2. If model discrepancy is modeled in a simple, uninformative way, then *interpolations* are unbiased but *extrapolative* predictions and parameter inference will still be biased.

3. In order to obtain 'realistic' learning about model parameters, it is important not just to incorporate model discrepancy but also to model carefully the available prior information about it.

1.3 Thesis contributions

In this thesis, we extend the framework of optimal Bayesian experimental design for nonlinear systems to include nuisance parameters by reformulating the optimization problem as a maximization of the information gain in a subset of model parameters, i.e., the parameters of interest. In addition to developing a new form of the expected utility of an experiment, we also developed a novel adaptive importance sampling scheme to improve the efficiency of the Monte Carlo estimator for the expected utility.

By incorporating nuisance parameters into the design objective, we can easily incorporate 'nonparametric' model error into the experimental design problem by augmenting simulation models with additional nuisance parameters from finite-dimensional representations of nonparametric model discrepancy. In addition to the information-theoretic design objective mentioned above, we also developed alternate criteria for experimental design that accounts for model misspecification, i.e. by computing the expected error during parameter inference when data from one model is fitted using a different, misspecified model.

We begin in Chapter 2 by formulating the optimal experimental design problem in general terms and presenting a framework for solving it exactly, which we use towards a simple experiment with a linear response to motivate the principle findings

above. We also discuss how to incorporate model discrepancy into the optimal design framework. Then, in Chapter 3, we present numerical methods for efficiently solving problems involving nonlinear experiments. In Chapter 4, we apply the numerical methods to a nonlinear design problem with model discrepancy. Finally, in chapter 5, we summarize the findings and discuss possible extensions of our optimal experimental design framework.

2

Optimal Experimental Design

Optimal experimental design requires an appropriate design criterion and objective function that describes the expected value of data from a specific experiment. The design criterion is determined by the specific application, i.e., what the user intends to do with the data from an experiment. For example, if the goal is to infer a physical constant, the objective function should favor experiments that produce data which result in smaller uncertainties in inferred parameters and penalize those that result in larger uncertainties. On the other hand, if our goal was to use experimental data to choose between several models, we would like to use an objective function that favors experiments that produce data that is useful for discriminating between models. In general, the objective function for optimal design should be determined by the specific goals of that experiment.

In this work, we examine the optimal design of experiments when the experimental goal is the inference of a finite number of model parameters, e.g., for model calibration. To this end, we develop an expected utility framework where the utility function describes the information gain in the parameters of interest from an experiment, i.e., we want to find an experimental design that maximizes the expected information gain. In particular, parameters of interest refer to a *subset* of model parameters, since in many scenarios, models have physical parameters and tuning parameters, but we may only wish to reduce our uncertainties in the physical parameters, e.g., for use in other models or for prediction. The concept of targeting a subset of parameters is also discussed

in controls literature as *focused* active inference, as opposed to *unfocused* problem, which also places value on irrelevant states called 'nuisance variables.' Focused active inference on graphical Bayesian models is discussed in [28].

Furthermore, we focus our attention on simultaneous, or, batch experimental design, as opposed to sequential experimental design, where data from one experiment can be used to inform the design of the next. It is also possible to add physical and resource constraints to the design problem by augmenting the resulting optimization problem. However, for simplicity's sake, we will not discuss examples with such constraints, as constraints are usually problem-specific.

2.1 Bayesian experimental design

2.1.1 Bayesian parameter inference

We choose to formulate our experimental design framework in a Bayesian setting for several reasons. First, by treating all uncertain model parameters as random variables, Bayesian parameter inference offers a straightforward method of performing inference in when the data is noisy or incomplete. Furthermore, the same Bayesian framework gives us access to a full description of uncertainty in model predictions from uncertainties in the parameters and models.

The Bayesian paradigm [29] treats unknown parameters as random variables. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where Ω is a sample space, $\mathcal{F}$ is a σ-field, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$. Let the vector of real-valued random variables $\boldsymbol{\theta} : \Omega \to \mathbb{R}^{n_\theta}$ and $\boldsymbol{\eta} : \Omega \to \mathbb{R}^{n_\eta}$ denote the uncertain *parameters of interest* and *nuisance parameters*, respectively. Both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are parameters to be conditioned on experimental data. $\boldsymbol{\theta}$ is associated with a measure μ on $\mathbb{R}^{n_\theta}$, such that $\mu(A) = \mathbb{P}(\boldsymbol{\theta}^{-1}(A))$ for $A \in \mathbb{R}^{n_\theta}$. Likewise, $\boldsymbol{\eta}$ is associated with a measure ν on $\mathbb{R}^{n_\eta}$, such that $\nu(B) = \mathbb{P}(\boldsymbol{\eta}^{-1}(B))$ for $B \in \mathbb{R}^{n_\eta}$. We can then define $p(\boldsymbol{\theta}) = \mathrm{d}\mu \,/\, \mathrm{d}\boldsymbol{\theta}$ as the density of $\boldsymbol{\theta}$ and $p(\boldsymbol{\eta}) = \mathrm{d}\mu \,/\, \mathrm{d}\boldsymbol{\eta}$ as the density of $\boldsymbol{\eta}$, with respect to Lebesgue measure. For the present purposes, we will assume that such densities always exist, and that

$p(\boldsymbol{\theta}, \boldsymbol{\eta})$ denote an appropriate joint probability density of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Similarly, the data $\boldsymbol{y}$ can be treated a real-valued random variable $\boldsymbol{y} : \Omega \to \mathbb{R}^{n_y}$. Finally, let $\boldsymbol{d} \in \mathbb{R}^{n_d}$ denote the *design variables*, or experimental conditions. Hence, n_θ is the number of uncertain parameters of interest, n_η is the number of uncertain nuisance parameters, and n_d is the number of design variables. In the Bayesian setting, upon observing a realization of the data $\boldsymbol{y}$ after performing an experiment under conditions $\boldsymbol{d}$, we can update our knowledge about the model parameters using Bayes' rule:

$$\underbrace{p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{d})}^{\text{likelihood}} \overbrace{p(\boldsymbol{\theta}|\boldsymbol{d})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y}|\boldsymbol{d})}_{\text{evidence}}}. \tag{2.1}$$

Here $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{d})$ is the prior density, $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{d})$ is the likelihood function, $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d})$ is the posterior density, and $p(\boldsymbol{y}|\boldsymbol{d})$ is the evidence. It is reasonable to assume that prior knowledge on $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ do not vary with the experimental design, which lets us write the following simplification: $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{d}) = p(\boldsymbol{\theta}, \boldsymbol{\eta})$.

Figure 2-1: An illustration of Bayes' rule; the shrinkage in the posterior reflects the reduction in uncertainty due to information gained from data $\boldsymbol{y}$.

2.1.2 Expected utility framework

Using a decision-theoretic approach from Lindley [30], the objective for experimental design should have the following general form:

$$U(\boldsymbol{d}) = \int_{\mathcal{Y}} \int_{\boldsymbol{\Theta}} \int_{\mathbf{H}} u(\boldsymbol{y}, \boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\eta})\, p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{y}|\boldsymbol{d})\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\eta}\, \mathrm{d}\boldsymbol{y} \tag{2.2}$$

where $u(\boldsymbol{y}, \boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\eta})$ is a *utility function*, representing the value of observing a particular outcome of the experiment $\boldsymbol{y}$ under experimental conditions $\boldsymbol{d}$, given a particular value of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Since we do not know the outcome of the experiment *a priori*, we instead consider the *expected* utility $U(\boldsymbol{d})$, where the expectation is taken over the joint distribution of $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, and $\boldsymbol{y}$, and where $\boldsymbol{\Theta}$, $\mathbf{H}$ and $\mathcal{Y}$ are the supports of $p(\boldsymbol{\theta})$, $p(\boldsymbol{\eta}|\boldsymbol{\theta})$, and $p(\boldsymbol{y}|\boldsymbol{d})$, respectively.

We choose an information-theoretic design criterion [29], where $u(\boldsymbol{y}, \boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\eta})$ is the relative entropy, or Kullback-Leibler (KL) divergence, from the *marginal* posterior to the *marginal* prior, where the marginalization is over the nuisance parameters $\boldsymbol{\eta}$ In other words, the divergence is from the marginal prior on the parameters of interest

$$p_\theta(\boldsymbol{\theta}) = \int_{\mathbf{H}} p(\boldsymbol{\theta}, \boldsymbol{\eta})\ \mathrm{d}\boldsymbol{\eta}, \tag{2.3}$$

to the marginal posterior on the parameters of interest

$$p_\theta(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d}) = \int_{\mathbf{H}} p(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{y}, \boldsymbol{d})\, \mathrm{d}\boldsymbol{\eta}\,. \tag{2.4}$$

Our reason for marginalizing over the nuisance parameters is to eliminate the

For generic distributions A and B, the divergence from A to B is

$$\mathrm{D}_{\mathrm{KL}}\big(A \,||\, B\big) = \int_{\boldsymbol{\Theta}} p_A(\boldsymbol{\theta}) \ln\left[\frac{p_A(\boldsymbol{\theta})}{p_B(\boldsymbol{\theta})}\right] \mathrm{d}\boldsymbol{\theta} = \mathbb{E}_A\left[\ln \frac{p_A(\boldsymbol{\theta})}{p_B(\boldsymbol{\theta})}\right] \tag{2.5}$$

where p_A and p_B are probability densities, $\boldsymbol{\Theta}$ is the support of $p_B(\boldsymbol{\theta})$, and $0 \ln 0 \equiv 0$. The KL divergence is non-negative, non-symmetric, and reflects the *difference* in information carried by the two distributions. Then, for our specific case, the utility

function is

$$u(\boldsymbol{y}, \boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\eta}) \equiv \mathrm{D}_{\mathrm{KL}}\big(p_\theta(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d}) \,\big|\big|\, p_\theta(\boldsymbol{\theta})\big) = \int_{\boldsymbol{\Theta}} p_\theta(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}, \boldsymbol{d}) \ln\left[\frac{p_\theta(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}, \boldsymbol{d})}{p_\theta(\tilde{\boldsymbol{\theta}})}\right] \mathrm{d}\tilde{\boldsymbol{\theta}}, \tag{2.6}$$

where $\tilde{\boldsymbol{\theta}}$ is a dummy variable of integration. Since the KL-divergence and inner marginalization remove the dependence of the utility criterion on the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, we can apply the following simplification

$$u(\boldsymbol{y}, \boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\eta}) \to u(\boldsymbol{y}, \boldsymbol{d}), \tag{2.7}$$

which lets us write the expected utility of an experimental design (2.2) as a single expectation:

$$U(\boldsymbol{d}) = \int_{\mathcal{Y}} \int_{\boldsymbol{\Theta}} \int_{\mathbf{H}} u(\boldsymbol{y}, \boldsymbol{d})\, p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{y}|\boldsymbol{d})\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\eta}\, \mathrm{d}\boldsymbol{y} \tag{2.8}$$

$$= \int_{\boldsymbol{Y}} u(\boldsymbol{y}, \boldsymbol{d}) p(\boldsymbol{y}|\boldsymbol{d})\, \mathrm{d}\boldsymbol{y}\,. \tag{2.9}$$

Now, substituting the utility criterion from (2.6) into the objective function (2.9), we obtain

$$U(\boldsymbol{d}) = \int_{\boldsymbol{Y}} \int_{\boldsymbol{\Theta}} p_\theta(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}, \boldsymbol{d}) \ln\left[\frac{p_\theta(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}, \boldsymbol{d})}{p_\theta(\tilde{\boldsymbol{\theta}})}\right] \mathrm{d}\boldsymbol{\theta}\; p(\boldsymbol{y}|\boldsymbol{d})\, \mathrm{d}\boldsymbol{y} \tag{2.10}$$

To simplify notation, $\tilde{\boldsymbol{\theta}}$ in (2.10) is replaced by $\boldsymbol{\theta}$, yielding

$$\begin{aligned} U(\boldsymbol{d}) &= \int_{\boldsymbol{Y}} \int_{\boldsymbol{\Theta}} p_\theta(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d}) \ln\left[\frac{p_\theta(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d})}{p_\theta(\boldsymbol{\theta})}\right] \mathrm{d}\boldsymbol{\theta}\; p(\boldsymbol{y}|\boldsymbol{d})\, \mathrm{d}\boldsymbol{y} \\ &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{d}}\Big[\mathrm{D}_{\mathrm{KL}}\big(p_\theta(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d}) \,\big|\big|\, p_\theta(\boldsymbol{\theta})\big)\Big], \end{aligned} \tag{2.11}$$

which means that the expected utility $U(\boldsymbol{d})$ is the *expected information gain* in the parameters of interest $\boldsymbol{\theta}$, irrespective of the information gain in the nuisance parameters $\boldsymbol{\eta}$. Intuitively, the expected information gain describes the average difference in entropy

in the marginal of $\boldsymbol{\theta}$ when updating our beliefs when observing data. We also note that $U(\boldsymbol{d})$ is equivalent to the *mutual information*[1] between the parameters of interest $\boldsymbol{\theta}$ and the data $\boldsymbol{y}$ given the design $\boldsymbol{d}$.

Finally, the expected utility must be maximized over the design space $\mathcal{D}$ to find the optimal experimental design

$$\boldsymbol{d}^* = \arg\max_{\boldsymbol{d}\in\mathcal{D}} U(\boldsymbol{d}). \tag{2.12}$$

In practice, most experimental design problems do not yield themselves to a fully analytic treatment of the expected utility, which motivates us to develop numerical approximations instead. In Chapter 3, we introduce Monte Carlo estimators of the expected utility and introduce the algorithms that allow us to perform optimization with a noisy objective function.

2.1.3 Alternative optimality criteria

Although applying the expected utility framework above results in experimental designs that maximize expected information gain, we wish to develop alternate criteria to asses the quality of inference for a chosen design. This is especially important for assess the quality of parameter inference when the model producing the data is different from the model used for parameter inference, e.g. in cases of model misspecification.

The criteria we use to measure inference quality is the *posterior expected loss*, also known as the Bayes' risk when the loss function is the mean squared error. In other words, let $\boldsymbol{\theta}^{\text{true}}$ represent the "true" value of $\boldsymbol{\theta}$, and let $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ be an estimator of $\boldsymbol{\theta}^{\text{true}}$.

[1] Using the rules of conditional probability, we can write

$$\begin{aligned} U(\boldsymbol{d}) &= \int_{\boldsymbol{Y}} \int_{\boldsymbol{\Theta}} p_\theta(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d}) \ln\left[\frac{p_\theta(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})}{p_\theta(\boldsymbol{\theta})}\right] \mathrm{d}\boldsymbol{\theta}, p(\boldsymbol{y}|\boldsymbol{d})\,\mathrm{d}\boldsymbol{y} \\ &= \int_{\boldsymbol{Y}} \int_{\boldsymbol{\Theta}} p_\theta(\boldsymbol{\theta},\boldsymbol{y}|\boldsymbol{d}) \ln\left[\frac{p_\theta(\boldsymbol{\theta},\boldsymbol{y}|\boldsymbol{d})}{p_\theta(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{d})}\right] \mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{y} \\ &= I(\boldsymbol{\theta};\, \boldsymbol{y}|\boldsymbol{d}), \end{aligned}$$

which is the mutual information between the parameters of interest and data, given the design.

The expectation of the squared loss function

$$L(\boldsymbol{\theta}, \boldsymbol{y}) = \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\boldsymbol{y})\|^2 \tag{2.13}$$

over the support of $\boldsymbol{\theta}$ and $\boldsymbol{y}$ is the Bayes' risk,

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{y}}\big[L(\boldsymbol{\theta}, \boldsymbol{y})\big] = \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{y}}\big[\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\boldsymbol{y})\|^2\big]. \tag{2.14}$$

Here, we let $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ be the minimum mean square error (MMSE) estimator of $\boldsymbol{\theta}$, which is the posterior mean, defined as

$$\begin{aligned} \boldsymbol{\theta}^{\mathrm{PM}}(\boldsymbol{y}, \boldsymbol{d}) &= \int_{\boldsymbol{\Theta}} \tilde{\boldsymbol{\theta}}\, p(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}, \boldsymbol{d})\, \mathrm{d}\tilde{\boldsymbol{\theta}} \\ &= \int_{\mathbf{H}} \int_{\boldsymbol{\Theta}} \tilde{\boldsymbol{\theta}}\, p(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}|\boldsymbol{y}, \boldsymbol{d})\, \mathrm{d}\tilde{\boldsymbol{\theta}}\, \mathrm{d}\tilde{\boldsymbol{\eta}}\,. \end{aligned} \tag{2.15}$$

Then, the Bayes' risk can be written as

$$\begin{aligned} \mathbb{E}_{\boldsymbol{y},\boldsymbol{\theta}}\, \|\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{PM}}(\boldsymbol{y}, \boldsymbol{d})\|^2 &= \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{Y}} \|\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{PM}}(\boldsymbol{y}, \boldsymbol{d})\|^2\, p(\boldsymbol{y}, \boldsymbol{\theta}|\boldsymbol{d})\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{Y}} \|\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{PM}}(\boldsymbol{y}, \boldsymbol{d})\|^2\, p(\boldsymbol{\theta})\, \check{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{d})\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\boldsymbol{\theta}\,. \end{aligned} \tag{2.16}$$

At this point, it is important to note that the Bayes' risk as formulated above is the criteria for Bayesian A-optimal design [7]. In the last line of Equation (2.16) we introduced the notation $\check{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{d})$, which for the Bayes' risk, is the likelihood distribution that corresponds to the posterior distribution in Equation (2.15). However, by modifying the Bayes' risk formulation by allowing $\check{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{d})$ to be the likelihood function of another distribution which could correspond to a higher fidelity model that shares[2] the same parameters $\boldsymbol{\theta}$, we can use this modified Bayes' risk as an integrated measure of the quality of parameter inference under model misspecification, i.e., we assume that the model used to compute the posterior mean in Equation (2.15) is misspecified, and the model corresponding to the likelihood $\check{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{d})$ is the "true"

[2]In the sense that the parameters $\boldsymbol{\theta}$ have similar physical meaning.

model.

2.1.4 Linear Gaussian example

Now that we have a formulation for optimal Bayesian experimental design with nuisance parameters, we will apply it to a toy problem that we can solve analytically. We can characterize experiments by using the following observation model

$$\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{d}) + \boldsymbol{\epsilon}, \tag{2.17}$$

where $\boldsymbol{G}$ represents a deterministic mathematical or computer model, and $\boldsymbol{\epsilon}$ is additive observation noise. This framework can be generalized to other noise models, e.g., multiplicative noise by using an appropriate likelihood function $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{d})$.

The need for incorporating nuisance parameters in the experimental design process can be illustrated by the following example. Consider the design of an experiment with an overall algebraic observation model given by

$$\boldsymbol{y}(\theta, \eta, d) = G(\theta, \eta, d) + \boldsymbol{\epsilon} \tag{2.18}$$

$$\begin{bmatrix} y_1(\theta, \eta, d) \\ y_2(\theta, \eta, d) \end{bmatrix} = \begin{bmatrix} \sigma_\epsilon \sqrt{kd} & 0 \\ 0 & \sigma_\epsilon \sqrt{k(1-d)} \end{bmatrix} \begin{bmatrix} \theta \\ \eta \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}, \quad k > 0 \tag{2.19}$$

with i.i.d. $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, \sigma_\epsilon)$, and design space $\boldsymbol{d} \in [0, 1]$. The prior distributions on θ and η are i.i.d. $\theta \sim \mathcal{N}(0, 1)$ and $\eta \sim \mathcal{N}(0, 1)$. The posterior $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d}) \sim \mathcal{N}(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Sigma}_{\text{post}})$ is a multivariate normal with mean and covariance

$$\boldsymbol{\mu}_{\text{post}} = \begin{bmatrix} \frac{\sqrt{kd}}{\sigma_\epsilon + kd\sigma_\epsilon} y_1 \\ \frac{\sqrt{k(1-d)}}{\sigma_\epsilon + k(1-d)\sigma_\epsilon} y_2 \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\text{post}} = \begin{bmatrix} \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + kd\sigma_\epsilon^2} & 0 \\ 0 & \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + k(1-d)\sigma_\epsilon^2} \end{bmatrix}$$

The structure of the model parameterizes a trade-off between learning about θ and learning about η. As d increases, we obtain a greater "signal-to-noise ratio" when targeting θ, while the "signal-to-noise ratio" of measuring η decreases. This trade-off can be visualized by observing anisotropic scaling of the posterior distribution in the

θ direction or the η direction when changing $\boldsymbol{d}$, as illustrated in Fig. 2-2. From the figure, it is obvious that the design that minimizes the uncertainty in the parameter of interest, θ, is the design corresponding to $d = 1$. We will apply the expected utility framework from §2.1.2 using two design criteria: information gain in both parameters (θ, η) and information gain in θ alone. We will show that using the first criteria results in a sub-optimal design, and that the second criteria results in an optimal design.

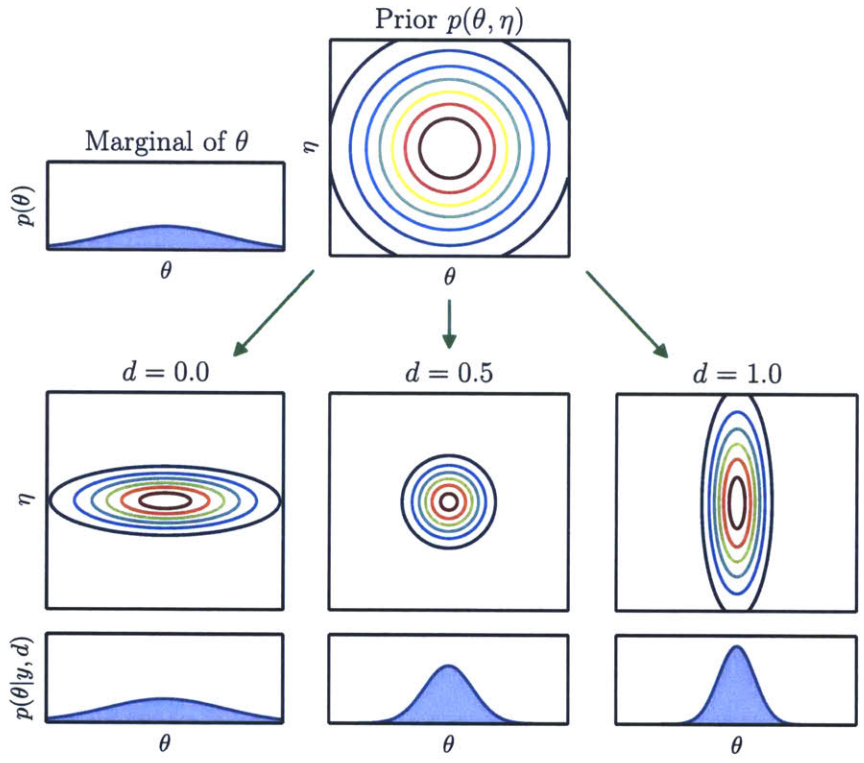


Figure 2-2: Joint and marginal posteriors of the 2D linear Gaussian example for $d = 0, 0.5, 1.0$.

Since we are working with a linear Gaussian model, we can obtain analytical results for the KL divergence and expected information gain for both θ and η, or for the marginal of θ. The KL divergence between two k-dimensional multivariate normal distributions $\mathcal{N}_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ is

$$\mathrm{D}_{\mathrm{KL}}(\mathcal{N}_0 \,\|\, \mathcal{N}_1) = \frac{1}{2}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0\right) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - k - \ln\left(\frac{\det \boldsymbol{\Sigma}_0}{\det \boldsymbol{\Sigma}_1}\right)\right].$$

If we let $G_{11} = \sigma\sqrt{kd}$ and $G_{22} = \sigma\sqrt{k(1-d)}$, the KL divergence from the joint prior to the joint posterior is

$$\mathrm{D_{KL}}\big(p(\theta,\eta|\boldsymbol{y},d)\,\|\,p(\theta,\eta)\big) = \frac{1}{2}\Bigg(\frac{G_{11}^2 y_1^2}{(G_{11}^2+\sigma_\epsilon^2)^2} + \frac{\sigma_\epsilon^2}{G_{11}^2+\sigma_\epsilon^2} + \frac{G_{22}^2 y_2^2}{(G_{22}^2+\sigma_\epsilon^2)^2} + \frac{\sigma_\epsilon^2}{G_{22}^2+\sigma_\epsilon^2} - 2 - \ln\left[\frac{\sigma_\epsilon^2}{(G_{11}^2+\sigma_\epsilon^2)(G_{22}^2+\sigma_\epsilon^2)}\right]\Bigg) \tag{2.20}$$

We can then take the expectation over $p(\boldsymbol{y}|\boldsymbol{d})$, which is also a multivariate normal with mean and covariance

$$\boldsymbol{\mu_y} = \begin{bmatrix}0\\0\end{bmatrix}, \quad \boldsymbol{\Sigma_y} = \begin{bmatrix}G_{11}^2+\sigma_\epsilon^2 & 0\\ 0 & G_{22}^2+\sigma_\epsilon^2\end{bmatrix}, \tag{2.21}$$

which gives us the expected information gain in both θ and η

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{d}}\big[\mathrm{D_{KL}}\big(p(\theta,\eta|\boldsymbol{y},d)\,\|\,p(\boldsymbol{\theta},\boldsymbol{\eta})\big)\big] &= -\frac{1}{2}\ln\left[\frac{\sigma_\epsilon^4}{(G_{11}^2+\sigma_\epsilon^2)(G_{22}^2+\sigma_\epsilon^2)}\right]\\ &= -\frac{1}{2}\ln\left[\frac{\sigma_\epsilon^2}{(1+kd)(1+k(1-d))}\right].\end{aligned} \tag{2.22}$$

It is obvious from visual inspection of the expected information gain on the entire design space $d \in [0,1]$, in Fig. 2-3) that the design that maximizes the expected information gain (the optimal design) is $d^* = \frac{1}{2}$. Notice that maximizing the expected information gain in Equation (2.22) is equivalent to minimizing the quantity in the denominator of the logarithm:

$$\arg\min_d -\frac{1}{2}\ln\left[\frac{\sigma_\epsilon^2}{(1+kd)(1+k(1-d))}\right] \quad \Rightarrow \quad \arg\min_d \det(\boldsymbol{\Sigma}_{\text{post}})$$

which is the equivalent of minimizing the determinant of the posterior covariance matrix, i.e., classical D-optimal design. The equivalence of Bayesian alphabetic optimality criteria to classical alphabetic optimality for linear Gaussian models is discussed in [7]. However, note that this equivalence does not hold when the objective is information gain in a *subset* of model parameters.

Now, we repeat the same procedure, but use the KL divergence from the marginal

prior to the marginal posterior, i.e.,

$$\mathrm{D_{KL}}\big(p_\theta(\theta|\boldsymbol{y},d)\,\|\,p_\theta(\theta)\big)=\frac{1}{2}\bigg(-\frac{G_{11}^2(G_{11}^2-y_2^2+\sigma_\epsilon^2)}{(G_{11}^2+\sigma_\epsilon^2)^2}-\ln\bigg[\frac{\sigma_\epsilon^2}{G_{11}^2+\sigma_\epsilon^2}\bigg]\bigg)$$

We then take the expectation over $p(\boldsymbol{y}|\boldsymbol{d})$ which gives us the expected information gain in the θ-marginal

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{d}}\big[\mathrm{D_{KL}}\big(p_\theta(\theta|\boldsymbol{y},d)\,\|\,p_\theta(\boldsymbol{\theta})\big)\big]&=-\frac{1}{2}\ln\bigg[\frac{\sigma_\epsilon^2}{(G_{11}^2+\sigma_\epsilon^2)}\bigg]\\&=-\frac{1}{2}\ln\bigg[\frac{1}{1+kd}\bigg].\end{aligned}\tag{2.23}$$

From visual inspection of the expected information gain on the entire design space $d\in[0,1]$, in Fig. 2-3) we see that the optimal design is $d^*=1$.

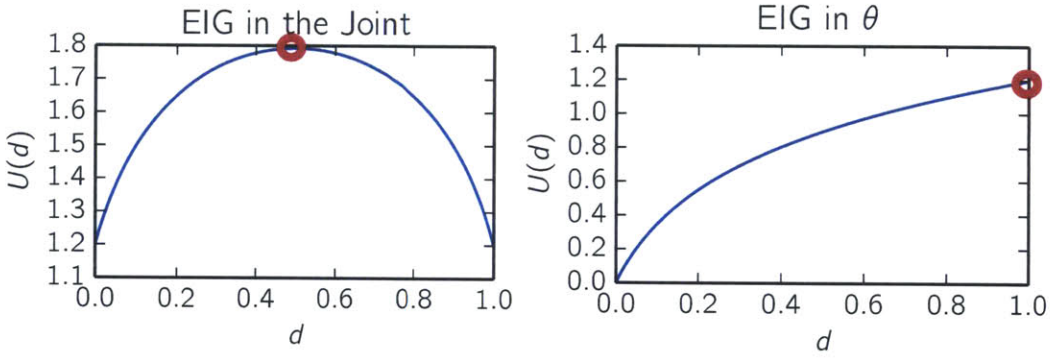


Figure 2-3: Expected information gain of the linear Gaussian example

We will also examine the quality of inference of the two designs found above using the Bayes' risk as defined in Equation (2.16). Since we only care about the parameters of interest, we need only find the mean squared error in the value of θ, which only depends on y_1:

$$\mathbb{E}_{\theta,y_1|d}\big[(\theta^{\mathrm{PM}}-\theta)^2\big]\tag{2.24}$$

where $y_1\sim\mathcal{N}(G_{11}\theta,G_{11}^2+\sigma_\epsilon^2)$ and θ^{PM} is the marginal posterior mean

$$\theta^{\mathrm{PM}}=\frac{\sqrt{kd}}{\sigma_\epsilon(1+kd)}y_1.\tag{2.25}$$

Then, the expectation for the expected mean squared error is taken over the prior $p(\theta) \sim \mathcal{N}(0, 1)$, which gives us the Bayes' risk

$$J(d) = \sigma_\epsilon^2 kd + \frac{kd + 2}{(kd + 1)^2}. \tag{2.26}$$

On the design space $d \in [0, 1]$, this quantity is minimized at $d = 1$, which coincides with the optimal design for expected information gain in θ alone. However, it is possible to choose k such that the $J(d)$ is minimized at $d < 1$. This is not unexpected, since $d = 1$ is the Bayesian D-optimal design, which is not always equivalent to the Bayesian A-optimal design.

To summarize, we have shown that in this simple two-dimensional linear Gaussian example, the expected utility framework developed in §2.1.2 correctly identifies the optimal design when the experimental goal is inference on a subset of model parameters. We will revisit this example in Chapter 4, where we develop numerical methods for evaluating the expected information gain and the Bayes' risk. The analytical results obtained here will be used as benchmarks for studying the bias and convergence of estimators and approximations.

2.2 Gaussian processes

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where Ω is a sample space, $\mathcal{F}$ is a σ-algebra over Ω, and $\mathbb{P}$ is a probability measure on $\mathcal{F}$. Also, let $D \subset \mathbb{R}^n$ be a bounded spatial domain. If $M(\boldsymbol{x}) : \Omega \to \mathbb{R}$ is a $\mathcal{F}$-measurable mapping for every $\boldsymbol{x} \in D$, then $M : \Omega \times D \to \mathbb{R}$ is a random field[3] We can interpret $M(\boldsymbol{x}, \omega)$, for $\omega \in \Omega$, as a collection of real-valued random variables indexed by $\boldsymbol{x} \in D$. Alternatively, one can view $M(\,\cdot\,)$ as a random variable taking values the space of all real-valued functions on D.

[3]Our presentation focuses on 'random fields,' where stochastic processes are indexed by a spatial coordinate. The results shown here are also applicable to process indexed by other coordinates, e.g., time. In our application, the domain D is usually the space of experimental designs, i.e., the space spanned by the control variables of the optimal design problem.

If, for any $n \geq 1$ we have

$$(M(\boldsymbol{x}_1), \ldots, M(\boldsymbol{x}_n)) \overset{\text{i.d.}}{=} (M(\boldsymbol{x}_1 + \boldsymbol{s}), \ldots, M(\boldsymbol{x} + \boldsymbol{s})), \tag{2.27}$$

where $\overset{\text{i.d.}}{=}$ denotes equality in distribution, $\boldsymbol{s}$ is a spatial shift, and $\{\boldsymbol{x}_i, \boldsymbol{x}_i + \boldsymbol{s}\}_{i=1}^n \in D$, then M is said to be stationary [16]. If in addition, all finite-dimensional distributions of M are multivariate normal, then M is a stationary *Gaussian process* (GP). Let $\boldsymbol{M}_{(n)} = (M(\boldsymbol{x}_1), \ldots, M(\boldsymbol{x}_n))$ denote the restriction of M to a finite set of indices. Then the characteristic function of $\boldsymbol{M}_{(n)}$ is

$$\phi_M(\boldsymbol{\lambda}) = \mathbb{E}[\exp(i\boldsymbol{\lambda}^T \boldsymbol{M}_{(n)})] = \exp\left(i\lambda^T \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\lambda}^T \boldsymbol{\Sigma} \boldsymbol{\lambda}\right), \qquad \boldsymbol{\lambda} \in \mathbb{R}^n, \tag{2.28}$$

where the mean is spatially invariant, $\boldsymbol{\mu} = \mu \mathbf{1}_n$, and entries of $\boldsymbol{\Sigma}$ are values of the covariance function C:

$$\Sigma_{ij} = C(\boldsymbol{x}_i, \boldsymbol{x}_j) \equiv \text{Cov}[M(\boldsymbol{x}_i), M(\boldsymbol{x}_j)] = \mathbb{E}[(M(\boldsymbol{x}_i) - \boldsymbol{\mu})(M(\boldsymbol{x}_j) - \boldsymbol{\mu})]. \tag{2.29}$$

Gaussian processes have finite second moments, i.e., $M(\boldsymbol{x}) \in L^2(\Omega)$ for every $\boldsymbol{x}$ [41]. If $\boldsymbol{\Sigma}$ is invertible, the finite-dimensional density of order n of the Gaussian process is then

$$p(\boldsymbol{m}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{m} - \boldsymbol{\mu})\right), \tag{2.30}$$

where $\boldsymbol{m} = (m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_n))$. If we further restrict C to depend only on the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, then the stationary GP is called *isotropic* [44].

2.2.1 Karhunen-Loève expansion

Let $M(\boldsymbol{x}, \omega)$ be a real-valued random field with finite second moments, mean $\mu(\boldsymbol{x})$, and a covariance function that is continuous on $D \times D$ with D bounded. Then M has

the following representation, termed a Karhunen-Loève (K-L) expansion [32]:

$$M(\boldsymbol{x},\omega) = \mu(\boldsymbol{x}) + \sum_{k=1}^{\infty} \sqrt{\lambda_k}\, c_k(\omega)\, \phi_k(\boldsymbol{x}). \tag{2.31}$$

The K-L expansion converges in the mean square sense for each $\boldsymbol{x} \in D$. If M is Gaussian and almost surely continuous, then convergence is uniform over D with probability one [1]. Here, λ_k and $\phi_k(\boldsymbol{x})$ are eigenvalues and eigenfunctions of the linear operator corresponding to the covariance kernel C:

$$\int_D C(\boldsymbol{x}_1,\boldsymbol{x}_2)\, \phi(\boldsymbol{x}_2)\, \mathrm{d}\boldsymbol{x}_2 = \lambda_k\, \phi_k(\boldsymbol{x}_1). \tag{2.32}$$

By the assumptions on M, the covariance kernel is symmetric and positive semidefinite, and thus by Mercer's theorem [16], we have

$$C(\boldsymbol{x}_1,\boldsymbol{x}_2) = \sum_{k=1}^{\infty} \lambda_k\, \phi_k(\boldsymbol{x}_1)\, \phi_k(\boldsymbol{x}_2), \tag{2.33}$$

where the eigenfunctions $\phi_k(\boldsymbol{x})$ are continuous and form a complete orthonormal system in $L^2(D)$. The random variables $c_k(\omega)$ are uncorrelated with zero mean and unit variance:

$$\mathbb{E}[c_k] = 0, \quad \mathbb{E}[c_j c_k] = \delta_{jk}. \tag{2.34}$$

These variables are in general non-Gaussian

$$c_k(\omega) = \frac{1}{\sqrt{\lambda_k}} \int_D (M(\boldsymbol{x},\omega) - \mu(\boldsymbol{x}))\, \phi_k(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{2.35}$$

but if M is also a Gaussian process, the c_k are Gaussian and independent, i.e., $c_k \sim \mathcal{N}(0,1)$. If $M(\,\cdot\,)$ is approximated by a K-term K-L expansion, then

$$M_K(\boldsymbol{x},\omega) = \mu(\boldsymbol{x}) + \sum_{k=1}^{K} \sqrt{\lambda_k}\, c_k(\omega)\, \phi_k(\boldsymbol{x}), \tag{2.36}$$

and the covariance function of M_K is simply

$$C_K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{k=1}^{K} \lambda_k \, \phi_k(\boldsymbol{x}_1) \, \phi_k(\boldsymbol{x}_2), \tag{2.37}$$

which converges uniformly to Equation (2.33) as $K \to \infty$. In particular, the total variance or "energy" of M_K is

$$\int_D \mathbb{E}[M_k(\boldsymbol{x}, \omega) - \mu(\boldsymbol{x})]^2 \, \mathrm{d}\boldsymbol{x} = \int_D C_K(\boldsymbol{x}, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \sum_{k=1}^{K} \lambda_k \tag{2.38}$$

following from the orthonormality of the $\{\phi_k(\boldsymbol{x})\}_{k=1}^{K}$. The truncation error is the sum of the truncated eigenvalues [13]:

$$\mathbb{E}\big[\|M - M^K\|_{L_2(D)}\big] = \sum_{k=K+1}^{\infty} \lambda_k.$$

We can approximate the eigenfunctions and corresponding eigenvectors $\{\phi_k\}_{k=1}^{K}$ and $\{\lambda_k\}_{k=1}^{K}$ by using a quadrature rule to replace the integral with a weighted sum:

$$\sum_{i=1}^{M} C(\boldsymbol{x}_i, \boldsymbol{x}) \, \phi(\boldsymbol{x}_i) w_i = \tilde{\lambda}_k \phi_k(\boldsymbol{x}), \quad \tilde{\lambda}_k \approx \lambda_k.$$

Using the same points, we can write

$$\sum_{i=1}^{M} C(\boldsymbol{x}_i, \boldsymbol{x}_j) \, \phi_k(\boldsymbol{x}_i) \, w_i = \tilde{\lambda}_k \phi_k(\boldsymbol{x}_j), \quad \forall \, j = 1, \ldots, M.$$

Now, we can define

$$\mathbf{C} = \begin{bmatrix} C(\boldsymbol{x}_1, \boldsymbol{x}_1) & \ldots & C(\boldsymbol{x}_1, \boldsymbol{x}_M) \\ \vdots & \ddots & \vdots \\ C(\boldsymbol{x}_M, \boldsymbol{x}_1) & \ldots & C(\boldsymbol{x}_M, \boldsymbol{x}_M) \end{bmatrix}, \quad \boldsymbol{\phi}_k = \begin{bmatrix} \phi_i(\boldsymbol{x}_1) \\ \vdots \\ \phi_i(\boldsymbol{x}_M) \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_M \end{bmatrix}$$

so we can solve the following eigenvalue problem for $(\tilde{\lambda}_k, \boldsymbol{\phi}_k)$ pairs,

$$\mathbf{CW}\boldsymbol{\phi}_k = \tilde{\lambda}_k \boldsymbol{\phi}_k.$$

2.2.2 Dimensionality reduction in inference problems

Consider an inference problem in which the unknown quantities comprise a real-valued field $M(\boldsymbol{x})$. As we will show in the next section, model discrepancy is often formulated as an additive random field, which must be inferred. In a computational setting, this field and the forward model must be discretized. If $M(\boldsymbol{x})$ can be adequately represented on a finite collection of points $\{\boldsymbol{x}_i\}_{i=1}^n \in D$, then we can write both the prior and the posterior densities in terms of $\boldsymbol{m}(M(\boldsymbol{x}_1), \ldots, M(\boldsymbol{x}_n))$, i.e., we can apply the Bayesian formulation from (2.1.1) and explore the posterior density of *vectorm* with Markov chain Monte Carlo (MCMC) [14]. The vector $\boldsymbol{m}$, however, will likely be high-dimensional (taking values in $\mathbb{R}^D$, the space of all real-valued functions on D). High dimensionality causes exploration of the posterior using MCMC to be more challenging, and therefore costly in terms of the number of forward model evaluations.

Instead of exploring the value of $M(\boldsymbol{x})$ on each of n index points, we can use the truncated K-L expansion to reduce the dimension of the posterior that needs to be explored [35]. Let $M(\boldsymbol{x}) \sim \mathrm{GP}(\mu, C)$ be endowed with a Gaussian process prior with mean $\mu(\boldsymbol{x})$ and covariance kernel $C(\boldsymbol{x}_1, \boldsymbol{x}_2)$. The corresponding K-term K-L representation of $M(\boldsymbol{x})$ is

$$M_K(\boldsymbol{x}, \omega) = \mu(\boldsymbol{x}) + \sum_{k=1}^{K} \sqrt{\lambda_k}\, c_k(\omega)\, \phi_k(\boldsymbol{x}), \tag{2.39}$$

with eigenvalues λ_k and eigenfunctions $\phi_k(\boldsymbol{x})$ satisfying (2.32). Realizations $M(\boldsymbol{x}, \omega)$ can be uniformly approximated by $M_K(\boldsymbol{x}, \omega)$, which implies that corresponding realizations $\boldsymbol{c}(\omega) \equiv (c_1(\omega), \ldots, c_K(\omega))$ can also be approximated with probability one. Thus, updating distributions of M, by conditioning on the data, is equivalent to

updating the joint distribution of the mode strengths c_k, i.e.,

$$M_K(\boldsymbol{x}, \omega) = M_K(\boldsymbol{x}, c(\omega)) = M_K(\boldsymbol{c}), \tag{2.40}$$

parameterizing M by the vector of weights $\boldsymbol{c}$. Components c_k are independent under the Gaussian process prior, with $c_k \sim \mathcal{N}(0,1)$. We thus truncate the KL expansion at K terms and write a posterior density for $\boldsymbol{c}$:

$$p(\boldsymbol{c}|\boldsymbol{y}) \propto p(\boldsymbol{y}, \boldsymbol{c}) \prod_{k=1}^{K} p(c_k). \tag{2.41}$$

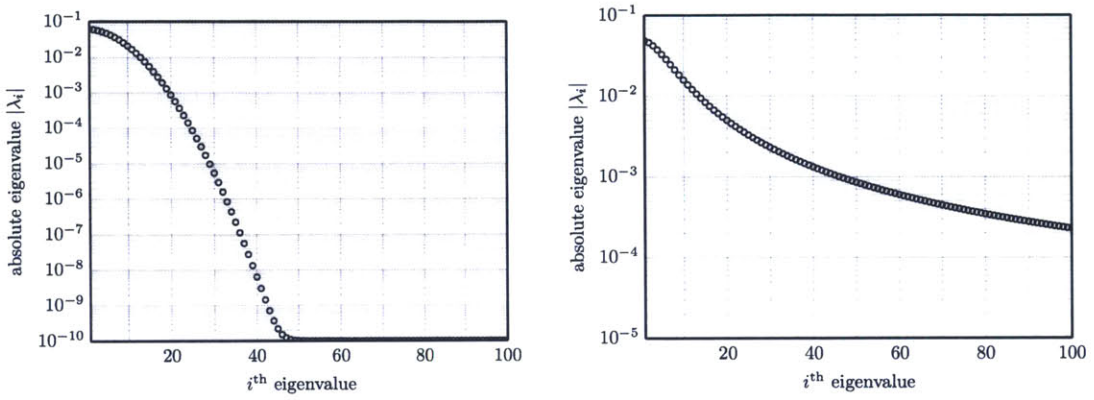
The inverse problem has been transformed to an inference problem on the weights c_k of a finite number of K-L modes. Note that the spatial discretization of $M(\boldsymbol{x})$ and of the forward model is now independent of the dimension of the posterior distribution. Here we have assumed the prior covariance to be completely known, thus ignoring hyperparameters in the expression for the posterior.

Truncating the K-L expansion in this context amounts to using a the truncated prior covariance kernel (2.37). Since the eigenvalues λ_k decay quickly (especially for square exponential kernels, see Figure 2-4a), a small number of terms may be sufficient to capture almost all of the prior covariance. The linear operator corresponding to the modified covariance kernel now has finite rank; $\phi_k(\boldsymbol{x})$ that are not eigenfunctions of this operator cannot contribute to the inverse solution [35].

Below in Figures 2.4–2.6 are plotted realizations and eigenvalue convergence of a mean-zero, stationary, isotropic Gaussian process with a covariance function

$$C(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sigma_C^2 \exp\left[\frac{1}{p}\left(\frac{|\boldsymbol{x}_1 - \boldsymbol{x}_2|}{L}\right)^p\right] \tag{2.42}$$

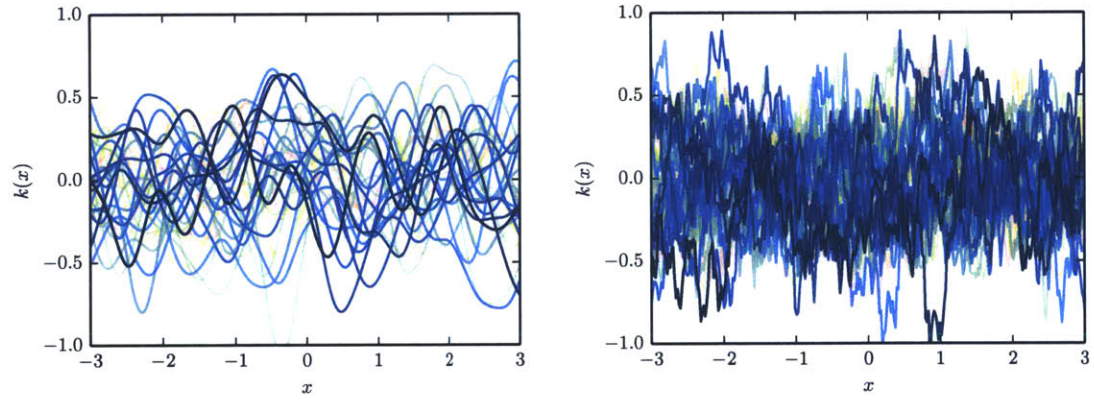
with total variance $\sigma_C^2 = 0.5$, and power $p = 2$ (square exponential) and $p = 1$ (exponential) kernels and correlation length $L = 0.3$.



(a) Square exponential kernel

(b) Exponential kernel

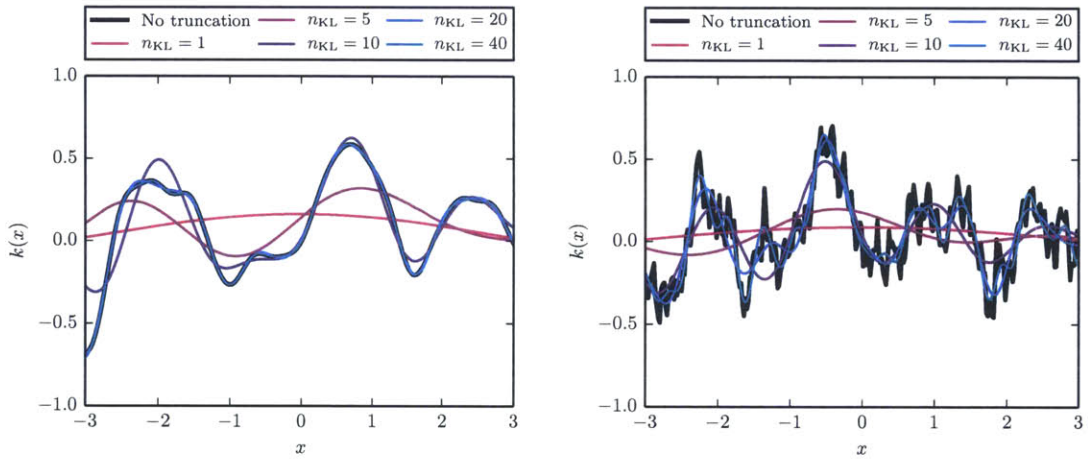
Figure 2-4: Eigenvalues of the discrete covariance matrix of a Gaussian process with covariance function given by Equation (2.42).



(a) Square exponential kernel

(b) Exponential kernel

Figure 2-5: Samples from a Gaussian process with covariance function given by Equation (2.42).



(a) Square exponential kernel

(b) Exponential kernel

Figure 2-6: K-L expansion of a Gaussian process with covariance function given by Equation (2.42).

2.3 Model Discrepancy

The method of Brynjarsdóttir and O'Hagan

Following [24, 5], for a general experiment the observation equation (2.17) is augmented with an additive discrepancy term $\delta(\cdot)$, which is function of the control parameters. In the experimental design context, the most relevant parameterization is $\delta(\boldsymbol{d})$, i.e., the discrepancy is defined as a function of the design variables.

$$\boldsymbol{y} = G(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{d}) + \delta(\boldsymbol{d}) + \boldsymbol{\epsilon} \tag{2.43}$$

Following Kennedy and O'Hagan ([24]), we represent the model discrepancy term $\delta(\boldsymbol{d})$ as a zero-mean Gaussian process (GP):

$$\delta(\boldsymbol{d}) \sim \mathrm{GP}(0, \sigma_C^2(\cdot\,|L)). \tag{2.44}$$

It is important to understand how such a representation might formulate prior knowledge about the discrepancy function δ. For example, consider the square exponential covariance function defined by Equation (2.42) with $p = 2$. Then, Equation (2.44) says that at any point $\boldsymbol{d}$, the prior probability distribution of $\delta(\boldsymbol{d})$ is normal with mean zero and variance σ_C^2. The zero mean implies that we do not have a prior expectation that $\delta(\boldsymbol{d})$ is more likely to be positive or more likely to be negative. The variance σ_C^2 expresses a prior belief that $\delta(\boldsymbol{d})$ is not likely to be outside the range $\pm 2\sigma_C$, so it measures the strength of prior information about $\delta(\boldsymbol{d})$. The fact that the variance is the same for all $\boldsymbol{d}$ implies that we do not have a prior expectation that $|\delta(\boldsymbol{d})|$ is likely to take larger values for some $\boldsymbol{d}$ than for others. The correlation function also expresses a prior belief that $\delta(d)$ will be a smooth function.

3

Numerical Methods

In §3.1, we formulate a numerical approximation of the expected utility (2.11) using Monte Carlo sampling. Then, in §3.2, we discuss stochastic optimization methods which are used for numerical optimization when the objective function is noisy. Finally, in §3.3, we introduce Markov chain Monte Carlo (MCMC) methods for Bayesian parameter estimation, which is used to evaluate the Bayes risk. The numerical tools developed in this chapter will allow us to bridge the gap between the analytical formulation in Chapter 2 to more realistic, nonlinear problems which do not admit a closed-form representation of the expected utility, such as those in Chapter 5.

3.1 Numerical evaluation of the expected utility

3.1.1 Monte Carlo estimation of the expected utility

The expected utility rarely has a closed-form representation, so we need to develop a numerical approximation for the expression in Equation (2.11). As written, Equation (2.11) contains nested posterior integrals, which can be difficult to evaluate in practice.

Using Bayes' rule, we can rewrite Equation (2.11) as

$$\begin{aligned} U(\boldsymbol{d}) &= \int_{\boldsymbol{Y}} \int_{\boldsymbol{\Theta}} p_\theta(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d}) \ln\left[\frac{p_\theta(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})}{p_\theta(\boldsymbol{\theta})}\right] \mathrm{d}\boldsymbol{\theta}\; p(\boldsymbol{y}|\boldsymbol{d})\,\mathrm{d}\boldsymbol{y} \\ &= \int_{\boldsymbol{Y}} \int_{\boldsymbol{\Theta}} \ln\left[\frac{p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})}{p(\boldsymbol{y}|\boldsymbol{d})}\right] p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})\, p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{y} \\ &= \int_{\boldsymbol{Y}} \int_{\boldsymbol{\Theta}} \Big\{ \ln\big[p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})\big] - \ln\big[p(\boldsymbol{y}|\boldsymbol{d})\big] \Big\}\, p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})\, p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{y}, \end{aligned} \tag{3.1}$$

where $p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})$ and $p(\boldsymbol{y}|\boldsymbol{d})$ are called the *marginal likelihood* and *evidence*, respectively, where the implicit marginalization is over the nuisance parameter $\boldsymbol{\eta}$, i.e., the marginal likelihood is

$$p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d}) = \int_{\mathbf{H}} p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{d})\, p(\boldsymbol{\eta}|\boldsymbol{\theta},\boldsymbol{d})\,\mathrm{d}\boldsymbol{\eta} \tag{3.2}$$

and the marginal evidence is

$$p(\boldsymbol{y}|\boldsymbol{d}) = \int_{\boldsymbol{\Theta}} \int_{\mathbf{H}} p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{d})\, p(\boldsymbol{\eta}|\boldsymbol{\theta},\boldsymbol{d})\,\mathrm{d}\boldsymbol{\eta}\, p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\,. \tag{3.3}$$

Then, Monte Carlo sampling is used to estimate the integral in Equation (3.1),

$$U(\boldsymbol{d}) \approx \frac{1}{N} \sum_{i=1}^{N} \Big\{ \ln\big[p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)},\boldsymbol{d})\big] - \ln\big[p(\boldsymbol{y}^{(i)}|\boldsymbol{d})\big] \Big\}, \tag{3.4}$$

where $\boldsymbol{\theta}^{(i)}$ and $\boldsymbol{\eta}^{(i)}$ are drawn from the joint prior $p(\boldsymbol{\theta},\boldsymbol{\eta})$; $\boldsymbol{y}^{(i)}$ are drawn from the conditional distributions $p(\boldsymbol{y}|\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)},\boldsymbol{\eta}=\boldsymbol{\eta}^{(i)},\boldsymbol{d})$; and N is the number of samples in the Monte Carlo estimate. The marginal likelihood and evidence evaluated at $\boldsymbol{y}^{(i)}$ typically do not have an analytical form. However, we can approximate them using importance sampling estimates. For the marginal likelihood, we can write

$$\begin{aligned} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)},\boldsymbol{d}) &= \int_{\mathbf{H}} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)},\boldsymbol{\eta},\boldsymbol{d})\, p(\boldsymbol{\eta}|\boldsymbol{\theta}^{(i)},\boldsymbol{d})\,\mathrm{d}\boldsymbol{\eta} \\ &\approx \frac{1}{M_1} \sum_{j=1}^{M_1} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)},\boldsymbol{\eta}^{(i,j)},\boldsymbol{d}) \end{aligned} \tag{3.5}$$

where the $\boldsymbol{\eta}^{(i,j)}$ are drawn from the conditional distribution $p(\boldsymbol{\eta}|\boldsymbol{\theta}^{(i)})$, and M_1 is the number of samples in the inner Monte Carlo estimator. Similarly, the evidence evaluated at $\boldsymbol{y}^{(i)}$ can be approximated by the following importance sampling estimate,

$$\begin{aligned} p(\boldsymbol{y}^{(i)}|\boldsymbol{d}) &= \int_{\boldsymbol{\Theta}} \int_{\mathbf{H}} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{d})\, p(\boldsymbol{\eta}|\boldsymbol{\theta}, \boldsymbol{d})\, \mathrm{d}\boldsymbol{\eta}\; p(\boldsymbol{\theta}|\boldsymbol{d})\, \mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{M_2} \sum_{k=1}^{M_2} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i,k)}, \boldsymbol{\eta}^{(i,k)}, \boldsymbol{d}) \end{aligned} \tag{3.6}$$

where the $\boldsymbol{\theta}^{(i,k)}$ and $\boldsymbol{\eta}^{(i,k)}$ are drawn from the joint prior $p(\boldsymbol{\theta}, \boldsymbol{\eta})$, and M_2 is the number of samples in this second inner Monte Carlo estimator. Substituting Equations (3.5) and (3.6) into Equation (3.4) yields the following Monte Carlo estimator for the expected utility:

$$\hat{U}(\boldsymbol{d}) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \ln \left[\frac{1}{M_1} \sum_{j=1}^{M_1} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)}, \boldsymbol{\eta}^{(i,j)}, \boldsymbol{d}) \right] - \ln \left[\frac{1}{M_2} \sum_{k=1}^{M_2} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i,k)}, \boldsymbol{\eta}^{(i,k)}, \boldsymbol{d}) \right] \right\}. \tag{3.7}$$

The variance and bias of this type estimator is discussed in [42] and satisfy

$$\mathrm{Var}[\hat{U}] \propto \frac{A(\boldsymbol{d})}{N} + \frac{B(\boldsymbol{d})}{NM_1} + \frac{C(\boldsymbol{d})}{NM_2}, \tag{3.8}$$

$$\mathrm{Bias}[\hat{U}] \propto \frac{D(\boldsymbol{d})}{M_1} + \frac{E(\boldsymbol{d})}{M_2}. \tag{3.9}$$

Since the variance and bias asymptotically approach zero with an increasing number of samples, we can say that this estimator the expected utility $\hat{U}(\boldsymbol{d})$ is *consistent*. Furthermore, the variance is dominated by the number of samples in the outer Monte Carlo estimator, and the bias is dominated by the number of samples in the inner Monte Carlo estimator.

3.1.2 Adaptive importance sampling

In the previous section, we naïvely used the prior distributions on $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ as our biasing distributions for the importance sampling estimates of the marginal likelihood

$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{d})$ and the evidence $p(\boldsymbol{y}|\boldsymbol{d})$. In general, the marginal likelihood and evidence are large in regions where the posterior density is large. Since the posterior is often more concentrated than the prior—especially when the data is informative—using the prior as a biasing distribution will result in low sampling efficiency since most samples from the prior distribution will lie in regions where the posterior density is small. It is well known that the statistical error of an importance sampling estimator reduces to zero if the density of the biasing distribution is chosen to be the target density on the support of the target distribution. In practice, however, this is an impossible choice, since the posterior density can only be evaluated pointwise up to a constant factor, so sampling from the posterior distribution is very difficult. However, if we can sample from an accurate approximation of the posterior distribution and construct the approximation in an efficient manner, i.e. without additional forward model evaluations[1], we can enjoy the variance reduction achieved by importance sampling with a good biasing distribution.

To this end, we wish to use the samples and forward model evaluations from the outer Monte Carlo estimator to approximate of each of the N posteriors biasing distributions, which we will use to construct biasing distributions for each of the N inner importance sampling estimates of the marginal likelihood and evidence. Since we wish to avoid additional evaluations of the forward model when "adapting" the biasing distribution, we cannot use adaptive importance sampling methods which iteratively update the biasing distribution and incur more than one model evaluation for effective sample to remain unbiased [6, 39]. Instead, our proposed method adaptively constructs the biasing distribution *once* for each inner importance sampling estimate so that the inner importance sampling estimators remain unbiased. In other words, even if the estimators for the posterior mean and covariance are biased—which are used to construct the biasing distributions for the inner importance sampling estimators—the inner importance sampling estimators will remain unbiased since they are using fixed biasing distributions.

[1]The efficiency in terms of model evaluations per effective sample is very important, since the process of constructing the approximate posterior and sampling from the "adaptive" biasing distribution occurs N times.

To find the approximate posterior distribution, we will use self-normalized importance sampling estimates of the posterior mean and covariance. For this discussion, we use the following observational model where forward model evaluations $G(\cdot)$ are corrupted by additive[2] noise $\boldsymbol{\epsilon}$,

$$\boldsymbol{y}^{(i)}|\boldsymbol{d} = G(\boldsymbol{\theta}^{(i)}, \boldsymbol{\eta}^{(i)}, \boldsymbol{d}) + \boldsymbol{\epsilon}, \tag{3.10}$$

and assume that the likelihood function $p(\boldsymbol{y}^{(j)}|\boldsymbol{\theta}^{(i)}, \boldsymbol{\eta}^{(i)}, \boldsymbol{d})$ can be computed cheaply for each $\boldsymbol{y}^{(j)} \sim p(\boldsymbol{y}|\boldsymbol{\theta}^{(j)}, \boldsymbol{\eta}^{(j)}, \boldsymbol{d})$ since the forward model evaluated at each of the outer N samples of $(\boldsymbol{\theta}^{(i)}, \boldsymbol{\eta}^{(i)})$ and $(\boldsymbol{\theta}^{(j)}, \boldsymbol{\eta}^{(j)})$ has been precomputed in the outer Monte Carlo estimator, so evaluating the likelihood function amounts to computing the density function of the noise, e.g., a squared exponential in the case of additive Gaussian noise.

Let $\boldsymbol{z} \equiv \left(\boldsymbol{\theta}, \boldsymbol{\eta}\right)^T$ so that the joint posterior can be written as

$$p(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{y}^{(i)}, \boldsymbol{d}) \Leftrightarrow p(\boldsymbol{z}|\boldsymbol{y}^{(i)}, \boldsymbol{d}).$$

The posterior mean $\boldsymbol{\mu}_{\text{post}}^{(i)} \equiv \mathbb{E}[p(\boldsymbol{z}|\boldsymbol{y}^{(i)}, \boldsymbol{d})]$ can be estimated using a self-normalized importance sampling estimator with the prior $p(\boldsymbol{z})$ as the biasing distribution. The prior density appears in the expectation when we apply Bayes' rule to the posterior:

$$\begin{aligned}\boldsymbol{\mu}_{\text{post}}^{(i)} &= \int_{\boldsymbol{Z}} \boldsymbol{z}\, p(\boldsymbol{z}|\boldsymbol{y}^{(i)}, \boldsymbol{d})\, \mathrm{d}\boldsymbol{z} \\ &= \int_{\boldsymbol{Z}} \boldsymbol{z}\, \frac{p(\boldsymbol{y}^{(i)}|\boldsymbol{z}, \boldsymbol{d})}{p(\boldsymbol{y}^{(i)}, \boldsymbol{d})} p(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}\,. \end{aligned} \tag{3.11}$$

The normalizing constant $p(\boldsymbol{y}^{(i)}, \boldsymbol{d})$ can be approximated by the following importance

[2]While an additive Gaussian error model is used here, applications with more systematic error exist and will generally use non-additive and/or non-Gaussian error models.

sampling estimate,

$$p(\boldsymbol{y}^{(i)}) = \int_{\boldsymbol{Z}} p(\boldsymbol{y}^{(i)}|\boldsymbol{z},\boldsymbol{d})\, p(\boldsymbol{z})\ \mathrm{d}\boldsymbol{z}$$
$$\approx \sum_{\ell=1}^{N} p(\boldsymbol{y}^{(i)}|\boldsymbol{z}^{(\ell)},\boldsymbol{d}) \tag{3.12}$$

with $\boldsymbol{z}^{(\ell)}$ drawn from the prior $p(\boldsymbol{z})$. Substituting this approximate normalizing constant into Equation (3.11) yields the self-normalized importance sampling estimator of the posterior mean with importance weights $\omega^{(k)}$ defined below

$$\hat{\boldsymbol{\mu}}_{\text{post}}^{(i)} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{z}^{(k)} \underbrace{\frac{p(\boldsymbol{y}^{(i)}|\boldsymbol{z}^{(k)},\boldsymbol{d})}{\sum_{\ell=1}^{N} p(\boldsymbol{y}^{(i)}|\boldsymbol{z}^{(\ell)},\boldsymbol{d})}}_{\omega^{(k)}} \tag{3.13}$$

where $\boldsymbol{z}^{(k)}$ and $\boldsymbol{z}^{(\ell)}$ are drawn from the prior $p(\boldsymbol{z})$. The posterior covariance $\boldsymbol{\Sigma}_{\text{post}}^{(i)} \equiv \text{Cov}[p(\boldsymbol{z}|\boldsymbol{y}^{(i)})]$ can also be estimated using the same biasing distribution and importance weights:

$$\boldsymbol{\Sigma}_{\text{post}}^{(i)} = \int_{\boldsymbol{Z}} (\boldsymbol{z}-\boldsymbol{\mu}_{\text{post}}^{(i)})(\boldsymbol{z}-\boldsymbol{\mu}_{\text{post}}^{(i)})^T\, p(\boldsymbol{z}|\boldsymbol{y}^{(i)},\boldsymbol{d})\ \mathrm{d}\boldsymbol{z}$$
$$= \int_{\boldsymbol{Z}} (\boldsymbol{z}-\boldsymbol{\mu}_{\text{post}}^{(i)})(\boldsymbol{z}-\boldsymbol{\mu}_{\text{post}}^{(i)})^T \frac{p(\boldsymbol{y}^{(i)}|\boldsymbol{z},\boldsymbol{d})}{p(\boldsymbol{y}^{(i)},\boldsymbol{d})}\, p(\boldsymbol{z})\ \mathrm{d}\boldsymbol{z},$$

and substituting $\boldsymbol{\mu}_{\text{post}}^{(i)}$ with $\hat{\boldsymbol{\mu}}_{\text{post}}^{(i)}$ from Equation (3.11), we obtain

$$\hat{\boldsymbol{\Sigma}}_{\text{post}}^{(i)} = \sum_{k=1}^{N} (\boldsymbol{z}^{(k)} - \hat{\boldsymbol{\mu}}_{\text{post}}^{(i)})(\boldsymbol{z}-\hat{\boldsymbol{\mu}}_{\text{post}}^{(i)})^T \underbrace{\frac{p(\boldsymbol{y}^{(i)}|\boldsymbol{z}^{(k)},\boldsymbol{d})}{\sum_{\ell=1}^{N} p(\boldsymbol{y}^{(i)}|\boldsymbol{z}^{(\ell)},\boldsymbol{d})}}_{\omega^{(k)}} \tag{3.14}$$

Returning to the original variables $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, the self-normalized importance weights $\omega^{(k)}$ can be written as the normalized likelihood of the k^{th} sample from the outer Monte Carlo estimator:

$$\omega^{(k)} = \frac{p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(k)},\boldsymbol{\eta}^{(k)},\boldsymbol{d})}{\sum_{\ell=1}^{N} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(\ell)},\boldsymbol{\eta}^{(\ell)},\boldsymbol{d})}, \qquad k = 1,\ldots,N. \tag{3.15}$$

Notice that the self-normalized importance sampling estimators for the mean and covariance derived above are simply the weighted sample mean and the weighted sample covariance, where the normalized weights $\omega^{(k)}$ are the self-normalized likelihoods of each of the N samples from the outer Monte Carlo estimator, i.e.,

$$\hat{\boldsymbol{\mu}}_{\text{post}} = \sum_{k=1}^{N} \omega^{(k)} \boldsymbol{z}^{(k)}, \tag{3.16}$$

$$\hat{\boldsymbol{\Sigma}}_{\text{post}} = \sum_{k=1}^{N} \omega^{(k)} \left(\boldsymbol{z}^{(k)} - \hat{\boldsymbol{\mu}}_{\text{post}}\right)\left(\boldsymbol{z}^{(k)} - \hat{\boldsymbol{\mu}}_{\text{post}}\right)^{T}. \tag{3.17}$$

This is an intuitive result, since samples from the prior weighted by their likelihoods should approximate samples from the posterior distribution, and normalizing the likelihood weights ensures that the total "number" of samples is conserved.

An importance sampling scheme for the evidence

Now, let $q_{\text{evd}}(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{y}^{(i)})$ be a probability distribution that has mean and covariance $\hat{\boldsymbol{\mu}}_{\text{post}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{post}}$, with nonzero density on the support of the prior, e.g., a multivariate normal distribution[3] $\mathcal{N}(\hat{\boldsymbol{\mu}}_{\text{post}}, \hat{\boldsymbol{\Sigma}}_{\text{post}})$. Using $q_{\text{evd}}(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{y}^{(i)})$ as a biasing distribution with importance weights

$$w_{\text{evd}}^{(k)} = \frac{p(\boldsymbol{\theta}^{(k)}, \boldsymbol{\eta}^{(k)})}{q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\eta}^{(k)}|\boldsymbol{y}^{(i)})}, \tag{3.18}$$

[3]When the target distribution is a Gaussian distribution, a common tactic is to take the biasing distribution to be a student's t distribution (or multivariate equivalent). Such a biasing distribution has heavier tails than the target distribution. The reverse practice, of using Gaussian biasing distribution q for a student's t nominal distribution, can easily lead to infinite variance [40].

we can write the following importance estimator for the evidence $p(\boldsymbol{y}^{(i)}|\boldsymbol{d})$ from Equation (3.4):

$$\begin{aligned} p(\boldsymbol{y}^{(i)}|\boldsymbol{d}) &= \int_{\boldsymbol{\Theta}} \int_{\mathbf{H}} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{d})\, p(\boldsymbol{\theta}, \boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\eta}\, \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} \int_{\mathbf{H}} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{d}) \frac{p(\boldsymbol{\theta}, \boldsymbol{\eta})}{q_{\text{evd}}(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{y}^{(i)})}\, q_{\text{evd}}(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{y}^{(i)})\, \mathrm{d}\boldsymbol{\eta}\, \mathrm{d}\boldsymbol{\theta} \\ &\approx \frac{1}{M_2} \sum_{k=1}^{M_2} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(k)}, \boldsymbol{\eta}^{(k)}, \boldsymbol{d})\, w_{\text{evd}}^{(k)}, \end{aligned} \tag{3.19}$$

where $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\eta}^{(k)}$ are sampled from the biasing distribution $q_{\text{evd}}(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{y}^{(i)})$.

An importance sampling scheme for the marginal likelihood

For the marginal likelihood $p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)}, \boldsymbol{d})$, where the marginalization is over the nuisance parameters $\boldsymbol{\eta}$, we devise another importance sampling estimator. Although the marginal likelihood is equivalent to the posterior conditioned on $\boldsymbol{\theta} = \boldsymbol{\theta}^{(i)}$, we cannot efficiently obtain an approximation of this distribution by simply reweighting the prior samples since the conditional distribution lives on a lower dimensional subspace, or manifold, of the prior support, so the number of prior samples on lie on that subspace is generally zero or very small. Instead, we make the assumption that the posterior is *smooth*, so that the posterior density associated with samples that are *close* to the manifold can be informative towards the value of the posterior density on the manifold. Then we can use the conditional mean and covariance of the now smooth (by assumption) posterior distribution to construct a biasing distribution on the lower dimensional subspace.

For example, if we chose approximate the posterior using a multivariate normal distribution[4] to satisfy the smoothness condition. First, we partition the mean vector

[4]In practice, the family of distribution should match that of the biasing distribution used for the importance sampling estimator of the evidence. Here, we choose to use a multivariate normal distribution since evaluating the conditional is relatively straightforward. For other choices of distributions, see Footnote 3.

and covariance matrix as

$$q \sim \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \quad \text{with} \quad \boldsymbol{\mu}_q = \begin{bmatrix} \boldsymbol{\mu}_\theta \\ \boldsymbol{\mu}_\eta \end{bmatrix}, \quad \boldsymbol{\Sigma}_q = \begin{bmatrix} \boldsymbol{\Sigma}_\theta & \boldsymbol{\Sigma}_{\mathrm{CC}} \\ \boldsymbol{\Sigma}_{\mathrm{CC}}^T & \boldsymbol{\Sigma}_\eta \end{bmatrix} \tag{3.20}$$

where $\boldsymbol{\Sigma}_{\mathrm{CC}}$ is the (non-symmetric) cross-covariance matrix between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, which has n_θ rows and n_η columns. Then, the conditional distribution we wish to extract is given by

$$\boldsymbol{\eta}|\boldsymbol{y}^{(i)}, \boldsymbol{\theta}^{(i)} \sim \mathcal{N}\big(\boldsymbol{\mu}_\eta + \boldsymbol{\Sigma}_{\mathrm{CC}}^T \boldsymbol{\Sigma}_\theta^{-1}(\boldsymbol{\theta}^{(i)} - \boldsymbol{\mu}_\theta),\ \boldsymbol{\Sigma}_\eta - \boldsymbol{\Sigma}_{\mathrm{CC}}^T \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\Sigma}_{\mathrm{CC}}\big). \tag{3.21}$$

Thus, the biasing distribution $q_{\mathrm{ml}}(\boldsymbol{\eta}|\boldsymbol{y}^{(i)}, \boldsymbol{\theta}^{(i)})$ for the importance estimator of the marginal likelihood should have the following mean and covariance:

$$\boldsymbol{\mu}_{q_{\mathrm{ml}}} = \boldsymbol{\mu}_\eta + \boldsymbol{\Sigma}_{\mathrm{CC}}^T \boldsymbol{\Sigma}_\theta^{-1}(\boldsymbol{\theta}^{(i)} - \boldsymbol{\mu}_\theta), \tag{3.22}$$

$$\boldsymbol{\Sigma}_{q_{\mathrm{ml}}} = \boldsymbol{\Sigma}_\eta - \boldsymbol{\Sigma}_{\mathrm{CC}}^T \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\Sigma}_{\mathrm{CC}}. \tag{3.23}$$

Note that this result is only valid if we assume the approximate posterior can be represented by a smooth multivariate normal distribution. More generally—but still assuming that the posterior is smooth—consider the biasing distribution $q_{\mathrm{ml}}(\boldsymbol{\eta}|\boldsymbol{y}^{(i)}, \boldsymbol{\theta}^{(i)})$ as the $\boldsymbol{\theta} = \boldsymbol{\theta}^{(i)}$ conditional of the biasing distribution for the estimator of the evidence evidence $q_{\mathrm{evd}}(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{y}^{(i)}, \boldsymbol{d})$. The associated importance weights are

$$w_{\mathrm{ml}}^{(j)} = \frac{p(\boldsymbol{\eta}^{(j)}|\boldsymbol{\theta}^{(i)})}{q_{\mathrm{ml}}(\boldsymbol{\eta}^{(j)}|\boldsymbol{y}^{(i)}, \boldsymbol{\theta}^{(i)})}, \tag{3.24}$$

which lets us write the following importance sampling estimator:

$$\begin{aligned} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)}, \boldsymbol{d}) &= \int_{\mathbf{H}} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)}, \boldsymbol{\eta}, \boldsymbol{d})\, p(\boldsymbol{\eta}|\boldsymbol{\theta}^{(i)})\, \mathrm{d}\boldsymbol{\eta} \\ &= \int_{\mathbf{H}} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)}, \boldsymbol{\eta}, \boldsymbol{d}) \frac{p(\boldsymbol{\eta}|\boldsymbol{\theta}^{(i)})}{q_{\mathrm{ml}}(\boldsymbol{\eta}|\boldsymbol{y}^{(i)}, \boldsymbol{\theta}^{(i)})} q_{\mathrm{ml}}(\boldsymbol{\eta}|\boldsymbol{y}^{(i)}, \boldsymbol{\theta}^{(i)})\, \mathrm{d}\boldsymbol{\eta} \\ &\approx \frac{1}{M_1} \sum_{j=1}^{M_1} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)}, \boldsymbol{\eta}^{(j)}, \boldsymbol{d})\, w_{\mathrm{ml}}^{(j)}, \end{aligned} \tag{3.25}$$

where $\boldsymbol{\eta}^{(j)}$ are sampled from the biasing distribution $q_{\mathrm{ml}}(\boldsymbol{\eta}|\boldsymbol{y}^{(i)}, \boldsymbol{\theta}^{(i)})$. Now that we have improved importance estimators for the marginal likelihood (3.25) and the evidence (3.19), we can rewrite the Monte Carlo estimator for the expected utility in Equation (3.7) as

$$\begin{aligned} \hat{U}(\boldsymbol{d}) = \frac{1}{N} \sum_{i=1}^{N} \Bigg\{ & \ln\left[\frac{1}{M_1} \sum_{j=1}^{M_1} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)}, \boldsymbol{\eta}^{(j)}, \boldsymbol{d})\, w_{\mathrm{ml}}^{(j)}\right] \\ & - \ln\left[\frac{1}{M_2} \sum_{k=1}^{M_2} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i,k)}, \boldsymbol{\eta}^{(k)}, \boldsymbol{d})\, w_{\mathrm{evd}}^{(k)}\right] \Bigg\}. \end{aligned} \tag{3.26}$$

where $\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(i)}$ and $\boldsymbol{y}^{(i)}$ are still drawn from the prior and conditional likelihood, respectively, but now $\boldsymbol{\eta}^{(j)} \sim q_{\mathrm{ml}}(\boldsymbol{\eta}|\boldsymbol{y}^{(i)}, \boldsymbol{\theta}^{(i)})$ and $(\boldsymbol{\theta}^{(k)}, \boldsymbol{\eta}^{(k)}) \sim q_{\mathrm{evd}}(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{y}^{(i)})$ are drawn from their respective biasing distributions. Since our construction of the importance sampling estimators are themselves unbiased, the overall estimator remains asymptotically unbiased. However, the asymptotic behavior of the overall estimator bias is no longer fully characterized by Equation (3.9). The outer estimator's bias now has a stronger dependence on the number of outer Monte Carlo samples N, since N controls the quality of the approximate distributions used to construct the biasing distributions for the inner estimators. In addition, the variance introduced by the randomness in the parameterization of the biasing distribution used in the inner estimators also contributes to the overall variance of the estimator, i.e., for small N, we expect the variance of the overall estimator to be larger when using the adaptive importance sampling scheme. From Figures 3-2 and 3-1, especially in

3-1, we can see that estimating the off-diagonal terms of the covariance matrix (seen in the marginal slices) is very difficult, even at moderate dimensions. Additionally, the posterior contracts significantly, especially in the θ-marginal, which reduces the effective sample size. Clearly, choosing an appropriate cutoff for the effective sample size is problem-specific. In the next section, we will examine the effect of different cutoffs on the performance of the adaptive importance sampling scheme.

3.1.3 Robust adaptive importance sampling

It is easy to see that the adaptive importance sampling scheme is not very robust—especially for very small N. Consider the case where $N = 1$; attempting to estimate the posterior covariance using a single sample in Equation (3.17) will result in a singular covariance matrix. Even for non-trivial cases, it is not impossible to have little or no overlap between the prior samples and the posterior for a given sample. Consider the case where the prior is very diffuse and the likelihood is very concentrated. Now assume one of the prior samples lies in the tail of the prior distribution; it is unlikely that the other $N - 1$ prior samples lie in the neighborhood around this "outlier." The resulting effective sample size for estimating the posterior mean and covariance is very close to one. More formally, the effective sample size for self-normalized importance sampling estimators is given in [31] as

$$\text{ESS} = \frac{1}{\sum_{k=1}^{N} (\omega^{(k)})^2}, \tag{3.27}$$

which can be used to evaluate the quality of an importance sampling estimate, i.e., the effective sample size ranges from 1 (poor) to N (good). A straightforward method to increase the robustness of the adaptive importance sampling estimator is to introduce a minimum effect sample size, which can be used as a cutoff for reverting to using the original naïve approach of sampling from the prior. An illustration of the adaptive importance sampling failing can be seen in Figure 3-1, where $N = 100$. In fact, due to the extreme anisotropy of the joint biasing distribution, the conditional biasing distribution has a covariance matrix that is is no longer positive definite due to

numerical error when applying Equation (3.23) to find the conditional distribution. Applying the adaptive importance sampling to the same problem, but with $N = 1000$ samples, we observe much better results in Figure 3-2 where the density contours of the biasing distribution (in red) closely match those of the posterior (in blue). To better

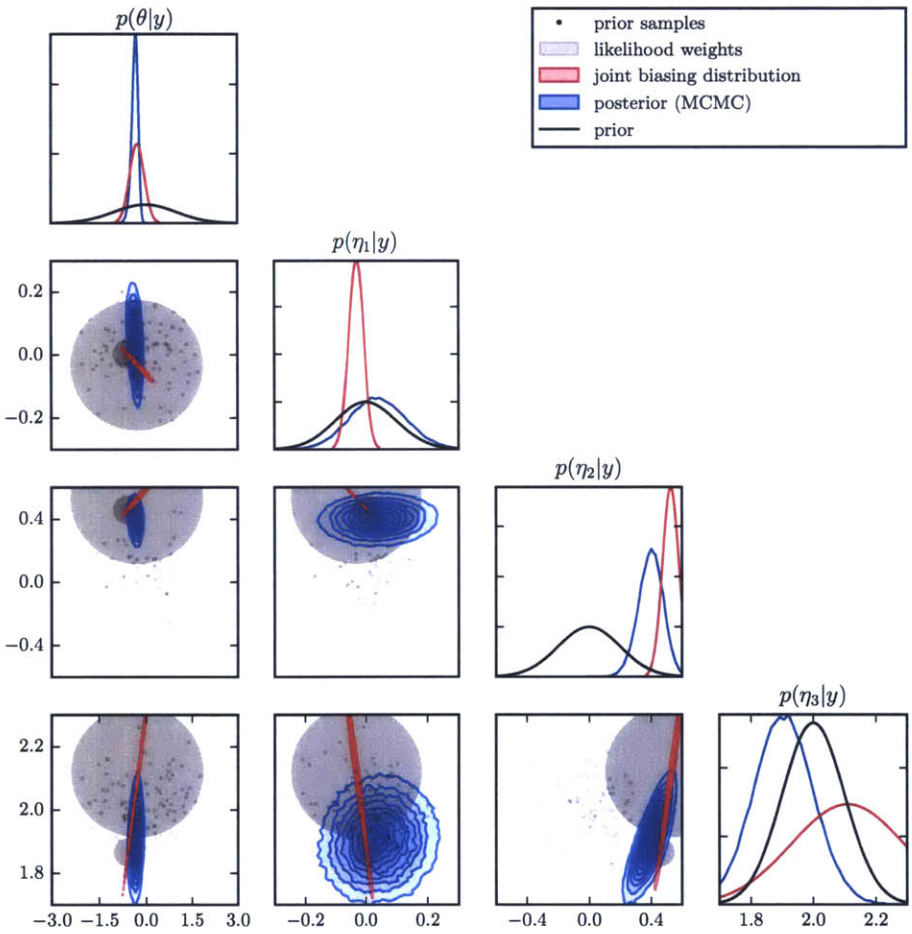


Figure 3-1: A small number of prior samples ($N = 100$) results in very few effective samples and poor estimation of the posterior mean and covariance. See §4.2 for the problem formulation.

understand how to choose a minimum effective sample size, it is important to examine the distribution of effective sample size for different numbers of outer Monte Carlo samples N. To this end, we computed histograms of the ESS for 10^5 realizations of N

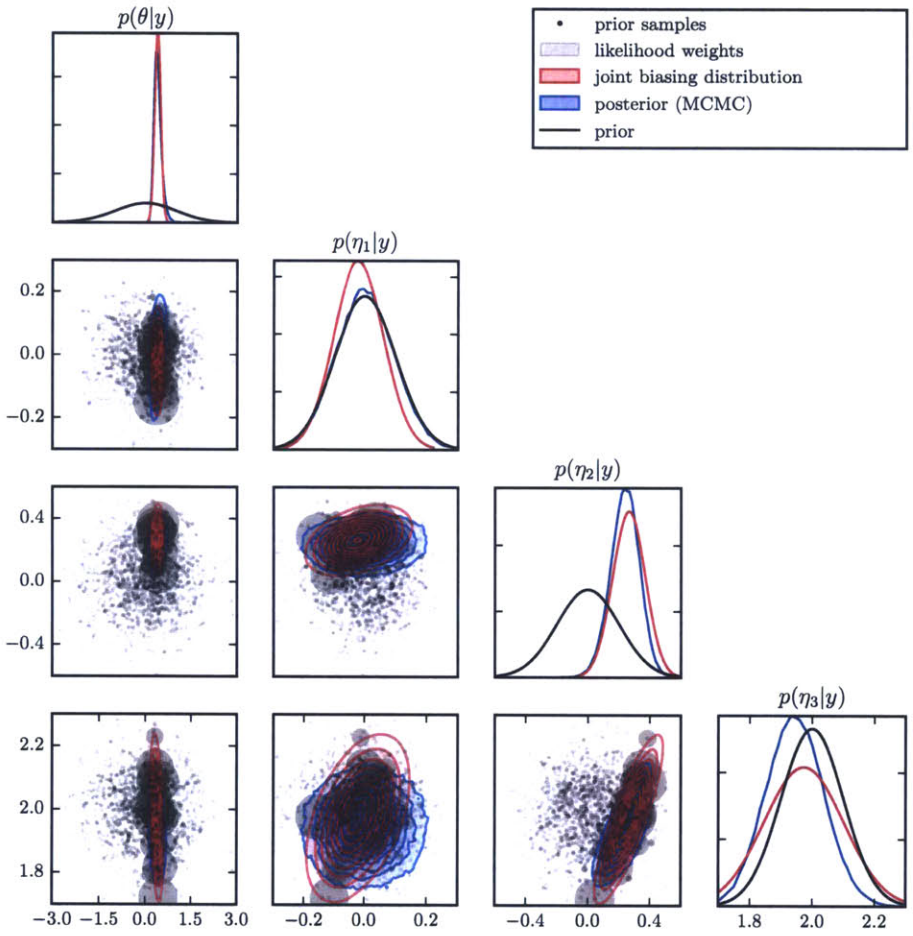
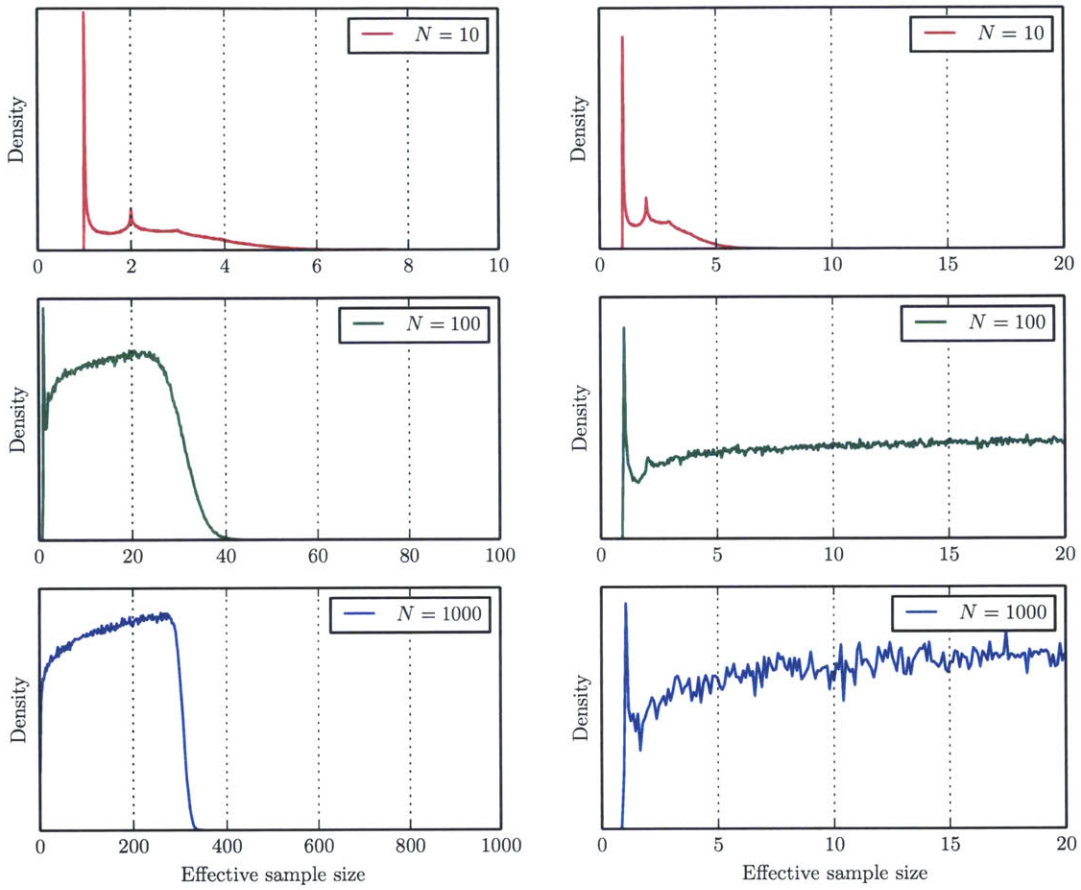


Figure 3-2: With a large number of prior samples ($N = 1000$), the biasing distribution is very close to the posterior (evaluated using MCMC). See §4.2 for the problem formulation.

samples for $N = 10, 100, 1000$, which are plotted in Figure A-1. We also computed



(a) Horizontal axis is range of ESS.

(b) Shared horizontal axis.

Figure 3-3: Distribution of effective sample size of the linear Gaussian example for $N = 10, 100, 1000$.

the failure (estimated covariance matrix is singular) and rejection (ESS below cutoff) percentages for different cutoffs of the effective sample size as a function of N, which are summarized in Table 3.1. As expected, we observe that as we increase the minimum cutoff, the number of failed adaptive sampling iterations decrease. However, increasing the cutoff negates the benefits of the adaptive importance sampling, since every rejection results in sampling from the prior. Ideally, we want to choose a cutoff that is just enough to reduce the failure percentage to nearly zero, so that we can maximize the effectiveness of the adaptive importance sampling scheme. Since the percentage of failures decreases with increasing N, we propose using a minimum cutoff that also

decreases as N increases. A strategy for choosing the minimum cutoff based on N that has performed well in numerical tests is:

$$\text{ESS} \geq 1 + \frac{100}{N}. \tag{3.28}$$

The form of this strategy is motivated by the rapid decrease in failure rate with increasing N, and the fact that the effective sample size never drops below one, so the minimum cutoff should asymptotically approach one. At the same time, we observe that

N	No cutoff		ESS ≥ 2		ESS ≥ 5		ESS ≥ 20		ESS $\geq 1 + \frac{100}{N}$	
	%cut	%fail	%cut	%fail	%cut	%fail	%cut	%fail	%cut	%fail
20	0.0	27.4	55.5	3.6	91.0	0.0	91.1	0.0	92.0	0.0
40	0.0	6.8	26.5	0.6	88.5	0.0	95.4	0.0	67.2	0.0
80	0.0	2.4	9.6	0.2	54.0	0.0	98.0	0.0	12.2	0.1
160	0.0	1.0	3.7	0.1	17.8	0.0	98.4	0.0	2.6	0.2
320	0.0	0.5	1.7	0.0	6.7	0.0	67.6	0.0	0.8	0.1
640	0.0	0.2	0.8	0.0	3.0	0.0	20.3	0.0	0.3	0.1
1280	0.0	0.1	0.4	0.0	1.4	0.0	7.8	0.0	0.1	0.0
2560	0.0	0.0	0.2	0.0	0.7	0.0	3.5	0.0	0.0	0.0

Table 3.1: Failure and rejection percentages for different cutoffs of the effective sample size as a function of the number of outer Monte Carlo samples N. Reported values are averages over 10^4 replicates of the expected utility estimator for the Mössbauer experimental design problem. See §4.2 for the problem formulation.

For an integrated measure of the quality of the Monte Carlo schemes for the expected utility $\hat{U}(\boldsymbol{d})$, we compute T replicates for each estimator of the expected utility, and compare the estimated bias and variance of $\hat{U}(\boldsymbol{d})$ across the T replicates with the exact expected utility $U_{\text{exact}}(\boldsymbol{d})$ computed from Equation (2.23). To reduce the variance across different importance sampling schemes, we use the same random numbers for the t^{th} replicate when evaluating $\hat{U}(\boldsymbol{d})$, i.e., for each of the $t = 1, 2, \ldots, T$ replicates, a new set of N samples and noises $\boldsymbol{\epsilon}$ are drawn from their respective priors, which are then used in each of evaluation of $\hat{U}(\boldsymbol{d})$, e.g., with and without the adaptive importance sampling scheme, with different cutoffs for the minimum effective sample size, etc. The mean $\hat{m}_T$ and variance $\hat{s}_T^2$ estimators of the expected utility estimator

$\hat{U}(\boldsymbol{d})$ are

$$\hat{m}_T = \frac{1}{T}\sum_{t=1}^{T}\hat{U}_t(\boldsymbol{d}), \qquad \hat{s}_T^2 = \frac{1}{T-1}\sum_{t=1}^{T}\left(\hat{U}_t(\boldsymbol{d}) - \hat{m}_T\right)^2$$

Using the standard errors of mean and variance estimators outlined in [2], the estimator for the bias and and the corresponding standard error is

$$\text{Bias}[\hat{U}(\boldsymbol{d})] \approx \hat{m}_T - U_{\text{exact}}(\boldsymbol{d}), \quad \sigma_{\text{Bias}} = \frac{\hat{s}_T}{T}. \tag{3.29}$$

Likewise, the standard error of the estimator variance is

$$\sigma_{\text{Var}} = \frac{\sqrt{2}}{T-1}\hat{s}_T^2. \tag{3.30}$$

First, we would like to verify that the asymptotic behavior of the bias and variance

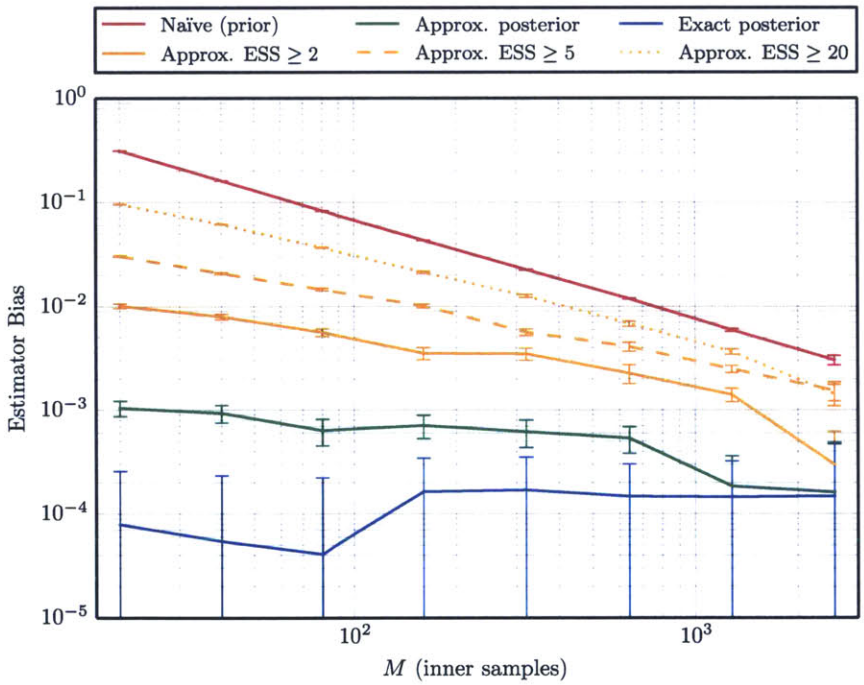


Figure 3-4: Bias of expected utility estimator $\hat{U}$ for the linear Gaussian example as a function of M, the number of inner samples, with a fixed number of outer samples $N = 1280$.

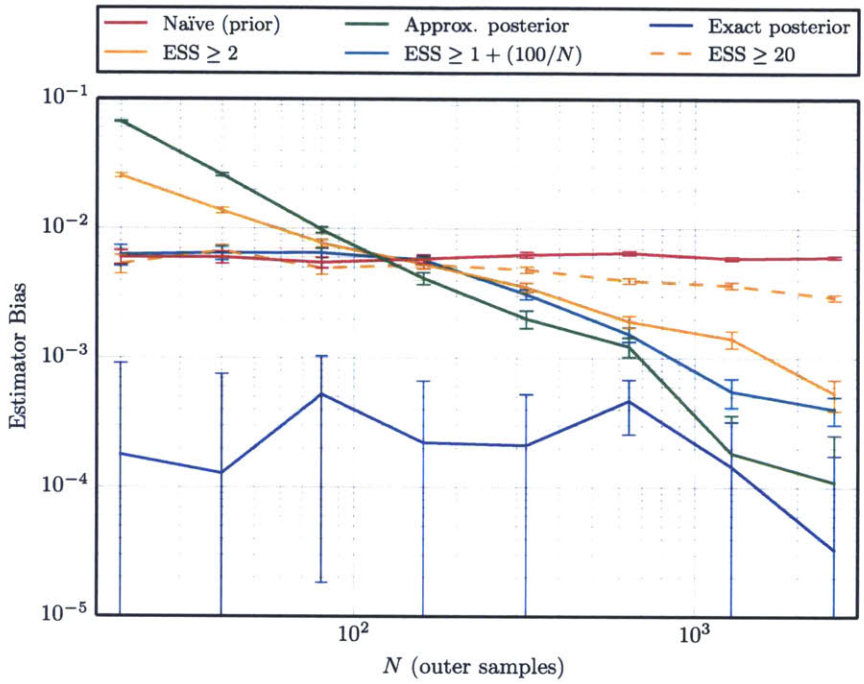


Figure 3-5: Bias of expected utility estimator $\hat{U}$ for the linear Gaussian example as a function of N, the number of outer samples, with a fixed number of inner samples $M = 1280$.

in Equations (3.9, 3.8) hold for the case where we naïvely use the prior as the biasing distribution. In Figure 3-4, we see perfect $1/M$ scaling of the estimator bias, which agrees with Equation (3.9). In Figure 3-5, the estimator bias does not seem to be noticeably affected by increasing the number of outer Monte Carlo samples. This is expected, since the bias is dominated by the first $1/M$ term in Equation (3.9). Then, the variance of $\hat{U}$ is also shown in Figure 3-8 to decrease exactly as $1/N$, which also agrees with Equation (3.8). Finally, the $1/(NM)$ terms in Equation (3.8) contribute to the variance Figure 3-7, but decrease more slowly than $1/M$.

Now, we will compare the convergence behavior of the adaptive importance sampling method with different minimum effective sample size cutoffs. First, we will examine the case where there is no cutoff, whose trends are labeled as 'Approx. posterior' and colored in green. Since we have the analytic form of the posterior distribution for the 2D linear Gaussian example, we can also evaluate the expected utility using the analytic posterior as the biasing distributions for the internal importance sampling

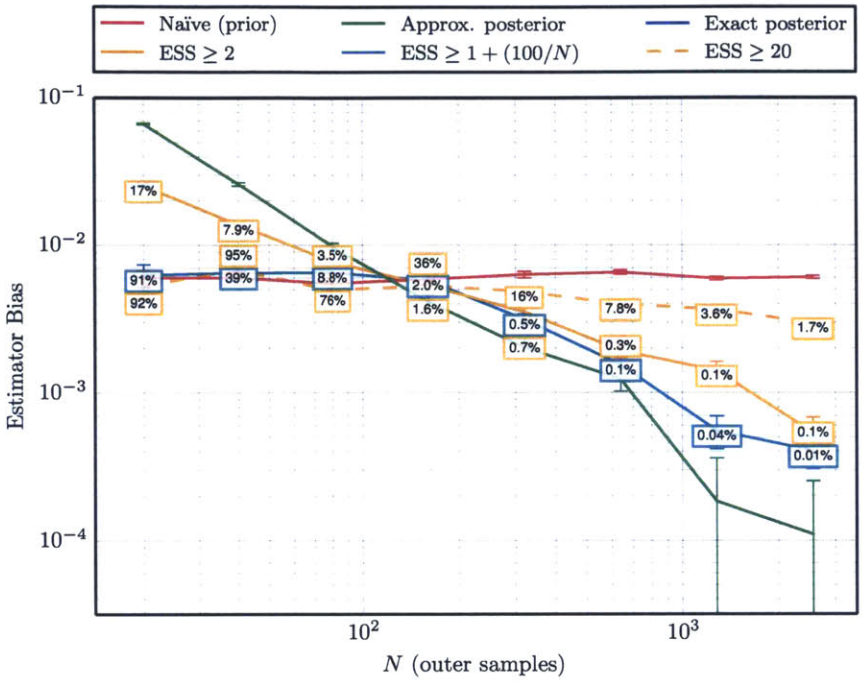


Figure 3-6: Bias of expected utility estimator $\hat{U}$ for the linear Gaussian example as a function of N, the number of outer samples, with a fixed number of inner samples $M = 1280$. The percentage of iterations where adaptive importance sampling was rejected by the minimum effective sample size cutoff is overlaid.

estimators. This effectively establishes a lower bound on the bias and variance, which we will compare our adaptive importance sampling scheme against.

The most important trend to notice is the dramatic decrease in estimator bias that is almost independent of the number of inner Monte Carlo samples, which can be observed in Figure 3-4. Even with $M \sim 10^1$, the estimator bias is reduced by several orders-of-magnitude, even against 10 to 100 times more inner samples using samples from the prior. This is potentially the two order-of-magnitude gain in computational efficiency that is needed to "full blown search" feasible, according to [42]. However, these gains are only observed for large numbers of outer Monte Carlo samples N. Looking at Figure 3-5, we see that there is a crossover point around $N = 100$, such that when $N < 100$, the bias of $\hat{U}$ evaluated using adaptive importance sampling is greater than the bias from the naïve sampler. Though, to appreciate the reduction in estimator bias, one needs an estimator with low variance, and to achieve low

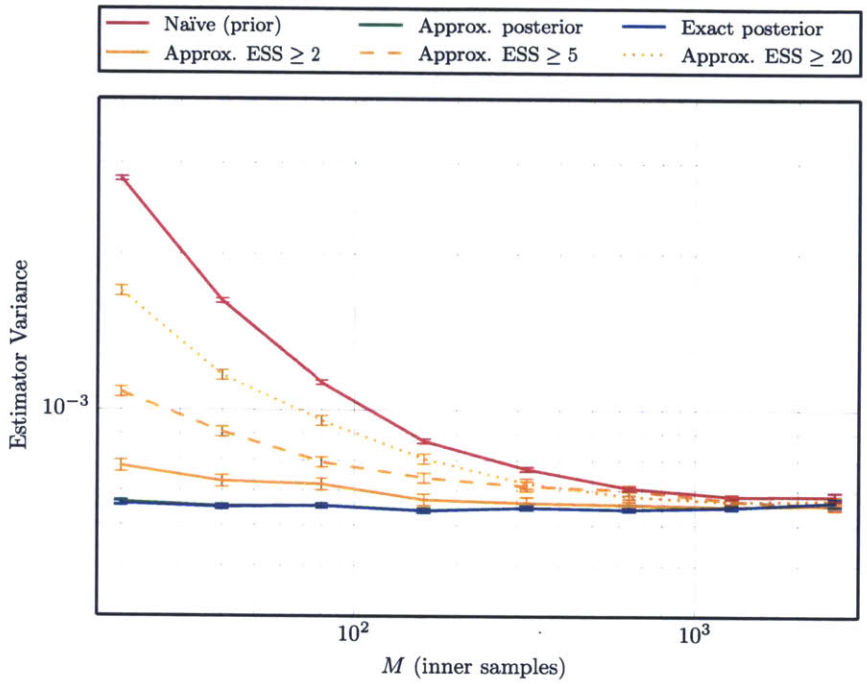


Figure 3-7: Variance of expected utility estimator $\hat{U}$ for the linear Gaussian example as a function of M, the number of inner samples, with a fixed number of outer samples $N = 1280$.

variance, we require a large number of outer Monte Carlo samples. Thus, in practice, it is unlikely that we evaluate $\hat{U}$ with N below the crossover point. Another subtle observation regarding the crossover point is the increase in variance for $N < 100$, which can be seen in Figure 3-8. However, the $1/N$ trend in overall estimator variance still dominates. Finally, the adaptive importance sampling eliminates the sub-$1/M$ variance contribution from the $1/(NM)$ term seen in Figure 3-7.

Now we focus on the effect of enforcing a minimum effective sample size when using the adaptive importance sampling scheme. Plotted in orange with different line styles are thresholds enforced for all values of N. In teal, is the strategy for adapting the threshold based on the value of N, detailed in Equation (3.28). In general, the more strict the minimum threshold, the more inner estimates are made using samples from the prior. This trend which was discussed earlier in Table 3.1 can be seen clearly in Figure 3-4. On the other hand, in Figure 3-9, the behavior is not as straightforward. For larger ESS cutoffs, the adaptive scheme never does worse than sampling from the

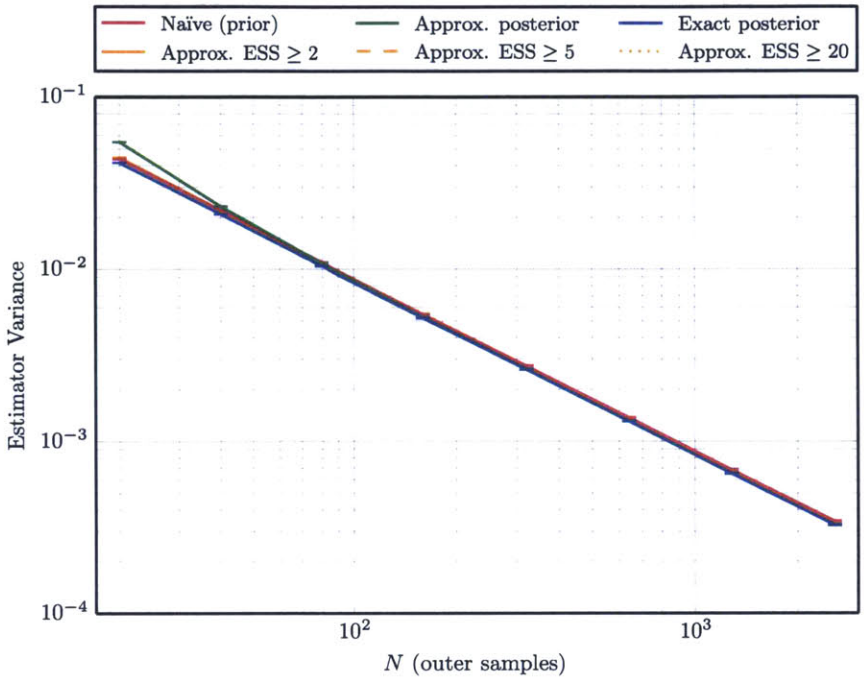


Figure 3-8: Variance of expected utility estimator $\hat{U}$ for the linear Gaussian example as a function of N, the number of outer samples, with a fixed number of inner samples $M = 1280$.

prior, which avoids the crossover seen when not enforcing an ESS cutoff with small N. However, the bias reduction for large N is diminished. The cause is subtle: if we reject using adaptive importance sampling because the effective number of prior samples is small, then it is very likely that the importance sampling estimator using prior samples will not perform well either, since its effective sample size will *also* be small. The strategy of enforcing ESS $\geq 1 + (100/N)$ is effective at both a) not performing worse than the naïve sampler, and b) not reducing the effectiveness of the adaptive importance sampler at larger N. This is most clear in Figure 3-5.

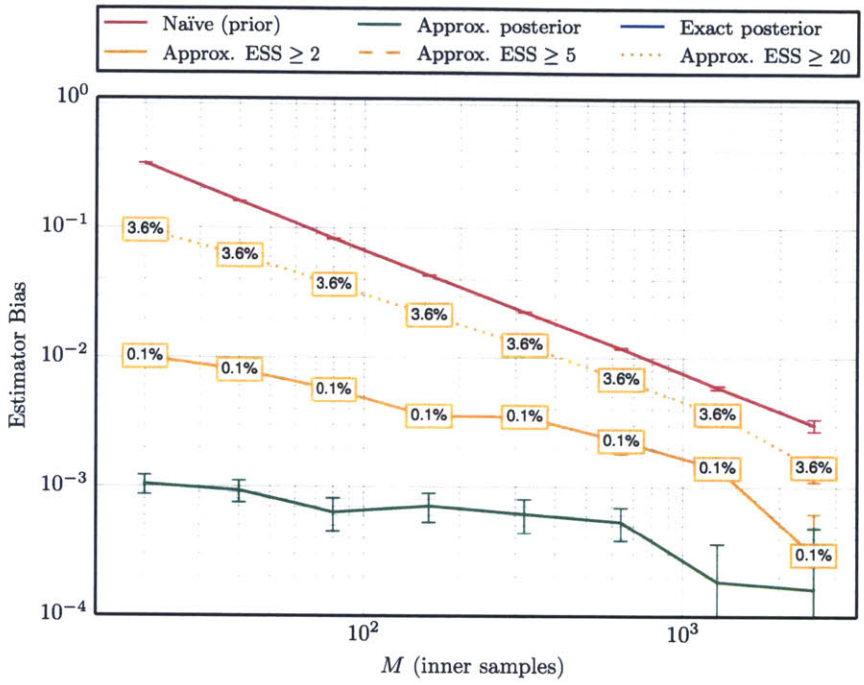


Figure 3-9: Bias of expected utility estimator $\hat{U}$ for the linear Gaussian example as a function of M, the number of inner samples, with a fixed number of outer samples $N = 1280$, with overlay of percent rejected based on ESS cutoff.

Method	**Bias**	**Variance**	**Robustness**
Naïve (prior)	Decreases as $1/M$	Decreases as $1/(NM)$,	Very robust.
No ESS cutoff	Decreases as $1/N$. Can be 10-100x smaller than naïve for large N.	Decreases as $1/N$. No dependence on M.	Can fail when N is small, or $(\theta^{(i)}, \eta^{(i)})$ is an 'outlier.'
ESS ≥ 2	Similar to 'no cutoff,' but less bias reduction at large N (2-10x).	Similar to 'no cutoff,' but now depends on $1/M$.	Slightly lower failure rate, but same failure conditions as no cutoff.
ESS ≥ 20	Similar to 'naïve,' small bias reduction at large N (1-2x).	Similar to 'naïve.'	Very robust.
ESS $\geq 1+100/N$	Not larger than 'naïve,' for small N, similar to 'no cutoff' at large N.	Similar to 'no cutoff.'	Mostly robust. Can fail $\sim 0.1\%$ if $100/N$ is too aggressive.

Table 3.2: Summary of estimator properties for the different proposed importance sampling schemes: naïve, adaptive importance sampling with different fixed ESS cutoff, and with an adaptive ESS cutoff.

3.2 Stochastic approximation methods

Now that the expected utility $U(\boldsymbol{d})$ can be estimated at any value of the design variables, we turn to the optimization problem (2.12). Maximizing U via a grid search over $\mathcal{D}$ is impractical, since the number of grid points grows exponentially with dimension. Since only a Monte Carlo estimate $\hat{U}(\boldsymbol{d})$ of the objective function is available, another naiïve approach would be to use a large sample size (N, M_1, M_2) at each $\boldsymbol{d}$ and then apply a deterministic optimization algorithm, but this is still too expensive, and even with large sample sizes, $\hat{U}(\boldsymbol{d})$ is essentially non-smooth. Instead, we would like to use only a *few* Monte Carlo samples to evaluate the objective at any given $\boldsymbol{d}$, and thus we need algorithms suited to noisy objective functions. Here, we focus our discussion on the simultaneous perturbation stochastic approximation (SPSA), which was successfully used by [23] for a similar objective.

SPSA, proposed by Spall [48, 47] is a stochastic approximation method that is similar to a steepest-descent method using finite difference estimates of the gradient, except that SPSA only uses two random perturbations to estimate the gradient regardless of the problem's dimension (instead of the $2n_d$ evaluations for a full finite-difference scheme):

$$\boldsymbol{d}_{k+1} = \boldsymbol{d}_k - a_k \boldsymbol{g}_k(\boldsymbol{d}_k) \tag{3.31}$$

$$\boldsymbol{g}_k(\boldsymbol{d}_k) = \frac{\hat{U}(\boldsymbol{d}_k + c_k\boldsymbol{\Delta}_k) - \hat{U}(\boldsymbol{d}_k - c_k\boldsymbol{\Delta}_k)}{2c_k} \begin{bmatrix} \Delta_{k,1}^{-1} \\ \Delta_{k,2}^{-1} \\ \vdots \\ \Delta_{k,n_d}^{-1} \end{bmatrix}, \tag{3.32}$$

where k is the iteration number,

$$a_k = \frac{a}{(A+k+1)^\alpha}, \quad c_k = \frac{c}{(k+1)^\gamma}, \tag{3.33}$$

and a, A, α, c, and γ are algorithm parameters with recommended values available in [47]. $\boldsymbol{\Delta}_k$ is a random vector whose entries are i.i.d. draws from a symmetric distribution

with finite inverse moments; here, we choose $\Delta_{k,i} \sim$ Bernoulli(0.5). Common random numbers are also used to evaluate each pair of estimates $\hat{U}(\boldsymbol{d}_k + c_k \boldsymbol{\Delta}_k)$ and $\hat{U}(\boldsymbol{d}_k - c_k \boldsymbol{\Delta}_k)$ at a given $\boldsymbol{d}_k$, in order to reduce variance in estimating the gradient [26].

An intuitive justification for SPSA is that error in the gradient "averages out" over a large number of iterations [48]. Convergence proofs with varying conditions and assumptions can be found in [20]. Randomness introduced through the noisy objective $\hat{U}$ and the finite-difference-like perturbations allows for a global convergence property [34]. Constraints in SPSA are handled by projection: if the current position does not remain feasible under *all* possible random perturbations, then it is projected to the nearest point that does satisfy this condition.

Using the algorithms described in this section, we can approximately solve the stochastic optimization problem posed in Equation (2.12) and obtain the best experimental design.

3.3 Markov-chain Monte Carlo (MCMC)

To evaluate the posterior mean used in the Bayes' risk expression in Equation 2.16, we need to be able to draw samples from the posterior distribution. The posterior can be evaluated pointwise up to a constant factor, but computing it on a grid is clearly impractical as the number of dimensions increases. A more economical method is to generate independent samples from the posterior, but given the arbitrary form of this distribution (particularly for nonlinear models), direct Monte Carlo sampling is usually not possible. Instead, we use Markov chain Monte Carlo (MCMC) sampling, which only requires pointwise evaluations of the *unnormalized* posterior density.in order to construct a Markov chain whose stationary and limiting distribution is the posterior distribution. Samples generated in this way are correlated, so the effective sample size is smaller than the number of MCMC steps. The resulting samples can then be used in various ways, e.g., to evaluate marginal posterior densities, or to

approximate posterior expectations

$$\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d}}[f(\boldsymbol{\theta})] = \int_{\Theta} f(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})\ \mathrm{d}\boldsymbol{\theta} \tag{3.34}$$

with the n_M-sample average is

$$\langle f_{n_M} \rangle = \frac{1}{n_M} \sum_{t=1}^{n_M} f(\boldsymbol{\theta}^{(t)}), \tag{3.35}$$

where $\boldsymbol{\theta}^{(t)}$ are samples extracted from the chain. For example, the minimum mean square error (MMSE) estimator is simply the mean of the posterior, while the corresponding Bayes risk is the posterior variance (approximately the trace of the sample covariance matrix), both of which can be estimated using MCMC.

A very simple and powerful MCMC method is the Metropolis-Hastings (MH) algorithm, first proposed in 1953 by Metropolis *et al.*[36], and later generalized by Hastings in [19]. Additional details about the algorithm can be found in [14]. Two improvements to the MH algorithm are delayed rejections (DR) [15] and adaptive Metropolis (AM) [18]; combining these lead to the DRAM algorithm of Haario *et al.* [17]. Although there are many other MCMC algorithms that may be more efficient at the cost of being more complex, DRAM is sufficient for our needs, since evaluating the Bayes' risk through posterior integration is not the focus of this optimal design framework.

Even with efficient proposals, MCMC typically requires a large number of samples (tens of thousands or even millions) to compute posterior estimates with acceptable accuracy. Since each MCMC step requires a forward model evaluation to compute the posterior density, in practice, surrogate models, such as polynomial chaos approximations [23] can result in huge computational savings. Since the examples here are very inexpensive, we do not need to construct surrogate models.

4

Example: Mössbauer Spectroscopy

4.1 Motivation and introduction

The Mössbauer effect refers to recoil-free nuclear resonance fluorescence, which involves the resonant and recoil-free emission and absorption of gamma radiation by atomic nuclei bound in a solid. Generally speaking, if the recoil energy of a gamma ray is below the lowest allowed vibrational mode (ground state energy) of the atomic lattice in which the emitting nucleus is bound, the entire lattice recoils as a single object, and since the atomic lattice is much more massive than the emitting nuclei, conservation of momentum means the emitting nuclei essentially undergoes no recoil. When emitted gamma rays essentially carry all of the energy of the atomic nuclear de-excitation that produced them, the energy is sufficient to excite the same energy state in a second immobilized nucleus of the same type. This recoil-free emission and re-absorption of gamma rays is the foundation of Mössbauer spectroscopy [37].

Mössbauer spectroscopy is responsible for the discovery of the isomer shift, with the first measurement in gamma spectroscopy reported in 1960 [25], two years after its first experimental observation in atomic spectroscopy. The measurement of the isomer shift provides valuable information regarding the electronic and chemical structure of the lattice in which emitting and absorbing nuclei are embedded. The isomer shift is illustrated in Figure 4-1.

The Mössbauer isomeric shift is seen in gamma ray spectroscopy when one compares

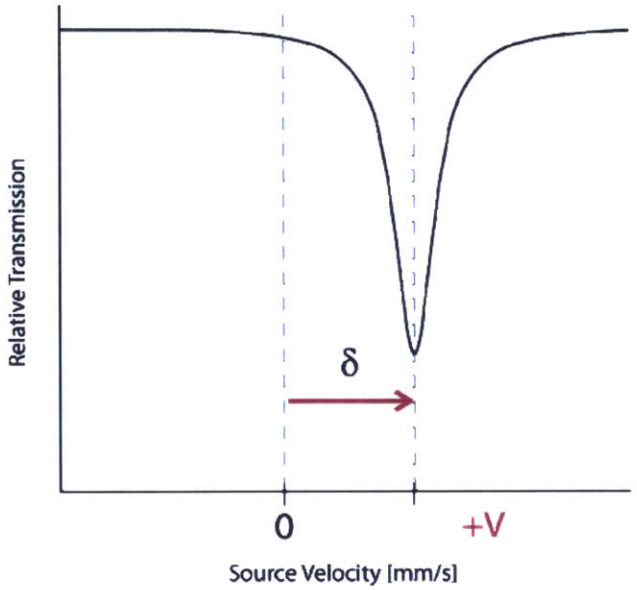


Figure 4-1: Isomeric shift of the nuclear energy levels and corresponding spectrum.

two different nuclear isomeric states in two different physical or chemical environments, and is due to the combined effect of the recoil-free Mössbauer transition between the two nuclear isomeric states and the transition between two atomic states in those two environments. By measuring this shift one obtains important and extremely precise information, both about the nuclear isomer states and about the physical, chemical or biological environment of the atoms [45]. These shifts are incredibly small, usually on the order of 10^{-8} eV, which is roughly 12 orders of magnitude smaller than the energy of the emitted gamma rays themselves. Mössbauer spectroscopy involves moving the gamma ray source relative to the absorber. The change in the energy of the emitted gamma ray from Doppler shifting can be precisely controlled by the relative velocity of the source and absorber. In practice, shifts of 10^{-8} eV correspond to velocities on the order of mm/s.

The general experimental setup for Mössbauer spectroscopy involves a radioisotope source, a drive system to move the source (either with constant acceleration or constant velocity), an absorber, and a gamma ray detector. A schematic is shown in Figure 4-2. Operation the Mössbauer drive at constant velocity (motion at a fixed velocity for a long time compared to the period between detector counts) has several advantages

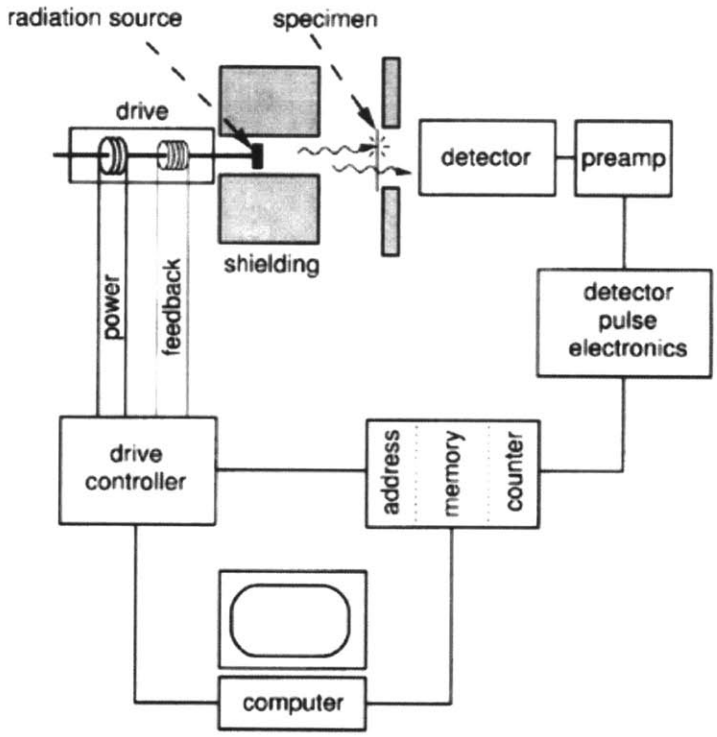


Figure 4-2: Transmission Mössbauer spectrometer. The radiation source sends gamma rays to the right through a collimator into a detector. An electromagnetic drive is operated with feedback control by comparing a measured velocity signal with a desired reference waveform. Counts from the detector are accumulated in a multichannel scaler. Each time interval corresponds to a particular velocity of the radiation source [12].

over operation at constant acceleration. Perhaps the most important is the increased efficiency of data acquisition, since integration time may be spent only at the velocities of interest, which, in general, are not uniformly or symmetrically distributed about zero. For example, the isomer shift frequently introduces a large displacement of the pattern; in some cases the entire region of interest lies entirely in a velocity range of one sign. Another important advantage is the ease with which high accuracy of velocity measurement and control can be obtained. Simple optical methods of measuring distance combined with electronic measurement of time provide an accurate absolute measurement of velocity which is quite independent of the main drive system, thus eliminating problems of calibration and drift [10]. If one uses long counting times and small velocity steps to obtain high precision, the total time of the run can take days, so careful selection of velocities at which to acquire data is critical for a successful experiment.

4.2 Problem formulation

The experimental design problem is to select a finite number of velocities $(d_1, d_2, \ldots, d_{n_d})$ for high precision measurement of the isomeric shift δ. We use the standard parameterization of the absorption peak as having a Lorentzian profile [21], such that the the number of detector counts y_i at velocity d_i is described by the following nonlinear model:

$$y_i = h_0 - \frac{h\,\gamma^2}{(\delta - d_i)^2 + \gamma^2} + \epsilon_i, \qquad i = 1, 2, \ldots, n_d \tag{4.1}$$

where h_0 is the gain of the detector, h is the height of the absorption peak, γ is a parameter specifying the width, and $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ characterizes the observation error. Clearly, δ is our parameter of interest, and the remaining variables (γ, h, h_0) are nuisance parameters. To keep our notation consistent with our framework for Bayesian optimal experimental design, let $\theta = (\delta)$ and $\boldsymbol{\eta} = (\log \gamma, \log h, h_0)$. We assign

the following priors on these model parameters:

$$\begin{aligned}
\theta &\sim \mathcal{N}(0, 1) \\
\eta_1 &\sim \mathcal{N}(0, 0.1^2) \\
\eta_2 &\sim \mathcal{N}(0, 0.2^2) \\
\eta_3 &\sim \mathcal{N}(2, 0.1^2). \\
\epsilon_i &\sim \mathcal{N}(0, 0.05^2)
\end{aligned}$$

We imbue the γ and h with log-normal priors, since the height and width of the absorption peak should be non-negative. Since this model is chosen for demonstrative purposes only, we have conveniently rescaled the parameters to be of similar order. In practice, prior distributions should be specified by combining existing data with expert opinion, e.g., via prior elicitation methods.

Substituting θ and $\boldsymbol{\eta}$ into Equation (4.1), we get

$$\begin{aligned}
y_i &= \eta_3 - \frac{\exp(2\eta_1 + \eta_2)}{(\theta - d_i)^2 + \exp(2\eta_1)} + \epsilon_i, \qquad i = 1, 2, \ldots, n_d \\
&\downarrow \\
\boldsymbol{y} &= G(\theta, \boldsymbol{\eta}, \boldsymbol{d}) + \boldsymbol{\epsilon},
\end{aligned} \tag{4.2}$$

which is defined using the same notation as the observation model (3.10) we used earlier to develop the numerical methods for evaluating the expected utility. By plotting several Mössbauer spectra simulated from the prior distribution in Figure 4-3, we observe a representative sample of absorption peaks.

4.3 Model discrepancy

4.3.1 Sources of model discrepancy

Possible sources of model discrepancy for this model include incomplete physics, e.g. higher order effects and thermal line broadening [46] and incomplete characterization

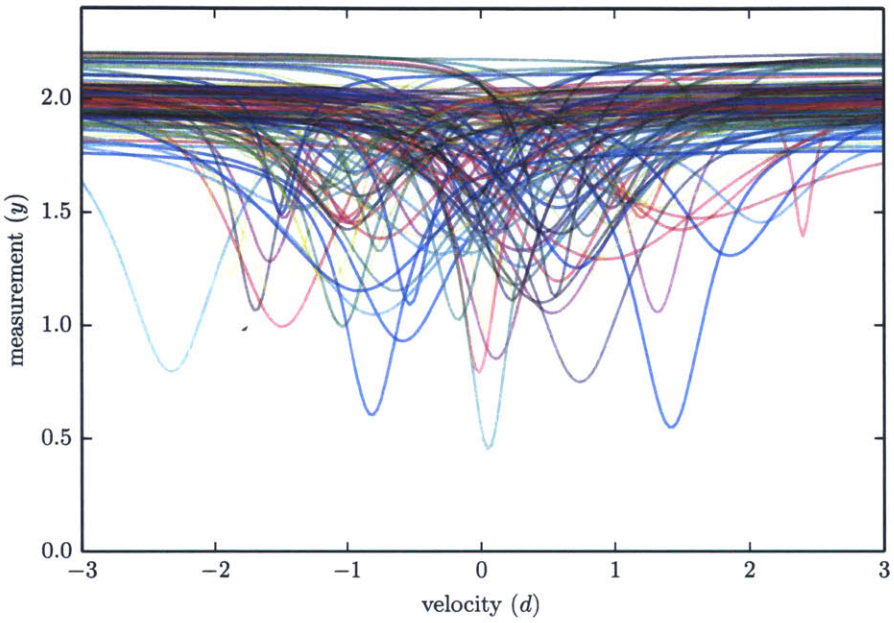


Figure 4-3: Simulated Mössbauer spectra from the model specified in Equation 4.2

of the instrumentation, which include rate-related gain shift from quenching and other systematic errors not included in the statistical model above. An example of of higher-order effects is illustrated in Figure 4-4, where hyperfine splitting is temperature dependent. Line broadening can be modeled by convolving the absorption spectrum with a Gaussian.

Another source of discrepancy is that the observation error is, in fact, Poisson distributed, since the observations are detector counts, i.e., the variance of the observation noise is proportional to the detector count, so our choice of modeling observation error with an additive Gaussian introduces some model discrepancy.

4.3.2 Characterization of model discrepancy

Following [24], we represent the model discrepancy term $\delta(d)$ as a zero-mean Gaussian process,

$$\delta(\,\cdot\,) \sim \mathrm{GP}\big(0, C(d_i, d_j)\big), \tag{4.3}$$

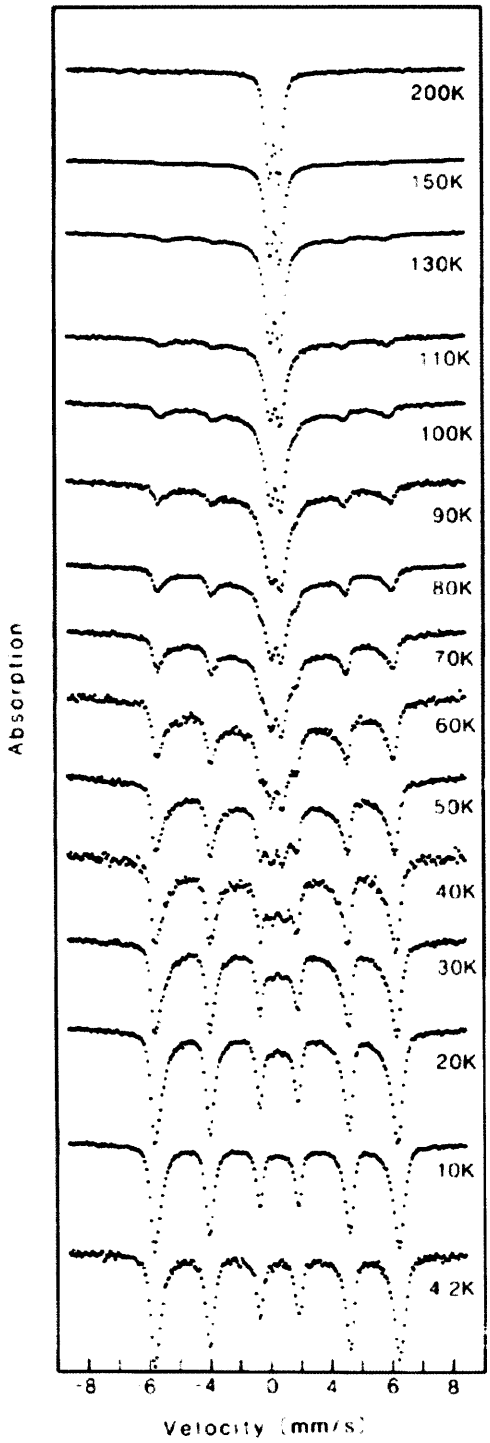


Figure 4-4: Mössbauer spectra from a specimen of haemosiderin, showing the effects of superparamagnetism with increasing temperature [4].

with the squared exponential correlation function

$$C(d_i, d_j\,; \sigma_C, L) = \sigma_C^2 \exp\left(-\left(\frac{d_i - d_j}{L}\right)^2\right). \tag{4.4}$$

It is important to understand how such a representation might formulate prior knowledge about the discrepancy function δ. A GP is a probability distribution for a function. Equation (4.3) says that at any point d the prior probability distribution of $\delta(d)$ is normal with mean zero and variance σ_C^2 from Equation (4.4). The zero mean implies that we do not have a prior expectation that $\delta(d)$ is more likely to be positive or more likely to be negative. The variance σ_C^2 expresses a prior belief that $\delta(d)$ is not likely to be outside the range $\pm 2\sigma_C$, so it measures the strength of prior information about $\delta(d)$. The fact that the variance is the same for all d implies that we do not have a prior expectation that $|\delta(d)|$ is likely to take larger values for some x values than for others. The correlation function expresses a prior belief that $\delta(d)$ will be a smooth function.

Without additional prior information about the model discrepancy, we cannot apply additional constraints to $\delta(d)$ in the manner of [5] and can only stick to the original model discrepancy formulation by Kennedy and O'Hagan. However, given the knowledge from the previous section on possible sources of model discrepancy, we can apply constraints on the discrepancy function $\delta(d)$ that reflect the physics that have not been accounted for in the model.

The model discrepancy induced by not considering thermal line broadening (from Doppler shift due to thermal vibration) can be eliminated by introducing additional nuisance parameters directly to the forward model by assigning parameters to a Gaussian convolution kernel that is applied to $G(\theta, \boldsymbol{\eta}, \boldsymbol{d})$ to result in $\tilde{G}(\theta, \tilde{\boldsymbol{\eta}}, \boldsymbol{d})$, where the tilde indicate the modified variables. However, this approach adds additional complexity in a very nonlinear way to our model, whereas if we are already using a Gaussian process to describe the model discrepancy, we can make sure that the correlation length is chosen such that $\delta(d)$ can capture variations on the length-scale corresponding to the amount of thermal broadening we expect for the temperature at

which the experiment will be run.

For rate-related gain shift, an additive model discrepancy term is not sufficient to capture the discrepancy that is a function of the underlying rate, which in our model is the output of the forward model evaluation $G(\theta, \boldsymbol{\eta}, \boldsymbol{d})$. Since we opted to not cancel out one of our model parameters by dividing the detector gain, the posterior predictives of our model still converge to the data if the rate-related gain shift is linear. However, the quenching schemes of most gamma ray detectors are very nonlinear, especially near the detector's saturation point [27].

4.4 Optimal designs

We apply the Monte Carlo estimator of the expected utility developed in Chapter 3 to the Mössbauer experimental design problem. First, we consider a simple setup, with 3 measurement points. For visualization purposes, we set $d_3 = 0$, so that we can plot the 2D expected utility surface. This setup is not entirely arbitrary. A measurement point at zero velocity is easy to set up, and is usually the first data point measured in practice. Intuitively, we expect there to be many symmetries in the expected utility surface. Since the 'order' of each measurement does not matter, we expect symmetry along the $d_1 = d_2$ axis. Furthermore, as the measurement points d_i overlap, the measurements become redundant, so we should expect less expected information gain. The degeneracy of repeated measurements should manifest as valleys along the $d_1 = d_2$, $d_1 = d_3$, $d_2 = d_3$ on the expected utility surface. Finally, since we have assigned a normal distribution to the parameter of interest, most of the realizations will be close to the origin, so we expect that measurements made near the area of most variation in the model output should be more informative than those made in the tail regions where there is less variation. Since the Mössbauer forward model is very cheap to evaluate, we can exhaustively evaluate the expected utility on a grid, which is shown in Figure 4-5.

Then, we apply the SPSA algorithm with uniformly sample starting locations to observe the convergence properties of the stochastic optimization for the Mössbauer

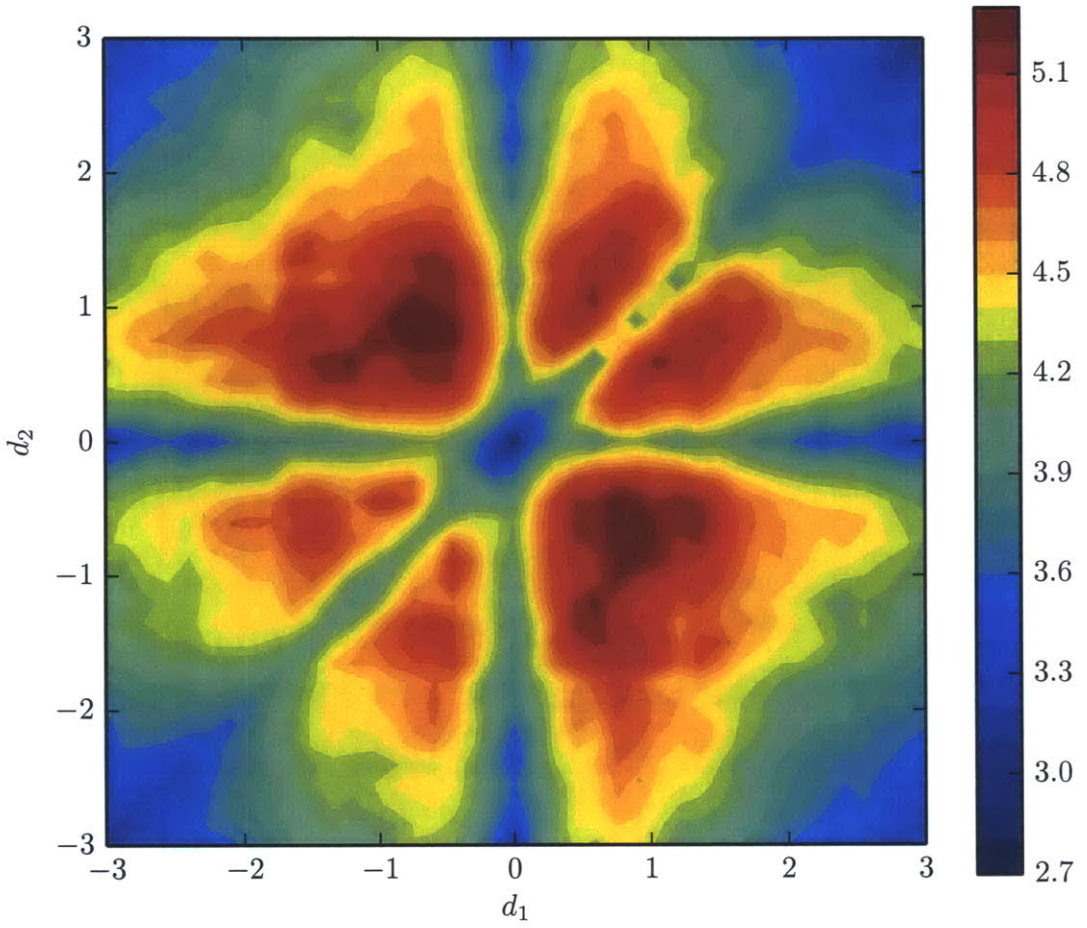


Figure 4-5: Expected utility surface for the Mössbauer experiment with 3 design points, one fixed at $d_3 = 0$. The expected utility estimator was evaluated on a 41×41 mesh with $N = 2000$, $M_1 = M_2 = 200$, with ESS ≥ 1.05.

example. The trajectories for 50 attempts are plotted in Figure 4-6. Even with only 500 iterations, the SPSA algorithm is able to locate the local minima. On some trajectories, we can also observe hopping between local minima. The distribution of stopping points at 500 iterations is shown in Figure 4-7.

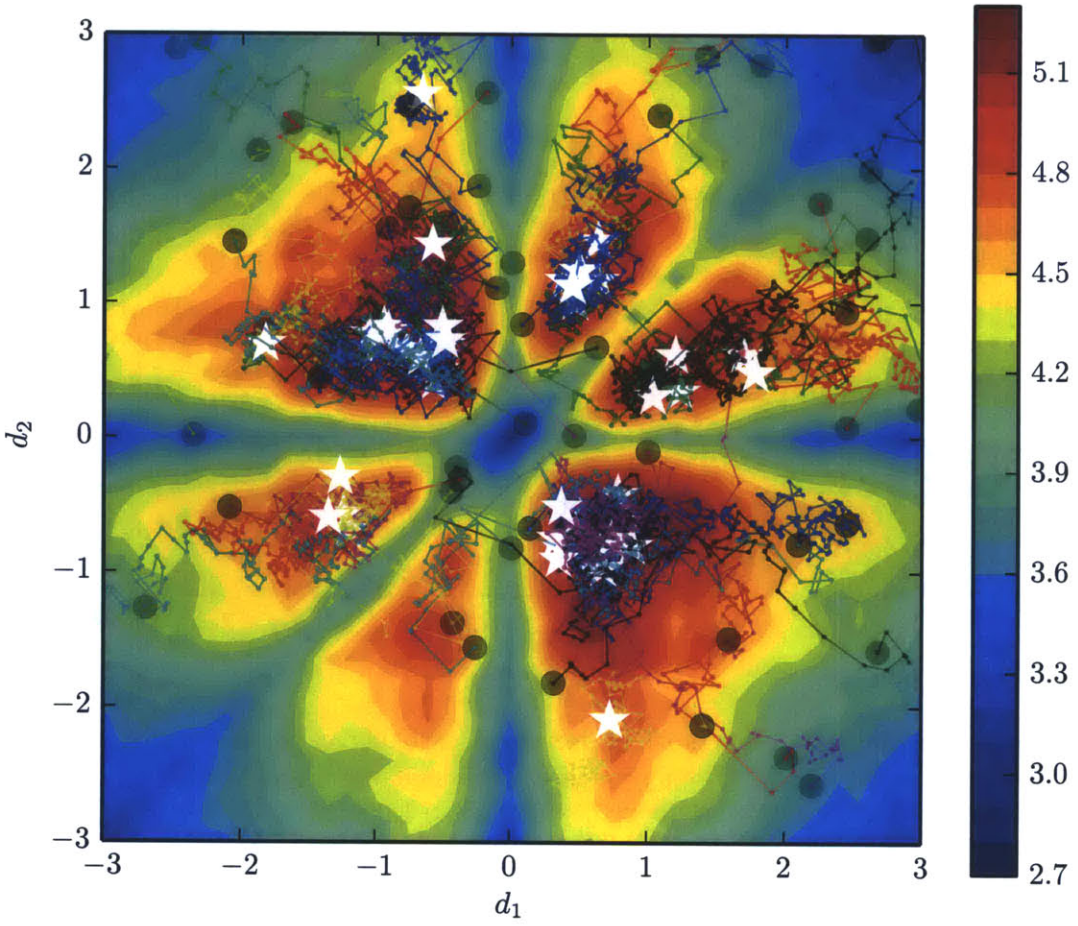


Figure 4-6: 50 SPSA Optimization trajectories overlaid on the expected utility surface for the Mössbauer experiment with 3 design points, one fixed at $d_3 = 0$. Black circles are starting locations and white stars are ending locations. Trajectories limited to 500 iterations.

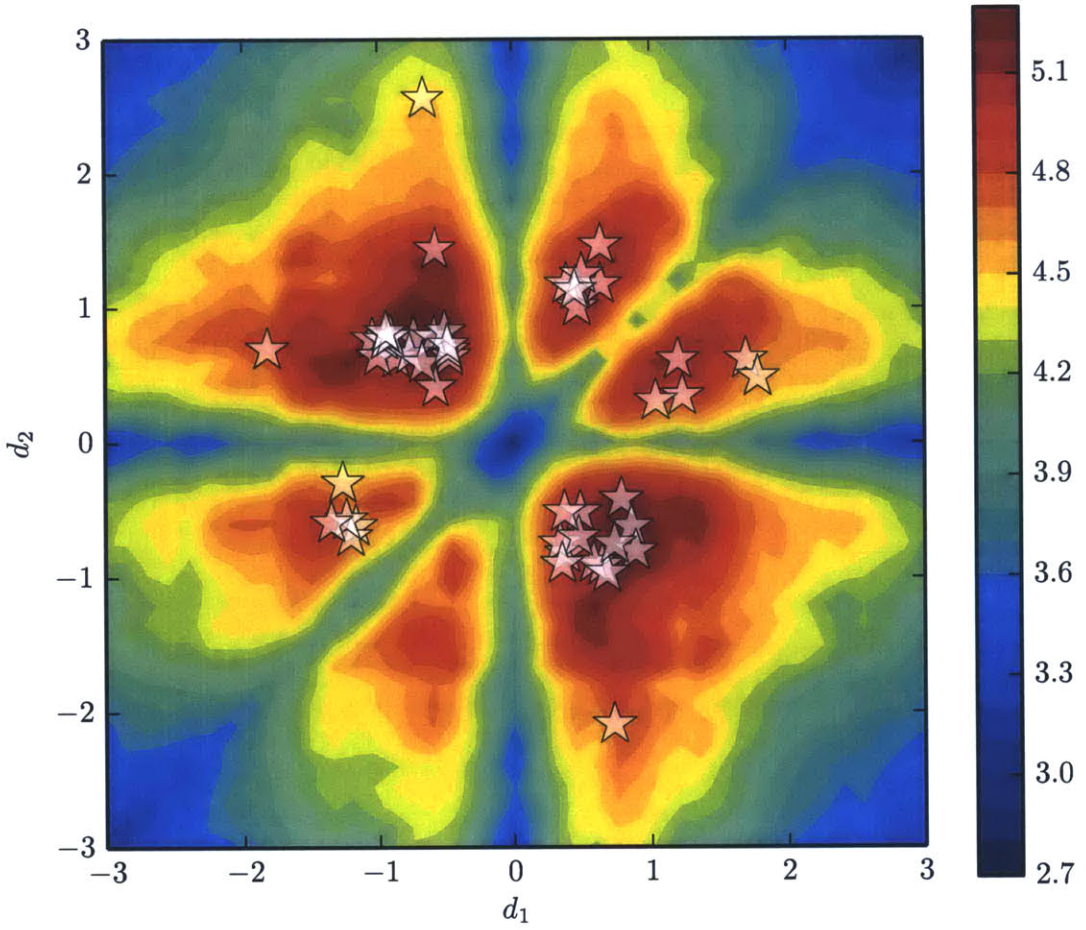
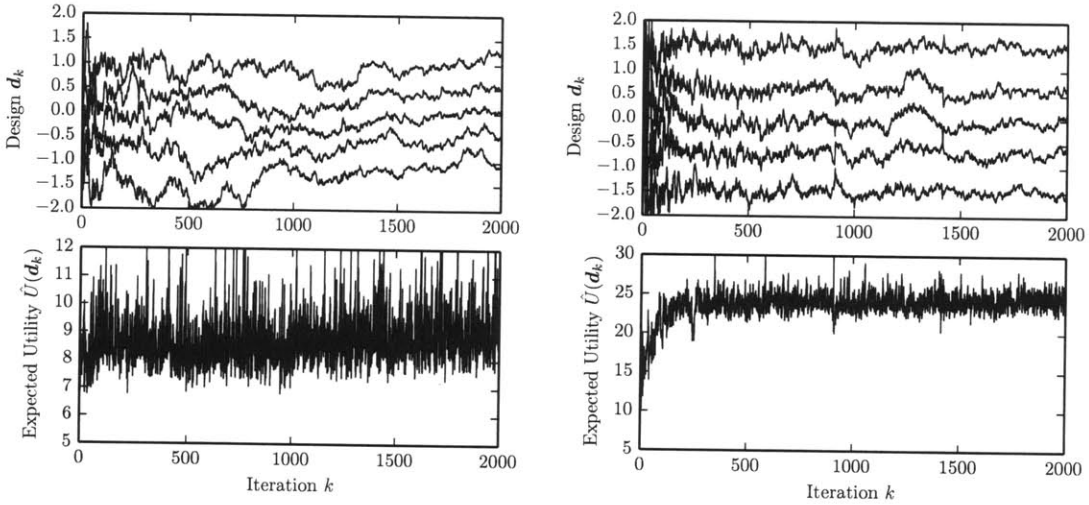


Figure 4-7: 50 SPSA Optimization endpoints overlaid on the expected utility surface for the Mössbauer experiment with 3 design points, one fixed at $d_3 = 0$.Trajectories limited to 500 iterations.

The next step is to study the effects of incorporating model discrepancy through an additive Gaussian process term. We choose to use a square exponential kernel since the expected sources of model discrepancy all have relatively smooth features. The length scale is chosen to be of the same order as the FWHM of the Lorentzian profile of the absorption peak, which with the log-normal prior results in $L \sim 0.3$. The total variance of the Gaussian process is chosen to be several times larger than the noise, so that it can be identified.

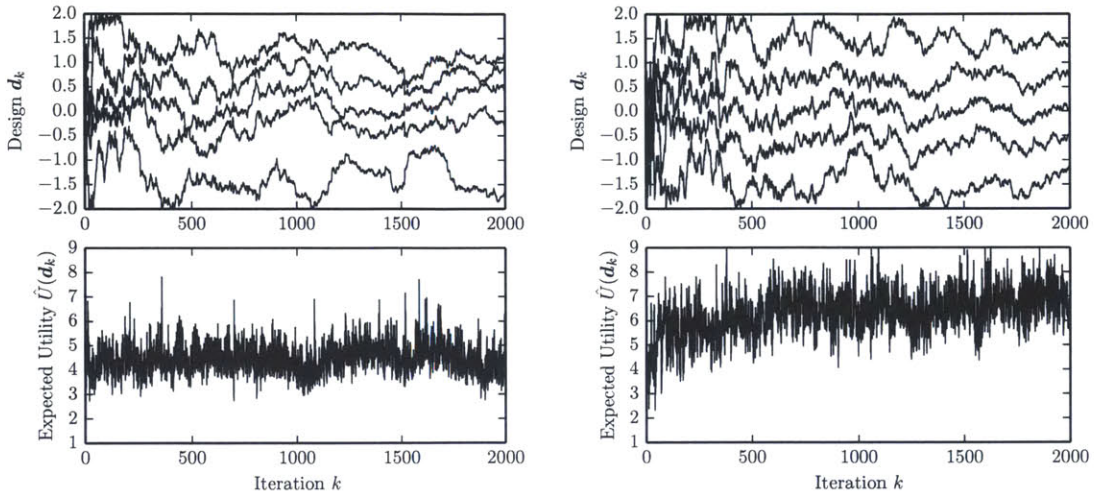


(a) With $n_{\mathrm{KL}} = 0$

(b) With $n_{\mathrm{KL}} = 4$

Figure 4-8: SPSA optimization trajectories for optimal experiments for inference on all model parameters.

To characterize the effect of model discrepancy, we apply the modified Bayes' risk formulation from Section 2.1 by using several models for $\check{p}$, the likelihood function from a 'higher-fidelity' model. The objective is to see which design (and using which model for inference) results in the lowest modified Bayes' risk in the presence of model discrepancy. The results are summarized in Table 4.1. For this table, every cell correspondes to a Monte Carlo estimate of the expected posterior mean squared error, where the posterior mean is computed using MCMC, where each chain is 10^6 samples and the outer Monte Carlo estimate uses 6×10^3 samples. The notation GP(0.3) means that the model discrepancy is modeled using a Gaussian process with correlation length $L = 0.3$. The total variance is kept the same between all models. The Gaussian



(a) With $n_{\mathrm{KL}} = 0$ (b) With $n_{\mathrm{KL}} = 4$

Figure 4-9: SPSA optimization trajectories for optimal experiments for inference on only the parameter of interest, θ.

profile (emulating thermal line broadening) is accomplished by replacing the forward model with a square-exponential function with the same height and FWHM as the Lorentzian forward model, and using the same prior distributions on the width, height, and horizontal shift.

		Model Discrepancy			
Design	Infer. model	None	$\delta \sim \mathrm{GP}(0.3)$	$\delta \sim \mathrm{GP}(0.1)$	Gaussian prof.
$\boldsymbol{d}^*_{\mathrm{joint}}$	No GP	2.4×10^{-2}	2.8×10^{-2}	3.0×10^{-2}	4.5×10^{-2}
	GP(0.3)	3.1×10^{-2}	3.0×10^{-2}	3.3×10^{-2}	3.9×10^{-2}
$\boldsymbol{d}^*_{\mathrm{marg}}$	No GP	2.4×10^{-2}	3.0×10^{-2}	2.9×10^{-2}	4.0×10^{-2}
	GP(0.3)	3.0×10^{-2}	2.6×10^{-2}	2.6×10^{-2}	3.4×10^{-2}

Table 4.1: Expected posterior mean squared error under model misspecification (modified Bayes' risk). Red entries highlight lowest error.

The message that the results of Table 4.1 convey is that using the design optimal for inference in only the parameters of interest, coupled with an additive Gaussian process discrepancy term is the most robust experimental design in the presence of the three different types of discrepancy. Although we did not target for this type

of optimality in our optimization scheme, this result is still optimal in this modified Bayes' risk sense.

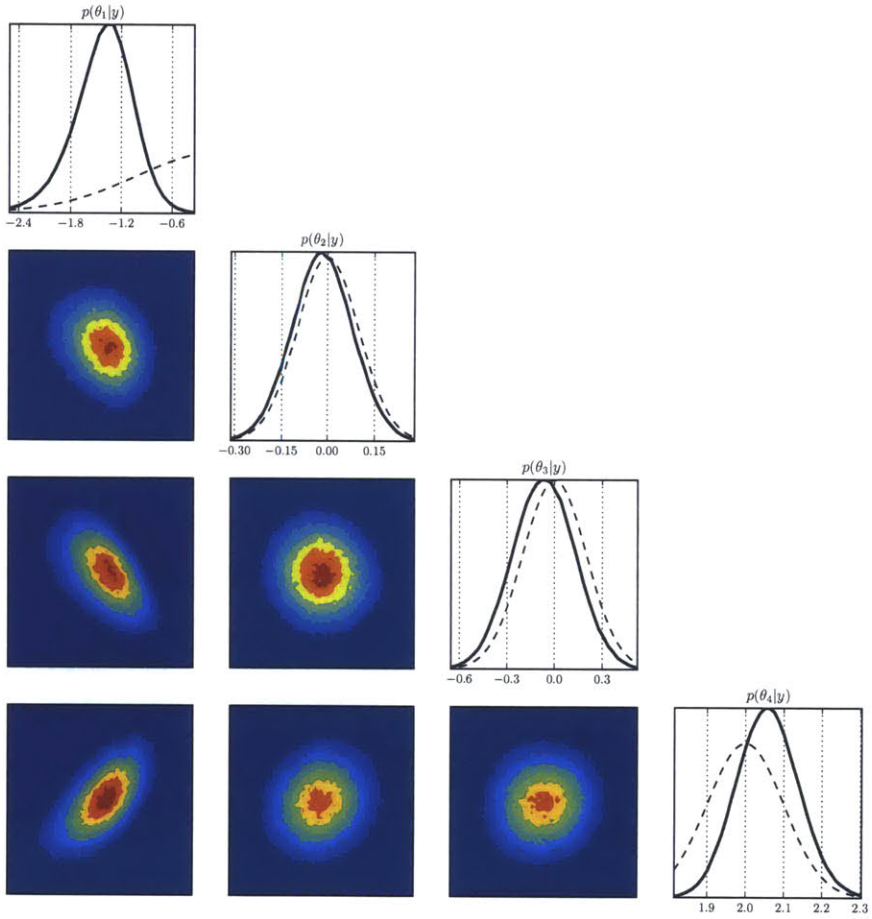


Figure 4-10: MCMC computed posteriors for the Mössbauer problem with 10^6 samples.

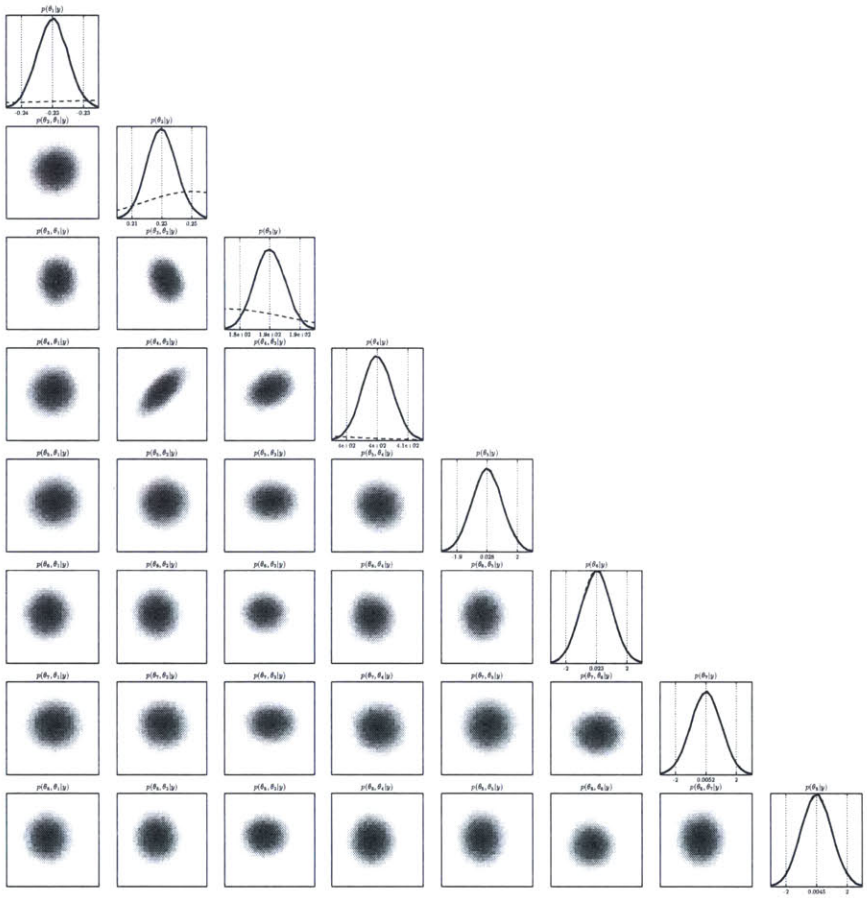


Figure 4-11: MCMC computed posteriors for the Mössbauer problem with with additional nuisance parameters used to parameterize the 4 K-L modes of the model discrepancy.

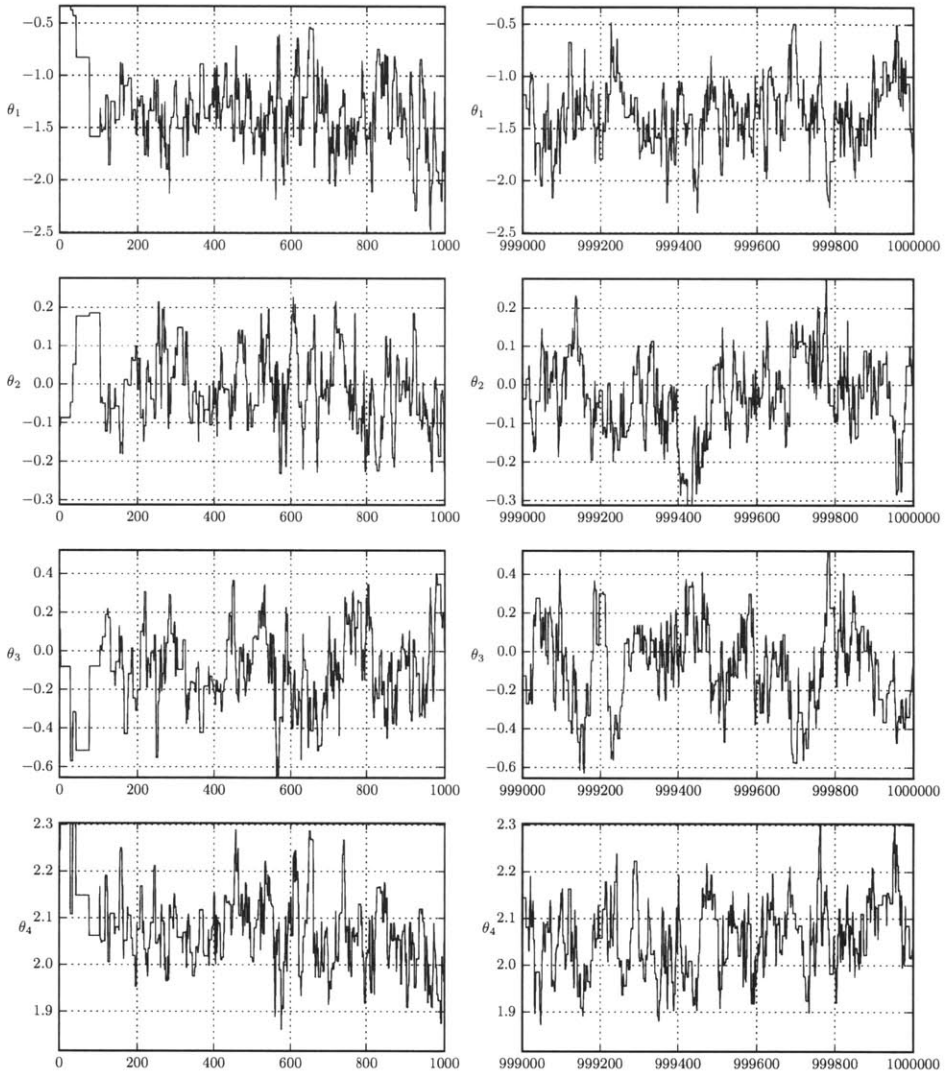


Figure 4-12: MCMC traces demonstrating the adaptation of the DRAM method in the first several hundred iterations. The MCMC is targeting the posterior distribution for the Mössbauer problem without discrepancy.

5

Conclusions and Future Work

5.1 Summary of results

This thesis has explored the challenges arising from a general nonlinear formulation of optimal Bayesian experimental design with nuisance parameters. In particular, we employed an objective that reflects the *expected information gain* in a subset of model parameters due to an experiment, and formulated numerical methods required to solve the optimization problem approximately. We naturally incorporated finite-dimensional representations of Gaussian processes used to describe model discrepancy into the optimal design framework, by adding nuisance parameters. We successfully applied the methods to linear and nonlinear problems, in addition to evaluating the performance of optimal designs when the model misspecification is known. The optimal design for information gain in only the parameters of interest, with consideration for model discrepancy performed the best when compared with designs chosen by other criteria. Furthermore, the adaptive importance sampling scheme used to approximate the expected utility is potentially orders of magnitude more efficient than the current state-of-the-art.

5.2 Future work

The models used as examples were relatively simple, without highly nonlinear dynamics. The performance of the methods developed in this work applied on more complicated dynamical systems is an important area of future study. In more complicated models, there are also more types of model discrepancy, e.g., discretization error and numerical solver error that were not discussed here. Another source of model error that was not discussed here is prior misspecification, which can also be treated in the modified Bayes' risk framework.

Although using the effective sample size as an indicator of robustness for the Monte Carlo estimator was shown to work well for the simple examples here, how well it performs with posteriors that are less Gaussian-like, and how to choose an optimal strategy for varying the cutoff for different models, or for different dimensions are more questions that should be answered.

The discussion here has focused only on batch or open-loop experimental design, where the parameters for all experiments are chosen before data are actually collected. An important target for future work is rigorous sequential or closed-loop design, where the data from one set of experiments are used to guide the choice of the next set of measurements. Here we expect the sampling methods for expected information gain will continue playing a important role.

Appendix A

Additional results

A.1 Adaptive importance sampling for the Mössbauer experiment

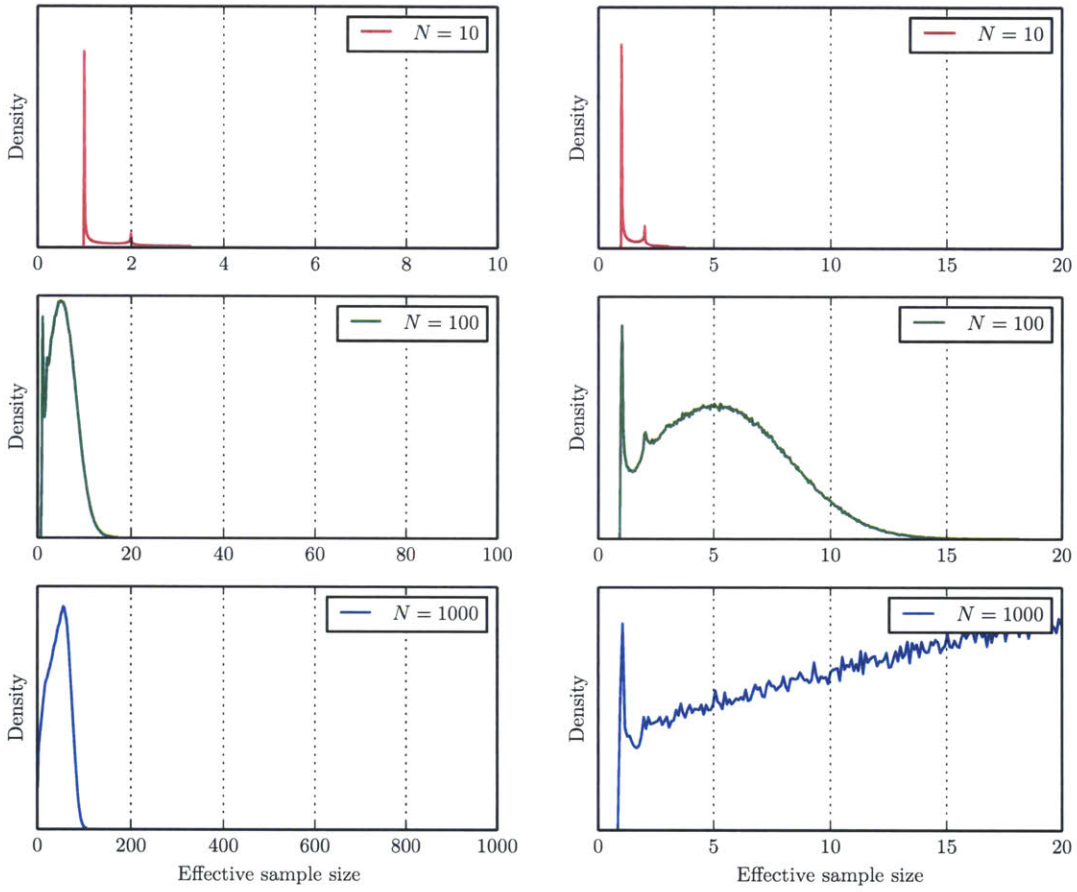


Figure A-1: Distribution of effective sample size for the Mössbauer example for $N = 10, 100, 1000$.

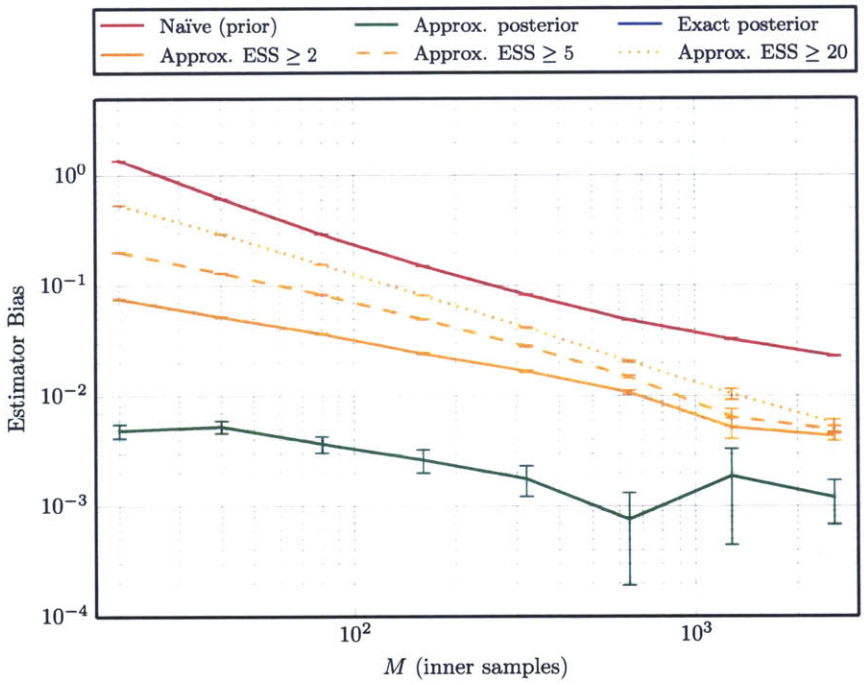


Figure A-2: Bias of expected utility estimator $\hat{U}$ for the Mössbauer example as a function of M, the number of inner samples, with a fixed number of outer samples $N = 1280$.

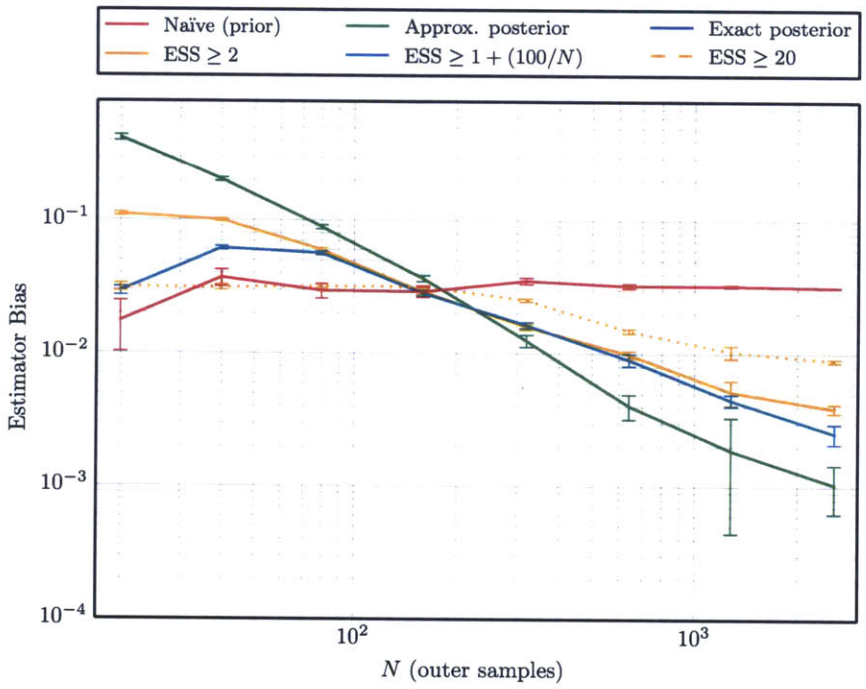


Figure A-3: Bias of expected utility estimator $\hat{U}$ for the Mössbauer example as a function of N, the number of outer samples, with a fixed number of inner samples $M = 1280$.

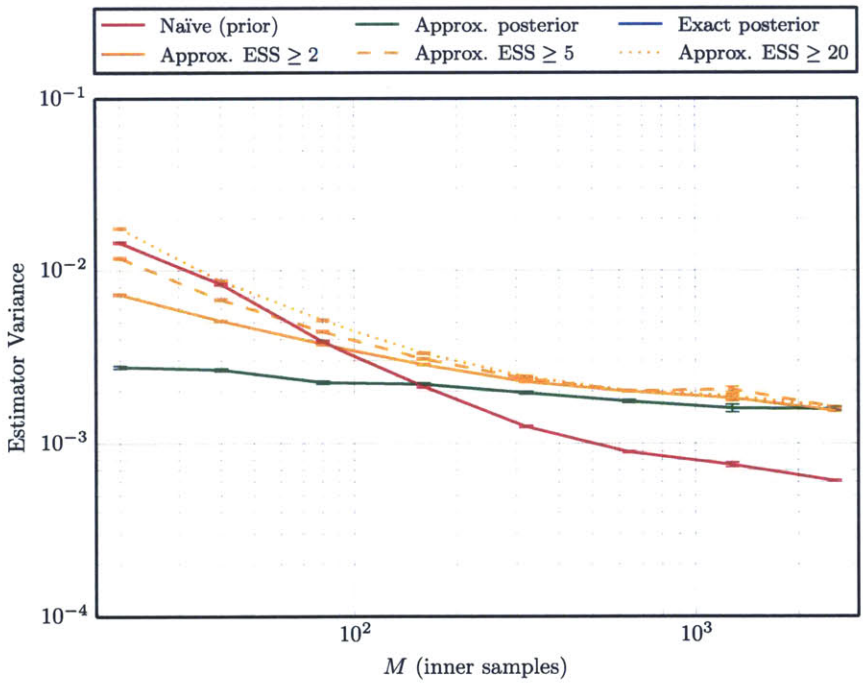


Figure A-4: Variance of expected utility estimator $\hat{U}$ for the Mössbauer example as a function of M, the number of inner samples, with a fixed number of outer samples $N = 1280$.

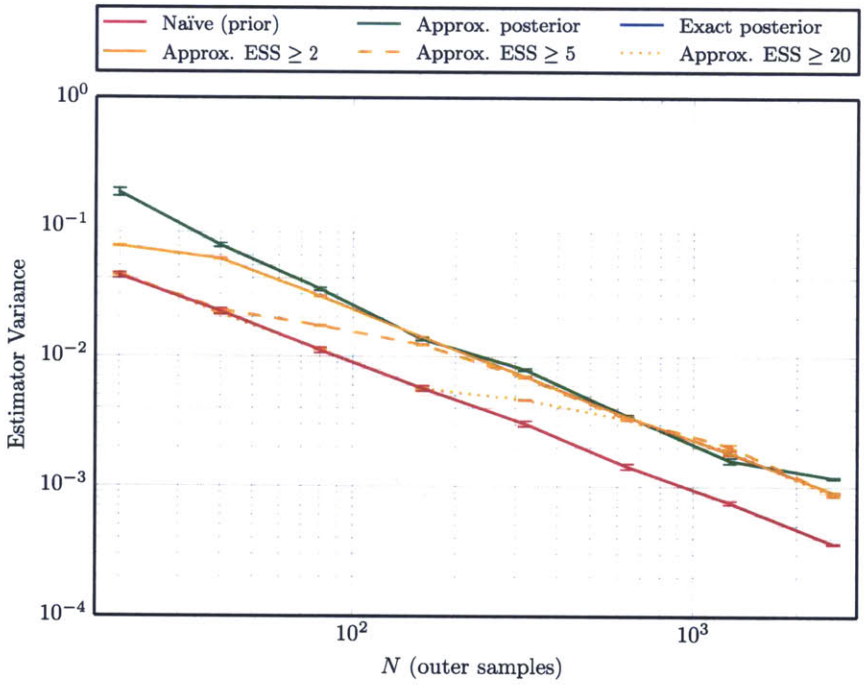


Figure A-5: Variance of expected utility estimator $\hat{U}$ for the Mössbauer example as a function of N, the number of outer samples, with a fixed number of inner samples $M = 1280$.

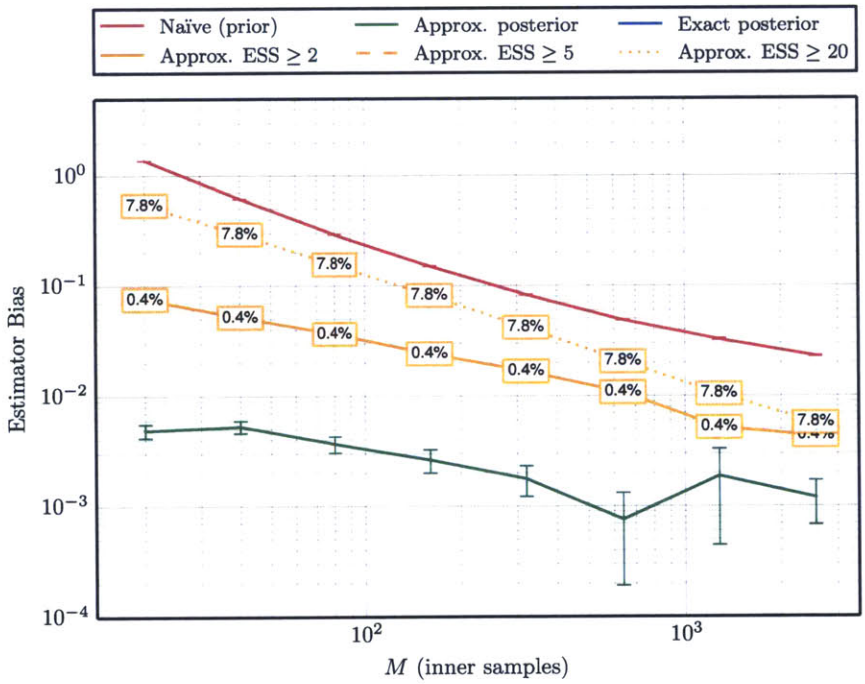


Figure A-6: Bias of expected utility estimator $\hat{U}$ for the Mössbauer example as a function of M, the number of inner samples, with a fixed number of outer samples $N = 1280$, with overlay of percent rejected based on ESS cutoff.

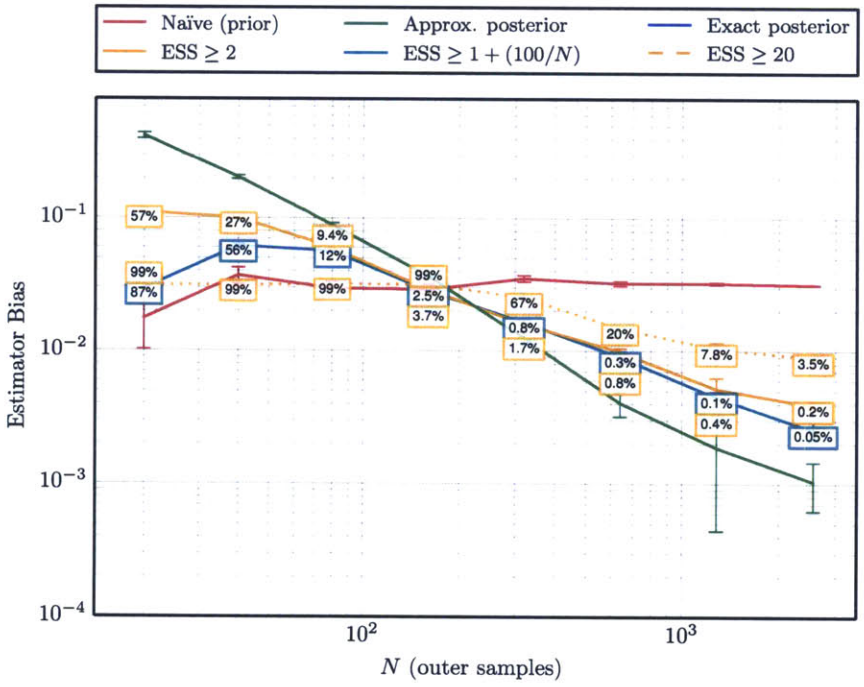


Figure A-7: Bias of expected utility estimator $\hat{U}$ for the Mössbauer example as a function of N, the number of outer samples, with a fixed number of inner samples $M = 1280$, with overlay of percent rejected based on ESS cutoff.

Bibliography

[1] R.J. Adler and J.E. Taylor. *Random Fields and Geometry.* Springer Monographs in Mathematics. Springer, 2009.

[2] Sangtae Ahn and Jeffrey A. Fessler. Standard errors of mean, variance, and standard deviation estimators, 2003.

[3] A. C. Atkinson and A. N. Donev. *Optimum Experimental Designs, with SAS.* Oxford Statistical Science Series. Oxford University Press, 2007.

[4] SH Bell, MP Weir, DP Dickson, JF Gibson, GA Sharp, and TJ Peters. Mössbauer spectroscopic studies of human haemosiderin and ferritin. *Biochimica et biophysica acta*, 787(3):227, 1984.

[5] Jennỳ Brynjarsdóttir and Anthony OÊijHagan. Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, 30(11):114007, 2014.

[6] Christian G. Bucher. Adaptive sampling–an iterative fast Monte Carlo procedure. *Structural Safety*, 5(2):119 – 126, 1988.

[7] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

[8] Yunfei Chu and Juergen Hahn. Integrating parameter selection with experimental design under uncertainty for nonlinear dynamic systems. *AIChE Journal*, 54(9):2310–2320, 2008.

[9] Merlise A. Clyde. Experimental design: A Bayesian perspective. International Encyclopedia of Social and Behavioral Sciences, April 2001.

[10] P. Flinn. Constant-velocity Mössbauer drive systems. In Irwin J. Gruverman, editor, *Mössbauer Effect Methodology*, pages 75–85. Springer US, 1965.

[11] Ian Ford, D.M̃. Titterington, and Christos Kitsos. Recent advances in nonlinear experimental design. *Technometrics*, 31(1):49–60, 1989.

[12] Brent Fultz. Mössbauer spectrometry. In Elton Kaufmann, editor, *Characterization of Materials.* John Wiley, New York, 2011.

[13] R.G. Ghanem and P.D. Spanos. *Stochastic Finite Elements: A Spectral Approach.* Civil, Mechanical and Other Engineering Series. Dover Publications, 2003.

[14] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Interdisciplinary Statistics. Chapman & Hall/CRC, 1996.

[15] Peter J. Green and Antonietta Mira. Delayed rejection in reversible jump Metropolis-Hastings. *Biometrika*, 88(4):1035–1053, 2001.

[16] M. Grigoriu. *Stochastic Calculus: Applications in Science and Engineering.* Springer, 2002.

[17] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, December 2006.

[18] Heikki Haario, Eero Saksman, and Johanna Tamminen. Adaptive proposal distribution for random walk Metropolis algorithm. Technical report, University of Helsinki, 1999.

[19] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[20] Y. He, M.C̃. Fu, and S.Ĩ. Marcus. Convergence of simultaneous perturbation stochastic approximation for nondifferentiable optimization. *IEEE Transactions on Automatic Control*, 48(8):1459–1463, 2003.

[21] J. Hesse. Simple arrangement for educational Mössbauer-effect measurements. *American Journal of Physics*, 41(1):127–129, 1973.

[22] Dave Higdon, James Gattiker, Brian Williams, and Maria Rightley. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103(482):570–583, 2008.

[23] X. Huan and Y. M. Marzouk. Simulation-based optimal bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 232(1):288âĂŞ317, 2013.

[24] Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[25] O. C. Kistner and A. W. Sunyar. Evidence for quadrupole interaction of ^{57m}Fe, and influence of chemical binding on nuclear gamma-ray energy. *Phys. Rev. Lett.*, 4:412–415, Apr 1960.

[26] Nathan L. Kleinman, James C. Spall, and Daniel Q. Naiman. Simulation-based optimization with stochastic approximation using common random numbers. *Management Science*, 45(11):1570–1578, 1999.

[27] G.F. Knoll. *Radiation Detection and Measurement.* Wiley, 2000.

[28] Daniel S. Levine. *Focused Active Inference.* PhD thesis, Massachusetts Institute of Technology, 2014.

[29] D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

[30] D. V. Lindley. *Bayesian Statistics, A Review.* Society for Industrial and Applied Mathematics (SIAM), 1972.

[31] J.S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer Series in Statistics. Springer, 2008.

[32] M. Loeve. *Probability Theory II.* F.W.Gehring P.r.Halmos and C.c.Moore. Springer, 1978.

[33] Thomas J. Loredo and David F. Chernoff. Bayesian adaptive exploration. In *Statistical Challenges of Astronomy*, pages 57–70. Springer, 2003.

[34] John L. Maryak and Daniel C. Chin. Global random optimization by simultaneous perturbation stochastic approximation. *Johns Hopkins APL Technical Digest*, 25(2):91–100, 2004.

[35] Y. M. Marzouk and H. N. Najm. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *Journal of Computational Physics*, 228(6):1862–1902, 2009.

[36] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[37] Rudolf L. Mossbauer. Kernresonanzfluoreszenz von Gammastrahlung in ^{191}ir. *Zeitschrift für Physik*, 151(2):124–143, 1958.

[38] Peter Müller. Simulation based optimal design. In *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, pages 459–474. Oxford University Press, 1998.

[39] Man-Suk Oh and James O Berger. Adaptive importance sampling in Monte Carlo integration. *Journal of Statistical Computation and Simulation*, 41(3-4):143–168, 1992.

[40] Art B. Owen. *Monte Carlo theory, methods and examples.* 2013.

[41] Charles J. Stone Paul G. Hoel, Sidney C. Port. *Introduction to stochastic processes.* Houghton Mifflin Company, 1972.

[42] Kenneth J. Ryan. Estimating expected information gains for experimental designs with application to the random fatigue-limit model. *Journal of Computational and Graphical Statistics*, 12(3):585–603, September 2003.

[43] Paola Sebastiani and Henry P. Wynn. Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(1):145–157, 2000.

[44] Matthias Seeger. Gaussian processes for machine learning. *International Journal of Neural Systems*, 14(02):69–106, 2004.

[45] G.K. Shenoy. Mössbauer-effect isomer shifts. In Gary J. Long, editor, *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, volume 1 of *Modern Inorganic Chemistry*, pages 57–76. Springer US, 1984.

[46] R. H. Silsbee. Thermal broadening of the Mössbauer line and of narrow-line electronic spectra in solids. *Phys. Rev.*, 128:1726–1733, Nov 1962.

[47] James C. Spall. Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Transactions on Aerospace and Electronic Systems*, 34(3):817–823, 1998.

[48] James C. Spall. An overview of the simultaneous perturbation method for efficient optimization. *Johns Hopkins APL Technical Digest*, 19(4):482–492, 1998.

[49] G. Terejanu, R. R. Upadhyay, K. Miki, and J. Marschall. Bayesian experimental design for the active nitridation of graphite by atomic nitrogen. *Experimental Thermal and Fluid Science*, 36:178–193, 2012.

[50] Jojanneke van den Berg, Andrew Curtis, and Jeannot Trampert. Optimal nonlinear Bayesian experimental design: an application to amplitude versus offset experiments. *Geophysical Journal International*, 155(2):411–421, November 2003.