Bayesian optimal experimental design for the study of natural phenomena

Lida Mavrogonatou

A thesis submitted to the

School of Mathematics and Statistics

University of Glasgow

for the degree of

Doctor of Philosophy

August 2020

Abstract

Modern science has been progressively moving towards the study of increasingly complex structures, investigating not only their individual components but also their interactions, dependencies and co-existence as a whole. This thesis is concerned with optimal experimental design methodology for the study of such phenomena.

A decision-theoretic framework for optimal experimental design is adopted in this thesis. The employed methods operate based on an optimality criterion, quantifying the benefit incurred from each alternative experimental design — commonly known as the *expected utility*. An analytical expression is, in most studies of interest, not available for this quantity and so estimation techniques are typically required for its evaluation.

Currently, existing estimation methods fail to adequately address issues arising in optimal design problems within a modern scientific framework. This is predominantly attributed to the considerable computational cost incurred by consideration of mathematical models sophisticated enough to adequately capture the complexity of the studied structures. In face of this restriction, researchers often resort to consideration of rather simplistic models, hindering the progress towards a more realistic representation and better understanding of such systems.

Efficient methodology for evaluation of the expected utility constitutes the first main contribution of this thesis. The presented approach adopts a flexible, non-parametric framework combined with variational approximation techniques that translate the initial evaluation problem to an alternative, more tractable problem, solution of which is achieved through more efficient and computationally inexpensive procedures. A problem shift is thus achieved

under which, estimation of the expected utility is accomplished through its corresponding dual problem. This alternative representation is shown to incur considerable computational gains compared to traditionally adopted approaches without compromising the quality of the produced estimates.

The proposed estimator paves the way to an autonomous, comprehensive framework for the optimal study of complex phenomena within a realistic time frame, currently posing an ongoing challenge. Establishment of such a setup composes the second main contribution of this thesis. The proposed framework attempts to emulate a typical research scheme of closed-loop data collection, knowledge update and optimal decision making which, combined with instrument control software, facilitates modern scientific studies under minimal human input. The class of Bayesian optimisation algorithms is finally considered, allowing for truly optimal decision making during the established procedure. This class of algorithms, although particularly well-suited to optimal experimental design problems, has been given little consideration in the relevant literature. Their integration to the proposed framework, thus, constitutes an additional contribution of this thesis.

Application of the adopted experimental design framework is examined in three increasingly challenging case studies, addressing a broad range of issues typically encountered in optimal design problems. The first study explores the optimal experimental design for a model discrimination problem adopting a set of simpler, polynomial models. An initial assessment of the proposed estimator and a comparison with the currently adopted methodology is performed, under a setup where application of the latter is not hindered by the incurred computational complexity. The subsequent two cases represent real-life problems of optimal experimental design for model inference in Systems Biology and Spectroscopy, employing models under which, traditionally adopted methods can become from highly inefficient to intractable and thus alternative approaches are needed for the study of such phenomena.

Declaration of Authorship

I, Lida Mavrogonatou, hereby declare that this thesis titled 'Bayesian optimal experimental design for the study of natural phenomena' and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly attributed.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr Vladislav Vyshemirsky. I am grateful for the opportunity to work on this project, for the calm and wise support I received and the very enjoyable meetings.

I would like to thank my examiners, Dr Timothy Heaton and Dr Mu Niu for taking the time to review this thesis and for their very helpful suggestions. Thank you also to Dr Nema Dean for convening the examination.

I would like to thank Dr Ludger Evers for the helpful discussions.

I am grateful for the supportive environment provided by the department of Mathematics and Statistics and for people that despite their busy schedules were always happy to help without having to.

Thank you to my family for all their support, guidance and encouragement.

Thank you to my friends: Craig, Marnie, Vinny and Sam. My PhD years would not have been the same without you.

Most importantly, Alan and Tom, I could not have asked for better company during these years. Thank you for all the adventures, the awful jokes and the amazing loukouri.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Decision making underlies any action faced with uncertainty. Due to the inherent complexity of modern phenomena, it can, nevertheless, be a challenging and unintuitive task. The work presented in this thesis aims to establish a comprehensive framework for the scientific study of such systems, ensuring optimal decision making under uncertainty. Focus is placed on studies where observation of the examined phenomenon is achieved through experiments. Additional challenges arise under this setup as dependence of the employed procedures on time and resource intensive tasks often places limitations on their conduct. Optimal decision making while adhering to the imposed restrictions, commonly referred to as optimal design, is a predominant topic of study in this thesis.

A modern direction towards a more complex, systematic view of studied phenomena has been appearing in numerous scientific disciplines such as Biology (Klipp et al., 2016), Ecology (Ricklefs et al., 1993), Physics (Pickup et al., 2005). This approach is thought to be key for better understanding and analysing such intricate, dynamic systems. Conveyance of these structures relies on highly sophisticated mathematical models, inducing computationally intensive procedures that, often obstruct the optimal design of experiments under traditionally adopted methodologies due to the incurred, unrealistic waiting times (Drovandi and Pettitt, 2013; Overstall et al., 2019). As a result, current studies are often limited to suboptimal experimental designs (Ryan et al., 2014; Ryan, 2003) or to consideration of rather simplistic models that do not adequately capture the behaviour of the examined system (Long et al., 2013; Overstall et al., 2017). Such issues

are more closely considered in the subsequent chapters, outlining the need for a more efficient experimental design framework that allows the truly optimal study of modern scientific systems. Accomplishing this task is essential to progressing our understanding of natural phenomena.

The work presented in this thesis aims to advance existing experimental design methodology towards the truly optimal study of natural phenomena through the following main contributions:

- development of efficient estimation methodology for evaluation of the expected utility, an optimality criterion that constitutes a vital component of the adopted decision-theoretic framework. Estimation of this quantity is typically necessary, however, traditionally adopted methodology has been shown to become highly inefficient under the targeted class of problems (Ryan et al., 2016). The proposed approach attempts to overcome these issues through a combination of variational approximation and non-parametric techniques, providing a highly efficient and computationally tractable estimator that — unlike existing methodology — does not rely on cumbersome and computationally demanding procedures. The presented methodology constitutes a novel contribution to the research field of optimal experimental design as the proposed estimator has never been considered in this context before. The significant advancement incurred from its employment has the potential to transform modern scientific research, allowing the truly optimal study of natural phenomena through experiments within realistic timelines.

- Establishment of a comprehensive framework for optimal experimental design on the basis of the proposed efficient estimator. The adopted setup provides an automated, closed-loop process composed of the stages typically employed in a research study: data collection, knowledge update and decision making. At each new cycle of the proposed procedure, experimental data obtained up to that point are incorporated into the study, guiding future decisions. This setup allows the full exploitation and most efficient allocation of experimental resources as the produced data serve two purposes: 1) answering different research questions of interest such as model inference or prediction tasks as per their initially intended use. 2) In addition,

already obtained data can be incorporated into the study during the knowledge update stage of the adopted procedure and thus facilitate better-informed optimal decisions regarding subsequent experiments. The proposed sequential and adaptive process poses a further contribution of this thesis towards the development of a truly optimal experimental design framework for the study of natural phenomena.

- Integration of efficient optimisation procedures, through the class of Bayesian optimisation algorithms, for maximisation of the expected utility over the experimental design space in order to identify the set of optimal experimental conditions. Although similar work exists in the literature, Bayesian optimisation algorithms have only been considered to a very small extent within the optimal experimental design literature despite their remarkable suitability to this class of problems. In addition, this class of optimisation algorithms has never been considered within the adopted sequential, adaptive framework and in conjunction with the proposed variational estimator. Unlike alternative optimisation procedures, Bayesian optimisation provides a highly efficient setup, accomplishing a systematic and timely search of the design space, producing truly optimal solutions. On the contrary, an incomplete search is likely to induce suboptimal experimental designs, having failed to consider a sufficiently broad range of possible options.

The following outline is adopted: Chapter 2 provides an introduction to the components and formulation of a decision problem, introducing notions such as the utility and expected utility of a particular decision. Optimal experimental design problems are subsequently presented from a decision-theoretic perspective and are particularly examined in the context of model discrimination problems.

A review of currently adopted methodologies for the solution of optimal design problems is presented in Chapter 3 and common, associated challenges. Particular focus is placed on issues arising in studies of complex phenomena due to their reliance on computationally demanding mathematical models. The inability of existing methods to tackle ongoing challenges and the need for more efficient approaches, better-suited to optimal experimental design problems under this setup are outlined.

This problem is further considered in Chapter 5, wherein a novel estimation method,

addressing ongoing issues, is proposed. The adopted estimator is highly efficient and reliant on less computationally intensive procedures than traditionally adopted methods, without compromising the quality of the produced estimates. The presented approach has not been previously considered in the context of optimal experimental design.

The proposed estimator paves the way to a fully automated optimal experimental design framework for the scientific study of natural phenomena within realistic time scales, an endeavour that had been previously hindered by the challenges associated with currently adopted methodologies. The considered framework is examined in Chapter 6, establishing a closed-loop setup of data collection, knowledge update and optimal decision making for the study of modern phenomena through experiments that requires minimal input from the researcher. As review of a considerably large collection of potential decisions may, often, be required for optimal decision making, the efficient class of Bayesian optimisation algorithms and its integration to experimental design problems are also presented therein.

Chapter 7 proceeds to examine implementation of the considered methodologies for optimal experimental design on a commonly employed benchmark study of model discrimination. Simple polynomial models are adopted in this study, allowing an initial assessment of the competing estimators on a case example where an arbitrarily accurate representation of the true estimated value can be established. This comparison is typically not possible under more complex models.

Optimal experimental design for model inference in Systems Biology is subsequently considered in Chapter 8. Description of the observed system relies on sophisticated and therefore, computationally demanding models. This case study constitutes an initial case example under which, traditionally adopted methods typically fail. The improvement incurred from implementation of the proposed approach in comparison with the existing methodology is explored. The proposed estimation method is subsequently exploited to establish a sequential and adaptive framework, providing an efficient and fully automated setup for the study of biochemical systems.

Application of the proposed framework to another real-life problem is considered in Chapter 9, studying the kinetics of fluorescent molecules in heterogeneous environments. The studied problem represents a case example in which the computational cost of tradi-

tionally adopted methods is prohibitive and allows assessment of the examined methodology on a high dimensional output space.

The thesis concludes with a summary and a brief discussion on possible directions of future work in Chapter 10.

Chapter 2

Background

This chapter provides an introduction to optimal experiment design from a decision-theoretic perspective and sets the foundation for the methodological work presented in this thesis.

2.1 Introduction to decision theory

The complexity of modern phenomena presents future decisions with numerous possibilities. The challenge of discriminating among potential decisions under uncertainty introduced by factors that are unknown but, nonetheless, impact the observed outcome forms a **decision problem**.

Decision problems appear in many forms: they may describe a resource allocation task, the treatment assignments in clinical trials or the experimental conditions entailing the study of natural phenomena. This section provides an introduction to the fundamental ideas involved in the formulation and solution of a decision problem: Section 2.1.1 examines the individual elements composing a decision problem, Section 2.1.2 presents an optimality criterion for comparison of potential outcomes while a criterion quantifying the benefit associated with a particular decision is established in Sections 2.1.3 and 2.1.4.

2.1.1 Decision problems

Regardless of the system under study, consideration of a decision problem relies on interpretation of potential outcomes along the lines of an assumed underlying structure. Such assumptions are often expressed in the form of a mathematical model $f(\cdot \mid \boldsymbol{\theta}, \boldsymbol{d})$ where $\boldsymbol{\theta}$ and $\boldsymbol{d}$ represent parameters determining the behaviour of the system.

The distinction between parameters $\boldsymbol{\theta}$ and $\boldsymbol{d}$ reflects their distinct nature. The set of **model parameters** $\boldsymbol{\theta}$ is introduced to address the uncertainty around different aspects of the studied system and includes factors that may impact its behaviour but are beyond ones control. On the other hand, $\boldsymbol{d}$ refers to the controllable factors which can be viewed as a collection of potential **decisions** associated with different observed outputs. Particularly, considering the previously examined example in which an observed system is studied through experiments, model parameters $\boldsymbol{\theta}$ may refer to unknown factors such as the interactions between the system components while $\boldsymbol{d}$ to controllable elements, for instance the temperature under which a system is observed.

In light of an observed phenomenon captured in the form of dataset $\boldsymbol{y}$, $f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{d})$ quantifies the degree to which these observations agree with ones imposed assumptions. In a statistical context, evaluation of $f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{d})$ is often referred to as the **likelihood** of observing dataset $\boldsymbol{y}$ under the assumed structure f and parameters $\boldsymbol{\theta}$ and $\boldsymbol{d}$.

The following key components thus form the basis of any decision problem:

- an **observed dataset** $\boldsymbol{y} \in \mathcal{Y}$,
- **model** $f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{d})$ parametrised by $\boldsymbol{\theta} \in \Theta$,
- a **decision** $\boldsymbol{d} \in \mathcal{D}$.

Each decision $\boldsymbol{d}$ and choice of model parameters $\boldsymbol{\theta}$ are associated with a distinct output $\boldsymbol{y}$ and so from a decision-theoretic perspective, each potential scenario $(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})$ can be viewed as a unique **event** $\boldsymbol{e}$ from the event space $\mathcal{E} \subseteq \Theta \times \mathcal{Y} \times \mathcal{D}$.

In its simplest form, an **optimal decision problem** seeks to identify the decision $\boldsymbol{d}$ that instigates the most beneficial event $\boldsymbol{e}$. This decision is considered as the optimal one and will be denoted by $\boldsymbol{d}^*$. Under uncertainty, a decision $\boldsymbol{d}$ may, however, be mapped to numerous possible parameters $\boldsymbol{\theta} \in \Theta$ and thus multiple potential outcomes $\boldsymbol{y} \in \mathcal{Y}$. The

problem is, in this case, restated to define $\boldsymbol{d}^*$ as the decision that is, on average, associated with the most beneficial events $\boldsymbol{e} = (\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})$. A formal expression of this formulation is established in Section 2.1.3, however a global optimality criterion enabling quantification and thus comparison of the benefit incurred from potential events is essential for this definition and, therefore, introduced in the following section.

2.1.2 Optimality assessment

Assumption of an optimal decision $\boldsymbol{d}^*$ implies the existence of an ordering between potential events, establishing that: in light of two events $\boldsymbol{e} = (\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})$ and $\boldsymbol{e}' = (\boldsymbol{\theta}', \boldsymbol{y}', \boldsymbol{d}')$, decision $\boldsymbol{d}$ under model parameters $\boldsymbol{\theta}$ and observed data $\boldsymbol{y}$ is more beneficial than decision $\boldsymbol{d}'$ under $\boldsymbol{\theta}'$ and resulting $\boldsymbol{y}'$ in the context of the problem under study. This statement will be symbolically represented by relation $\succ$ and so the representation $\boldsymbol{d} \succ \boldsymbol{d}'$ indicates that decision $\boldsymbol{d}$ incurs higher benefit over decision $\boldsymbol{d}'$. Under the assumed order, the optimal decision is such that fulfils:

$$\boldsymbol{d}^* \succeq \boldsymbol{d}, \text{ for all } \boldsymbol{d} \in \mathcal{D} \ . \tag{2.1}$$

Establishing an order among potential decisions requires the consideration of multiple contributing factors and so comparison of different alternatives is a complex and often unintuitive procedure. The problem is simplified with the definition of a global criterion, often referred to as the **utility function** u, that quantifies this ranking through the relationship:

$$\boldsymbol{d} \succeq \boldsymbol{d}' \iff u(\boldsymbol{e}) \geq u(\boldsymbol{e}') \ . \tag{2.2}$$

In other words, the more beneficial the decision of a particular event $\boldsymbol{e}$ is, the higher its corresponding utility $u(\boldsymbol{e})$ will be too.

Relationship (2.2) maps each decision $\boldsymbol{d}$ to only one event $\boldsymbol{e}$ thus failing to address the presence of uncertainty in most decision problems. This issue is formally addressed in Sections 2.1.3 and 2.1.4. Definition of the utility function is subjective and relies of the decision-maker's interpretation of the 'benefit' and the targeted objective as illustrated in Section 2.2.3.

2.1.3 Expected utility from a Bayesian perspective

The previously introduced utility function is essential to the formulation of optimal decision problems, establishing a ranking between potential events, incurred by their corresponding decisions. However, model parameters are unknown factors of a decision problem with each potential realisation of $\boldsymbol{\theta}$ leading to a distinct event $\boldsymbol{e}$. As a result, a decision may in reality be linked to numerous potential events. The optimal decision, under this setup, is defined as the decision associated with the events incurring the highest average benefit. This section formalises this convention by introducing the idea of the **expected utility** in a Bayesian context. This criterion is briefly explored in Section 2.1.4 from a frequentist viewpoint.

In a Bayesian framework, uncertainty over $\boldsymbol{\theta}$ is captured by an assumed distribution with corresponding density function $p(\boldsymbol{\theta})$ that reflects the decision-maker's prior beliefs over the possible values of parameter vector $\boldsymbol{\theta}$. The assumed distribution is subjective, often decided based on previous information and expert knowledge without relying on observation of $\boldsymbol{y}$ and is referred to as the **prior distribution**.

Similarly, the distribution over $\Theta \times \mathcal{Y}$ corresponding to a particular decision $\boldsymbol{d}$ is known as the **joint distribution**, denoted by P and has a corresponding probability density function (p.d.f.) $p(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})$.

In summary, assessing the average benefit incurred from a decision $\boldsymbol{d}$ relies on consideration of the following characteristics associated with each potential event $(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})$:

- the utility $u(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})$ of the event,
- the probability $p(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})$ of its occurrence.

A summary of this process is subsequently obtained in the form of the expected utility of each decision $\boldsymbol{d}$:

$$\begin{aligned} U(\boldsymbol{d}) &= \mathbb{E}\left[u\left(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d}\right)\right] \\ &= \int u\left(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d}\right) p\left(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d}\right) \, \mathrm{d}\left(\boldsymbol{\theta}, \boldsymbol{y}\right) \\ &= \int u\left(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d}\right) f\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{d}\right) p(\boldsymbol{\theta}) \, \mathrm{d}\left(\boldsymbol{\theta}, \boldsymbol{y}\right) \ . \end{aligned} \tag{2.3}$$

While the term $p(\boldsymbol{\theta})$ is technically chosen to reflect the choice of $\boldsymbol{d}$, it is generally assumed that the prior assumptions are the same for each case. However, if there is evidence to support the contrary, then expression $p(\boldsymbol{\theta} \mid \boldsymbol{d})$ should be used instead.

It may also be, in some cases, sensible to assume that not all events occur at the same cost. This is often addressed with the use of a cost function $c(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})$ quantifying the resources required for observation of event $(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})$. Expression (2.3) can in such cases be modified accordingly and so the expected utility takes the following form:

$$U(\boldsymbol{d}) = \mathbb{E}\left[u(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d}) - c(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})\right] .$$

Henceforth, the cost function will be omitted unless there are reasons to believe that the cost among different events varies. In any case, the optimal decision will be chosen as the one incurring the highest utility over $\mathcal{D}$ or:

$$\boldsymbol{d}^* = \arg\max_{\mathcal{D}} U(\boldsymbol{d}) . \tag{2.4}$$

Decision-making may often be achieved in more than one stages where newly obtained information is used to assist in future decisions taking the form of an iterative procedure. These are often referred to as **adaptive** or **sequential designs**. After observation of the system, new data can be incorporated through the likelihood into the **posterior distribution** of $\boldsymbol{\theta}$ via Bayes' theorem:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{d}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{d}) p(\boldsymbol{\theta})}{p(\boldsymbol{y} \mid \boldsymbol{d})} . \tag{2.5}$$

The posterior distribution represents ones updated beliefs in light of new knowledge. The term $p(\boldsymbol{y} \mid \boldsymbol{d})$ will be referred to as the **marginal likelihood** of $\boldsymbol{y}$ and provides a summary of the likelihood of $\boldsymbol{y}$ averaged over Θ:

$$p(\boldsymbol{y} \mid \boldsymbol{d}) = \int f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{d}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} . \tag{2.6}$$

Consideration of a prior distribution over $\boldsymbol{\theta}$ has been much criticised in the sense that choice of it is subjective and construction of the prior is artificial rather than naturally

appearing in a problem. As Robert (2007) stresses on these claims, incorporating prior probabilities into a problem under study intends to summarise the current knowledge one has (or does not have) about the parameters rather than to implicitly state that the parameters have been generated from the prior distribution. It is also important to consider that in a decision-theoretic framework one is often dealing with problems under uncertainty and so subjective thinking is the only tool available until more information becomes known. In fact, the ability to incorporate ones informed views and logical reasoning seems like a natural process in any real-life decision-making problem. Most importantly, choice of the utility function is itself subjective and relies on the decision-maker's critical thinking and interpretation of the problem under study. More specifically, as Jaynes (2003) and numerous other authors demonstrate, choice of the prior can have as much of an effect in the resulting optimal decision as that of a utility function and so considering the idea of prior distributions arbitrary and invalid should motivate one to reject decision theory altogether.

Although these arguments are well-known among the scientific community, many researchers appear sceptical towards the Bayesian viewpoint and prefer alternative approaches. The frequentist approach is briefly considered in Section 2.1.4.

2.1.4 Expected utility from a frequentist perspective

In Section 2.1.3, uncertainty over $\mathcal{E}$ was addressed, in a Bayesian context, by assigning a prior distribution to the unknown parameters and marginalising the utility over the corresponding space to obtain the expected utility of a particular decision $\boldsymbol{d}$. In a frequentist framework, rather than considering a range of possible values for $\boldsymbol{\theta}$, one 'optimal' value $\boldsymbol{\theta}^*$ is used instead. Optimality is defined in terms of the likelihood function and so, $\boldsymbol{\theta}^* = \arg\max_{\Theta} f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{d})$. The optimal $\boldsymbol{\theta}^*$ is commonly referred to as the **Maximum Likelihood Estimate** (MLE). Under this setup, the expected utility finds expression in:

$$U(\boldsymbol{d}) = \int u(\boldsymbol{\theta}^*, \boldsymbol{y}, \boldsymbol{d})\, f(\boldsymbol{y} \mid \boldsymbol{\theta}^*, \boldsymbol{d})\, \mathrm{d}\boldsymbol{y} \ , \tag{2.7}$$

Considering expressions (2.3) and (2.7), two possible interpretations of the expected utility arise respectively:

- from a Bayesian viewpoint, it expresses the weighted average of the utility function over space $\mathcal{Y}$, *averaged over all possible values* of $\boldsymbol{\theta}$,

- from a frequentist viewpoint, the expected utility represents the weighted average of the utility function over space $\mathcal{Y}$, *evaluated at one particular value* of $\boldsymbol{\theta}$ that is considered as the optimal one.

As opposed to the Bayesian framework, a frequentist approach operates under the assumption of a 'true' value of $\boldsymbol{\theta}$ which, therefore, attempts to find. The expected utility can in that sense be conceived as the benefit of making a decision $\boldsymbol{d}$ repeatedly under the true value $\boldsymbol{\theta}^*$. As a result, $\boldsymbol{\theta}^*$ in expression (2.7) is fixed and there is no longer a need for marginalisation over the space Θ.

Contrarily, as previously discussed in Section 2.1.3, definition of the expected utility from a Bayesian standpoint treats $\boldsymbol{\theta}$ as an unknown quantity and quantifies the uncertainty around its potential values by imposing a prior distribution on them. A value of $\boldsymbol{\theta}$ associated with an observation $\boldsymbol{y}$ that agrees with the assumed underlying structure f will be assigned a higher probability and similarly a value generating predictions that appear unlikely under f will be assigned a lower probability. When making a decision both values will be, nonetheless, taken into account with their corresponding probability acting as a weight on the amount of influence each value of $\boldsymbol{\theta}$ should have in reaching a decision. However, when adopting a frequentist approach, the average utility expresses the utility function evaluated over different noisy versions of a 'true' dataset $\boldsymbol{y}$ weighed by the frequency with which each version is observed.

As briefly discussed in the subsequent Chapter 3, optimal experimental design within a frequentist framework is, in its current state, restrictive and difficult to generalise to a broad class of problems, particularly those targeted in this thesis. On the contrary, the preferred Bayesian methods offer a particularly flexible and efficient setup for addressing such cases, as further demonstrated in the remaining of this thesis.

2.1.5 Existence of the expected utility as a decision-theoretic criterion

This section presents an axiomatic framework, ensuring the existence of the utility function along the lines of Section 2.1.2. The examined framework regulates the decision space, imposing a strict ordering among potential decisions. As property (2.2) would be meaningless without the existence of such an order, this framework is essential for guaranteeing the validity of the utility function as a decision-theoretic criterion and as a consequence, of the expected utility $U(\boldsymbol{d})$ under any of the two specifications examined in Sections 2.1.3 and 2.1.4.

Two different approaches are examined in this section: the ordering conditions established by Abraham Wald (Wald, 1950) and a more general framework laid by Frederic Bohnenblust, Lloyd Shapley and Seymour Sherman and later extended by David Blackwell in Blackwell (1953). The two approaches were independently developed during the same time. Although there is a significant overlap between the two views both explore areas not covered by their alternatives. The work presented in this thesis borrows elements from both frameworks: while Wald's system is predominantly followed, certain desirable intrinsic properties induced from Blackwell's interpretation are integrated too.

Wald's framework Wald's approach establishes an imposed ordering among possible events, expressed by Conditions 1-4. Such a setup guarantees the existence of $U(\boldsymbol{d})$ as an optimality criterion, as subsequently stated in Definition 1.

Under the assumption of the distinct events $\boldsymbol{e}$, $\boldsymbol{e}'$, $\boldsymbol{e}'' \in \mathcal{E}$ with associated decisions $\boldsymbol{d}, \boldsymbol{d}', \boldsymbol{d}'' \in \mathcal{D}$ and distributions P, P', P'' respectively, conditions 1-2 ensure the existence of a strict ordering among any two decisions thus imposing a clear distinction between them.

Condition 1. *Exactly one of the following three relations must hold:* $\boldsymbol{d} \prec \boldsymbol{d}'$, $\boldsymbol{d} = \boldsymbol{d}'$, $\boldsymbol{d} \succ \boldsymbol{d}'$.

Condition 2. *If* $\boldsymbol{d} \preceq \boldsymbol{d}'$ *and* $\boldsymbol{d}' \preceq \boldsymbol{d}''$, *then* $\boldsymbol{d}' \preceq \boldsymbol{d}''$.

Condition 3 ensures that when an order is established between two decisions $\boldsymbol{d}$ and $\boldsymbol{d}'$, then this order is not affected in light of a new, unrelated decision $\boldsymbol{d}''$.

Condition 3. $\boldsymbol{d} \prec \boldsymbol{d}'$, *if and only if,* $a\boldsymbol{d} + (1-a)\boldsymbol{d}'' \prec a\boldsymbol{d}' + (1-a)\boldsymbol{d}''$, *for any* $a \in (0,1)$.

Lastly, Condition 4 suggests that no ordering can be so extreme that it remains unaffected by any transformation applied to it. In other words no event can have such low (or high) utility that no matter how great an incentive (or hindrance) one assigns to it, it is still considered of the lowest (or highest) benefit.

Condition 4. *If* $\boldsymbol{d} \preceq \boldsymbol{d}' \preceq \boldsymbol{d}''$, *then there exist* $a \in (0,1)$ *and* $b \in (0,1)$ *such that:* $b\boldsymbol{d} + (1-b)\boldsymbol{d}'' \preceq \boldsymbol{d}' \preceq a\boldsymbol{d} + (1-a)\boldsymbol{d}''$.

The following definition emerges from the fulfilment of Conditions 1-4:

Definition 1. *(DeGroot, 1970) Let* $\boldsymbol{d}$ *and* $\boldsymbol{d}'$ *be two decisions with associated parameter spaces* $\Theta \times \mathcal{Y}$, $\Theta' \times \mathcal{Y}'$ *and distributions* P, P' *respectively. it will be said that, decision* $\boldsymbol{d}$ *is* ***preferred over*** *decision* $\boldsymbol{d}'$, *or* $\boldsymbol{d} \succeq \boldsymbol{d}'$, *if, and only if,* $U(\boldsymbol{d}) \geq U(\boldsymbol{d}')$.

Definition 1 indicates that, the relationship between alternative decisions is fully captured by their corresponding expected utilities.

Bohnenblust, Shapley and Sherman's framework A more general framework was developed independently by Bohnenblust, Shapley and Sherman in an unpublished work and was further developed by Blackwell (1953). The methodology focuses on ranking of experiments and underlies but is not limited to Wald's decision-theoretic approach. Bohnenblust et. al. use the loss incurred by the occurrence of a particular event as their ranking metric. A direct correspondence between loss and utility functions is established by defining the loss as the negative utility. Under Wald's decision-theoretic framework, optimality is, in that case, achieved when the minimum expected loss is attained. Bohnenblust, Shapley and Sherman propose that:

Definition 2. *(Bohnenblust, Shapley and Sherman, unpublished work) Decision* $\boldsymbol{d}$ *with corresponding space* $(\Theta, \mathcal{Y})$ *is* ***more informative*** *than decision* $\boldsymbol{d}'$ *with corresponding space* $(\Theta', \mathcal{Y}')$, *if for every* $(\boldsymbol{\theta}', \boldsymbol{y}') \in (\Theta', \mathcal{Y}')$ *incurring loss* $l(\boldsymbol{\theta}', \boldsymbol{y}', \boldsymbol{d}')$ *there exists a vector* $(\boldsymbol{\theta}, \boldsymbol{y}) \in (\Theta, \mathcal{Y})$ *such that* $l(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d}) = l(\boldsymbol{\theta}', \boldsymbol{y}', \boldsymbol{d}')$. *Relationship* $\overset{(\text{BSS})}{>}$ *will be used to denote that one decision is more informative than another under the given definition and so the previous sentence can be restated as:* $\boldsymbol{d} \overset{(\text{BSS})}{>} \boldsymbol{d}'$.

In a continuation of this work, Blackwell poses the question of sufficiency of the benefit incurred from a particular decision $\boldsymbol{d}$ in order to deem occurrence of decision $\boldsymbol{d}'$ non-important.

An illustration of 'sufficiency' from a Bayesian perspective is provided in the following example: when considering an estimation problem a decision $\boldsymbol{d}$ is sufficient if for every a priori distribution of parameters, the a posteriori distribution under $\boldsymbol{d}$ is the same as under the union of every possible decision $\boldsymbol{d} \in \mathcal{D}$. In other words, decision $\boldsymbol{d}$ captures as much information as would have been provided by every potential $\boldsymbol{d} \in \mathcal{D}$. Blackwell (1951) further proves that $\boldsymbol{d} \overset{(\text{BSS})}{>} \boldsymbol{d}'$ implies that $\boldsymbol{d} \overset{(\text{B})}{>} \boldsymbol{d}'$, where relationship $\overset{(\text{B})}{>}$ represents an ordering under Blackwell's extended framework.

The following result arising from Blackwell's axiomatic framework will play an important role in laying the foundation for the proposed methodology presented in Chapter 6. Blackwell (1953) shows that, sufficiency of a decision $\boldsymbol{d}$ can be translated in terms of the variability present in the corresponding distribution. Particularly, $\boldsymbol{d}$ is sufficient for decision $\boldsymbol{d}'$ when the corresponding distribution P is more variable than P'. Variability of a distribution is reflected through (2.3), with Blackwell introducing the additional assumption of a convex utility function. The idea behind this is that, as convex functions obtain larger values over extreme regions, the resulting expected utility $U(\boldsymbol{d}) = \mathbb{E}[\varphi(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{d})]$ acts as a measure of dispersion of distribution P. Certain choices of φ result to a well-known dispersion measure known as the $f-$divergence (Ali and Silvey, 1966). Under this setup, decision $\boldsymbol{d}$ inholding lower $f-$divergence compared to decision $\boldsymbol{d}'$ implies that there exists a set of prior distributions such that $\boldsymbol{d} \overset{(\text{BSS})}{>} \boldsymbol{d}'$ and therefore $\boldsymbol{d} \overset{(\text{B})}{>} \boldsymbol{d}'$. In addition, convexity of φ ensures that the returned optimal point will be global rather than local. The idea of using $f-$divergences as interpretations of the expected utility is revisited in Chapter 5.

2.2 Introduction to experimental design

This section introduces a class of decision problems concerned with the optimal design of experiments. Section 2.2.1 focuses on the individual components of an experimental design problem while Section 2.2.2 examines its connection to decision problems and

employs the previously introduced formulation to tackle them. In Section 2.2.4 focus is placed on optimal experimental design targeting model discrimination problems.

2.2.1 Experimental design

In studies of natural phenomena, direct observation of a studied system is often not possible and so inference based on observations obtained from their different components through experiments holds an essential part. The obtained knowledge may subsequently be used to facilitate statistical analysis such as parameter inference, hypothesis testing, model selection or prediction.

Modern scientific research is often concerned with analysis of complex phenomena characterised by numerous components and their interactions. Obtaining measurements from the different components typically relies on experimental procedures that are associated with numerous limiting conditions such as employment of resource intensive practices (Ryan, 2003) or adherence to regulatory requirements (Overstall et al., 2019), placing constraints on the observations available for consideration. Optimal experiment design methods can thus be employed to optimally allocate the available resources under these restrictions.

In a setup similar to the decision-theoretic framework of Section 2.1, the set of parameters consists of controllable and random unknown components. The latter group is assumed to be beyond the experimenter's control and are incorporated into the study as model parameters $\boldsymbol{\theta} \in \Theta$. On the contrary, the former refers to **experimental conditions** that are purposely set to specific values, with interest lying on observation of their impact to the studied phenomenon. Optimal experimental design, thus, seeks the condition under which the most beneficial impact relative to the studied objective is achieved while accounting for the uncertainty present through $\boldsymbol{\theta}$. The vector of experimental parameters will be denoted by $\boldsymbol{\delta} \in \Delta$ while the experimental condition incurring the highest benefit, also referred to as the **optimal experimental condition**, by $\boldsymbol{\delta}^*$.

2.2.2 Optimal experimental design as a decision problem

The foundations of the experimental design framework, from a decision-theoretic perspective were first laid by Wald (1950) and Schlaifer and Raiffa (1961) and were further established by the review of Lindley (1972).

Summarising from the previous section, an experiment refers to the process of observing a produced output related to a particular set of experimental conditions and model parameters. Each distinct experiment can thus be thought of as an *event* $\boldsymbol{e} = (\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\delta})$. The output $\boldsymbol{y}$ depends on choice of experimental condition $\boldsymbol{\delta}$ and so, selection of a particular $\boldsymbol{\delta}$ poses a *decision*.

Adopting the decision-theoretic framework of Section 2.1, the benefit incurred from a particular experiment will be quantified through a utility function u evaluated for each experiment $(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\delta})$. Finally, to account for the uncertainty over $\Theta \times \mathcal{Y}$ the expected utility $U(\boldsymbol{\delta})$ in (2.3) is used to assess the optimality of a potential experimental condition $\boldsymbol{\delta}$.

Unfortunately, analytical evaluation of $U(\boldsymbol{\delta})$ is only possible in simple problems, for example, those considering conjugate priors and so a considerable part of the relevant literature is devoted to the problem of efficient expected utility estimation. Chapter 3 explores available estimation methods as well as challenges involved in existing approaches while Chapter 5 proposes a new approximation method tailored to model selection problems that has not been previously considered in the context of experimental design.

2.2.3 Common utility functions

The subjectivity in choice of the utility function and its dependence on ones objectives has been previously considered in Section 2.1.2. Some examples of utility functions, commonly used in experiment design are provided below:

- The **0-1 utility** is frequently adopted in the context of hypothesis testing. Assuming that the truth for a hypothesised situation is known, the 0-1 utility function indicates whether, under a particular event $e = (\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\delta})$, the favoured hypothesis $H(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\delta})$ represents the true H_{true} or not. A mathematical expression of this

procedure takes the form:

$$u_{0-1}(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\delta}) = \mathbb{I}[\, H(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\delta}) = H_{\text{true}} \,], \tag{2.8}$$

where $\mathbb{I}$ represents the indicator function, assigning the value of 1 when H_{true} is successfully identified and 0 otherwise. The expected utility, thus, effectively represents the probability of a hypothesis test being successful under a particular experimental condition $\boldsymbol{\delta}$ and is as such sought to be maximised.

- The **relative entropy**, employed in parameter inference problems quantifies the additional new information, provided from an observed dataset $\boldsymbol{y}$. This criterion was considered by Lindley (1956) for design of experiments and is defined as:

$$u_{RE}(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\delta}) = \log \frac{p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\delta})}{p(\boldsymbol{\theta})} \ . \tag{2.9}$$

  Under the assumption of equal costs for observation of each produced output $\boldsymbol{y}$, interest lies in maximising the information gain. Notably, the expected utility induced by u_{RE} is equivalent to the Kullback-Leibler divergence between the posterior and prior distribution of $\boldsymbol{\theta}$. A slightly modified version of this utility function can be used in prediction problems to quantify the amount of additional information obtained from observing a new dataset compared the present state.

- A different perspective for parameter inference problems quantifies gain in the form estimate accuracy in light of a dataset $\boldsymbol{y}$. The utility quantifying this improvement is known as the **posterior precision**, expressed as:

$$u_{PP}(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\delta}) = \frac{1}{\text{Var}\,[\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\delta}]} \tag{2.10}$$

  As the variance of the posterior distribution decreases, the higher the utility of the corresponding experiment will be.

2.2.4 Optimal experimental design for model discrimination problems

As demonstrated in Section 2.2.3, treatment of a design problem is highly dependent on the objective of the study. Atkinson (2008) and numerous other authors have previously considered the idea of a utility suitable to multiple objectives but it is often shown that a global design strategy is suboptimal. The scope of this thesis is predominately focused on model discrimination problems which are employed by the case studies presented in Chapters 7, 8 and 9. As a consequence, emphasis is henceforth placed on utility functions and ongoing challenges associated with optimal experimental design in the context of this class of problems.

Model discrimination refers to the process of constructing some initial hypotheses, formalised by mathematical models, and assessing their ability to capture the observed system behaviour. The hypothesis under which the model predictions agree the most with the actual measurements of the observable components describes more suitably the studied phenomenon.

Formulation of experimental design problems has, so far, incorporated the following key components: the observed experimental output $\boldsymbol{y}$, the unknown experimental parameters $\boldsymbol{\theta}$ and the experimental conditions $\boldsymbol{\delta}$. As multiple models are adopted in model selection problems, an additional parameter m is included as an indicator of the model under consideration from a set of competing models $\mathcal{M}$. The probability distribution associated with each candidate model will be denoted by P_m. This term has been omitted in previous sections as the consideration of more than one models was not required. The optimal experimental condition $\boldsymbol{\delta}^*$ is expressed in the form of (2.4) where:

$$U(\boldsymbol{\delta}) = \sum_{m \in \mathcal{M}} \left[\int u(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\delta}, m)\, f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\delta}, m)\, p(\boldsymbol{\theta})\, \mathrm{d}(\boldsymbol{\theta}, \boldsymbol{y}) \right] p(m) \,, \tag{2.11}$$

for a suitable utility function.

In the Bayesian experimental design literature, a common expression of the utility function for model discrimination problems is the **Shannon entropy** (Box and Hill,

1967), defined as:

$$u\left(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\delta}, m\right) = \sum_{\substack{m' \in \mathcal{M}, \\ m' \neq m}} \log \frac{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m)}{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m')} \; p(m') \;, \tag{2.12}$$

where $p(\boldsymbol{y} \mid \boldsymbol{\delta}, m)$ expresses the marginal likelihood previously introduced in (2.6).

This choice of utility function possesses several desirable properties. The marginal likelihood ratio in (2.12) is a commonly used criterion in model selection problems, often referred to as the Bayes' factor (Kass and Raftery, 1995). The Bayes' factor $B_{\boldsymbol{\delta}}(m, m') = \frac{p(\boldsymbol{y}|\boldsymbol{\delta},m)}{p(\boldsymbol{y}|\boldsymbol{\delta},m')}$ represents the evidence in favour of model m given data $\boldsymbol{y}$ at experimental condition $\boldsymbol{\delta}$. The expected utility can thus be interpreted as the average evidence provided by the data in support of m when $\boldsymbol{y}$ has actually been generated by model m (see equation (2.13)). Hence, the highest the evidence in favour of the resulting data, the more preferable the corresponding experimental condition will be. Interestingly, the corresponding expected utility bares meaningful information not only for comparison with alternative experimental conditions but for the usefulness of the design itself. Following Jeffreys (1961)'s guidelines, one can conclude not only the optimal condition but how useful the dataset obtained at this condition is. This is particularly useful for cases when the design is performed sequentially and early stopping is an option. When considering the next stage of the design it may be concluded that the data obtained in $\boldsymbol{\delta}^*$, although optimal, do not offer any significant new information to the study.

Drawbacks are mostly attributed to the computational burden involved in the use of the Shannon entropy (Ryan et al., 2014) as will be further demonstrated in Chapter 3. Chapter 5 presents an efficient approximation method that is shown to tackle currently persisting challenges.

Alternative utilities include the **total separation** proposed by Roth (1967) that is defined as the absolute distance between the posterior predictive means of the competing models. Although this utility offers the desirable characteristic of relatively low computational requirements, it has a reportedly high chance of generating misleading results by only extracting information from the mean and thus ignoring the overall shape of the predictive distributions. For this reason, this alternative will not be considered further in this thesis.

Under (2.12), the expected utility of a particular experimental condition takes the form:

$$
\begin{aligned}
U(\boldsymbol{\delta}) &= \sum_{m\in\mathcal{M}} \left\{ \int \left[\sum_{\substack{m'\in\mathcal{M},\\ m'\neq m}} \log \frac{p(\boldsymbol{y}\mid\boldsymbol{\delta},m)}{p(\boldsymbol{y}\mid\boldsymbol{\delta},m')}\; p(m')p(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\delta},m) \right] \mathrm{d}(\boldsymbol{\theta},\boldsymbol{y}) \right\}\; p(m) \\
&= \sum_{m\in\mathcal{M}} \sum_{\substack{m'\in\mathcal{M},\\ m'\neq m}} \left\{ \int \left[\log \frac{p(\boldsymbol{y}\mid\boldsymbol{\delta},m)}{p(\boldsymbol{y}\mid\boldsymbol{\delta},m')} p(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\delta},m) \right] \mathrm{d}(\boldsymbol{\theta},\boldsymbol{y}) \right\}\; p(m')\; p(m) \\
&= \sum_{m\in\mathcal{M}} \sum_{\substack{m'\in\mathcal{M},\\ m'\neq m}} \left\{ \int \left[\log \frac{p(\boldsymbol{y}\mid\boldsymbol{\delta},m)}{p(\boldsymbol{y}\mid\boldsymbol{\delta},m')} p(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\delta},m) \right] \mathrm{d}(\boldsymbol{\theta},\boldsymbol{y}) \right\}\; p(m')\; p(m) \\
&= \sum_{m\in\mathcal{M}} \sum_{\substack{m'\in\mathcal{M},\\ m'\neq m}} \left\{ \int \log \frac{p(\boldsymbol{y}\mid\boldsymbol{\delta},m)}{p(\boldsymbol{y}\mid\boldsymbol{\delta},m')} \left[\int p(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\delta},m)\mathrm{d}\boldsymbol{\theta} \right] \mathrm{d}\boldsymbol{y} \right\}\; p(m')\; p(m) \\
&= \sum_{m\in\mathcal{M}} \sum_{\substack{m'\in\mathcal{M},\\ m'\neq m}} \left\{ \int \log \frac{p(\boldsymbol{y}\mid\boldsymbol{\delta},m)}{p(\boldsymbol{y}\mid\boldsymbol{\delta},m')} p(\boldsymbol{y}\mid\boldsymbol{\delta},m)\mathrm{d}\boldsymbol{y} \right\}\; p(m')\; p(m) \\
&= \sum_{m\in\mathcal{M}} \sum_{\substack{m'\in\mathcal{M},\\ m'\neq m}} \mathrm{KL}_{\boldsymbol{\delta}}(m,m')\; p(m')\; p(m)\;, \qquad (2.13)
\end{aligned}
$$

where:

$$
\mathrm{KL}_{\boldsymbol{\delta}}(m,m') = \int \log \frac{p(\boldsymbol{y}\mid\boldsymbol{\delta},m)}{p(\boldsymbol{y}\mid\boldsymbol{\delta},m')} p(\boldsymbol{y}\mid m,\boldsymbol{\delta})\; \mathrm{d}\boldsymbol{y}\;. \qquad (2.14)
$$

The term defined in (2.14) represents the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the predictive distributions of the competing models at experimental condition $\boldsymbol{\delta}$. Intuitively, the larger the discrimination between predictive distributions the easier the classification of a newly obtained dataset will be.

Consideration of divergence metrics for optimal design is commonly encountered in the relevant literature. Vanlier et al. (2014) adopt a weighted KL divergence between the predictive distributions of competing models while Bingham and Chipman (2002) consider the Hellinger distance instead in a similar setup. The class of estimators proposed in this thesis is suitable to both choices of utility functions.

Summary

This chapter provides an introduction to optimal experiment design problems from a decision-theoretic perspective. Sections 2.1 and 2.2 set the foundations for the formulation of design problems, introducing their individual components: the set of potential experimental conditions (decisions), a set of models under consideration with their corresponding set of model parameters and the resulting output. Optimality assessment of each experimental condition is subsequently addressed through the notions of the utility function and the expected utility, examined in Sections 2.1.2, 2.1.3 and 2.1.4. The subjectivity involved in the choice of utility function is discussed in Section 2.1.2 while alternative options tailored to different objectives are examined in Section 2.2.3. Section 2.2.4 considers the special case of experimental design for studies that employ model discrimination methods. Such case examples constitute the main interest of this thesis and will be further considered in Chapters 8 and 9. Utility functions tailored to these problems are discussed, with particular emphasis on the Shannon entropy and alternative divergence measures.

The succeeding chapter provides a review on the commonly adopted methodology for experiment design optimisation focusing on two ongoing challenges: estimation of the expected utility when an expression for it is not available in closed form and efficient optimisation of the expected utility over the design space.

Chapter 3

Review of Bayesian optimal experimental design methods

This chapter provides a review of the most current and commonly adopted methodologies for tackling experimental design problems and relevant ongoing issues. A brief summary of the earlier work and the challenges that have since then arisen are initially considered. Two key problems are identified: the estimation of the expected utility when an analytical expression is not available and the efficient maximisation of this quantity over the design space. Throughout this chapter, the presented methodologies are predominantly discussed from the perspective of model discrimination problems while emphasis is placed on cases when model evaluations are computationally demanding.

3.1 Early work and key challenges

During the early years of Bayesian experimental design (from Wald's seminal work in 1950 (Wald, 1950) until the 1980s) implementation of the decision-theoretic approach, introduced in Chapter 2, targeted mainly simpler problems for which an analytical expression of the expected utility exists. One of the main contributors to the developed methodology for these problems was Kiefer (1959). As the frequentist viewpoint was prevalent in the field at that time, the Bayesian approach was considered by many simply as an extension of the frequentist case (Pukelsheim, 1980). Eventually, the distinction

between Bayesian and frequentist methods became evident when interest was extended to more complex problems with the former adapting flexibly while the latter restricting to identification of only locally optimal designs. Extension to a wider class of problems even in a Bayesian setup, nonetheless, entails two main challenges:

1) Evaluation of the expected utility: an analytical expression for the expected utility is typically not available and so numerical approximation methods are required instead. Shortcomings of currently adopted methods are predominantly attributed to their computational complexity, their inability to scale up to high dimensions or the restrictive assumptions they impose. The majority of the existing work tackles the estimation problem using Monte-Carlo integration based approaches which are introduced in Section 3.2.1. Numerous adaptations have been proposed, however, reportedly, none lowers the computational burden in such a degree that makes it suitable for consideration of complex systems similar to those studied in this thesis.

2) Optimisation over the design space: maximisation of the expected utility becomes increasingly challenging as the number of designs under consideration grows. Traditional grid search or random search approaches quickly add up to an infeasible number of calculations thus becoming highly inefficient, often resulting to highly suboptimal designs.

Currently adopted methods can be divided into three main categories based on the challenge being targeted: estimation of the expected utility, optimisation over the design space or both. The most common and relevant methodologies from the former category are considered in Section 3.2 while the remaining two are discussed in Section 3.3.

3.2 Evaluation of the expected utility

Early work considered relatively simplistic approaches for estimation of the expected utility such as normal approximations to the expected utility through the expected Fisher information matrix (Atkinson and Donev, 1992; Silvey, 1980). Alternative methods include discretisation of the prior and consequently averaging over the utility evaluated

at each distinct point or numerical integration (D'Argenio, 1990; Pronzato and Walter, 1985). However, such approaches have often been shown to either assume an oversimplified structure for the studied system or to quickly add up to an infeasible number of calculations as the dimensionality of the problem increases (Chaloner and Verdinelli, 1995; Ryan, 2003).

The idea of Monte-Carlo estimation is presented in Section 3.2.1 as it is currently prevalent in the Bayesian experimental design literature. Its implementation for expected utility estimation is subsequently considered while challenges relevant to model selection problems are addressed in Section 3.2.2. A brief overview of alternative estimation methods is provided in Section 3.2.3 and ongoing challenges associated with Monte-Carlo methods are discussed in Section 3.2.4.

3.2.1 Monte-Carlo integration

A brief introduction to Monte-Carlo methods and their role in Bayesian experimental design is provided in this section. The approach is initially presented in its general form while its application to expected utility estimation is considered subsequently.

Introduction to Monte-Carlo integration Monte-Carlo integration (MCI), first introduced by Metropolis and Ulam (1949), is an estimation method for integrals of the form:

$$I = \int \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \ , \tag{3.1}$$

where $\eta : \mathcal{X} \to \mathbb{R}$ represents a function with associated probability distribution P and corresponding probability density function p.

Under the examined approach, quantity (3.1) can be approximated by the average of function η evaluated at a large collection of N samples from $\mathcal{X}$ that are distributed according to P, taking the form:

$$\hat{I} \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\eta(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)} \ . \tag{3.2}$$

The idea behind estimator (3.2), is that the term $\hat{I}$ converges asymptotically to the true

quantity I, or formally:

$$\lim_{N\to\infty} \hat{I} = I \ . \tag{3.3}$$

In practice, an approximation of the ratio in expression (3.2) can be achieved by drawing a sample $\boldsymbol{x}_i$ according to P and calculating the corresponding $\eta(\boldsymbol{x}_i)$. The Monte-Carlo estimate is subsequently obtained by repeating this process N times and averaging over the resulting evaluations. Due to 3.3, the produced estimates fall arbitrarily close to the true value for a sufficiently large N. Two main properties of the Monte-Carlo estimator are noted below:

Property 1. *(Unbiased)* $\mathbb{E}\left[\hat{I}\right] = I$

Property 2. *(Consistent)* $\mathrm{Var}\left[\hat{I}\right] \xrightarrow{N\to\infty} 0$.

Proofs for Properties 1 and 2 are provided in Appendix A.

Monte-Carlo integration for estimation of the expected utility Monte-Carlo integration has been commonly employed in optimal experimental design problems for estimation of the expected utility (Ryan et al., 2016). This section is concerned with application of Monte-Carlo methods for expected utility estimation in the context of model discrimination problems, employing the Shannon entropy utility function (2.12). Under this setup, the expected utility finds expression in (2.13) that, for convenience, is restated below:

$$U(\boldsymbol{\delta}) = \sum_{m\in\mathcal{M}} \sum_{\substack{m'\in\mathcal{M},\\ m'\neq m}} \left\{ \int \log \frac{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m)}{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m')} p(\boldsymbol{y} \mid \boldsymbol{\delta}, m) \, \mathrm{d}\boldsymbol{y} \right\} \, p(m') \, p(m) \ .$$

Along the lines of (3.2), the corresponding Monte-Carlo estimator obtains the form:

$$\hat{U}(\boldsymbol{\delta}) = \sum_{m\in\mathcal{M}} \sum_{\substack{m'\in\mathcal{M},\\ m'\neq m}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[\log p(\boldsymbol{y}_i^{P_m} \mid \boldsymbol{\delta}, m) - \log p(\boldsymbol{y}_i^{P_m} \mid \boldsymbol{\delta}, m') \right] \right\} \, p(m') \, p(m) \ , \tag{3.4}$$

where $\boldsymbol{y}_i^{P_m} \sim p(y \mid \delta, m)$, $i = 1, \ldots, N$. Samples $\boldsymbol{y}_i^{P_m}$ can be acquired through $\left\{ \left(\boldsymbol{\theta}_i^{P_m}, \boldsymbol{y}_i^{P_m} \right) \right\}_{i=1}^{N}$ by disregarding the term $\boldsymbol{\theta}_i^{P_m}$.

Popularity of Monte-Carlo methods can be attributed to their wide applicability and

ease of implementation. Application of Monte-Carlo integration to optimal design problems is, however, often hindered by the considerable computational burden involved in obtaining sufficiently good approximations (Ryan et al., 2014; Ryan, 2003). Due to Property 2, the estimator is only asymptotically unbiased, meaning that a large sample size N is necessary to achieve unbiasedness in the obtained estimates. Moreover, evaluation of expression (2.11) of the expected utility, requires further evaluation of another integral for each of the N samples which may rely on MCI or alternative approximation methods. Thus, evaluation of $U(\boldsymbol{\delta})$ for only one potential experimental condition $\boldsymbol{\delta}$ requires $\mathcal{O}(cN^2)$ operations for a sufficiently large N and a constant c that accounts for the multiple models under consideration.

Under expression 3.4 of the expected utility, evaluation of the marginal likelihood $p(\boldsymbol{y}_i^{P_m} \mid \boldsymbol{\delta}, m)$ is required for each $\boldsymbol{y}_i^{P_m}$, $i = 1, \ldots, N$, $m \in \mathcal{M}$, which, often, relies on additional approximation methods. This issue is addressed in the subsequent Section 3.2.2.

3.2.2 Estimation of the marginal likelihood

An analytical expression of the marginal likelihood is, under most model specifications, not available. Additional estimation methods are, therefore, required for approximation of this quantity. In its simplest form, estimation can be achieved through an additional Monte-Carlo estimator under which, the marginal likelihood, previously defined in (2.6) and summarised below:

$$p(\boldsymbol{y}_i \mid \boldsymbol{\delta}, m) = \int p(\boldsymbol{y}_i \mid \boldsymbol{\theta}, \boldsymbol{\delta}, m) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \ ,$$

takes the form:

$$\hat{p}(\boldsymbol{y}_i \mid \boldsymbol{\delta}, m) = \sum_{j=1}^{M} p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_j, \boldsymbol{\delta}, m) \ , \tag{3.5}$$

where $\{\boldsymbol{\theta}_j\}_{j=1}^{M}$ represent samples from the prior distribution of Θ corresponding to model m. This approach is typically avoided because the prior distribution of model parameters completely disregards the properties of the likelihood surface, extracting the same amount of information from different samples regardless their corresponding probability

(Vyshemirsky and Girolami, 2008). As a result, a significantly large sample size M is required in order to adequately capture the shape of the approximated surface.

A modification addressing this issue considers the same estimator (3.5) with the inner term being evaluated at samples $\{\boldsymbol{\theta}_j \mid \boldsymbol{y}_i\}_{j=1}^{M}$ instead. Knowledge from the observed data is, thus, used to draw focus on higher density regions. Unfortunately, this revision comes at a high computational cost as for each observation $\boldsymbol{y}_i$, evaluation of the corresponding posterior distribution $\Theta \mid \boldsymbol{y}_i$ is required, typically obtained through sampling algorithms.

An alternative class of approaches includes the commonly adopted estimation methods of annealed importance sampling (Neal, 2001) and thermodynamic integration (Friel and Pettitt, 2008). In a comparison study of alternative marginal likelihood estimation methods, Vyshemirsky and Girolami (2008) show that, this class outperforms competing approaches such as the previously examined Monte-Carlo estimators. An additional advantage of these estimators is that Sequential Monte Carlo (SMC; Del Moral et al. (2006)) sampling techniques lend themselves to their implementation as they can be naturally obtained as by-products of SMC algorithms. An introduction to SMC methods is provided in Appendix B.

Briefly, sampling from the posterior distribution using thermodynamic integration is achieved by linking the prior and the targeted posterior distribution through a sequence of S intermediate un-normalised distributions Q_b, for $0 \leq b \leq 1$. Assuming that the intermediate distributions are fairly close to each other, this setup allows a smoother transition from the prior to the target distribution which tackles issues arising when the discrepancy between them is high. An estimate of the marginal likelihood is then obtained through:

$$\hat{p}(\boldsymbol{y}_i \mid \boldsymbol{\delta}, m) = \int_0^1 \sum_{j=1}^{\tilde{M}} p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_{j,b}, \boldsymbol{\delta}, m) \, \mathrm{d}b \, , \tag{3.6}$$

where $\{\boldsymbol{\theta}_{j,b}\}_{j=1}^{\tilde{M}}$ are samples from the intermediate distribution Q_b. SMC algorithms lend themselves to implementation of the examined estimator due to the reliance of both on samples from the sequence of S intermediate distributions. Once the population of samples $\{\boldsymbol{\theta}_{j,b}\}_{j=1}^{\tilde{M}}$ is available at each stage b of the SMC algorithm it can be further used for evaluation of the corresponding sum within (3.6).

Estimators of the same class have also been considered in McGree et al. (2012) and

Drovandi et al. (2013) in the context of Bayesian experimental design. McGree et al. (2012) report this approach to be computationally intensive and still of order $\mathcal{O}(cN^2)$, where N is the number of samples from the predictive distribution of each model $m \in \mathcal{M}$, and resort to parallel implementation in order to produce estimates within realistic timescales. This approach, combined with the Monte-Carlo estimator expressed in (3.4), is further considered in the case studies examined in Chapters 7 and 8 for evaluation of the expected utility as a representation of the state-of-the-art of currently employed methodologies for optimal experimental design.

3.2.3 Approximate methods

An alternative viewpoint is provided by Long et al. (2013) who propose an approximate approach for evaluation of the marginal likelihood through Laplacian approximations. This approach completely avoids posterior sampling, thus incurring significant computational savings, however imposes the assumption of normality which can often be restrictive and, potentially, unrealistic. Ryan et al. (2016) report that, Laplace approximations perform well in practice, particularly when large amounts of data are available, however suffer from the curse of dimensionality and are, therefore, restricted to only targeting low-dimensional problems. Given the considerable restrictions this approach imposes, it will not be considered further in this thesis.

Drovandi and Pettitt (2013) propose another approximate method that employs approximate Bayesian computation (ABC) techniques to avoid evaluation of the likelihood. This approach is mainly targeted to intractable likelihood problems which are not encountered in this thesis and are thus not of direct interest.

3.2.4 Current challenges

The main shortcoming of Monte-Carlo based approaches lies in their high computational complexity. As briefly demonstrated in this section, assessing the expected utility for only one experimental condition $\boldsymbol{\delta}$ relies on $\mathcal{O}(cN^2)$ operations where a sufficiently large N is required for the estimation bias to be deemed negligible. Each operation involves model evaluations (accounting for the intermediate stages of the SMC algorithms) which

is aggravated further in cases that evaluation of the studied models is itself demanding. Unfortunately, this is often the case in the studies of natural phenomena as highly sophisticated models are typically adopted in order to realistically capture the system behaviour. In particular, evaluation of a model describing a simple biochemical system, examined in Chapter 8, requires $1.12 \cdot 10^{-4}$ seconds. As a result, the time required for approximation of the expected utility corresponding to one experimental condition considering only this one operation is:

$$1.12 \cdot 10^{-4} \text{ s} \times N^2 \text{ samples } (100^2) \times \text{intermediate stages } (40) \times \text{models } (2) \approx \mathbf{748 \text{ min}}.$$

This unrealistic time frame discourages any attempt for complete exploration of the design space, limiting the study to consideration of a small number of designs on a pre-defined grid which may often be highly suboptimal if the chosen points lie quite far from the optimal condition. On the contrary, more efficient exploration of the design space can be achieved through various available optimisation algorithms, establishing a response adaptive setup under which function evaluations drive better informed future decisions. Such algorithms will be further considered in Section 3.3 and Chapter 6.

Response adaptive algorithms are further, frequently, incorporated in sequential designs. Under this setup, knowledge obtained from the observed system during previous stages informs subsequent decisions regarding future experiments, establishing a closed-loop framework of experimentation, knowledge update and optimal decision making. Unfortunately, the long waiting times associated with the current estimation methods, involved in the decision making stage are not well-suited to such an exchange and thus simpler but potentially suboptimal designs are often considered instead.

Under an even worse scenario, obtaining one evaluation from the models studied in Chapter 9 requires on average 11.6 seconds and thus estimation of the expected utility for one experimental condition through Monte-Carlo based methods takes up to 15 hours using parallel computing. Given the current methodology, optimal experimental design for the study of such phenomena poses an intractable problem and thus consideration of more efficient approximation methods is imperative. Indeed, in their recent review on modern Bayesian methods for optimal design, Ryan et al. (2016) point out the need for improved methods that will allow consideration of "problems in which the likelihood is

intractable or computationally prohibitive to calculate".

Chapter 5 presents a class of methods that provide highly efficient estimators of the expected utility through variational approximation techniques. Chapters 8 and 9, provide a comparison of the proposed estimator with the traditional Monte-Carlo based approaches for Bayesian experimental design targeted at model selection problems.

3.3 Optimisation over the design space

The preceding section provided an overview of related methodology for evaluation of the expected utility. Efficient estimation is certainly necessary but, nonetheless, does not itself solve the overall optimisation problem of (2.4). Optimisation of the expected utility over the design space Δ relies on evaluation of it for multiple experimental conditions and subsequently the comparison between them. Unfortunately, evaluation of the expected utility of numerous conditions on a finely defined grid and subsequent optimisation through deterministic comparison is often intractable. This section is, therefore, concerned with efficient optimisation methods that inevitably, often, address the estimation challenge too.

3.3.1 Approximation of the utility surface

Müller and Parmigiani (1995) propose approximation of the expected utility surface by fitting a curve through the observed utility of randomly drawn points according to the prior distribution. Optimisation over the approximate surface is subsequently achieved deterministically as summarised in Algorithm 1.

Algorithm 1 Approximation of the expected utility surface through curve fitting (Müller and Parmigiani, 1995)

1: Select an initial set of experimental conditions $\boldsymbol{\delta} = \{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_T\}$.
2: For each $\boldsymbol{\delta}_i$, draw a sample $(\boldsymbol{\theta}_i, \boldsymbol{y}_i) \sim P$ and evaluate $u_i = u(\boldsymbol{\theta}_i, \boldsymbol{y}_i, \boldsymbol{\delta}_i)$.
3: Fit a j-dimensional curve where j is the dimension of each experimental condition $\boldsymbol{\delta}_i$ through points $(u_i, \boldsymbol{\delta}_i)$.
4: Use the obtained curve to evaluate the maximum expected utility deterministically.

For more complex and potentially noisy problems, the authors recommend drawing multiple samples of $(\boldsymbol{\theta}, \boldsymbol{y})$ for each $\boldsymbol{\delta}$ and fitting a curve through the average utility

instead in order to mitigate the higher levels of error present. Choice of the curve can vary depending on the complexity of the problem, for example Müller and Parmigiani (1995) fit a non-linear regression model when optimising the sample size of a binomial model while Overstall and Woods (2017) use a more flexible Gaussian process (GP; Rasmussen and Williams (2006)) regression model to target more challenging problems.

An advantage of this approach is that evaluation of the utility for each given $\boldsymbol{\delta}$ is only required for a small number of samples from $\Theta \times \mathcal{Y}$ unlike the typical Monte-Carlo evaluation. This already incurs considerable computational savings whereas further evaluations of the utility are no longer necessary once an approximation through the fitted curve is available.

This approach has been mainly used to tackle low-dimensional problems (Kuo et al., 1999) as curve fitting may be challenging in higher-dimensional spaces. Moreover, in cases of a flatter utility surface and especially in the presence of high noise levels, a large number of utility evaluations may be required regardless. In particular, Overstall and Woods (2017) state that even by using their proposed approach, finding Bayesian optimal designs using Monte-Carlo integration is confined to simpler problems and focusing on parameter estimation with only one assumed model. An extension of this approach is further considered in Chapter 6.

3.3.2 Approximation of an augmented expected utility surface through sampling techniques

An alternative class of approaches arose with the work of Müller (1999) who treats expression (2.11) as an un-normalised probability density function of the form:

$$h(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\delta}, m) \propto u(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\delta}, m) p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\delta}, m) p(\boldsymbol{\theta}, m) \tag{3.7}$$

and employ sampling methods to obtain observations from $(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\delta})$ under h. Due to relationship (3.7), the marginal likelihood of h at point $\boldsymbol{\delta}$ will be proportional to the expected utility at that point. To avoid potential issues such as a flat utility surface or high error levels the following simulation-annealing inspired modification has also been

suggested in the relevant literature:

$$h(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_J, \boldsymbol{\delta}, m) \propto \prod_{j=1}^{J} u(\boldsymbol{\theta}_j, \boldsymbol{y}_j, \boldsymbol{\delta}, m) p(\boldsymbol{y}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\delta}, m) p(\boldsymbol{\theta}_j) \ , \tag{3.8}$$

where $h(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_J, \boldsymbol{\delta}) \propto U(\boldsymbol{\delta})^J$. $U(\boldsymbol{\delta})^J$ represents an augmented expected utility surface which facilitates identification of its mode.

Algorithm 2 MH algorithm for sampling from the augmented expected utility surface (Müller, 1999)

1: Select an initial experimental condition $\boldsymbol{\delta}_1$.
2: Draw J samples $(\boldsymbol{\theta}_j, \boldsymbol{y}_j \mid \boldsymbol{\delta}_1) \sim P$ and evaluate $u_1 = \prod_{j=1}^{J} u(\boldsymbol{\theta}_j, \boldsymbol{y}_j, \boldsymbol{\delta}_1) p(\boldsymbol{y}_j \mid \boldsymbol{\theta}_j) p(\boldsymbol{\theta}_j)$.
3: Set $i = 1$.
Repeat steps 4-8 until convergence:
4: Propose new condition $\boldsymbol{\delta}^* \sim Q_{\boldsymbol{\delta}_i}$.
5: Draw J samples $(\boldsymbol{\theta}_j, \boldsymbol{y}_j \mid \boldsymbol{\delta}_i) \sim P$ and obtain the corresponding product of utilities u^*.
6: Compute the acceptance probability:

$$a = \min \left\{ 1, \frac{u_i^* \cdot q(\boldsymbol{\delta}_i \mid \boldsymbol{\delta}^*)}{u_i \cdot q(\boldsymbol{\delta}^* \mid \boldsymbol{\delta}_i)} \right\}$$

7: Set $\boldsymbol{\delta}_{i+1} = \boldsymbol{\delta}^*$ and $u_{i+1} = u^*$ with probability a, otherwise set $\boldsymbol{\delta}_{i+1} = \boldsymbol{\delta}_i$ and $u_{i+1} = u_i$.
8: $i \to i + 1$.

Algorithm 2 summarises the process of using a Metropolis-Hastings (MH) MCMC sampler to approximate the expected utility surface. Function $Q_{\boldsymbol{\delta}_i}$ denotes the probability distribution of the newly proposed steps while $q(\cdot \mid \boldsymbol{\delta}_i)$ its corresponding probability density function at experimental condition $\boldsymbol{\delta}_i$. The MH algorithm encourages exploration in areas of $\boldsymbol{\delta}$ incurring higher expected utility while preventing visits to regions with low expected utility. However, practice has shown that at its current state the method fails to scale up to dimensions higher than 4 while convergence issues have also been reported (Ryan et al., 2016). To tackle these challenges Amzal et al. (2006) alternatively employ SMC methods for sampling from the augmented surface instead. This method is applied to determine the optimal dose of a caffeine treatment to preterm infants with respiratory issues using a compartmental time series model. Unlike MCMC algorithms, samples obtained through SMC methods have been proven to converge to the distribution of interest (Del Moral et al., 2006) thus resolving any potential convergence issues. In addition, SMC samplers have shown to work well for multimodal surfaces as they enable

efficient exploration of the space. However this approach is still restricted to lower-dimensional problems and the high computational cost may hinder the consideration of highly complex models.

3.3.3 Additional optimisation algorithms

Huan and Marzouk (2013) used standard stochastic optimisation algorithms such as the Nelder-Mead non-linear simplex to maximise the expected information gain for kinetic models described by a system of ordinary differential equations. Similarly to approaches discussed in Sections 3.3.1 and 3.3.2, this class of optimisation algorithms can be used to locate the global optimum however, unlike MCMC or SMC methods, provide no uncertainty quantification. To avoid evaluation of the expected utility at each step of the optimisation algorithm, polynomial chaos surrogate models (Ghanem and Spanos, 2003) were employed, relying on a set of initial estimations of the expected utility through Monte-Carlo integration.

3.3.4 Potential extensions

Surface approximation using a model fitted through a small sample of observations from the expected utility has many appealing properties: unlike grid search methods, optimisation is possible based on only a few evaluations of the expected utility and unlike surface approximation through sampling, sampling is not required thus keeping the computational cost at realistic levels while also avoiding time-consuming parameter tuning. Lastly, unlike traditional optimisation algorithms, it can be easily adapted to provide uncertainty quantification in the produced output by employing a probabilistic model for approximation of the expected utility surface.

Nevertheless, several shortcomings may hinder the optimisation performance of these methods: the produced optimum may be highly suboptimal if none of the initially chosen designs fall sufficiently close to the true optimum. Performance of the optimiser is thus sensitive to the initial design. In addition, further refinement of the approximation is possible but, again, choice of subsequent designs for observation may be arbitrary. Most importantly, there is a lack of a systematic and comprehensive framework that could

potentially increase the efficiency of the optimiser and ensure optimality of the produced result.

An attempt to address the need for a more autonomous framework is made in Chapter 6 through a class of optimisation algorithms, commonly known as Bayesian Optimisation (BO; Močkus (1975)). Broadly speaking, BO algorithms can be viewed as an extension of the curve fitting approaches discussed in Section 3.3.1, however providing a more efficient and automated framework. The similarity of the two approaches lies predominantly in the approximation of the target surface through a curve or potentially more flexible models. BO algorithms extend on the methods of 3.3.1 by establishing a sequence of observing the unknown function and re-fitting the approximating surface until the optimal design has been found rather than relying on a deterministic search for it. A more efficient and comprehensive exploration of the design space is thus possible, ensuring truly optimal proposed designs.

As will be further discussed in Chapter 6 BO algorithms lend themselves to optimisation of the expected utility over the design space. Although previously considered in the Bayesian experiment design literature (Kleinegesse and Gutmann, 2019; von Kügelgen et al., 2019), this class of methods have not gained enough popularity given its suitability but also high efficiency and automation compared to currently considered approaches.

Summary

This chapter reviews previous work on optimal experiment design focusing on two key objectives: 1) estimation of the expected utility (Section 3.2) and 2) optimisation over the design space (Section 3.3). Issues specific to discrimination problems such as estimation of the marginal likelihood are also considered (Section 3.2.2) with emphasis on cases when model evaluation is particularly demanding. Advantages of the considered methods as well as prevailing challenges are highlighted throughout this chapter.

The succeeding chapters consider methodology for tackling these challenges: Chapter 5 proposes an efficient estimation procedure adopting variational approximation techniques that is shown to reduce the number of required operations substantially compared to currently adopted methods. Chapter 6 incorporates the proposed estimator into an

efficient optimisation procedure from the class of Bayesian optimisation algorithms along the lines of the approximation approaches discussed in Section 3.3.1.

Chapter 4

Variational approximation methods

This chapter provides a brief overview to variational approximation methods, particularly focusing on a class of procedures targeted at estimation problems. This class of methods is subsequently adopted in Chapter 5 for efficient estimation of the expected utility in the context of experimental design optimisation, tackling the currently prevalent issues, previously discussed in Chapter 3.

4.1 Overview

Variational methods refer to a class of deterministic approximation procedures that translate an initial, potentially complex problem into a theoretically simpler and more tractable one, constituting a generalised expression of the problem at hand. This generalisation is achieved though the incorporation of additional parameters, known as the **variational parameters** (Jordan et al., 1999) which are optimised to approximate the initial problem of interest as closely as possible. Although variational methods originate in the field of convex analysis, they have been gaining increasing popularity in recent years in the machine learning literature where they have been applied to various inference (Blei et al., 2017) and estimation problems (Nguyen et al., 2010; Ruderman et al., 2012).

In the former case, variational methods provide an alternative approach to approximate Bayesian inference, especially in settings that exact sampling methods may be too

resource intensive (Blei et al., 2017; Titsias, 2009). Briefly, a posterior density of interest is approximated by an alternative function from a proposed class of potential functions through an optimisation procedure that aims to minimise the KL divergence between itself and the exact quantity.

Perhaps less popularly, variational methods have also been employed for estimation purposes such as that of the Mutual Information criterion (Belghazi et al., 2018; Song and Ermon, 2019) or of the divergence between two distributions (Nguyen et al., 2010; Ruderman et al., 2012). The variational methods examined in this thesis belong to the latter branch. More specifically, a technique that translates the initial evaluation problem into an alternative optimisation problem is employed for this purpose, commonly known as the **duality principle**. The mutual element in this class, enabling this transition, is the idea of the **Fenchel transform** which was introduced in Rockafellar (1970) and is presented in Section 4.2. Optimisation is subsequently performed over an appropriately chosen class of functions which is briefly introduced in Section 4.3. The approximation methodology presented in this chapter is subsequently adopted in Chapter 5 for efficient evaluation of the expected utility.

4.2 Fenchel transform

This section introduces the idea of the Fenchel transform that allows generalisation of an initial problem through its variational expression. An introduction to the notions of the inner product and norm is required before proceeding to the definition of the Fenchel transform and are, therefore, provided in Definitions 3 and 4, respectively. The Fenchel transform is subsequently presented in Definition 5 .

Definition 3 (Inner product)**.** *Let $\mathcal{V}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is said to be an* ***inner product*** *on $\mathcal{V}$ if:*

1. $\langle a_1 \boldsymbol{v} + a_2 \boldsymbol{v}', \boldsymbol{v}'' \rangle = a_1 \langle \boldsymbol{v}, \boldsymbol{v}'' \rangle + a_2 \langle \boldsymbol{v}', \boldsymbol{v}'' \rangle$, $a_1, a_2 \in \mathbb{R}$, $\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{v}'' \in \mathcal{V}$
2. $\langle \boldsymbol{v}, \boldsymbol{v}' \rangle = \langle \boldsymbol{v}', \boldsymbol{v} \rangle$
3. $\langle \boldsymbol{v}, \boldsymbol{v} \rangle \geq 0$ *and* $\langle \boldsymbol{v}, \boldsymbol{v} \rangle = 0$ *if and only if* $\boldsymbol{v} = 0$.

Definition 4 (Norm). *The corresponding* ***norm*** *of* $\mathcal{V}$ *can subsequently be defined by the inner product as:* $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$.

Definition 5 (Fenchel transform). *Let* $h : \mathcal{G} \to [-\infty, +\infty]$ *be a convex function on a compact domain* $\mathcal{G}$. *The conjugate function* $h^* : \mathcal{G}^* \to [-\infty, +\infty]$, *where* $\mathcal{G}^*$ *is also a compact domain, is defined as:*

$$h^*(g) = \sup_{\chi \in \mathcal{G}^*} \{ \langle \chi, g \rangle - h(\chi) \} \quad \text{for } g \in \mathcal{G} \ , \tag{4.1}$$

where $\langle \cdot, \cdot \rangle : \mathcal{G} \times \mathcal{G}^* \to \mathbb{R}$ *is the inner product restricted on* $\mathcal{G} \times \mathcal{G}^*$[1]. *The Fenchel transform refers to the operation* $h \to h^*$.

Conveniently, the conjugate function $h^*(g)$ finds expression as the supremum of the family of affine continuous functions of the form $\{ \langle \cdot, g \rangle - h(\cdot) \}_{\chi \in \mathcal{G}^*}$ and is as such convex and lower semi-continuous.

The conjugate function h^{**} of h^* is commonly referred to as the **biconjugate** of h. An important result, presented in Theorem 1, states that the biconjugate h^{**} is equal to h when h is a closed convex function.

Theorem 1 (Fenchel-Moreau theorem;Rockafellar (1970)). *The biconjugate function* h^{**} *of* h^* *is equal to* h, *if* h *is convex and closed. If* h *is proper*[2], *then the property of closeness is equivalent with* h *being lower semi-continuous.*

Proof. A proof can be found in (Rockafellar, 1970), page 104. □

A combination of Theorem 1 and Definition 5 results in the following expression of function h in terms of its conjugate:

$$h(\chi) = \sup_{g \in \mathcal{G}} \{ \langle \chi, g \rangle - h^*(g) \} \quad \text{for } \chi \in \mathcal{G}^* \ , \tag{4.2}$$

for every proper and lower semi-continuous convex function h. This dual representation is the key to deriving the dual expression of the partial utility in Chapter 5. Optimisation of the dual problem over an appropriately chosen class of functions allows the efficient estimation of the initial quantity of interest through variational approximation.

[1] In the special case when $\mathcal{G} = \mathcal{G}^* = \mathbb{R}$, $< g, \chi >= g \cdot \chi$.

[2] In fact, the only closed, improper convex functions are the constant functions $-\infty$ and $+\infty$.

Example 1 provides guidance on application of the Fenchel transform for the special case of $h \equiv -\log$.

Example 1. *This example demonstrates the application of the Fenchel transform to an initial estimation problem for function* $h \equiv -\log$. *Under Definition 4.1, the Fenchel transform* h^* *of function* h *obtains the form:*

$$h^*(g) = \sup_{\chi} \{\chi g + \log(\chi)\} \ . \tag{4.3}$$

The supremum in (4.3) is subsequently obtained by solving $\frac{\partial[\chi g + \log(\chi)]}{\partial \chi} = 0$ *with respect to* χ*, producing:*

$$\begin{aligned} g + \frac{1}{\chi} &= 0 \\ \chi &= -\frac{1}{g} \ , \end{aligned} \tag{4.4}$$

which, substituted back to the initial term, results in:

$$h^*(g) = -1 - \log(-g) \ , \ g < 0 \tag{4.5}$$

Using Theorem 1, a variational representation of function h *can be formulated as:*

$$h(\chi) = \sup_{g} \{\chi g + 1 + \log(-g)\} \ , \ g < 0. \tag{4.6}$$

The two alternative problems arising from this expression are: the estimation problem $h(\chi)$ *for a given* χ *and the optimisation problem of function* $\gamma(\chi, g) = \chi g + 1 + \log(-g)$. *The optimal element* g*, maximising* γ *and henceforth denoted by* g_{opt}*, can be obtained in a similar manner as in (4.4), returning* $g_{opt} = -\frac{1}{\chi}$. *Figure 4.1 provides a visual representation of the Fenchel transform for* $h \equiv -\log$. *Both problems are depicted therein, 1) function* $h \equiv -\log$ *and, more specifically, its estimation at a given* χ *(in this example chosen as* 2*) and 2) function* γ *that is sought to be optimised. As demonstrated in Figure 4.1, the solution to the optimisation problem is exactly the same as the solution to the initial estimation problem (both -0.5).*

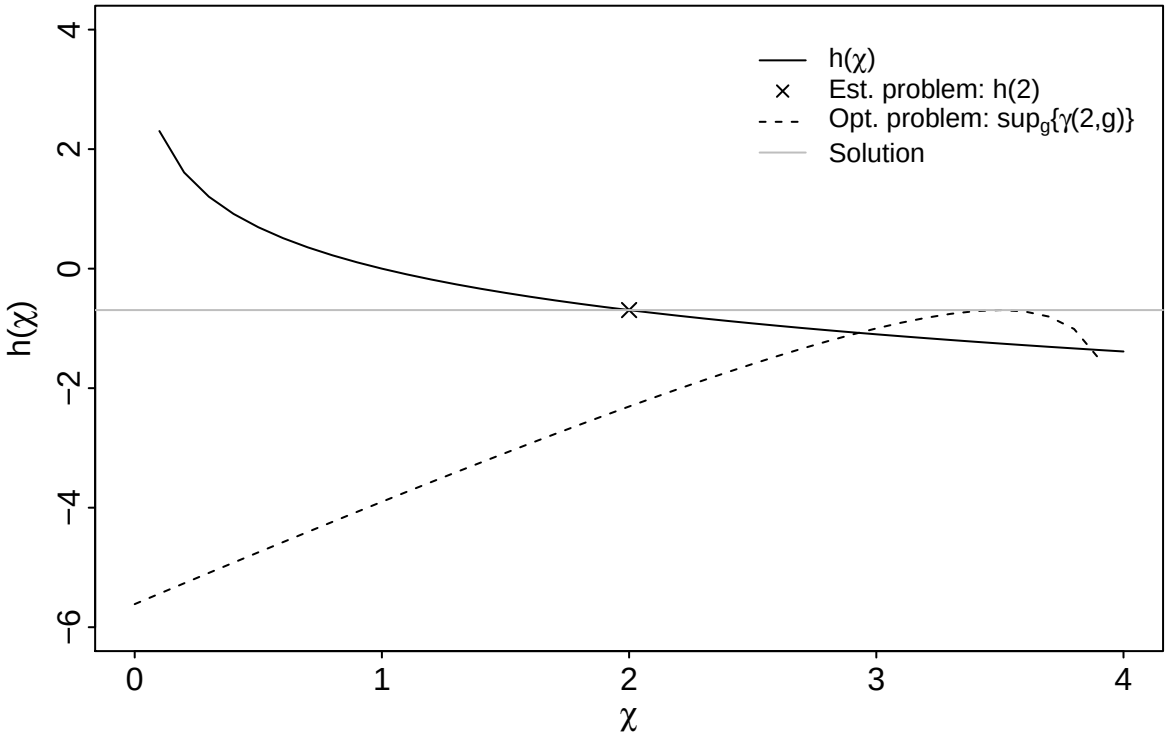


Figure 4.1: Illustration of the Fenchel transform for $h \equiv -\log$. The two alternative problems are showcased therein: function $h(\chi)$ and, more specifically, the initial, evaluation problem $h(\chi)$, for $\chi = 2$, marked by a solid black line and a cross respectively. The alternative optimisation problem $\sup_g \{\gamma(2, g)\}$ is indicated by a dashed black line. A tangent on the optimum g_{opt} of the induced optimisation problem $\sup_g \{\gamma(2, g)\}$, represented by a solid grey line, outlines that the solutions to the two problems indeed coincide.

△

Two additional notions are essential for deriving this result and are introduced in Definition 6 and Theorem 2, presenting the ideas of subgradient function and the infimal convolution theorem, respectively. The idea of the subgradient (Rockafellar, 1974) effectively constitutes a generalisation of the gradient to functions that are not necessarily differentiable, a more formal definition is provided in Definition 6. The subgradient exists under any case as opposed to the gradient that requires the function to be differentiable. However when that holds, the subgradient and the gradient are represented by exactly the same vector.

Definition 6 (Subgradient)**.** *Function z, is the* ***subgradient*** *of $s : X \to \mathbb{R}$ at point u, if:*

$$s(u') \geq s(u) + z(u' - u) \ , \quad \forall u' \in X \tag{4.7}$$

The set of all subgradients of s at point u is referred to as the ***subdifferential*** *of s at u and will be denoted by $\partial s(u)$.*

In general, equality in (4.2) is achieved if and only if the subdifferential contains an element of $\mathcal{G}$ (Rockafellar, 1970). When that is not the case, the solution to the optimisation problem acts as a lower bound to the solution of the estimation problem an so, the optimum of the dashed line in Figure 4.1 would appear lower than the solid grey line. In the case of Example 1 equality is, nevertheless, achieved at the supremum $g_{\text{opt}} \in \mathcal{G}$.

Lastly, the infimal convolution theorem, presented in Theorem 2, allows the decomposition of the conjugate of a sum of function to a sum of the conjugates of the individual functions. This result is essential for the derivation of the partial utility estimator of Chapter 5.

Theorem 2 (Infimal convolution theorem)**.** *For $h_1, \ldots, h_n : \mathbb{R} \to \mathbb{R}$ convex functions:*

$$(h_1 + \cdots + h_n)^* (v) = \inf_v \{h_1^*(v_1) + \cdots + h_n^*(v_n) \mid v_1 + \cdots + v_n = v\} \tag{4.8}$$

Proof. See Rockafellar (1974). □

As discussed in the introductory part of this chapter, the adopted class of variational approximation methods consists of two main steps: the initial part, covered in this section, establishes a dual representation of the initial, complex problem through the introduction of variational parameters. During the subsequent step, the variational parameters are optimised over a class of appropriately chosen class of functions to approximate the initial problem. Since the two alternative problems are equivalent only when $\mathcal{G}$ includes the supremum (as noted in the preceding paragraphs), care should be taken as for the function class to be rich enough in order to achieve or at least approximate the supremum as closely as possible while at the same time keeping the complexity of the optimisation problem at reasonable levels. The most commonly selected candidates, for this purpose, are the class of Reproducing Kernel Hilbert spaces (RKHS; Berlinet and Thomas-Agnan

(2011),Nguyen et al. (2010)) or classes parametrised by neural networks (Belghazi et al., 2018). The work presented in this thesis focuses on the former, an introduction of which is provided in the subsequent Section 4.3.

4.3 The class of reproducing kernel Hilbert spaces

Definition 7 (Hilbert space)**.** *A* ***Hilbert space*** $\mathcal{H}$ *is a complete space equipped with an inner product.*

Definition 8 (Reproducing kernel Hilbert space)**.** *An RKHS is a Hilbert space, equipped with inner product* $\langle \cdot, \cdot \rangle$ *where a unique* ***kernel function*** $K : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ *exists such that:*

- *for all* $v \in \mathcal{V}$, $K(\cdot, v) \in \mathcal{H}$,
- *for all* $v \in \mathcal{V}$ *and for all* $h \in \mathcal{H}$, $h(v) = \langle h, K(\cdot, v) \rangle$.

The latter attribute expresses the **reproducing property** of RKHS stating that: any element $h \in \mathcal{H}$ evaluated at point v can be expressed as a linear combination of other elements in $\mathcal{H}$ evaluated at this point. It is worth noting that, since $\mathcal{H}$ is a space of functions, $h(v)$ corresponds to a functional with assigned value v and, as opposed to vector evaluation h, refers to a specific evaluation of h. The mapping of a given v to a functional in $\mathcal{H}$ will be denoted as:

$$\Phi(v) = K(\cdot, v)$$

and so from Definition 8 follows directly that:

$$K(v, u) = \langle \Phi(v), \Phi(u) \rangle, \text{ for } v, u \in \mathcal{V} \ . \tag{4.9}$$

Evaluation $h(v)$ can equivalently be restated as a linear combination as:

$$h(v) = \sum_{i=1}^{n} h_i \Phi_i(v) = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ \vdots \end{bmatrix}^T \begin{bmatrix} \Phi_1(v) \\ \Phi_2(v) \\ \Phi_3(v) \\ \vdots \end{bmatrix} . \tag{4.10}$$

Relationship (4.10) essentially deconstructs $h(v)$ into a set of simpler basis functions $\Phi_1, \Phi_2, \ldots$. The basis functions are such that any element in $\mathcal{H}$ can be expressed as a distinct linear combination of them. An illustration of this representation is provided in Figure 4.2 where the target function follows a Gaussian distribution.

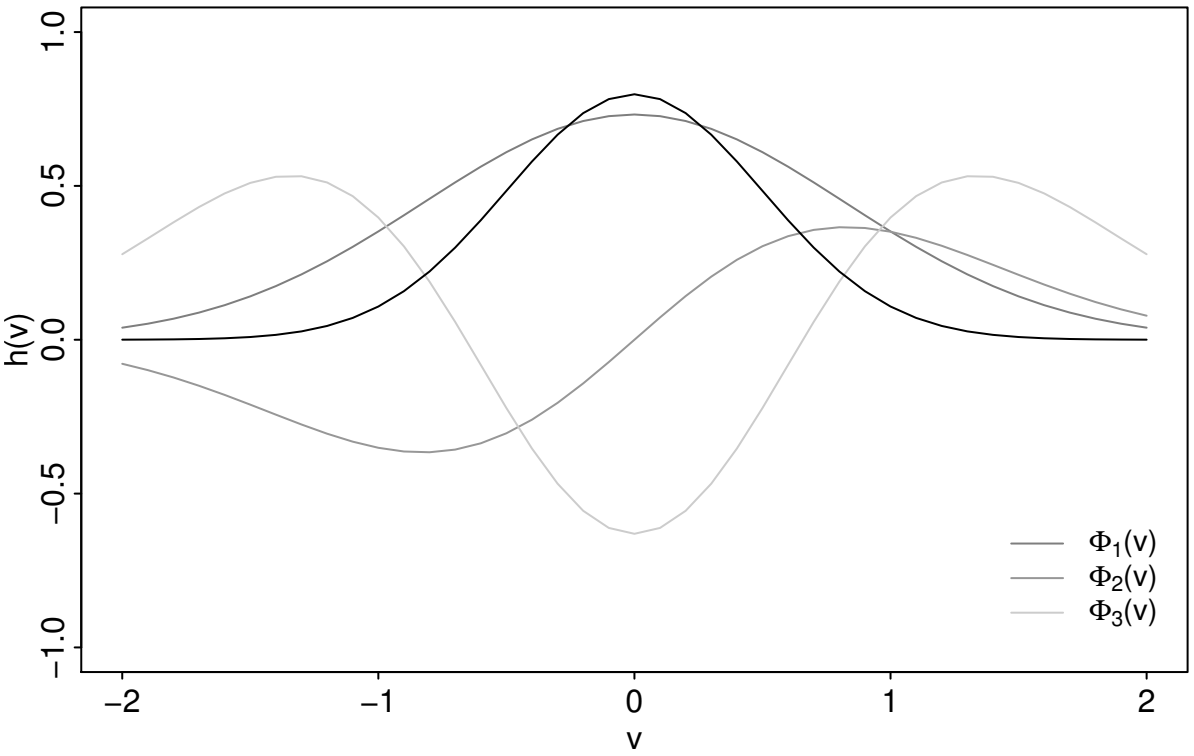


Figure 4.2: Illustration of relationship (4.10). The solid black line represents the target function $h(v)$ which can be expressed as a sum of a finite set of basis functions weighted by $h_1, h_2, h_3, \ldots$. In this plot, only the first three basis functions Φ_1, Φ_2, Φ_3 are shown.

Summary

This chapter provides an introduction to variational approximation methods targeted at estimation problems. The presented methodology establishes a dual representation of the initial, potentially complex problem, resulting in a theoretically simpler, more tractable problem. In the examined class of methods, this alternative representation is achieved through the Fenchel transform, introduced in Section 4.2 which further relies on optimisation over a class of candidate functions in order to approximate the initial problem at hand as closely as possible. The class of RKHS has been widely considered in

the literature for this purpose due to a number of desirable properties (further discussed in Chapter 5) and is also adopted in this thesis. A brief introduction to RKHS is provided in Section 4.3. In Chapter 5, the presented methodology is applied to efficiently solve the estimation problem of the expected utility for a given experimental condition, overcoming the currently persisting issues, introduced in Chapter 3.

Chapter 5

Variational estimation of the expected utility

This chapter proposes a novel methodology for Bayesian experiment design incorporating a class of variational approximation methods for efficient estimation of the expected utility. This approach is shown to overcome issues faced by current methods without compromising the quality of the resulting estimates. The proposed estimation procedure is applicable to a broad class of utility functions along the lines of the decision-theoretic framework of Blackwell (1951), introduced in Chapter 2, including the previously considered Shannon entropy.

5.1 Outline

The notion of the expected utility and its role in design optimisation was introduced in Chapter 2 while common estimation methods for evaluation of this quantity and associated challenges were discussed in Chapter 3. Motivated by the case studies, considered in the subsequent Chapters 7, 8 and 9, focus was placed on methodology targeting model selection problems, particularly involving models with computationally expensive likelihoods. This setup is frequently integrated in studies of modern, highly complex phenomena which constitute a focal objective in this thesis. This chapter presents a novel approximation approach, incorporating the variational approximation methodol-

ogy of Chapter 4 for evaluation of the expected utility in an effort to address issues that currently hinder the optimal solution of such problems. The proposed approach paves the way towards a complete and fully automated framework for the study of modern phenomena within realistic timelines, presently constituting an ongoing challenge.

Improvement over existing methods is attributed to the following key procedures: 1) consideration of an alternative representation of the initial, computationally challenging evaluation problem as a convex optimisation problem through the Fenchel transform, previously introduced in Chapter 4. This new expression no longer relies on the demanding marginal likelihood approximation, incurring a considerable decrease on the number of model evaluations required for estimation of the expected utility, thus leading to substantial computational savings. 2) Adoption of a flexible, non-parametric framework for estimation of this dual representation through the class of RKHS, previously presented in Chapter 4, allowing a systematic comparison of predictive distributions against each other. This approach achieves direct evaluation of the ratio, as opposed to the currently adopted Monte-Carlo based approach (presented in Chapter 3) that uses the marginal likelihood as a summary statistic, extracting information from each individual predictive sample and subsequently using it for comparison between their corresponding distributions. In subsequent chapters, this approach is shown to be highly inefficient, leading to considerable waiting times for optimisation of the expected utility. In addition, when estimation of each ratio term is achieved individually and subsequently substituted into the ratio for its evaluation, the error caused by the latter step is not taken into account during the initial step which may lead to less accurate estimates (Sugiyama et al., 2012).

The proposed methodology is applicable to previously considered utility functions including the Shannon entropy and mutual information but more importantly, extends to a broader class defined as:

$$\mathcal{F} = \left\{ \varphi\left(\frac{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m')}{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m)} \right);\ \varphi : \mathbb{R} \to \mathbb{R}, \text{ convex and lower semi-continuous} \right\}, \tag{5.1}$$

where the Shannon entropy is obtained under $\varphi \equiv -\log$. This class of functions adheres to the decision-theoretic framework of Blackwell (1951), introduced in Section 2.1.5, and thus, the induced expected utilities constitute a valid criterion for experimental design

optimisation. Under the general representation (5.1) of the utility, the expected utility of (2.11) finds expression in:

$$U(\boldsymbol{\delta}) = \sum_{m,m' \in \mathcal{M}} \mathrm{F}_{\varphi;\boldsymbol{\delta}}(m, m') p(m') p(m) \ , \tag{5.2}$$

where:

$$\mathrm{F}_{\varphi;\boldsymbol{\delta}}(m, m') = \int \varphi\left(\frac{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m')}{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m)}\right) p(\boldsymbol{y} \mid \boldsymbol{\delta}, m) \ \mathrm{d}\boldsymbol{y} \ . \tag{5.3}$$

The methodology proposed in this chapter is targeted at evaluation of the term $\mathrm{F}_{\varphi;\boldsymbol{\delta}}(m, m')$ which will be henceforth referred to as the **partial utility**. Estimation of the expected utility is thus reduced to summation of the partial utility over all possible pairs $m, m' \in \mathcal{M}$. It is worth noting that, in general, the model ordering should not be ignored as $\mathrm{F}_{\varphi;\boldsymbol{\delta}}(m, m') \neq \mathrm{F}_{\varphi;\boldsymbol{\delta}}(m', m)$.

The methodology examined in this chapter is, more generally, applicable to estimation of any quantity of the form:

$$F_{\varphi} = \int \varphi\left(\frac{q(\boldsymbol{y})}{p(\boldsymbol{y})}\right) p(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \ ,$$

where $p(\cdot)$ and $q(\cdot)$ are the densities corresponding to two distributions of interest P and Q respectively. In the particular case of (5.3), P and Q represent the predictive distributions of the two competing models m and m'. Function class F_{φ}, where φ convex, is commonly known as the class of $f-$divergences introduced by Ali and Silvey (1966) and frequently serve as information-theoretic metrics in numerous statistical applications such as independent component analysis (Comon, 1994), classification (Moreno et al., 2004), asymptotic analysis of hypothesis testing and more. As subsequently shown in Chapter 8, consideration of $f-$divergences other than the traditionally adopted KL divergence may incur additional benefits due to certain intrinsic properties such as symmetricity, boundedness and more.

This chapter introduces an efficient estimation procedure targetting quantities from the class $\mathrm{F}_{\varphi;\boldsymbol{\delta}}(m, m')$ and subsequent evaluation of the expected utility through (5.2). The following outline is established: Section 5.2 introduces the class of variational approximation methods and demonstrates derivation of the resulting estimator for evaluation of

the expected utility within the problems of interest. Algorithmic implementation of the proposed estimator for expected utility evaluation is provided in 5.3 with application to cases of commonly employed utility functions shown in Section 5.4. The chapter concludes with a discussion summarising the benefits and shortcomings associated with the proposed approach and their comparison to currently adopted methods in Section 5.5.

5.2 Estimation of the expected utility

This section presents an application of the variational approximation methodology, introduced in Chapter 4, for efficient evaluation of the expected utility. This class of methods has not been previously considered in the context of experimental design. The proposed methodology adopts a generalised version of the estimators considered in Nguyen et al. (2010) for the evaluation of $f-$divergences. This approach consists of two key steps: consideration of the dual representation of the initial evaluation problem through the Fenchel transform, as demonstrated in the following section and subsequent optimisation over an appropriately chosen class of functions, as shown in Section 5.2.2.

A variational expression of the partial utility expressed in (5.3) is obtained by transforming the ratio evaluation $\varphi\left(\frac{p(\boldsymbol{y}|\boldsymbol{\delta},m')}{p(\boldsymbol{y}|\boldsymbol{\delta},m)}\right)$ into its dual representation. More specifically, the alternative expression takes the form of (4.2) of h, for $h = \varphi$ and $\chi(\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\delta},m')}{p(\boldsymbol{y}|\boldsymbol{\delta},m)}$, leading to a variational representation of the partial utility, as shown in Corollary 1.

Corollary 1 (Dual representation of the partial utility)**.**

$$\begin{aligned} F_{\varphi;\boldsymbol{\delta}}(m,m') &\geq \sup_{g\in\mathcal{G}}\left\{\int g(\boldsymbol{y})p(\boldsymbol{y}\mid\boldsymbol{\delta},m')\,\mathrm{d}\boldsymbol{y} - \int \varphi^*\left(g\left(\boldsymbol{y}\right)\right)p(\boldsymbol{y}\mid\boldsymbol{\delta},m)\,\mathrm{d}\boldsymbol{y}\right\} \\ &= \sup_{g\in\mathcal{G}}\left\{\mathbb{E}_{P_{m'}}\left[g(\boldsymbol{y})\right] - \mathbb{E}_{P_m}\left[\varphi^*\left(g\left(\boldsymbol{y}\right)\right)\right]\right\}\ , \end{aligned} \tag{5.4}$$

where $\mathcal{G}$ represents a class of measurable functions, conjugate to $\mathcal{G}^*$, induced by χ.

Proof. The Fenchel transform is applied to evaluation $\varphi(\frac{p(\boldsymbol{y}|\boldsymbol{\delta},m')}{p(\boldsymbol{y}|\boldsymbol{\delta},m)})$, producing the following representation:

$$\varphi\left(\frac{p(\boldsymbol{y}\mid\boldsymbol{\delta},m')}{p(\boldsymbol{y}\mid\boldsymbol{\delta},m)}\right) = \sup_{g\in\mathcal{G}}\left\{g(\boldsymbol{y})\frac{p(\boldsymbol{y}\mid\boldsymbol{\delta},m')}{p(\boldsymbol{y}\mid\boldsymbol{\delta},m)} - \varphi^*\left(g(\boldsymbol{y})\right)\right\}\ . \tag{5.5}$$

Integration over $\mathcal{Y}$ to obtain $\mathrm{F}_{\varphi;\boldsymbol{\delta}}(m, m')$ combined with a few additional operations lead to expression (5.4) as demonstrated below:

$$\begin{aligned}
\mathrm{F}_{\varphi;\boldsymbol{\delta}}(m, m') &\geq \int \sup_{g\in\mathcal{G}} \left\{ g(\boldsymbol{y}) \frac{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m')}{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m)} - \varphi^*\left(g(\boldsymbol{y})\right) \right\} p(\boldsymbol{y} \mid \boldsymbol{\delta}, m)\, \mathrm{d}\boldsymbol{y} \\
&= \sup_{g\in\mathcal{G}} \left\{ \int \left[g(\boldsymbol{y}) \frac{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m')}{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m)} - \varphi^*\left(g(\boldsymbol{y})\right) \right] p(\boldsymbol{y} \mid \boldsymbol{\delta}, m)\, \mathrm{d}\boldsymbol{y} \right\} \\
&= \sup_{g\in\mathcal{G}} \left\{ \int g(\boldsymbol{y}) p(\boldsymbol{y} \mid \boldsymbol{\delta}, m') - \varphi^*\left(g(\boldsymbol{y})\right) p(\boldsymbol{y} \mid \boldsymbol{\delta}, m)\, \mathrm{d}\boldsymbol{y} \right\} \\
&= \sup_{g\in\mathcal{G}} \left\{ \int g(\boldsymbol{y}) p(\boldsymbol{y} \mid \boldsymbol{\delta}, m')\, \mathrm{d}\boldsymbol{y} - \int \varphi^*\left(g\left(\boldsymbol{y}\right)\right) p(\boldsymbol{y} \mid \boldsymbol{\delta}, m)\, \mathrm{d}\boldsymbol{y} \right\} ,
\end{aligned}$$

where the exchange between supremum and integral in the second line of equations is a valid operation because of continuity of the function within. The inequality, appearing in the first line of equations accounts for cases when the supremum over $\mathcal{G}$ is not attained and thus, the returned optimal value serves as a lower bound to quantity $\mathrm{F}_{\varphi;\boldsymbol{\delta}}(m, m')$. The conditions under which, equality in the induced relationship holds are discussed in the succeeding paragraphs. □

The variational representation of (5.4) expresses the partial utility for any convex φ of models m and m' as the maximum difference between the expected values of function g and its conjugate dual $\varphi^*(g)$ with respect to the compared predictive distributions P_m and $P_{m'}$ respectively. An important aspect of this new expression is that estimation of the partial utility does no longer rely on evaluation of the probability density functions of competing models, deeming the proposed methods particularly suitable to models with intractable or computationally demanding likelihoods.

The significance of the result expressed in (5.4) lies in the incurred problem shift: from the computationally challenging evaluation of the expected utility into an optimisation problem over a convex set. Provided that equality in (5.4) holds, the value attained from optimisation of the left hand side (LHS) term provides the answer to the initial evaluation problem in the right hand side. When, however, that is not the case the LHS will act as a lower bound of the quantity of interest. Expression (5.6) states the condition under which equivalence of the alternative problems holds.

Equality in (5.4) is achieved if and only if there exists a function $g_{opt} \in \mathcal{G}$ such that:

$$\varphi(s(\boldsymbol{y})) \geq \varphi\left(\chi(\boldsymbol{y})\right) + g_{opt}(\boldsymbol{y})\left(s(\boldsymbol{y}) - \chi(\boldsymbol{y})\right), \quad \forall s \in \mathbb{R} \tag{5.6}$$

Proof. Under the assumption of existence of a function g_{opt} such that (5.6) holds, the following holds true:

$$\begin{aligned}
\varphi(s(\boldsymbol{y})) &\geq \varphi\left(\chi(\boldsymbol{y})\right) + g_{opt}(\boldsymbol{y})\left(s(\boldsymbol{y}) - \chi(\boldsymbol{y})\right) && \Rightarrow \\
\varphi(s(\boldsymbol{y})) &\geq \varphi\left(\chi(\boldsymbol{y})\right) + g_{opt}(\boldsymbol{y})s(\boldsymbol{y}) - g_{opt}(\boldsymbol{y})\chi(\boldsymbol{y}) && \Rightarrow \\
\varphi(s(\boldsymbol{y})) - g_{opt}(\boldsymbol{y})s(\boldsymbol{y}) &\geq \varphi\left(\chi(\boldsymbol{y})\right) - g_{opt}(\boldsymbol{y})\chi(\boldsymbol{y}) && \Rightarrow \\
g_{opt}(\boldsymbol{y})s(\boldsymbol{y}) - \varphi(s(\boldsymbol{y})) &\leq g_{opt}(\boldsymbol{y})\chi(\boldsymbol{y}) - \varphi\left(\chi(\boldsymbol{y})\right) && \Rightarrow \\
\sup_{s}\left\{g_{opt}(\boldsymbol{y})s(\boldsymbol{y}) - \varphi(s(\boldsymbol{y}))\right\} &\leq g_{opt}(\boldsymbol{y})\chi(\boldsymbol{y}) - \varphi\left(\chi(\boldsymbol{y})\right) && \Rightarrow \\
\varphi^*\left(g(\boldsymbol{y})\right) &\leq g_{opt}(\boldsymbol{y})\chi(\boldsymbol{y}) - \varphi\left(\chi(\boldsymbol{y})\right) && \Rightarrow \\
\varphi\left(\chi(\boldsymbol{y})\right) &\leq g_{opt}(\boldsymbol{y})\chi(\boldsymbol{y}) - \varphi^*\left(g(\boldsymbol{y})\right), && (5.7)
\end{aligned}$$

where the definition of the Fenchel transform is used in the derivation of (5.7).

Combining relationships (5.5) and (5.7) the desired outcome is obtained:

$$\varphi\left(\chi(\boldsymbol{y})\right) = g_{opt}(\boldsymbol{y})\chi(\boldsymbol{y}) - \varphi^*\left(g_{opt}(\boldsymbol{y})\right),$$

proving that under g_{opt}, equality in (5.6) is achieved. □

In subsequent sections, choice of class $\mathcal{G}$ is performed with fulfilment of condition (5.6) in mind, deeming full exploitation of the dual representation (5.4) possible. Overall, care should be taken during this choice to ensure a sufficiently tight bound around the estimated quantity. Later sections provide examples of function classes that can be chosen to be rich enough to include g_{opt} while also leading to practical solutions.

Ruderman et al. (2012) propose a slightly modified representation of the lower bound in (5.4) that is shown to be overall tighter around the optimum. This approach is not considered further in this thesis for three reasons: 1) both approaches exhibit the same behaviour at the optimum with Ruderman et al. (2012)'s representation having the potential to achieve tighter bounds elsewhere. Since care is taken to ensure that g_{opt} belongs

to $\mathcal{G}$ and thus the optimum value is attained such an improvement is not of immediate interest. 2) Empirical results provided therein fail to consistently demonstrate an improved performance of their proposed estimator and lastly, 3) no theoretical properties of the alternative expression are provided (such as the convergence rate) as opposed to the estimator proposed in Nguyen et al. (2010).

Functions that fulfil relationship (5.6) are subgradients of φ at a given point, as stated in Definition 6, and so g_{opt} is the subgradient of φ at $\chi(\boldsymbol{y})$. Although not of direct interest in this thesis, it is worth noting that g_{opt} also provides an estimate for the ratio $\frac{p(\boldsymbol{y}|\boldsymbol{\delta},m')}{p(\boldsymbol{y}|\boldsymbol{\delta},m)}$.

5.2.1 Empirical estimation of the expected utility

As analytical evaluation of the integrals composing expression (5.4) is typically not possible, the following estimator is adopted:

$$\hat{F}_{\varphi;\boldsymbol{\delta}}(m,m') = \sup_{g\in\mathcal{G}} \left\{ \frac{1}{N}\sum_{k=1}^{N} g(\boldsymbol{y}_k^{P_{m'}}) - \frac{1}{N}\sum_{k=1}^{N} \varphi^*\left(g\left(\boldsymbol{y}_k^{P_m}\right)\right) - \frac{\rho}{2} I(g) \right\} , \tag{5.8}$$

where $\boldsymbol{y}^{P_m} = \left\{\boldsymbol{y}_k^{P_m}\right\}_{k=1}^{N}$ and $\boldsymbol{y}^{P_{m'}} = \left\{\boldsymbol{y}_k^{P_{m'}}\right\}_{k=1}^{N}$ are samples from distributions P_m and $P_{m'}$ corresponding to models m and m' respectively, obtained at experimental condition $\boldsymbol{\delta}$. An additional penalty term $I(g)$ is introduced in the proposed estimator in order to maintain a trade-off between two desired but conflicting properties of $\mathcal{G}$. On one hand, $\mathcal{G}$ is expected to be sufficiently rich so that it includes g_{opt} (see Section 4.2) while also not too large so that optimisation is achieved within reasonable timescales. Term $I(g)$ thus penalises functions with higher complexity while ρ acts as a weight to the imposed penalty.

5.2.2 Optimisation over the class of reproducing kernel Hilbert spaces

This section considers the evaluation of the expected utility using the estimator of (5.8) when the structure of a RKHS is imposed on class $\mathcal{G}$. This choice has been prominently adopted in such problems (Nguyen et al., 2010; Ruderman et al., 2012) as it provides a class of functions that is sufficiently rich to include g_{opt}, imposes minimal assumptions on

g while the complexity of the functions can be regulated by the norm corresponding to each RKHS class. In addition, exploitation of the class's inherent properties results in a simple optimisation problem relying only on evaluation of the Gramian matrix.

Reconsidering the proposed estimator of (5.8) within the class of RKHS, evaluations $g(\boldsymbol{y}_k^{P_{m'}})$ and $\varphi^*\left(g\left(\boldsymbol{y}_k^{P_m}\right)\right)$ can be flexibly expressed as inner products using the reproducing property, similarly to (4.10). More specifically:

$$g(\boldsymbol{y}_k^{P_{m'}}) = \left\langle g, \Phi(\boldsymbol{y}_k^{P_{m'}})\right\rangle \qquad \text{and} \qquad \varphi^*\left(g\left(\boldsymbol{y}_k^{P_m}\right)\right) = \varphi^*\left(\left\langle g, \Phi(\boldsymbol{y}_k^{P_m})\right\rangle\right) . \tag{5.9}$$

The penalty term $I(g)$, chosen as the induced norm $\|g\|_{\mathcal{H}}^2$ regulates the complexity of function g by penalising less smooth functions that tend to overfit the samples. Expression (5.8) thus takes the form:

$$\hat{F}_{\varphi;\boldsymbol{\delta}}(m, m') = \sup_{g\in\mathcal{G}}\left\{\frac{1}{N}\sum_{k=1}^{N}\left\langle g, \Phi(\boldsymbol{y}_k^{P_m})\right\rangle - \frac{1}{N}\sum_{k=1}^{N}\varphi^*\left(\left\langle g, \Phi(\boldsymbol{y}_k^{P_{m'}})\right\rangle\right) - \frac{\rho}{2}\|g\|_{\mathcal{H}}^2\right\} . \tag{5.10}$$

Proposition 1 provides an expression of (5.10) that is completely independent of mapping Φ. Definition of Φ is a common challenge in numerous applications that adopt kernel methods such as kernel principal component analysis (PCA ;Bishop (2006)) and kernel support vector machines (SVM; Burges et al. (1999)) and is typically avoided using property (4.9).

Proposition 1 (Partial utility estimator)**.** *An estimator of the partial utility $F_{\varphi;\boldsymbol{\delta}}(m, m')$ of models $m, m' \in \mathcal{M}$ with corresponding distributions P_m and $P_{m'}$, can be written as a sum of elements in the form:*

$$\hat{F}(m, m') = \sum_{k=1}^{N}\left[\tilde{a}_k\varphi'(N\tilde{a}_k) - \frac{1}{N}\varphi^*\left(\varphi'(N\tilde{a}_k)\right)\right], \tag{5.11}$$

where:

$$\tilde{\boldsymbol{a}} = \arg\inf_{\boldsymbol{a}=\{a_1\dots,a_N\}}\left\{\sum_{k=1}^{N}\left[a_k\varphi'(Na_k) - \frac{1}{N}\varphi^*\left(\varphi'(Na_k)\right)\right] + \frac{1}{2\rho}\sum_{k,l=1}^{N}a_k a_l K(\boldsymbol{y}_k^{P_{m'}}, \boldsymbol{y}_l^{P_{m'}})\right.$$
$$\left. - \frac{1}{N\rho}\sum_{k,l=1}^{N}a_l K(\boldsymbol{y}_k^{P_m}, \boldsymbol{y}_l^{P_{m'}}) + \frac{1}{2\rho N^2}\sum_{k,l=1}^{N}K(\boldsymbol{y}_k^{P_m}, \boldsymbol{y}_l^{P_m})\right\} , \tag{5.12}$$

φ^* the conjugate and φ' the derivative of function φ.

Proof. Defining the following terms:

$$\begin{aligned} T_{1,k}(g) &= -\frac{1}{N}\left\langle g, \Psi(\boldsymbol{y}_k^{P_{m'}})\right\rangle \\ T_{2,k}(g) &= \frac{1}{N}\varphi^*\left(\left\langle g, \Psi(\boldsymbol{y}_k^{P_{m'}})\right\rangle\right) \\ T_3(g) &= \frac{\rho}{2}\|g\|^2 \ , \end{aligned}$$

expression (5.10) can be restated as its dual representation through the following operations:

$$\begin{aligned} \hat{F}(m,m') &= \sup_{g\in\mathcal{G}}\left\{\langle 0, g\rangle - (T_{1,k}+T_{2,k}+T_3)(g))\right\} \\ &= (T_{1,k}+T_{2,k}+T_3)^*(0) \ . \end{aligned} \tag{5.13}$$

Application of Theorem 2 decomposes expression (5.13) into individual conjugate functions and so:

$$\hat{F}_{\varphi;\boldsymbol{\delta}}(m,m') = \inf_{v,u}\left\{\sum_{k=1}^N T_{1,k}^*(v_k) + \sum_{k=1}^N T_{2,k}^*(u_k) + T_3^*\left(-\sum_{k=1}^N v_k - \sum_{k=1}^N u_k\right)\right\} \ . \tag{5.14}$$

Evaluation of terms T_1^*, T_2^* and T_3^* can be simply achieved through the Fenchel transform, resulting in the following representations, a detailed derivation of which is provided in Appendix C.1:

$$T_{1,k}^*(v_k) = \begin{cases} 0, & \text{for } u_k = -\frac{1}{N}\Phi(\boldsymbol{y}_k^{P_m}) \\ +\infty, & \text{otherwise} \end{cases} \tag{5.15}$$

$$T_{2,k}^*(u_k) = \begin{cases} a_k\varphi'(Na_k) - \frac{1}{N}\varphi^*\left(\varphi'(Na_k)\right), & \text{for } v_k = a_k\Phi(\boldsymbol{y}_k^{P_{m'}}) \\ +\infty, & \text{otherwise} \end{cases} \tag{5.16}$$

$$T_3^*\left(-\sum_{k=1}^N v_k - \sum_{k=1}^N u_k\right) = \frac{1}{2\rho}\left\|-\sum_{k=1}^N v_k - \sum_{k=1}^N u_k\right\|_{\mathcal{H}}^2 \ . \tag{5.17}$$

Substituting (5.16), (5.15) and (5.17) into (5.14) results in expression:

$$
\begin{aligned}
&\hat{F}_{\varphi;\boldsymbol{\delta}}(m,m') = \\
&\quad = \inf_{v,u} \left\{ \sum_{k=1}^{N} a_k \varphi'(N a_k) - \frac{1}{N} \varphi^* \left(\varphi'(N a_k) \right) + \frac{1}{2\rho} \left\| - \sum_{k=1}^{N} a_k \Phi(\boldsymbol{y}_k^{P_{m'}}) + \frac{1}{N} \sum_{k=1}^{N} \Phi(\boldsymbol{y}_k^{P_m}) \right\|_{\mathcal{H}}^2 \right\} \\
&\quad = \inf_{v,u} \left\{ \sum_{k=1}^{N} a_k \varphi'(N a_k) - \frac{1}{N} \varphi^* \left(\varphi'(N a_k) \right) + \right. \\
&\qquad \left. \frac{1}{2\rho} \left[\sum_{k,l=1}^{N} a_k a_l K(\boldsymbol{y}_k^{P_{m'}}, \boldsymbol{y}_l^{P_{m'}}) - \sum_{k=1}^{N} \frac{a_k}{N} \sum_{l=1}^{N} K(\boldsymbol{y}_k^{P_{m'}}, \boldsymbol{y}_l^{P_m}) + \frac{1}{N^2} \sum_{k,l=1}^{N} K(\boldsymbol{y}_k^{P_m}, \boldsymbol{y}_l^{P_m}) \right] \right\},
\end{aligned}
$$

as required. Derivation of the final line makes use of the relationship $\left\langle \Phi(\boldsymbol{y}_k^{P_m}), \Phi(\boldsymbol{y}_P^l) \right\rangle = K\left(\boldsymbol{y}_k^{P_m}, \boldsymbol{y}_l^{P_m}\right)$.

Denoting as $\tilde{\boldsymbol{a}}$ the vector of points where the infimum is attained and letting $\rho \to 0$, expression (5.11) follows. □

5.2.3 Alternative function classes

Belghazi et al. (2018) consider an alternative class of functions to Section (5.2.2), that is parametrised by neural networks. Similarly to the class RKHS, their proposed class imposes no assumptions on the functions within while functions are highly flexible which allows them to approximate the estimated quantity with arbitrary accuracy. There is currently no indication to suggest that this alternative class achieves any improvement over RKHS and comparison of the two approaches is beyond the scope of this thesis, it was, therefore, not considered further.

5.3 Algorithmic representation

This section demonstrates an algorithmic implementation using the estimator presented in Proposition 1 for evaluation of the expected utility provided in Algorithm 3.

Algorithm 3 Estimation of the expected utility of $\boldsymbol{\delta}$ through variational methods.

1: Generate predictions $\boldsymbol{y}^{P_m} = \{\boldsymbol{y}_1^{P_m}, \ldots, \boldsymbol{y}_N^{P_m}\}$ from P_m for all $m \in \mathcal{M}$ at experimental condition δ
2: Obtain the conjugate φ^* and derivative φ' and substitute into (5.12)
3: **For** $m \in \mathcal{M}$ and $m' \in \mathcal{M}$ that $m \neq m'$ **do**
4: Use a suitable optimization algorithm to optimise $\hat{\boldsymbol{a}}$ for samples $\boldsymbol{y}^{P_m}$, $\boldsymbol{y}^{P_{m'}}$
5: Use $\tilde{\boldsymbol{a}}$ to evaluate $\hat{F}_{\varphi;\boldsymbol{\delta}}(m, m')$ according to (5.11)
6: **end for**
7: Average over $\hat{F}_{\varphi;\boldsymbol{\delta}}(m, m')$ for all $m, m' \in \mathcal{M}$, $m \neq m'$ to obtain an estimate of the expected utility $U(\boldsymbol{\delta})$ according to (2.3)

The procedure described in Algorithm 3 is expressed in a general form and is applicable for any choice of φ. Section 5.4 clarifies certain steps further by considering a specific choice of utility function.

5.4 Estimation of the KL divergence

This section considers the application of the proposed methodology for evaluation of the weighted KL divergence represented by the special case of (2.13) for choice of $\varphi_{\mathrm{KL}}(\chi(\boldsymbol{y})) = -\log(\chi(\boldsymbol{y}))$ given by expression (2.13). This expression of expected utility results from definition of the utility as the Shannon entropy. The motivation behind this application is to further clarify certain steps of Algorithm 3 for a specific case of φ that may currently appear to be quite generic. Steps for which no further clarification is deemed necessary are omitted.

- Step 2. The conjugate function of φ_{KL} is derived in Example 1 and given by:

$$\varphi^*\left(g(\boldsymbol{y})\right) = -1 - \log(-g(\boldsymbol{y}))\ ,\ g < 0$$

while $\varphi'\left(\chi\left(g(\boldsymbol{y})\right)\right) = -\dfrac{1}{\chi(g(\boldsymbol{y}))}$.

- Steps 4-5. Using the expressions derived in Step 2 and under Proposition 1, the KL

divergence estimator takes the form:

$$\hat{F}_{\varphi_{KL};\boldsymbol{\delta}}(m, m') = -\frac{1}{N}\sum_{k=1}^{N}\log(Na_k) \ ,$$

where:

$$\tilde{\boldsymbol{a}} = \arg\inf_{\boldsymbol{a}} \left\{ -\frac{1}{N}\sum_{k=1}^{N}\log(Na_k) + \frac{1}{2\rho}\sum_{k,l=1}^{N} a_k a_l K\left(\boldsymbol{y}_k^{P_{m'}}, \boldsymbol{y}_l^{P_{m'}}\right) \right.$$
$$\left. - \frac{1}{N\rho}\sum_{k,l=1}^{N} a_l K\left(\boldsymbol{y}_k^{P_m}, \boldsymbol{y}_l^{P_{m'}}\right) + \frac{1}{2\rho N^2}\sum_{k,l=1}^{N} K\left(\boldsymbol{y}_k^{P_m}, \boldsymbol{y}_l^{P_m}\right) \right\} . \qquad (5.18)$$

The optimal vector $\tilde{\boldsymbol{a}}$ can typically be obtained through simple optimisation routines as will be further discussed in Chapter 8. An additional example of $f-$divergence estimation using the proposed estimator is provided in Appendix C.2 for $\varphi_H(\chi(\boldsymbol{y})) = 2\sqrt{(\chi(\boldsymbol{y}))}$, $\chi > 0$. Choice of this function induces expression of the expected utility in terms of the Hellinger distance between predictive distributions corresponding to the competing models under study.

5.5 Discussion

The estimation process proposed in this chapter consists of two key steps. The initial step, presented in Sections 4.2 and 5.2.1, establishes a variational representation for the expected utility under the formulation induced by utility functions from the proposed function class $\mathcal{F}$. Exploitation of this dual expression provides an answer to the initial, computationally demanding evaluation problem through the solution of a less challenging optimisation problem. An empirical estimator of the expected utility is thus formulated as expression (5.8) which no longer relies on evaluation of the marginal likelihood but on a new function g from a measurable class of functions $\mathcal{G}$. This attribute makes the proposed approach particularly suited to the problems examined in this thesis where evaluation of the likelihood is computationally burdensome.

The second step of the estimation procedure relies on a clever choice of function class $\mathcal{G}$. Section 5.2.2 considers the class of RKHS resulting in a practical expression of the adopted estimator that relies only on evaluation of the Gram matrix using samples from

the predictive distributions of the competing models. The expression of each function evaluation within RKHS through a potentially infinite set of basis functions equips the resulting estimator with increased flexibility and efficiency, allowing it to extract information directly from the provided predictive samples. This property allows direct comparison of the competing distributions against each other without relying on an indirect representation of each considered sample through its corresponding marginal likelihood, employed by currently adopted Monte-Carlo based approaches. As a result, computationally challenging marginal likelihood estimation methods are no longer required.

Despite the various benefits incurred from approximation of a quantity of interest through variational techniques, some noteworthy challenges are also inherited. As shown in Section 4.2, evaluation of the expected utility is bounded below by its dual problem, expressed in (5.4). Equality is fulfilled under condition (5.6) in which case, solution of the corresponding optimisation problem provides the answer to the initial evaluation problem. In general, solution of the optimisation problem serves as a lower bound thus representing a worst-case scenario for the expected utility. The question is, therefore, raised regarding the proximity of the resulting approximation to the true value, often referred to as the tightness of the bound. In the context of this class of estimation problems, this issue has been commonly addressed through choice of the class function $\mathcal{G}$ ensuring the inclusion of sufficient options within to achieve approximations that are sufficiently close to the estimated value. Class functions allowing such conditions are considered in Section 5.2.2. In the more specific context of optimal experimental design for studies employing computationally demanding models, the examined issue poses a lesser challenge compared to those associated with the most commonly adopted Monte-Carlo based methods. Indeed, as shown in Chapter 3, Monte-Carlo estimators are, in general, not faced with such challenges as asymptotical convergence is guaranteed by Property 2. In other words, evaluation of the estimated function on an arbitrarily large population of N samples from the considered predictive distributions provides an estimate that is sufficiently close to the true value. However, as demonstrated in Chapters 8 and 6, the computational complexity involved in evaluation of the function of interest on each predictive sample may lead to corresponding waiting times of the order of hours and, in cases, days. Given that interest lies on completion of the optimal design study

within a realistic time frame, consideration of only a limited collection of samples is possible. Particularly, in the study examined in Chapter 5, restriction to a population of $50 - 100$ predictive samples is necessary for accomplishing this goal. Considering that this limited population lies far from infinity, Monte-Carlo estimators provide as much of an approximation as the proposed variational approach, inheriting the exact same issues. Nevertheless, the role of expected utility in optimal experimental design problems is to facilitate the comparison of alternative experimental conditions. Even when an exact value of the expected utility is not available, comparison of potential conditions under the alternative, worst case scenario is still informative to optimal design studies.

Furthermore, compared to Monte-Carlo based approaches the proposed approach provides a highly efficient framework for estimation of the expected utility. Using flexible, non-parametric methods, evaluation of N, computationally demanding marginal likelihood is avoided at the cost of an N-dimensional convex optimisation problem. This problem shift is shown to incur substantial computational savings as further supported by empirical comparison in Chapters 7 and 8. The rate of convergence is similar for the two approaches with order $\mathcal{O}(N^{-1/2})$ (Nguyen et al., 2010; Ryan, 2003).

Summary

This chapter introduces a novel methodology for estimation of the expected utility addressing challenges that have so far remained unresolved through currently adopted approaches. The proposed estimator allows consideration of a broad class of utility functions tailored to model discrimination problems including commonly employed utilities such as the Shannon entropy as well as other, less frequently considered functions that are however shown to possess further appealing properties improving the estimation performance. Efficacy of the proposed method is attributed to two key components: a dual representation, established in Section 4.2, allowing consideration of a theoretically simpler problem in place of the initial challenging evaluation problem. Focus is thus shifted to the solution of a convex optimisation problem over a class of measurable functions. Choice of the induced class constitutes the second key component and is considered in Section 5.2.2. A practical implementation of the proposed estimator is provided in Sections 5.3

and 5.4 including an algorithmic representation for evaluation of expected utility and an accompanying example of its application for a specific choice of utility function. The chapter concludes with a discussion reviewing the benefits and shortcomings associated with the presented methodology as well as the improvement it incurs over the currently adopted approaches.

The preceding Chapter 3 introduced two ongoing challenges in the experimental design literature: efficient evaluation of the expected utility is addressed in this chapter while the succeeding chapter focuses on optimisation of the expected utility over the design space incorporating the proposed estimator. Overcoming this challenges paves the way towards the fully automated study of complex phenomena that is also considered in Chapter 6.

Chapter 6

Sequential and adaptive experimental design

Methodology for efficient estimation of the expected utility was previously presented in Chapter 5, tackling one of the prevalent issues present in modern optimal experimental design problems. This chapter proceeds into considering a comprehensive framework for optimal experimental design on the basis of the proposed estimator, addressing common challenges such as expected utility estimation and efficient design optimisation. The adopted setup aims to deliver a fully automated solution to the study of natural phenomena through experiments, that requires minimal human input.

6.1 Sequential and adaptive designs

The importance of experimental data in modern scientific studies has been frequently outlined throughout the preceding chapters. Since the phenomena of interest are often not directly observable, the obtained information is vital for gaining insight into the inner workings of the studied system. Under the design setup considered so far, experimental data are used towards answering a particular objective. More specifically, within the class of problems targeted in this thesis, experimental data are incorporated into the model inference problem under study, providing evidence in support of one model against its rivals. Once the process of model selection has been completed, the experimental data

are not considered further. Such designs are commonly known as **static designs**.

Under a more sophisticated setup, already existing experimental data can be incorporated into the study, further informing future experimental design decisions. A sequential procedure is, thus, established under which, an update of the current knowledge occurs upon observation of the newly obtained experimental data, driving subsequent optimal design decisions. This, in turn, generates new experimental observations, leading to further updates, followed by proposition of their corresponding optimal designs and so on. This setup is commonly known as an **adaptive design** under which, observation of previous experiments impacts decisions on subsequent experiments.

This closed-loop procedure of data collection, knowledge update and optimal decision-making provides a fully automated and comprehensive setup for the study of scientific phenomena. The latter stage is predominantly addressed in this thesis, throughout Chapters 2-5. Knowledge update is, typically, achieved in a Bayesian context under which, experimental data are incorporated into the study through an update in the assumed distributions of unknown experimental parameters conditioned on the newly obtained information. This step can be achieved through the previously discussed SMC algorithm, provided in Appendix B or any other sampling algorithm when an analytical expression for the updated form is not available. The stage of data collection falls outwith the scope of this thesis. However it is worth noting that, availability of modern software, allowing instrument control for data collection at a selected experimental condition, facilitates complete process automation for the study of different phenomena through experiments.

An algorithmic representation of the examined procedure is provided in Algorithm 4. Focus in this thesis is placed on sequential designs under which, the most recent experimental observations are incorporated into decisions regarding only the very next experiment, an approach known as **myopic**. A more sophisticated, non-myopic design looks further than the immediately subsequent experiment and decisions are made accordingly, however such an approach will not be considered in detail in this thesis.

Algorithm 4 A fully automated framework for the study of modern phenomena through experiments

1: **while** (stopping conditions are not met) **do**

2: Generate samples $\boldsymbol{y}^{P_m} = \{\boldsymbol{y}_1^{P_m}, \ldots, \boldsymbol{y}_N^{P_m}\}$ from the prior predictive distributions for all $m \in \mathcal{M}$.

Make optimal decision

3: Acquire the optimal design $\boldsymbol{\delta}^*$ given $\boldsymbol{y}^{P_m}$, $m \in \mathcal{M}$.

4: Set $\boldsymbol{\delta}^*_{\text{seq}} \leftarrow \{\boldsymbol{\delta}^*\}$ during the initial cycle and $\boldsymbol{\delta}^*_{\text{seq}} \leftarrow \{\boldsymbol{\delta}^*_{\text{seq}}, \boldsymbol{\delta}^*\}$ onwards.

Observe studied system

5: Perform experiment at $\boldsymbol{\delta}^*$ and collect observed data $\mathbf{D}$.

Update current knowledge

6: Acquire the posterior predictive distribution $\boldsymbol{y}^{P_m|\mathbf{D}}$.

7: Set posterior as prior of the next stage $\boldsymbol{\theta}^{P_m} \rightarrow \boldsymbol{\theta}^{P_m|\mathbf{D}}$ and $\boldsymbol{y}^{P_m} \leftarrow \boldsymbol{y}^{P_m|\mathbf{D}}$, $m \in \mathcal{M}$.

8: **end while**

The efficient variational estimation approach, presented in Chapter 6 is essential for consideration of this framework within a realistic time frame. As previously discussed, the inherent complexity of the phenomena studied in this thesis, deem traditionally adopted optimal experimental design methods highly inefficient. This results in unrealistically long waiting times between the sequential stages of Algorithm 4. The proposed estimator is shown to address such issues, enabling an efficient transition from the optimal decision making stage to knowledge update and experimentation.

The performance of the procedure described in Algorithm 4 is, in addition, highly dependent on maximisation of the expected utility. This common challenge, associated with most experimental design problems, has been previously discussed in Chapter 3 wherein existing methodology for addressing such issues was presented. Briefly, two desirable conditions are targeted: 1) a comprehensive search of the design space, ensuring optimality of the proposed design. Sub-optimal designs may, alternatively, lead to considerable information loss and waste of experimental resources. 2) Completion of the optimisation stage within reasonable time scales. Unfortunately, a comprehensive coverage of Δ under an inefficient optimisation approach may lead to significantly long waiting times and so optimisation methods that achieve a balance between conditions 1 and 2 are necessary.

Under the sequential and adaptive setup of Algorithm 4, the shape of the targeted expected utility surface changes with each knowledge update and thus a new search is necessary at each cycle of the procedure. Challenges associated with inefficient optimisation approaches are, therefore, aggravated further under this setup, hindering the successful completion of the conducted study.

An efficient class of optimisation algorithms, addressing the examined issues is considered in the succeeding Section 6.2. The presented approach, in combination with the proposed variational estimator for evaluation of the expected utility are, in Chapters 8 and 9, shown to enable implementation of Algorithm 4 within a realistic time frame, providing a fully automated and efficient framework for the study of complex phenomena through experimentation.

6.2 Efficient search of the expected utility surface

Methodology for efficient optimisation of the expected utility is considered in this section. The presented approach is predominantly targeted to problems with continuous domains or, more generally, when deterministic comparison of every experimental condition in Δ is computationally infeasible. The class of Bayesian optimisation algorithms is employed for that purpose, providing a highly efficient framework which is particularly suited to problems employing computationally demanding functions. The presented methodology combined with the expected utility estimation procedure proposed in Chapter 5, provides a fully automated framework for the study of modern phenomena as demonstrated in this chapter.

Optimal experimental design finds application in a wide range of disciplines and may target various objectives (inverse problems, model selection, prediction). In any case, the question of interest remains unchanged and is summarised by:

$$\boldsymbol{\delta}^* = \arg\max_{\Delta} U(\boldsymbol{\delta}) \ , \tag{6.1}$$

where function $U(\boldsymbol{\delta})$ quantifies the expected utility incurred from a particular design $\boldsymbol{\delta}$ along the lines of the decision-theoretic context of Chapter 2. Although focus has been

so far drawn in definition and evaluation of the expected utility $U(\boldsymbol{\delta})$, optimisation of it over Δ, as stated by (6.1), remains unexplored.

In its simplest form, solution of problem (6.1) can be approached through evaluation of $U(\boldsymbol{\delta})$ — for every $\boldsymbol{\delta} \in \Delta$ under a discrete design space or a discretised subset under a continuous design space — and deterministic identification of the optimal condition $\boldsymbol{\delta}^*$. However, computation of the expected utility for every possible design quickly becomes infeasible as the size of designs under consideration increases, despite the computational improvement incurred by the proposed estimator. In addition, maximisation of the expected utility over a subset of experimental conditions may lead to highly suboptimal designs if the considered designs lie sufficiently far from the actual optimal solution.

Commonly adopted approaches, addressing this problem have already been discussed in Chapter 3. This chapter examines the class of Bayesian optimisation algorithms as an extension to the previously considered methodology. Although having been previously employed in design optimisation problems (Kleinegesse and Gutmann, 2019; von Kügelgen et al., 2019), Bayesian optimisation algorithms are not as widely adopted in this context despite their flexibility, efficiency and suitability to such problems.

Section 6.2.1 provides a general introduction to the Bayesian optimisation framework while its individual components are considered in Sections 6.2.2 and 6.2.3. Section 6.2.4 explores its role within optimal experiment design problems and establishes a complete and automated framework for studies of phenomena through experimentation.

6.2.1 Bayesian optimisation algorithms

The idea of Bayesian optimisation was introduced by Jonas Močkus in Močkus (1975, 1989), yet have only in recent years gained considerable attention. This can be predominantly attributed to an increase in both need and computational capacity: the former results from a general interest shift towards the study of more complex structures (Calder et al., 2006; Overstall and Woods, 2016) that are potentially costly to evaluate and may produce noisy observations thus requiring highly efficient search procedures. The latter is associated with the emergence of increasingly powerful computational resources that deem Bayesian optimisation algorithms executable within reasonable time frames.

Bayesian optimisation algorithms seek to locate the extrema of a particular function

of interest which will often be referred to as the **objective function**. As with any optimisation procedure, a sequence of steps is established during which the current evaluations of the objective function guides evaluations in future stages. The efficiency of Bayesian optimisation algorithms lies in their ability to minimise the required evaluations by employing an alternative function to serve in its place, also known as the **surrogate function**. Observation of this alternative function is typically achieved at a lower computational cost. The surrogate function incorporates observations from the objective function available up to the most recent optimisation step to improve its approximation performance and is subsequently used as a simulator of the expensive objective function in between optimisation steps. In a decision-theoretic setup, an additional optimisation procedure is introduced in between optimisation steps that locates the optimal set of points for the objective function to be observed next. Comparison of alternative sets is possible through an optimality criterion, commonly known as the **acquisition function**.

Two key components are, therefore, distinguished:

1. the *surrogate function* acting as a simulator of the objective function. At the initial step of the algorithm, the surrogate model reflects ones prior beliefs about the model and is sequentially refined as new observations become available, providing an increasingly improved approximation to the objective function. New information is incorporated through the likelihood function of the surrogate model into its posterior distribution which is subsequently used as a surrogate model in the succeeding optimisation step. This aspect attributes Bayesian optimisation algorithms their data efficiency. Commonly adopted optimisation algorithms, typically rely on the function evaluation and potentially some additional surface properties of their current state to inform their subsequent moves. Bayesian optimisation algorithms extract further information from every function evaluation observed up to that state thus guiding better informed decisions. Alternative surrogate models are further examined in Section 6.2.2.

2. The updated surrogate function, treated as the 'true' model, is used to inform selection of points at which observation of the objective function during the following optimisation step incurs the optimal results. Potential sets are ranked through the

acquisition function along the lines of the previously introduced utility function, although function definitions in the two cases typically differ reflecting their distinct objectives. Choice of the acquisition function establishes a desired level of trade-off between exploration and exploitation of the space ensuring both optimality and sufficient coverage of the area. Commonly adopted acquisition functions are presented in Section 6.2.3.

The two components produce and exchange information in an iterative way, as summarised in Algorithm 5. The following notation is adopted: at optimisation step i, the surrogate model approximating the expected utility surface is updated given the most current information $\mathcal{S}_{\text{current}} = \{s(\boldsymbol{x}^{(0)}), \ldots, s(\boldsymbol{x}^{(i-1)})\}$ where s represents the objective function and $\boldsymbol{x}^{(i)} \in \mathcal{X}$ the corresponding query point. The term $\alpha(\boldsymbol{x})$ will denote evaluation of the acquisition function at point $\boldsymbol{x}$. The optimisation procedure is repeated until a pre-defined stopping condition is fulfilled, typically set as a fixed number of steps or as a bound on the improvement acquired between two consecutive optimisation steps.

Algorithm 5 Bayesian Optimisation

1: Set $\mathcal{S}_{\text{current}} = \{s(\boldsymbol{x}^{(0)})\}$ for an initial point $\boldsymbol{x}^{(0)}$ and $i \leftarrow 0$.
2: **while** (stopping condition is not met) **do**
3: Update the surrogate function given data $\mathcal{S}_{\text{current}}$.
4: Optimise the acquisition function α to indicate new query points $\boldsymbol{x}^{(i)}$ such that:

$$\boldsymbol{x}^{(i)} = \arg\max_{\mathcal{X}} \alpha(\boldsymbol{x}) \tag{6.2}$$

and obtain evaluation of objective function $s(\boldsymbol{x}^{(i)})$.
5: Augment data $\mathcal{S}_{\text{current}} \leftarrow \{\mathcal{S}_{\text{current}}, s(\boldsymbol{x}^{(i)})\}$.
6: Set $i \leftarrow i + 1$.
7: **end while**
8: Acquire the optimal $\boldsymbol{x}^* = \arg\max_{\{\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(i-1)}\}} s(\boldsymbol{x})$.

Requirements for application of Bayesian optimisation algorithms is minimal as optimisation is possible simply based on pointwise (and potentially even noisy) evaluations

of the objective function and while additional surface properties such as gradient information can also be taken into account, their provision is not necessary.

In the context of experimental design problems, interest lies in maximisation of the expected utility U over the design space Δ which thus constitutes the objective function. As evaluation of the expected utility frequently incurs high computational cost as discussed in Chapter 3, Bayesian optimisation methods constitute an appealing option due to their efficiency with respect to the required function evaluations. The role of Bayesian optimisation in experiment design is revisited in Section 6.2.4.

The following sections 6.2.2 and 6.2.3 present commonly adopted surrogate models and acquisition functions respectively in a more general setup.

6.2.2 Surrogate models

This section provides a closer examination on the function of surrogate models and discusses commonly adopted choices.

Similarly to the previously considered methodology of Chapter 3, numerous models can serve as approximations to the expected utility surface. Bayesian optimisation relies on a probabilistic model which facilitates an update in the assumed distribution once new information is observed while also quantifying the uncertainty in the provided approximation. Parametric models constitute an appealing choice as they typically impose a simpler structure and involve less parameter tuning however their limited flexibility may fail to adequately capture the targeted surface. As a result, considerable focus is placed on non-parametric approaches (Ginsbourger et al., 2008; Grünewälder et al., 2010) and more specifically the framework of Gaussian Processes (GP) (Rasmussen and Williams, 2006) which is briefly considered in the succeeding paragraphs.

Gaussian processes Gaussian process models provide a flexible probabilistic framework over an unknown function of interest, following Definition 9.

Definition 9. *(Dudley, 2002) Let* $\mathbf{m} : \mathcal{X} \to \mathcal{R}$ *denote any real-valued function and* $\mathbf{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$ *a kernel function, as introduced in Chapter 5. A random function* $\mathbf{f} : \mathcal{X} \to \mathcal{R}$ *is said to be a Gaussian process with mean function* $\mathbf{m}$ *and covariance kernel* $\mathbf{K}$*, if for*

any finite set $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subseteq \mathcal{X}$ *and* n*, the random vector* $\mathbf{f} = \{\boldsymbol{f}(\boldsymbol{x}_1), \ldots, \boldsymbol{f}(\boldsymbol{x}_n)\}$ *follows a multivariate Normal distribution* $\mathcal{N}(\mathbf{m}, \mathbf{K})$*, where* $\mathbf{m}(\boldsymbol{x}) = (\mathbf{m}(\boldsymbol{x}_1), \ldots, \mathbf{m}(\boldsymbol{x}_n))$ *and* $\mathbf{K}(\boldsymbol{x}, \boldsymbol{x}) = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^n = [\mathrm{Cov}(\mathbf{f}(\boldsymbol{x}_i), \mathbf{f}(\boldsymbol{x}_j))]_{i,j=1}^n$*, where* $\boldsymbol{K}$ *represents the* $n \times n$ *variance-covariance matrix while* $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ *individual matrix entries.*

Bayesian Optimisation algorithms operate under the assumption of normally distributed function values $\boldsymbol{s}(\boldsymbol{x}) = \{s(\boldsymbol{x}_1), \ldots, s(\boldsymbol{x}_n)\}$ given $\mathbf{f}$, posing the following regression problem:

$$\begin{aligned} \mathbf{f} &\sim \mathcal{N}(\mathbf{m}, \mathbf{K}) \\ \boldsymbol{s} \mid \mathbf{f}, \sigma^2 &\sim \mathcal{N}\left(\mathbf{f}, \sigma^2 \mathbb{1}_n\right) , \end{aligned} \tag{6.3}$$

where σ^2 the error variance associated with observations $\boldsymbol{s}$ and $\mathbb{1}_n$ the $n \times n$ identity matrix.

In light of newly observed function evaluations $s(\boldsymbol{x})$ at $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, the prior assumption expressed in (6.3) can be updated to incorporate this information. Under the resulting posterior Gaussian process, $\mathbf{f}$ at an arbitrary new point $\boldsymbol{x}_{new}$ also follows a Normal distribution with mean $\mathbf{m}^*$ and variance σ^{*2} that obtain the form:

$$\begin{aligned} \mathbf{m}^*(\boldsymbol{x}_{new}) &= \mathbf{m}(\boldsymbol{x}_{new}) + \mathbf{k}(\boldsymbol{x}_{new}, \boldsymbol{x})^T [\mathbf{K}(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 \mathbb{1}_n]^{-1} (\boldsymbol{s}(\boldsymbol{x}) - \mathbf{m}(\boldsymbol{x})) \\ \sigma^{*2}(\boldsymbol{x}_{new}) &= k(\boldsymbol{x}_{new}, \boldsymbol{x}_{new}) - \mathbf{k}(\boldsymbol{x}_{new}, \boldsymbol{x})^T [\mathbf{K}(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 \mathbb{1}_n]^{-1} \mathbf{k}(\boldsymbol{x}_{new}, \boldsymbol{x}) , \end{aligned} \tag{6.4}$$

where $\mathbf{k}(\boldsymbol{x}_{new}, \boldsymbol{x})$ represents the column-vector of the covariance of $\boldsymbol{x}_{new}$ with each element of $\boldsymbol{x}$.

A collection of samples from the GP prior is depicted in the top plot of Figure 6.1 for $\mathbf{m}(\boldsymbol{x}) = 0$ for every $\boldsymbol{x}$ and $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\ell^2}\right)$, where $\|\cdot\|$ is the Euclidean norm and $\ell = 1$. The solid black line represents the prediction mean $\mathbf{m}$ while the grey lines represent 100 samples from the corresponding Gaussian process prior. The posterior GP in light of a newly observed dataset is subsequently presented in the bottom plot of Figure 6.1 in which the obtained observations are marked with red crosses.

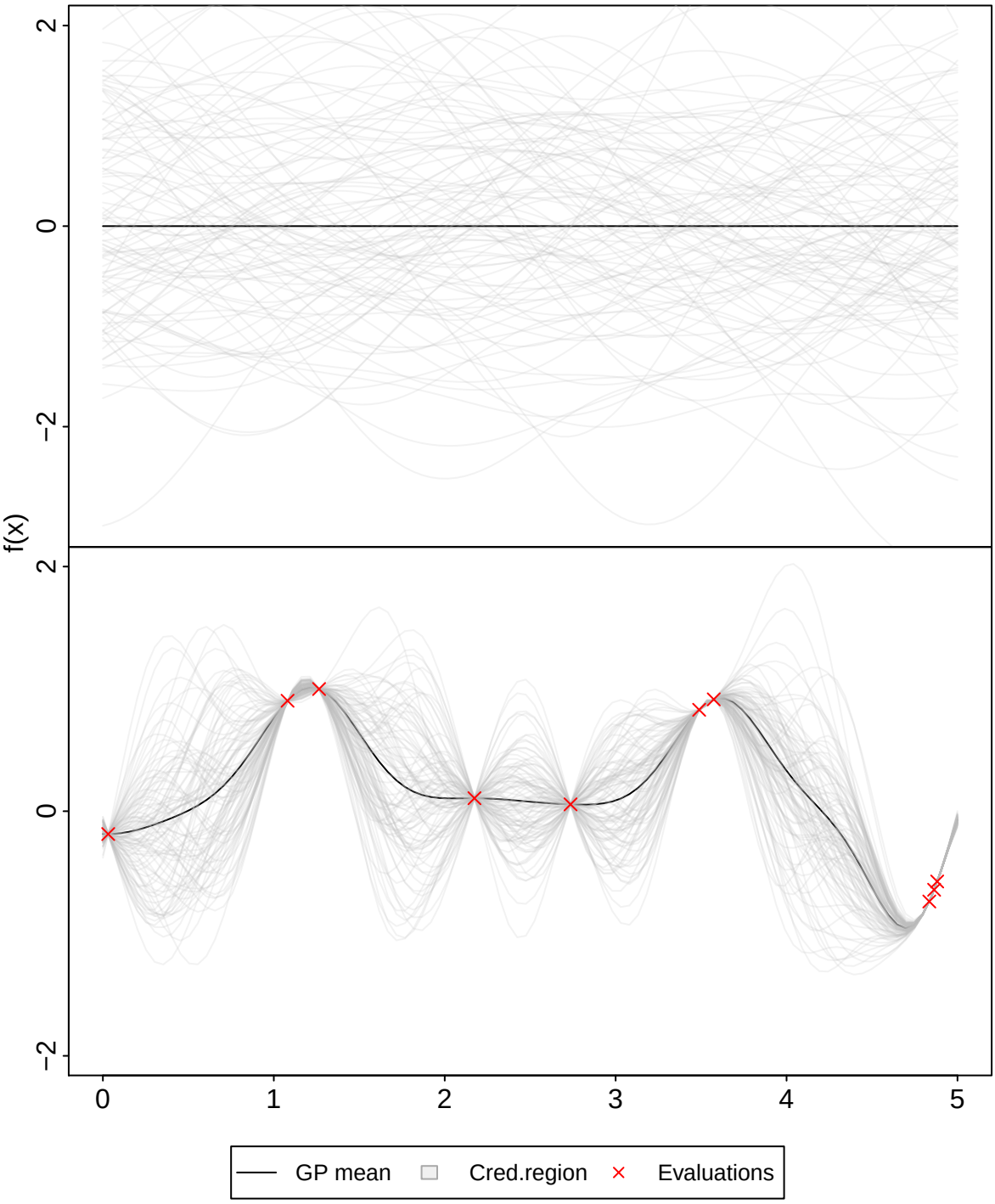


Figure 6.1: The top plot depicts a collection of 100 samples from the prior GP and the predictive mean, represented by grey and black solid lines respectively. Given a set of 10 observations, marked with red crosses, the GP can be updated to incorporate the new information, as shown in the bottom plot.

6.2.3 Acquisition functions

The idea of an acquisition function resembles the decision-theoretic utility function introduced in Section 5 aiming to quantify the utility of alternative sets of points at which the objective function is to be observed next. The acquisition function is thus maximised at each optimisation step as shown in Algorithm 5 with respect to a particular objective that is imposed through different definitions of the function, potential choices of which are discussed in succeeding paragraphs. In general, maximisation of the acquisition function seeks to maintain a desired trade-off between two actions: *exploration* of the design space and *exploitation* of current knowledge. Exploration ensures sufficient coverage of the design space driving queries towards points where the uncertainty in the surrogate model is relatively high. On the contrary, exploitation takes advantage of already existing information and thus motivates moves towards areas closer to the presently observed optimum.

Differing objectives give rise to a range of possible acquisition functions, a selection of commonly employed choices are considered in this section. A frequently adopted class of acquisition functions ranks new query points based on their potential improvement upon the current situation and are typically referred to as *Improvement-based*. An intuitive definition, introduced by Kushner (1964), considers the *Probability of Improvement*:

$$\alpha_{\mathrm{PI}}(\boldsymbol{x}) = Pr(\mathbf{f}(\boldsymbol{x}) > \mathbf{f}_{\max}) \tag{6.5}$$

where $\mathbf{f}_{\max}$ refers to the highest prediction observed up to the current stage of the Bayesian optimisation algorithm. Estimates of quantity 6.5 can be obtained by generating a collection of samples $\mathbf{f}(\boldsymbol{x})$ from the GP and subsequently summing over evaluations of the 0-1 utility function, as shown in Section 2.2.3, assigning value 1 if $\mathbf{f}(\boldsymbol{x})$ offers an improvement over $\mathbf{f}_{\max}$ and 0 otherwise.

Such choice incurs a considerable amount of information loss indicating only whether a case was successful or not while disregarding any information about the level of improvement. A modified expression, known as the *Expected Improvement*, allows for consideration of this information by incorporating the term $\mathbf{f}(\boldsymbol{x}) - \mathbf{f}_{\max}$ resulting in the alternative

definition:

$$\alpha_{\text{EPI}}(\boldsymbol{x}) = \mathbb{E}\left[(\mathbf{f}(\boldsymbol{x}) - \mathbf{f}_{\max})\right] \tag{6.6}$$

$$= (\mathbf{m}(\boldsymbol{x}) - \mathbf{f}_{\max})\,\Phi(z(\boldsymbol{x})) + \sigma(\boldsymbol{x})\phi(z(\boldsymbol{x}))\ , \tag{6.7}$$

where $z(\boldsymbol{x}) = \frac{\mathbf{m}(\boldsymbol{x}) - \mathbf{f}_{\max}}{\sigma^2(\boldsymbol{x})}$. Derivation from 6.6 to 6.7 is specific to using a Gaussian process as a surrogate model where, as shown in Section 6.2.2, the posterior distribution evaluated at any arbitrary point is normally distributed. Functions Φ and ϕ refer to the Gaussian c.d.f. and p.d.f. respectively. The EI incorporates both elements of exploration and exploitation. The former is encouraged through inclusion of the first term, reflecting the improvement upon the current maximum prediction, while the latter is captured through the second term that accounts for the uncertainty in a particular search area.

An alternative class operates under the optimistic policy of considering the best-case scenario in the face of uncertainty. Due to this property, members of this class are often called *optimistic policies*. The acquisition function, commonly known as the *Upper Confidence Bound*, is defined as:

$$\alpha_{\text{UCB}}(\boldsymbol{x}) = \mathbf{m}(\boldsymbol{x}) + \epsilon\sigma(\boldsymbol{x})\ , \tag{6.8}$$

where ϵ can be tuned for optimal performance according to the guidelines provided in Srinivas et al. (2010).

A comparison between alternative options in the context of optimal experiment design is briefly considered in Chapter 8.

6.2.4 Bayesian optimisation for optimal experimental design

This section explores the function of Bayesian optimisation in experiment design problems and establishes an automated framework for the study of different phenomena through experiments which constitutes a central objective of this thesis. Emphasis is particularly placed on phenomena characterised by highly complex structures that, as a result, require employment of more sophisticated and thus computationally demanding models. As previously discussed in preceding chapters, these problems incur numerous challenges when

employing commonly adopted methodologies. Chapter 5 addressed the issue of evaluating the benefit associated with potential experimental conditions within a realistic time frame by proposing a highly efficient estimation procedure. Exploration over the design space for optimisation of the expected utility is subsequently tackled in this chapter.

Section 6.2.1 provides a general outline of Bayesian optimisation algorithms while Sections 6.2.2 and 6.2.3 introduce their two key components: the surrogate model and acquisition function. This section focuses on the specific application of Bayesian optimisation within experimental design and more specifically for optimisation of the expected utility U over the design space, a summary of which is provided in Algorithm 6. The variational estimator of Chapter 5 is incorporated in the optimisation procedure for efficient evaluation of the expected utility at a particular experimental condition $\boldsymbol{\delta}$. In addition, the flexible probabilistic framework of Gaussian processes is adopted as an approximation to the expected utility surface.

Algorithm 6 Bayesian optimisation for maximisation of the expected utility over Δ

1: Set $\boldsymbol{s}_{\text{current}} = \{U(\boldsymbol{\delta}^{(0)})\}$ for an initial design $\boldsymbol{\delta}^{(0)}$ and $i \leftarrow 0$.
2: **while** (stopping condition is not met) **do**
3: Obtain the Gaussian process posterior given $\boldsymbol{s}_{\text{current}}$ following (6.4).
4: Optimise the acquisition function α to indicate new query points $\boldsymbol{\delta}^{(i)}$ such that:

$$\boldsymbol{\delta}^{(i)} = \arg\max_{\boldsymbol{\delta}} \alpha(\boldsymbol{\delta}) \tag{6.9}$$

and evaluate the corresponding expected utility $U(\boldsymbol{\delta}^{(i)})$ using Algorithm 3.
5: Augment data $\boldsymbol{s}_{\text{current}} \leftarrow \{\boldsymbol{s}_{\text{current}}, U(\boldsymbol{\delta}^{(i)})\}$.
6: Set $i \leftarrow i + 1$.
7: **end while**
8: Acquire the optimal $\boldsymbol{\delta}^* = \arg\max_{\{\boldsymbol{\delta}^{(0)},\ldots,\boldsymbol{\delta}^{(i-1)}\}} U(\boldsymbol{\delta})$.

Algorithm 6 offers an efficient framework for optimal decision making, constituting a vital stage of the sequential adaptive procedure proposed in this chapter and summarised in Algorithm 4. Implementation of Bayesian optimisation algorithms for search of the expected utility surface in the context of a real-life problem is examined in Chapter 8.

Summary

This chapter presents a comprehensive and fully automated setup for the study of modern phenomena, requiring minimal input from the experimenter. The class of Bayesian optimisation algorithms that are particularly well-suited to addressing optimal experimental design problems is also considered. The presented class can flexibly incorporate the previously proposed variational estimator, offering a highly efficient setup for optimisation of sequential and adaptive designs. Key components of this procedure, namely, the surrogate model and acquisition function are briefly examined while implementation of Bayesian optimisation algorithms for maximisation of the expected utility within an optimal experimental design problem is considered subsequently. Combination of the efficient estimation and Bayesian optimisation frameworks are shown to allow establishment of an autonomous and efficient setup for the study of highly complex structures expressed by computationally expensive models within realistic timelines. Its application, initially to a simple benchmark study of model discrimination followed by real-life studies from the fields of Systems Biology and Spectroscopy, is further considered in Chapters 7-9.

Chapter 7

Optimal experimental design for model discrimination

This chapter explores the application of methodologies, presented in Chapters 2-6, to a commonly employed benchmark study of model discrimination using experimental data. Performance of the proposed variational estimator and the traditionally adopted Monte-Carlo based approach for evaluation of the expected utility is assessed under a setup where an arbitrarily close representation of the true, estimated value is available. The chapter concludes with consideration of a sequential design setup on the basis of the examined estimators, providing a fully automated and efficient solution to the studied problem.

7.1 Problem setup

This chapter studies the optimal design of experiments targeting a model discrimination problem, initially considered by Box and Hill (1967). The study employs a collection of polynomial models, representing alternative hypotheses on an assumed underlying process, observed through an obtained experimental dataset. This section provides an introduction to the competing hypotheses and their formulation through statistical models in a Bayesian setting. As subsequently shown, under this particular model setup the study benefits from a closed form expression for the marginal likelihood, allowing

a representation of the expected utility that is considered to be arbitrarily close to the truth. The produced values will in Section 7.2 serve as a 'gold' standard for assessing the performance of the examined estimators.

Due to the simplicity of the adopted models, the studied case is not considered to be entirely representative of the class of problems targeted in this thesis. This is because application of standard Monte-Carlo approaches, although still costly, is not hindered by the issues associated with computationally intensive likelihoods, discussed in Chapter 3. Nevertheless, interest lies on establishing an initial assessment for the performance of the proposed estimator and a comparison with existing approaches. Unlike the real-life applications examined in the succeeding Chapters 8 and 9, assessment of the competing estimators is possible against the 'true' approximated value, providing useful insight on their performance.

7.1.1 Alternative hypotheses

The following four models were considered by Box and Hill (1967) for application of their examined optimal design methodology:

$$\begin{aligned}
&\textbf{Model 1:} && \eta_1(\beta_1, \xi) = \beta_{11}\xi \\
&\textbf{Model 2:} && \eta_2(\beta_2, \xi) = \beta_{20} + \beta_{21}\xi \\
&\textbf{Model 3:} && \eta_3(\beta_3, \xi) = \beta_{30} + \beta_{31}\xi + \beta_{32}\xi^2 \\
&\textbf{Model 4:} && \eta_4(\beta_4, \xi) = \beta_{41}\xi + \beta_{42}\xi^2 \ ,
\end{aligned} \tag{7.1}$$

where $\boldsymbol{\beta}_m = \left(\beta_{m0}, \ldots, \beta_{m(\kappa_m - 1)}\right)$ expresses the vector of κ_m parameters associated with model $m = 1, 2, 3, 4$. Parameter ξ denotes a controllable model input and is, in this case study, treated as the experimental condition. Thus the general notation $\boldsymbol{\delta}$ and the case-specific ξ will be, in this chapter, used interchangeably.

An equivalent vector-matrix notation restates the model outputs in (7.1) as:

$$\eta_m(\boldsymbol{\beta}_m, \xi) = \boldsymbol{\Xi}_m^T \boldsymbol{\beta}_m \ ,$$

for $m = 1, 2, 3, 4$, where $\boldsymbol{\Xi}_m$ is the $\kappa_m \times \lambda$ matrix of inputs, corresponding to model m, with

λ representing the dimensionality of the model output. In the considered study, $\lambda = 1$.

7.1.2 Statistical models

Under the assumption of normally distributed observed error, the likelihood of $\boldsymbol{y}$ under model m and corresponding model parameters $\boldsymbol{\theta}_m = (\boldsymbol{\beta}_m, \sigma_m^2)$ at experimental condition $\boldsymbol{\delta}$, is formulated as:

$$f(\boldsymbol{y} \mid m, \boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\delta}) = \mathcal{N}(\boldsymbol{y};\ \eta_m(\boldsymbol{\beta}_m, \boldsymbol{\delta}), \sigma_m^2)\ , \tag{7.2}$$

where $\mathcal{N}(\cdot\ ;\ \eta_m, \sigma_m^2)$ expresses the probability density function corresponding to the Normal distribution with mean η_m and error variance σ_m^2.

The unknown vector of model parameters θ_m is assigned prior distributions, reflecting initial beliefs on their values without taking into account any information from observation of the underlying process under study. The following prior distributions were adopted in this study:

$$\begin{aligned} \boldsymbol{\beta_m} \mid \sigma_m^2 &\sim \mathcal{MVN}(\boldsymbol{\mu}_{m0}, \boldsymbol{\Lambda}_{m0}\sigma_m^2) \\ \sigma_m^2 &\sim IG(a_{m0}, b_{m0})\ , \end{aligned} \tag{7.3}$$

where $\mathcal{MVN}$ represents the multivariate Normal distribution with κ_m-mean vector $\boldsymbol{\mu}_{m0}$ and variance-covariance matrix $\boldsymbol{\Lambda}_{m0}\sigma_m^2$, with $\boldsymbol{\Lambda}_{m0}$ being a $\kappa_m \times \kappa_m$ matrix. Parameter σ_m^2 is assigned an inverse Gamma distribution, denoted by IG and parametrised by a_{m0}, b_{m0} positive constants. As there was no indication on whether $\boldsymbol{\beta_m} \mid \sigma^2$ admits only positive or negative values, a prior centred around 0 was preferred and thus $\boldsymbol{\mu}_{m0} = \mathbf{0}$. A relatively wide spread on its potential values was imposed through σ^2 in order to account for multiple potential scenarios, leading to choices of hyperparameters $a_0 = 3$ and $b_0 = 1$ and $\boldsymbol{\Lambda}_{m0} = \mathbb{1}_{\kappa_m}$, for $\mathbb{1}_{\kappa_m}$ the $\kappa_m \times \kappa_m$ identity matrix.

The particular choice of prior distributions was preferred because of their property of conjugacy when considering regression models of the form (7.2), providing an analytical expression of the model parameter posterior distribution and the marginal likelihood. Availability of a closed form solution allows quick evaluation of the expected utility with-

out resorting to the computationally intensive approaches, discussed in Chapter 3. Under this setup, the marginal likelihood, previously defined in (2.6), obtains the following expression:

$$p(\boldsymbol{y} \mid \boldsymbol{\delta}, m) = St_{2a_{m0}}\left(\boldsymbol{y};\ \boldsymbol{\Xi}_m^T \boldsymbol{\mu}_{m0}, \frac{b_{m0}}{a_{m0}}\left(1 + \boldsymbol{\Xi}_m^T \boldsymbol{\Lambda}_{m0} \boldsymbol{\Xi}_m\right)\right) , \tag{7.4}$$

where $St(\cdot;\cdot)$ represents the probability density function of the Student-t distribution with $2a_{m0}$ degrees of freedom. Given an observed dataset $\tilde{y}$, the posterior marginal $p(\boldsymbol{y} \mid \tilde{\boldsymbol{y}}, \boldsymbol{\delta})$ follows expression (7.4), parametrised by the posterior hyperparameters $\boldsymbol{\mu}_{mn}, \boldsymbol{\Lambda}_{mn}, a_{mn}, b_{mn}$ conditioned on dataset $\tilde{\boldsymbol{y}}$, provided below:

$$\begin{aligned}
\boldsymbol{\mu}_{mn} &= \left(\boldsymbol{\Lambda}_{m0}^{-1} + \boldsymbol{\Xi}_m \boldsymbol{\Xi}_m^T\right)^{-1}\left(\boldsymbol{\Lambda}_{m0}^{-1}\boldsymbol{\mu}_{m0} + \boldsymbol{\Xi}_m \boldsymbol{y}\right) \\
\boldsymbol{\Lambda}_{mn} &= \left(\boldsymbol{\Lambda}_{m0}^{-1} + \boldsymbol{\Xi}_m \boldsymbol{\Xi}_m^T\right)^{-1} \\
a_{mn} &= a_{m0} + \frac{1}{2} \\
b_{mn} &= b_{m0} + \frac{1}{2}\left[\boldsymbol{\mu}_{m0}^T \boldsymbol{\Lambda}_{m0}^{-1} \boldsymbol{\mu}_{m0} + \boldsymbol{y}^2 - \boldsymbol{\mu}_{mn}^T \boldsymbol{\Lambda}_{mn}^{-1} \boldsymbol{\mu}_{mn}\right] .
\end{aligned}$$

Expression (7.4) allows convenient evaluation of the expected utility in the context of model discrimination problems, as demonstrated in the succeeding section. This representation is further used as a benchmark for comparison of two estimators under examination: the variational estimator, introduced in Chapter 5 and the commonly adopted Monte-Carlo based approaches, relying on approximation of the marginal likelihood, discussed in Chapter 3.

7.2 Evaluation of the expected utility

Following Box and Hill (1967), the set of experimental conditions under consideration was defined as $\Delta = \{0, 5, 10, 15\}$. An initial impression of the model prediction density plots based on samples from the prior distributions of model parameters and evaluated for each experimental condition in Δ is provided in Figure 7.1.

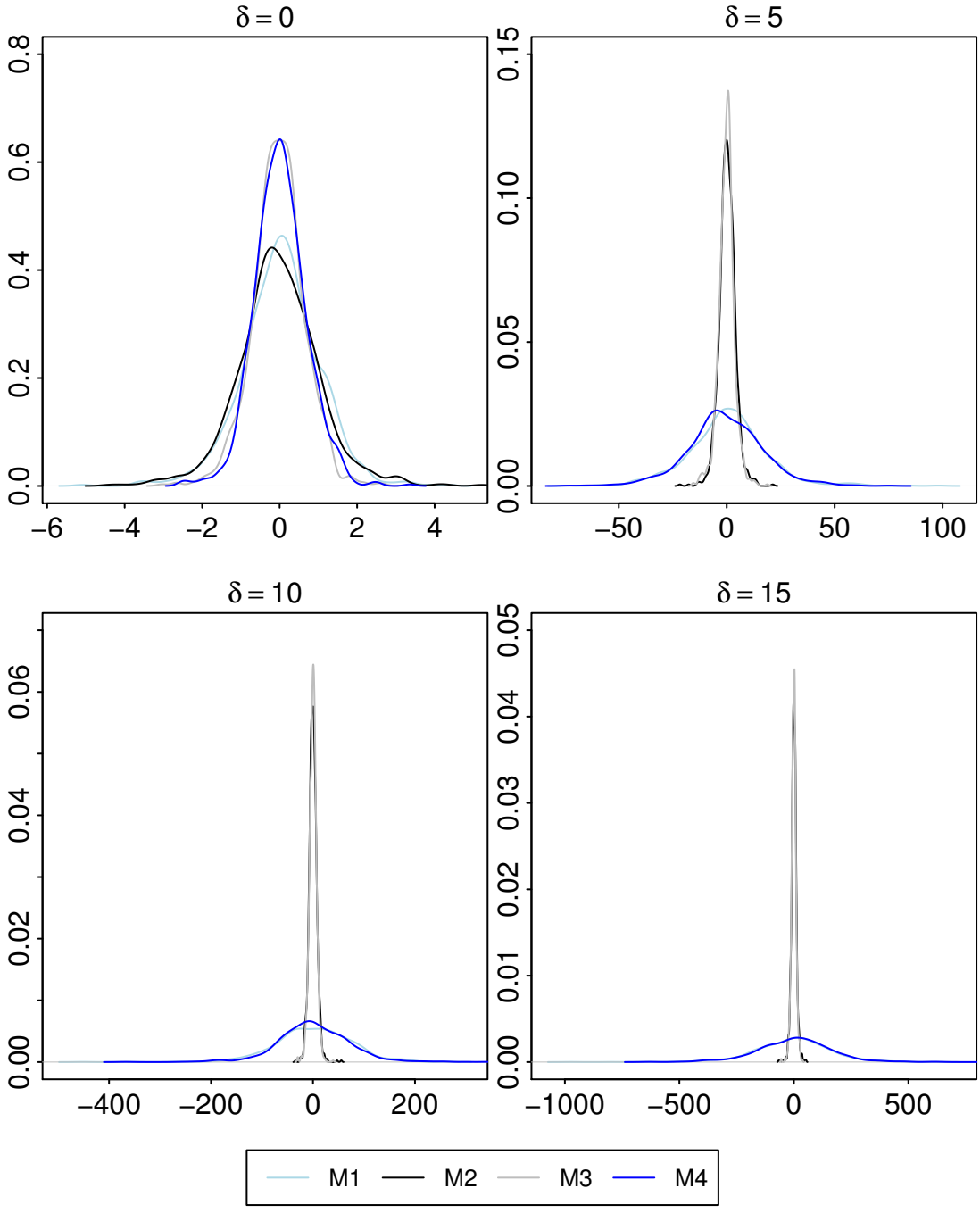


Figure 7.1: Density plots of model predictions based on samples from the prior distribution of model parameters associated with the studied experimental conditions in Δ.

7.2.1 Choice of utility

Following the experimental design framework introduced in Chapters 2 and 5, comparison of the candidate experimental conditions $\boldsymbol{\delta} \in \Delta$ required specification of the utility function, as a measure of optimality in the context of the studied problem. As interest, in this case study, lay on optimal experimental design for model inference, choice of the utility function was such that reflected the benefit an experimental condition incurred towards effective discrimination between the candidate models. The class of utility functions described by (5.2) was, therefore, adopted for this case study. The findings presented in this section were obtained under the particular choice of $\varphi(\chi) = -\log(\chi)$, inducing the most commonly adopted Shannon entropy, introduced in (2.13) and briefly restated below for convenience:

$$
\begin{aligned}
U(\boldsymbol{\delta}) &= \sum_{m \in \mathcal{M}} \sum_{\substack{m' \in \mathcal{M}, \\ m' \neq m}} \mathrm{KL}_{\boldsymbol{\delta}}(m, m')\ p(m')\ p(m) \\
&= \sum_{m \in \mathcal{M}} \sum_{\substack{m' \in \mathcal{M}, \\ m' \neq m}} \left[\int_Y \log \frac{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m)}{p(\boldsymbol{y} \mid \boldsymbol{\delta}, m')} p(\boldsymbol{y} \mid m, \boldsymbol{\delta})\ \mathrm{d}\boldsymbol{y} \right] p(m')\ p(m)\ ,
\end{aligned}
$$

where $\mathrm{KL}_{\boldsymbol{\delta}}(m, m')$ expresses the KL divergence between the predictive distributions corresponding to models m and m' at experimental condition $\boldsymbol{\delta}$. Under the induced expression, evaluation of the expected utility, effectively reduces to that of $\mathrm{KL}_{\boldsymbol{\delta}}(m, m')$ for all possible 12 pairs of models. It is worth noting that, as the KL divergence is not symmetric, separate evaluations for quantities $\mathrm{KL}_{\boldsymbol{\delta}}(m, m')$ and $\mathrm{KL}_{\boldsymbol{\delta}}(m', m)$ are required. The candidate models of set $\mathcal{M}$ were considered equiprobable and so the quantity $p(m') \cdot p(m) = \frac{1}{4 \cdot 3} = \frac{1}{12}$ for each pair of $m, m' \in \mathcal{M}$, where $m \neq m'$.

7.2.2 Evaluation based on a closed-form expression

Unfortunately, an analytical expression for the KL divergence is not available under the considered model setup. Efficient estimation can, nevertheless, be achieved through the closed form representation of the marginal likelihood provided in (7.4). Evaluation relies on adoption of a Monte-Carlo estimator for approximation of the integral involved in the expression of KL divergence, previously derived in (3.4). Under (7.4) the corresponding

Monte-Carlo estimator takes the form:

$$\begin{aligned}\hat{\mathrm{KL}}_{\boldsymbol{\delta}}(m,m') &= \frac{1}{N^*}\sum_{i=1}^{N^*}\left[\log p(\boldsymbol{y}_i^{P_m}\mid\boldsymbol{\delta},m)-\log p(\boldsymbol{y}_i^{P_m}\mid\boldsymbol{\delta},m')\right]\\ &= \frac{1}{N^*}\sum_{i=1}^{N^*}\left\{\log\left[St_{2a_{m0}}\left(\boldsymbol{y}_i^{P_m};\boldsymbol{\Xi}_m\boldsymbol{\mu}_{m0},\frac{b_{m0}}{a_{m0}}\left(1+\boldsymbol{\Xi}_m\boldsymbol{\Lambda}_{m0}\boldsymbol{\Xi}_m^T\right)\right)\right]\right.\\ &\qquad\left.-\log\left[St_{2a_{m'0}}\left(\boldsymbol{y}_i^{P_m};\boldsymbol{\Xi}_{m'}\boldsymbol{\mu}_{m'0},\frac{b_{m'0}}{a_{m'0}}\left(1+\boldsymbol{\Xi}_{m'}\boldsymbol{\Lambda}_{m'0}\boldsymbol{\Xi}_{m'}^T\right)\right)\right]\right\},\end{aligned}\tag{7.5}$$

where $\boldsymbol{y}_i^{P_m}$, $i=1,\ldots,N^*$ are samples from the predictive distribution corresponding to model m. Efficient evaluation of the marginal likelihood through its closed form expression allows consideration of an arbitrarily large population size N^* from the competing predictive distributions. As a result and due to Property 2 of Monte-Carlo estimators, stating its asymptotical convergence to the true value as the number of considered samples grows to infinity, the obtained estimate will be treated as a sufficiently close representation of the true value and, therefore, serve as a benchmark for comparison of the two compared estimators, considered in Sections 7.2.3 and 7.2.4.

7.2.3 Variational approximation

This section examines application of the variational estimator, proposed in Chapter 5, for evaluation of the expected utility at any given experimental condition $\boldsymbol{\delta}\in\Delta$. Estimation of $U(\boldsymbol{\delta})$ is achieved following Algorithm 3 for choice of $\varphi(\chi)=-\log(\chi)$, inducing the more specific formulation, discussed in Section 5.4. Unlike the traditionally adopted Monte-Carlo estimators of (7.5), the proposed estimator avoids evaluation of the marginal likelihood for each sample from the predictive distribution by efficiently comparing the collections of samples from the competing predictive distributions against each other. As described in Chapter 5, this is achieved through a dual representation of the initial, computationally challenging evaluation problem as the optimisation problem, summarised in Proposition 1. On the contrary, Monte-Carlo methods consider each sample from the population individually, information from which is extracted through the marginal likelihood. As discussed in Chapter 3 and further shown in Chapters 8 and 9, application of Monte-Carlo approaches quickly becomes intractable when evaluation of the marginal

likelihood is not available in closed form and so, more efficient estimation methods are required. A comparison of the two methods through empirical results is provided in Section 7.2.5.

Credible intervals summarising the obtained estimates under alternative sample sizes compared against the 'true' value are shown in Figure 7.2. The structure of the RKHS class induced by the Gaussian kernel was imposed on function class G. This is potentially the most commonly adopted choice of kernel assuming fairly smooth functions within the class, definitions of which takes the form:

$$K(\boldsymbol{y}^i, \boldsymbol{y}^j) = \exp\left\{-\frac{\|\boldsymbol{y}^i - \boldsymbol{y}^j\|^2}{s}\right\}, \tag{7.6}$$

where $\|\cdot\|$ represents the Euclidean metric in $\mathbb{R}^\lambda$ and $s > 0$. The term $K(\boldsymbol{y}^i, \boldsymbol{y}^j)$ is often interpreted as a similarity measure between $\boldsymbol{y}^i$ and $\boldsymbol{y}^j$. Under definition (7.6), parameter s regulates the assumed spread between them and should be, therefore, chosen within a range, representative of the overall considered distributions. To this end, s was chosen as the variance of the joint predictive distribution corresponding to m and m'.

An additional parameter determining the behaviour of the estimator is the penalty weight ρ that is present in the corresponding optimisation problem, as shown in Proposition 1. Choice of the value for ρ was based on findings of Gretton et al. (2007) and Ruderman et al. (2012) who show that, when $\rho = \mathcal{O}(N^{-1})$ the produced estimates are bounded away from infinity under a finite vector $\hat{a}$. Alternatively the authors claim that, penalty weights of a smaller order would attribute unnecessarily high significance to the penalty causing it to dominate over the remaining terms, thus resulting in convergence of the estimator to a positive constant that is not necessarily the true value. Larger order penalties would, on the other hand, underestimate the discrepancy between the two distributions setting the penalty as 0 even when that is not the case. Overall, it was found that $\rho = \frac{1}{N}$ and $\rho = \frac{0.1}{N}$ provided particularly accurate estimates with the latter providing better estimates in cases when the divergence between the two distributions was higher.

Overall, the estimator was found to perform particularly well when evaluating smaller divergences both in terms of accuracy and computational time. For instance, estimation of the expected utility for experimental condition $\boldsymbol{\delta} = 0$ was achieved in less than 10

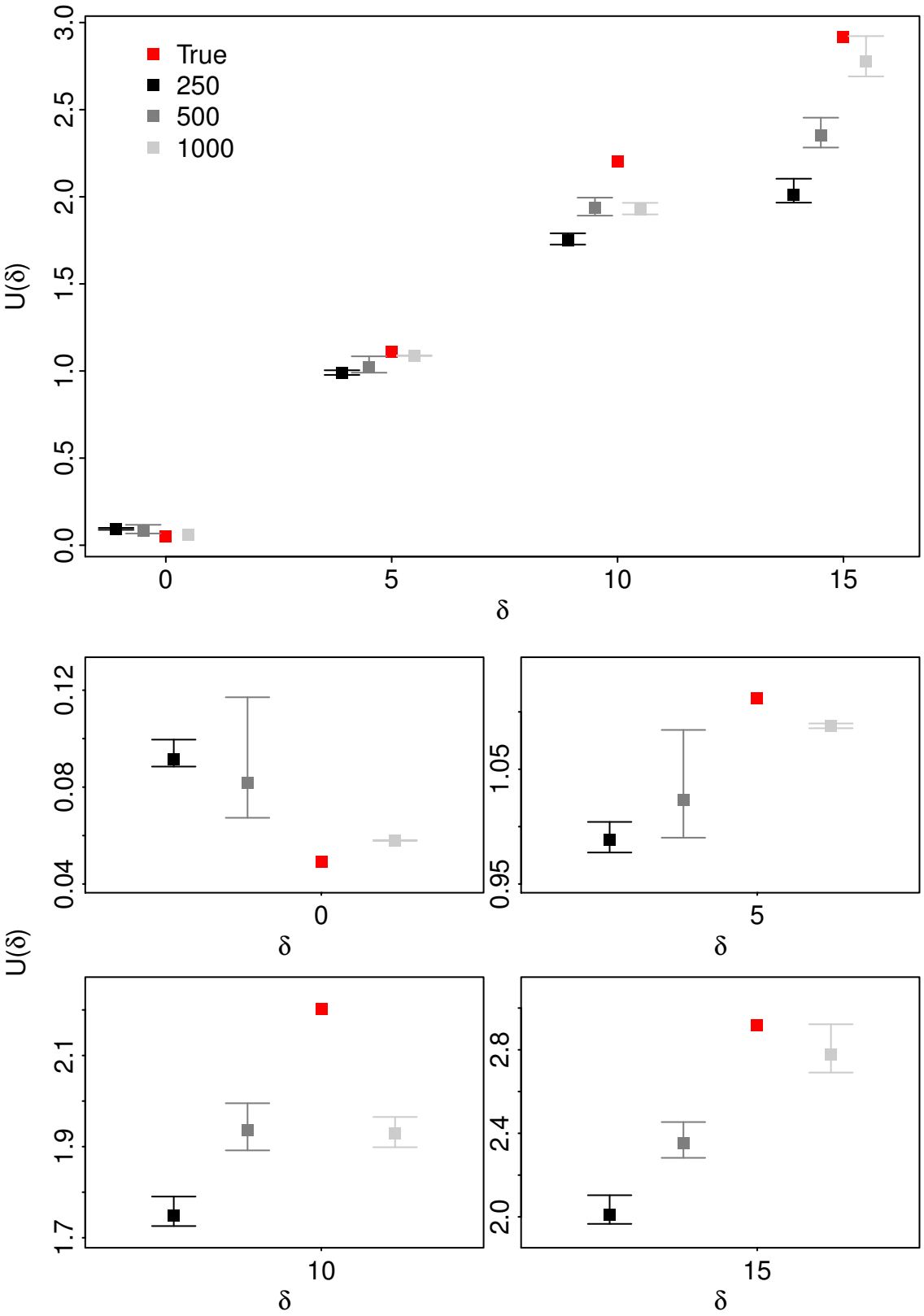


Figure 7.2: 95% credible intervals of expected utility estimates obtained through the proposed estimator using alternative sample sizes from the competing predictive distributions. Consecutive intervals in the top figure represent estimates produced at the exact same experimental condition, however, a jitter term has been added to values of $\boldsymbol{\delta}$ for clarity. The four bottom plots provide a closer look, showing sets of intervals presented in the top figure but for each experimental condition individually.

minutes for a chosen rate of $\rho = \frac{1}{N}$ and $N = 500$ samples from the competing distributions. Larger divergences, for example those observed at $\boldsymbol{\delta} = 5,\ 10$ required a lower rate of $\rho = \frac{0.1}{N}$, leading to slower convergence of the estimator and so evaluation of the corresponding expected utility ranged from 2 to 50 minutes depending on the difficulty of the induced optimisation problem.

Additional performance measures and their comparison against those obtained through the Monte-Carlo estimator and the almost closed form solution are provided in 7.2.5.

7.2.4 Monte-Carlo based methods

An alternative evaluation of the expected utility through the traditionally employed Monte-Carlo approach is examined in this section. The adopted estimator relies on the Monte-Carlo integral described in the introductory part of this section, however, an approximation of the marginal likelihood rather than the closed form expression of (7.4) is used instead. Performance of this alternative estimator is assessed since an analytical expression for the marginal likelihood is not available in most real-life applications as those examined in Chapters 8 and 9 and so alternative evaluation methods are required in such cases.

Following the work of Vyshemirsky and Girolami (2008), estimation of the marginal likelihood is achieved using the method of thermodynamic integration, discussed in 3.2.2. Adopting an importance sampling estimator, similar to Drovandi et al. (2013), estimates of the marginal likelihood result as a by-product of a Sequential Monte-Carlo sampling algorithm where samples obtained at each of the M intermediate stages are used for evaluation of the intermediate marginal likelihood, previously described in (3.6). A population of 1000 samples was used for the SMC algorithms as lower sample sizes were associated with population degeneracy issues while sampling was performed over 100 intermediate stages. The tempering schedule regulating the transition between subsequent intermediate distributions was chosen as:

$$b_i = \left(\frac{i}{M}\right)^4 , \tag{7.7}$$

following guidelines provided in Friel and Pettitt (2008) who suggest a relatively dense

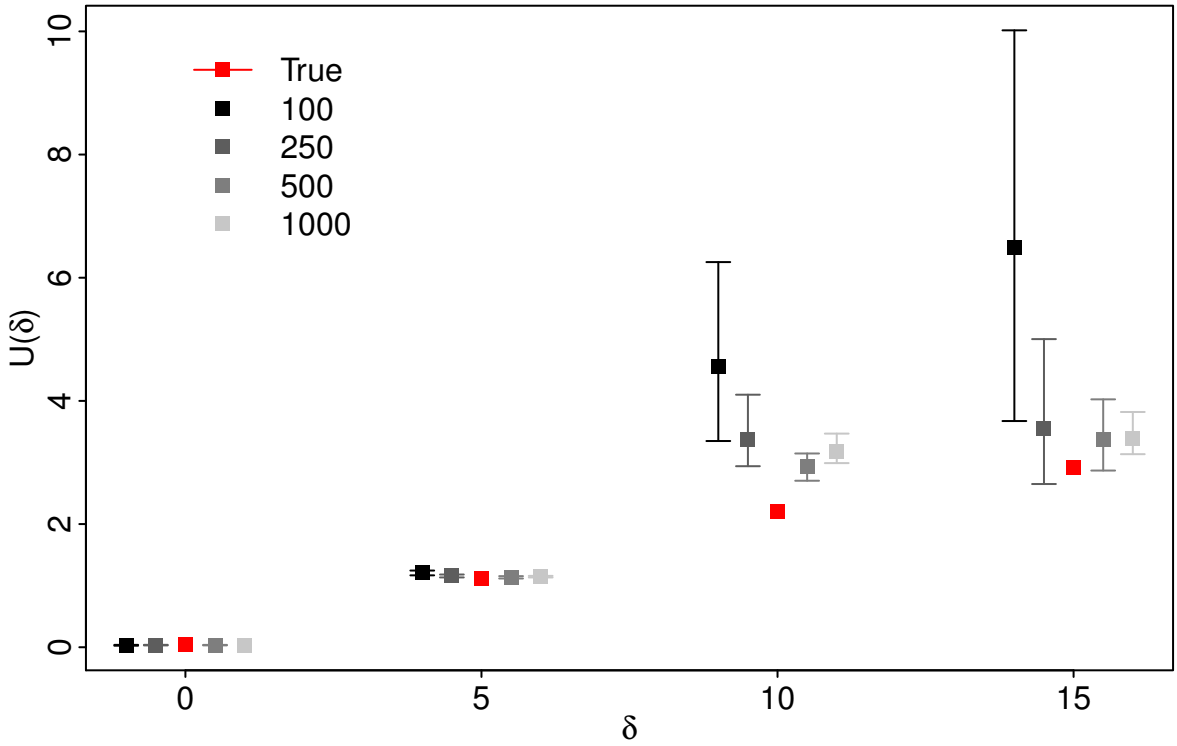


Figure 7.3: Credible intervals (95%) of expected utility estimates obtained through Monte-Carlo estimation. Estimates resulted from use of alternative sample sizes of populations from the competing predictive distributions are shown. Consecutive intervals represent estimates produced at the exact same experimental condition, however, a jitter term has been added to values of $\boldsymbol{\delta}$ corresponding to alternative sizes for clarity.

allocation of distributions during the initial sampling stages when the information gain is higher and increasingly scarce distributions thereafter.

A summary of estimates in the form of 95% credible intervals corresponding to alternative sizes of samples from the competing predictive distributions is provided in Figure 7.3, including a representation of the true value obtained through 7.5.

Estimation of the marginal likelihood corresponding to one prediction and one model required 35 seconds on average. As evaluation of the expected utility corresponding to one experimental condition $\boldsymbol{\delta}$ relies on examination of 12 pairs of models on a collection of N predictions, resulting in a total required time of 5 hours when considering 100 predictive samples. Unfortunately, as demonstrated in Figure 7.3 such low sample sizes generate highly biased and inaccurate estimates and thus larger samples need to be considered with corresponding evaluation times up to 50 hours (for $N = 1000$). Nevertheless, Monte-Carlo methods are easily parallelisable which may incur significant gains depending on the

computational capabilities available. Using a high performance distributed computing framework, estimation of the expected utility of one experimental condition based a population of 500 predictive samples was accomplished in less than 1 hour, however, it is worth noting that, such resources are not commonly accessible.

7.2.5 Comparison study

Further comparison between the alternative estimation methods, considered in Sections 7.2.3 and 7.2.4, is presented in this section. Additional performance criteria, complementing the findings of the preceding section are included in Table 7.1 and Figure 7.4. The former summarises estimates of the partial utility corresponding to each pair of competing models (m, m'), obtained by reproducing an experiment at condition $\xi = 0$ ten times based on a population of 500 prior predictive samples. Estimates are compared against the assumed true value, obtained through expression 7.1. It is worth noting that, negative estimates can be attributed to estimation error since the KL divergence admits only non-negative values.

(m,m')	**(1,2)**	**(1,3)**	**(1,4)**	**(2,1)**
True	0.0836	0.0836	0	0.1132
f**−d**	0.0813 (0.0661,0.1082)	0.0853 (0.0721,0.1213)	0.0404 (0.0202,0.0781)	0.0874 (0.0676,0.1286)
MC	0.1034 (0.1025,0.1055)	0.0751 (0.0742,0.0762)	2e-04 (-4e-04 ,9e-04)	0.0352 (0.0324,0.0362)
(m,m')	**(2,3)**	**(2,4)**	**(3,1)**	**(3,2)**
True	0	0.11324	0.103011	0
f**−d**	0.018 (0.0076,0.0324)	0.086 (0.0703,0.1469)	0.0944 (0.0736,0.1372)	0.021 (0.0085,0.0509)
MC	-0.0278 (-0.0303,-0.0258)	0.0323 (0.0288,0.0342)	0.0799 (0.0786,0.0814)	0.0288 (0.026,0.0305)
(m,m')	**(3,4)**	**(4,1)**	**(4,2)**	**(4,3)**
True	0.103011	0	0.078875	0.078875
f**−d**	0.1024 (0.0866,0.1628)	0.0316 (0.0222,0.0444)	0.0959 (0.0649,0.1589)	0.1023 (0.0847,0.1329)
MC	0.0774 (0.0752,0.0793)	1e-04 (-4e-04,6e-04)	0.1143 (0.1135,0.1148)	0.0865 (0.085,0.0886)

Table 7.1: Mean, lower and upper bounds of estimates representing the partial utility of each pair (m, m') obtained through 10 replications at experimental condition $\boldsymbol{\delta} = 0$ using a population of 500 samples from the corresponding prior predictive distributions. Three alternative evaluation methods are considered: Monte-Carlo relying on closed form expression of the marginal likelihood (True), variational approximation (f−d) and Monte-Carlo relying on estimation of the marginal likelihood through thermodynamic integration (MC).)

Figure 7.3 depicts 95% credible intervals of expected utility estimates obtained through each of the considered methods based on 1000 samples from each of the competing predictive distributions. As demonstrated therein, both estimators perform better when the divergence between distributions is small, potentially due to the low variability in the sample, allowing estimation by observation of less samples. It is further shown that, the error in estimates obtained through Monte-Carlo estimation is relatively higher than those corresponding to the proposed method which can be attributed to an additional error arising from estimation of the marginal likelihood. Therefore, the former approach, as also shown in succeeding case studies, requires larger sample sizes to provide estimates of quality similar to the latter. However, this demand is often hindered by the high computational cost associated with Monte-Carlo methods, as was shown in the preceding Section 7.2.4. This problem is aggravated further when models employing computationally demanding likelihoods as will be shown in the subsequent Chapters 8 and 9.

7.3 Sequential and adaptive design

Application of a sequential, response-adaptive setup to the studied model discrimination problem is considered in this section. The adopted framework, described in Algorithm 4, establishes a fully autonomous process of data collection, knowledge update and decision making, leading to more efficient and better-informed designs.

The sequential process, summarised in Figure 7.5, is composed of 4 stages during which, each experimental condition was allowed to only be considered once. In other words, once a condition $\boldsymbol{\delta} \in \Delta$ is identified as the optimal, it is assumed to offer no information gain in subsequent stages and thus assigned an expected utility of 0[3]. With respect to the stopping conditions described in steps 2 and 4 of Algorithm 4, an experimental budget equal to the number of studied conditions was assumed while designs incurring an expected utility of 0.5 (following the guidelines of Jeffreys (1961)) were deemed beneficial enough to encourage further experimentation.

[3]For the particular definition of utility function, considered in this case study, the value 0 is the minimum of the induced expected utility.

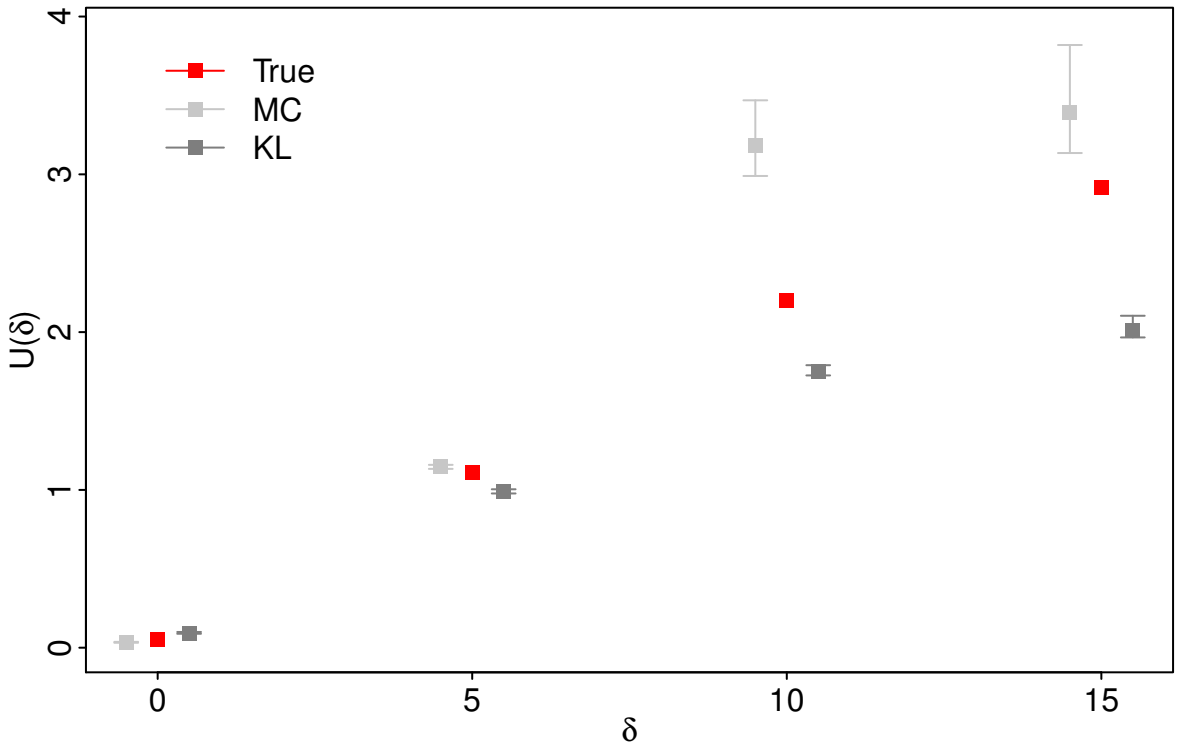


Figure 7.4: Credible intervals (95%) of the expected utility estimates corresponding to the studied experimental conditions of Δ obtained through the alternative competing methods. Results are compared against the assumed true value marked with red lines and symbols.

The process is initiated at stage 0, during which model predictions rely solely on preliminary information captured through the prior distribution of model parameters. Evaluation of the expected utility incurred from each considered experimental condition allows identification of the optimal condition $\boldsymbol{\delta}^*$ for observation of the studied system at stage 1. Having recorded the experimental dataset $\mathbf{D}$ at the stage 1, the prior distribution is refined to reflect the new knowledge and the updated predictive distributions are used for determination of the optimal condition for experimentation at the subsequent stage. In this case study, experimental data were simulated from Model 3. This iterative procedure continues until either the available experimental resources do not permit further experimentation or until the information gain incurred from additional experiments is deemed insufficient. In this example, an experimental budget of 4 experiments was assumed causing the process to terminate once this target was reached.

As Figure 7.5 suggests, the information gain from observing the optimal experiment

is maximised at stage 1 during which the very first experiment of the study is performed. Subsequent experiments incur decreasing benefit although still informative enough to justify data collection at these points with the exclusion of the final stage where the estimated expected utility of the optimal experimental condition was found to be 0.023 and thus, not worthy of consideration.

Unlike non-adaptive designs, each current stage of Algorithm 4 uses the already observed experimental data to inform subsequent decisions. As a consequence, the expected utility surface changes at each sequential stage in light of the newly observed experimental data. Maximisation of the expected utility surface is, thus, performed anew at each cycle. An illustration of the expected utility surface, refined to incorporate the information collected at the data collection stage, is provided in Figure 7.6.

In this case study, a design space, consisting of only 4 experimental conditions was considered. Maximisation of the expected utility is under such cases feasible simply through evaluation of the expected utility corresponding to each potential experimental condition followed by deterministic comparison of the obtained values in order to determine the maximum. As will be demonstrated in the succeeding Chapter 8, this approach does not necessarily incur optimal maximisation over a continuous design space when review of a considerable number of potential experimental conditions is required. The class of Bayesian optimisation algorithms, introduced in Chapter 6, can be employed in such cases for efficient optimisation of the expected utility. Implementation of this approach is demonstrated in the succeeding Chapter 8.

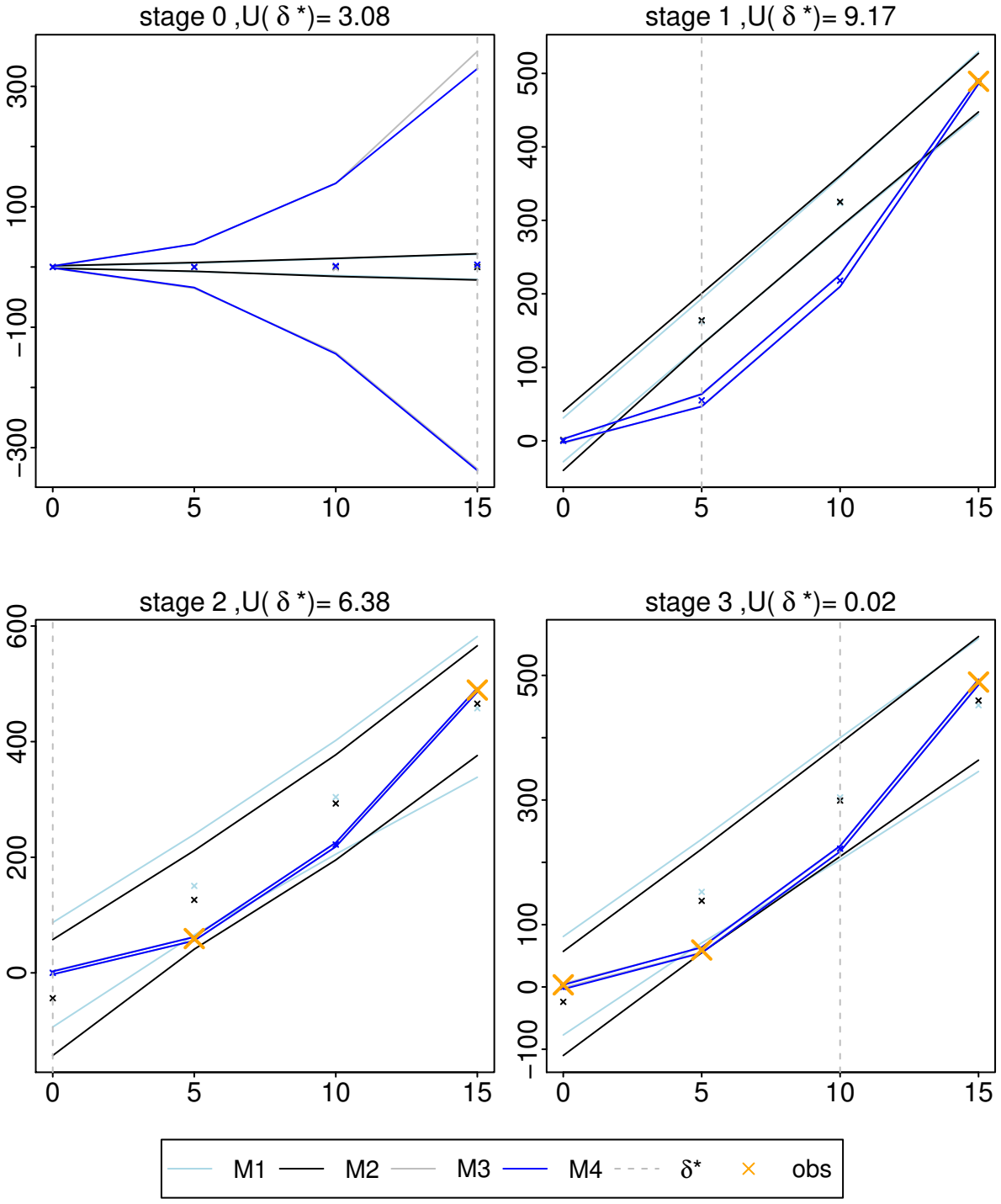


Figure 7.5: A sequential framework for optimal experimental design. Each subfigure depicts predictions made with each of the four competing models (y-axis) at each potential experimental condition δ (x-axis). The transition from one subfigure to the next represents the change in model predictions as new information becomes available at each stage of the procedure. The coloured lines represent the lower and upper bounds of the prediction intervals for each model while crosses denote their means calculated at each experimental condition of Δ.

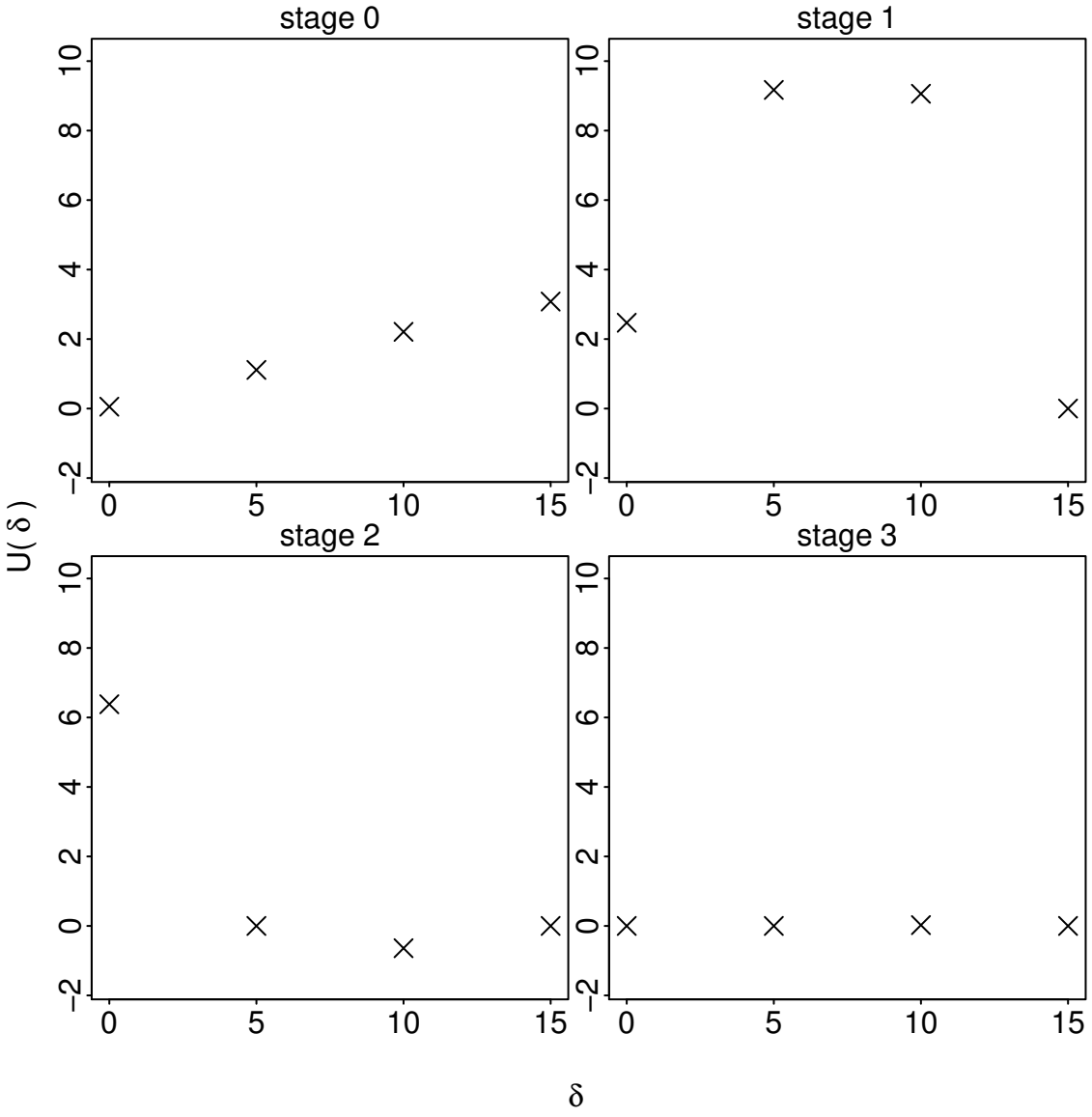


Figure 7.6: Sequential update of the expected utility surface in light of newly observed experimental data. Each subfigure corresponds to a distinct stage of the sequential procedure, depicting the expected utility (y-axis) of each of the potential experimental conditions δ (x-axis). Once the system is observed at a given experimental condition δ its expected utility becomes noticeably small, indicating that no further information collected at this δ will be substantial.

Summary

This chapter presents some initial findings on optimal experimental design through the proposed methodology and its comparison to the commonly adopted Monte-Carlo ap-

proach for evaluation of the expected utility in the context of a simple model discrimination problem. The considered study benefits from availability of a closed form solution for evaluation of the marginal likelihood thus producing estimates that are assumed to be arbitrarily close to the true value through Monte-Carlo integration based on a sufficiently large population of samples from the competing predictive distributions. Comparison of the produced estimates against this representation of the true value, confirms that both methods perform particularly well as was demonstrated through various summaries. The main difference of the two estimation methods lie in their corresponding computational time with the proposed estimator requiring significantly less time for the conclusion of a particular study of interest. As previously discussed in Chapter 5, this can be attributed to the reliance of Monte-Carlo estimation to sampling from the posterior distribution of model parameters for each of a large collection of samples from the predictive distributions of the competing models. On the contrary, the variational estimator avoids sampling and only relies on optimisation of a convex problem over an $N-$simplex where N expresses the number of samples from the predictive distributions. The difference between the competing methods becomes more apparent when considering models with computationally demanding likelihoods, a problem that is examined more closely in Chapters 8 and 9.

Having explored the problem of estimation of the expected utility, a response-adaptive design for model discrimination was subsequently explored. The proposed sequential procedure allows full process automation in such studies, establishing a closed-loop setup of data collection, knowledge update and decision making. This framework is studied further in the context of the real-life applications, presented in Chapters 8 and 9.

Chapter 8

Optimal experimental design for the study of biochemical networks

This chapter explores the application of optimal experimental design for inference in Systems Biology. As direct observation of the composition and dynamics of biological systems is typically not possible, analysis through wet lab experiments is often necessary for their study. Unfortunately, the considerable cost and time demand of these experiments places limiting conditions on their conduct. Optimal experimental design addresses this class of problems, allowing the study of such phenomena in the most beneficial manner while adhering to imposed limitations. In this chapter, design optimisation is achieved using a range of methodologies, introduced in Chapters 2-6. Ongoing challenges, discussed therein, are addressed using the variational approximation approach of Chapter 5 for efficient evaluation of the expected utility and a comparison with currently adopted methods is provided in the context of a real-life application. An autonomous and efficient framework for the study of this problem is subsequently considered on the basis of the proposed method.

8.1 Experimental design in Systems Biology

Systems Biology is a scientific discipline concerned with the study of biological systems, encompassing computational and experimental procedures. It provides a natural transi-

tion from component-level modelling to understanding the functionalities and dynamic behaviour of cells as parts of a system (Kriete and Eils, 2013). This new representation is thought to be key for improving current diagnostics and treatments thus transforming existing medical practices.

Integration of expert knowledge from multiple disciplines is essential for this endeavour. Biologists contribute a deeper insight into the unique properties of cells and are in a position to conduct lab experiments that provide useful information on the studied biological system. Modellers rely on such observations to form an initial set of hypotheses and subsequently test them against the obtained knowledge. The process is often cyclical in the sense that, prior hypotheses are updated in the light of new experimental data which subsequently guide decisions on future experiments. Completion of this exchange may be achieved when the limiting conditions are met or if the study can be concluded with sufficient confidence.

Undoubtedly, experimental data constitute a vital component of this procedure. In fact, a large collection of publications (Liepe et al., 2012; Vyshemirsky and Girolami, 2008; Wilkinson, 2007) is devoted to performing modelling, inference or prediction tasks in Systems Biology based on observation of a system through experiments. The methodology proposed in Chapters 5 and 6 finds direct application in challenges related to the experimental design process for data collection.

Planning of experiments often relies on decision making that has an immediate effect on the subsequent system behaviour. This may concern any controllable conditions such as the times of observation, the pattern of stimulation or choice of the directly observed quantities. Certain decisions may instigate behaviours that are more favourable to the study of a particular objective. Unfortunately, the cost and time requirements involved in performing an experiment for observation of a biochemical system, often, place limitations on the number of potential conditions that can be investigated. The process of identifying the most favourable conditions under the imposed restrictions, forms an optimal experimental design problem along the lines of Chapter 2. Following the notation introduced therein, external factors controlled by the experimenter will be treated as the experimental conditions $\boldsymbol{\delta}$ while the remaining system components will constitute the experimental parameters $\boldsymbol{\theta}$. The succeeding Section 8.2 introduces a case example

of an optimal design problem for discrimination between two artificial models, offering alternative views to the structure of a particular biochemical network of interest. Sections 8.3 and 8.4 showcase application of the proposed methodology in order to identify experimental conditions under which model selection is performed with the highest efficiency. Section 8.5 revisits the sequential process of data collection, knowledge update and decision making described in previous paragraphs. As shown therein, integration of the proposed, efficient estimator allows the automation of such an exchange and its completion within realistic time frames which is, otherwise, not possible using currently adopted approaches.

8.2 Case study

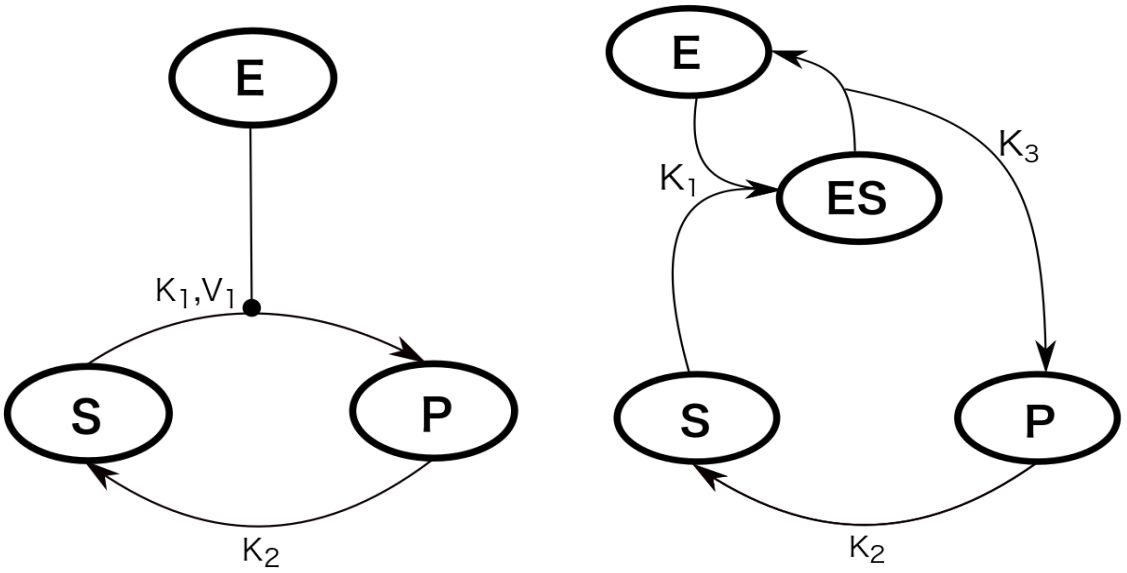
This section introduces a case study in Systems Biology that seeks to infer the structure of a biochemical system through model selection using experimental data.

8.2.1 Alternative hypotheses

Models expressing two alternative dynamic structures of an enzymatic activation system are considered in this study. The competing hypotheses are depicted in Figure 8.1 with ellipses representing proteins, present in the system and the directed arrows connecting them denoting the reactions between them. The rates at which reactions occur are expressed by the kinetic parameters written along the arrows while arrows ending with a black dot indicate enzymatic behaviour. The examined systems as well as similar structures and their corresponding representations can be found in Vyshemirsky and Girolami (2008) and Lawrence et al. (2010).

Hypothesis 1 assumes a simpler behaviour where substrate S is converted into product P by the action of enzyme E. Hypothesis 2 adopts a similar structure incorporating the additional component ES as a by-product of the reaction between proteins E and S.

As the dynamic behaviour of biological systems is monitored over time, mathematical expression for it is typically provided by systems of differential equations (o.d.e.) with time t acting as an independent variable, the reaction rates $\{V_1, K_1, K_2, K_3\}$ as model parameters and the concentration of proteins $\{S, P, E, ES\}$ as dependent variables.



(a) Hypothesis 1 **(b)** Hypothesis 2

Figure 8.1: Graphical representation of two competing hypotheses representing alternative structures of a biochemical system under study.

The o.d.e. system expressing the structure depicted in Figure 8.1a takes the form:

$$\textbf{Model 1:} \qquad \begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\frac{V_1 \cdot S \cdot E}{S + K_1} + K_2 \cdot P \\ \frac{\mathrm{d}P}{\mathrm{d}t} &= \frac{V_1 \cdot S \cdot E}{S + K_1} - K_2 \cdot P \\ \frac{\mathrm{d}E}{\mathrm{d}t} &= 0 \ , \end{aligned} \tag{8.1}$$

where terms S, P and E represent the concentration of the corresponding proteins at a given time point and V_1, K_1, K_2 denote the kinetic parameters regulating the interactions between them. In this case study, the initial values of protein concentration were considered known, however in a different scenario they can be flexibly optimised instead. The initial values of protein concentration were set to:

$$S\,|_{t=0} = 1 \qquad P\,|_{t=0} = 0 \qquad E\,|_{t=0} = 0.01 \ .$$

The alternative hypothesis, depicted in Figure 8.1b, finds mathematical expression in

the following system of o.d.e.s:

$$\textbf{Model 2:} \qquad \begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -K_1 \cdot S \cdot E + K_2 \cdot P \\ \frac{\mathrm{d}ES}{\mathrm{d}t} &= K_1 \cdot S \cdot E - K_3 \cdot ES \\ \frac{\mathrm{d}P}{\mathrm{d}t} &= -K_2 \cdot P + K_3 \cdot ES \\ \frac{\mathrm{d}E}{\mathrm{d}t} &= -K_1 \cdot S \cdot E + K_3 \cdot ES \end{aligned} \tag{8.2}$$

where S, ES, P, E are concentrations of proteins composing the system and K_1, K_2, K_3 the kinetic parameters. The initial concentration, in this case, was set to:

$$S\,|_{t=0} = 1 \qquad ES\,|_{t=0} = 0 \qquad P\,|_{t=0} = 0 \qquad E\,|_{t=0} = 0.01\ .$$

An illustration of the system behaviour under the two alternative hypotheses is provided in Figure 8.2, where the concentration of its components is depicted over time with $t = 0$ s marking the time of system activation. The predictions were generated using fixed choices of initial concentrations and model parameters.

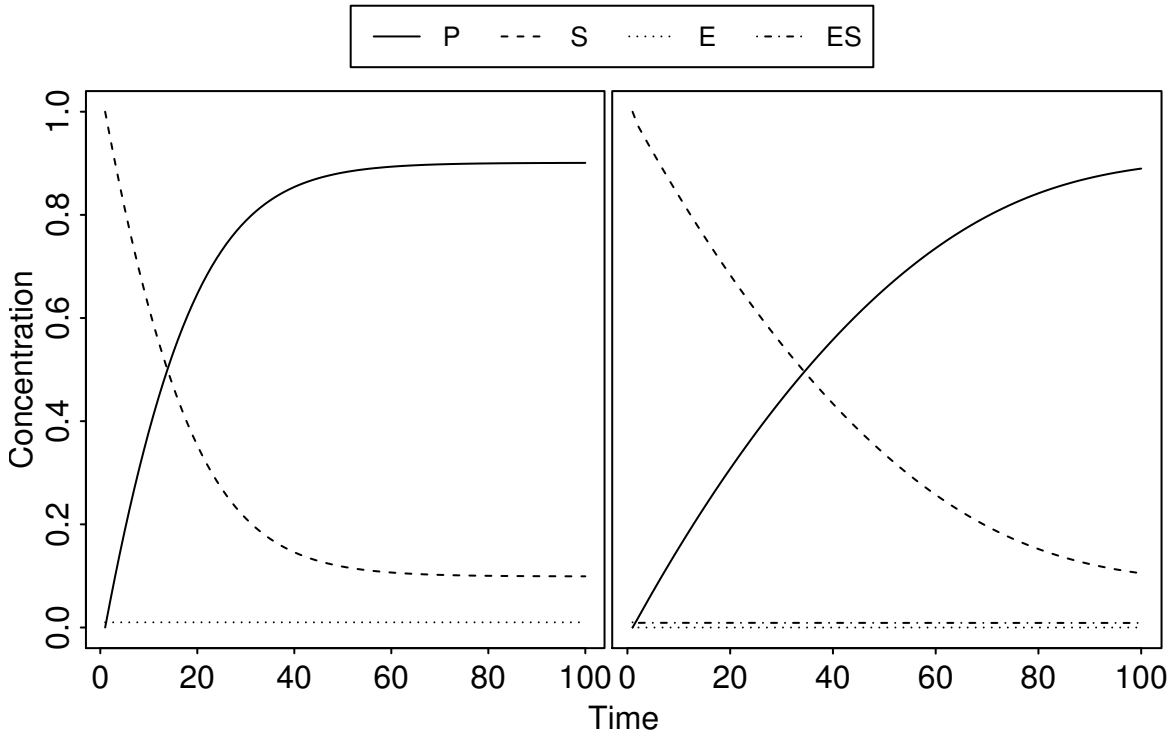


Figure 8.2: Concentration predictions for the proteins composing the studied system under Hypothesis 1 (left) and Hypothesis 2 (right) over a time period of 100 seconds.

Overall, the initial concentrations were assumed to be known and fixed, however, the kinetic parameters were considered unknown and were as such assigned a prior distribution in a Bayesian context. The particular form of the prior distributions, adopted for this case study, is presented in Section 8.2.2 along with a specification for the likelihood function of a given model prediction under an observed dataset. As a result, time remains as the only controllable factor impacting the observed output and was thus treated as the experimental condition $\boldsymbol{\delta}$. The general notation $\boldsymbol{\delta}$ and the problem-specific t will be, throughout this chapter, used interchangeably. The succeeding section discusses the statistical models employed in this case study in more detail.

8.2.2 Statistical models

Following the notation adopted in previous chapters, the observed dataset will be denoted by $\boldsymbol{y} = \{\boldsymbol{y}_{\boldsymbol{\delta}_1}, \ldots, \boldsymbol{y}_{\boldsymbol{\delta}_K}\}$, where $\boldsymbol{y}_{\boldsymbol{\delta}_i}$ represents the protein concentration corresponding to experimental condition $\boldsymbol{\delta}_i$, $i = 1, \ldots, K$. The vector of model parameters $\boldsymbol{\theta}_m$, in this context, refers to the kinetic parameters of the system and the observation error variance, which will be denoted by $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}$ and σ^2 respectively. As a result, $\boldsymbol{\theta}_m = \{\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}, \sigma^2\}$, where $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m_1}} = \{V_1, K_1, K_2\}$ and $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m_2}} = \{K_1, K_2, K_3\}$.

To test the competing hypotheses, the corresponding model outputs are compared against experimental data. Treating one of the models as the 'true' model, simulated data served, in this case study, as artificial experimental observations, an approach that has been widely adopted (Mendes et al., 2003; Vyshemirsky and Girolami, 2008) for testing and comparing methodologies in a controlled environment. In this study, Model 1 was considered as the 'true' model and focus was placed on measurements from protein P at a selection of experimental conditions as this component is present in both competing models and is sensitive to changes in the model parameters which prevents possible identifiability issues.

The likelihood function is subsequently employed for this comparison, quantifying the probability of the assumed model reproducing the observed data. Under the assumption of normally distributed observation errors, the likelihood of dataset $\boldsymbol{y}$ given model m

parametrised by the corresponding vector $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}$ is formulated as:

$$f(\boldsymbol{y} \mid m, \tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}, \sigma^2, \boldsymbol{\delta}) = \mathcal{N}(\boldsymbol{y}; \text{ode}_m(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}, \boldsymbol{\delta}), \sigma^2) \ , \tag{8.3}$$

where $\text{ode}_m(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}, \boldsymbol{\delta})$ expresses the solution of the o.d.e. system, quantifying the concentration of P at a particular experimental condition $\boldsymbol{\delta}$ according to Model m with parameters $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}$. A uniform prior distribution (U) was assigned to the model parameters as a rough range was known for each of them and there was no reason to assume higher probability for particular values while an inverse Gamma distribution (IG) was assumed for σ^2 to ensure it is always assigned positive values. In summary:

$$\begin{aligned} V_1 &\sim U(0, 20) \\ K_1 &\sim U(0, 20) \\ K_2 &\sim U(0, 10) \\ K_3 &\sim U(0, 10) \\ \sigma^2 &\sim \text{IG}(4, 2) \ . \end{aligned} \tag{8.4}$$

These choices of hyperpriors were found to provide a wide enough coverage of potential model parameter values. Given that some existing knowledge was incorporated through the prior distributions they can be considered as weakly informative.

As described in Section 8.1, learning by experimentation is an iterative procedure. In the initial stage, the predictive distributions are typically chosen to be sufficiently wide covering a broad range of potential observations. Due to the significant overlap, further illustrated in Figure 8.3, identification of an optimal experimental condition, incurring considerable evidence in the context of model discrimination was not possible. The utility surface corresponding to the predictive distributions depicted therein was found to be fairly flat with the returned optimum not exceeding the value of 0.1 which, following (Jeffreys, 1961), is insignificant. Common practices (Silk et al., 2014), in this case, resort to random selection or rely on expert knowledge for choice of the initial experimental condition. The information $\mathbf{D}$ acquired from the preliminary experiment can be flexibly incorporated into the study by updating expression (8.3) into the posterior predictive

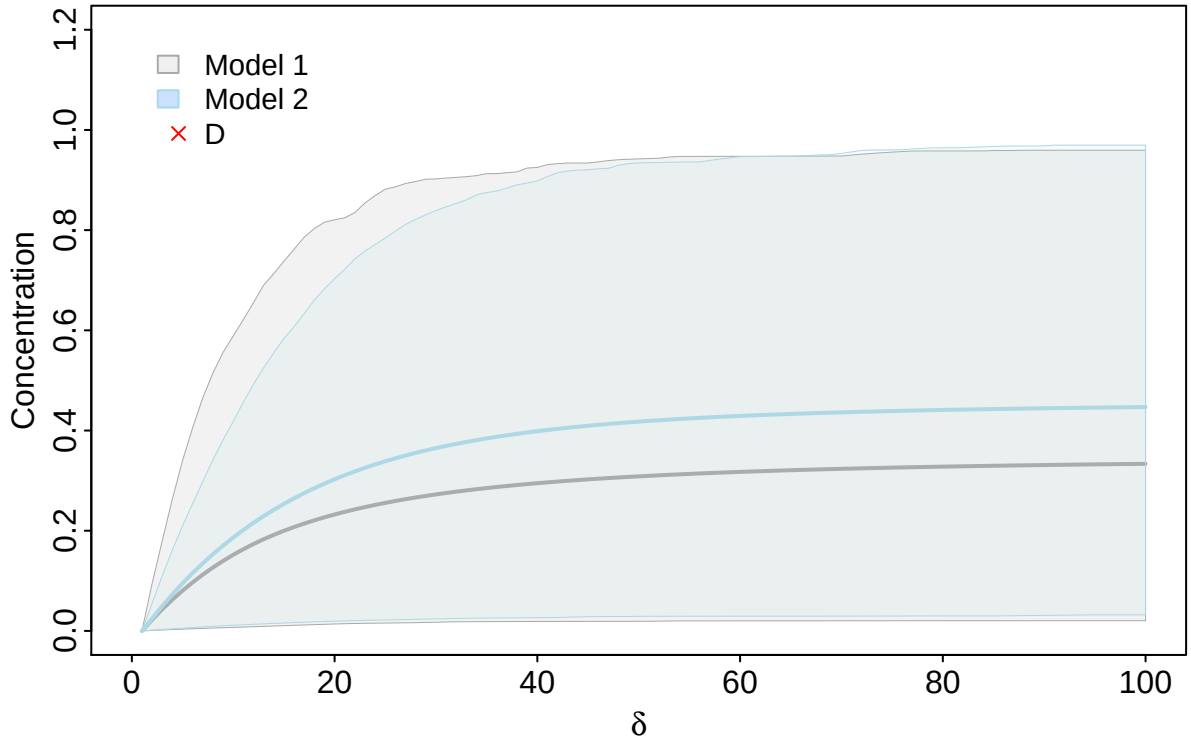


Figure 8.3: Predictions corresponding to different values of model parameters sampled from the prior distribution for the two candidate models.

distribution formulated as:

$$p(\boldsymbol{y} \mid m, \boldsymbol{\delta}, \mathbf{D}) = \int p(\boldsymbol{y} \mid m, \boldsymbol{\theta}_m, \sigma^2, \boldsymbol{\delta}, \mathbf{D}) p(\boldsymbol{\theta}_m \mid \mathbf{D}) p(\sigma^2 \mid \mathbf{D}) \, \mathrm{d}\sigma^2 \mathrm{d}\boldsymbol{\theta}_m \, . \tag{8.5}$$

This knowledge is further refined in light of new experimental results at subsequent stages. At each stage the posterior distribution conditioned on the most current data will serve as the prior distribution for the subsequent stage. This response-adaptive setup is revisited in Section 8.5.

Opting for an initial condition where the shift between predictive means appears to be comparatively high, previous knowledge was simulated from Model 1 at $\boldsymbol{\delta} = \{100\}$s, obtaining 3 replicates with added noise in order to reproduce the error present in real-life scenarios. An expression for the posterior distribution of model parameters was not available in closed form and so samples from it were obtained using Sequential Monte-Carlo methods, as described in Appendix B. The posterior density plots are provided in Figures 8.4 and 8.5 for Models 1 and 2 respectively. These samples were subsequently

used to generate observations from the posterior predictive distributions of the two models which are depicted in Figure 8.6. A knowledge update is evident therein, when compared with the prior predictive distribution of Figure 8.3, as the predictive distributions are refined to reflect observation of the experimental data (marked by red crosses).

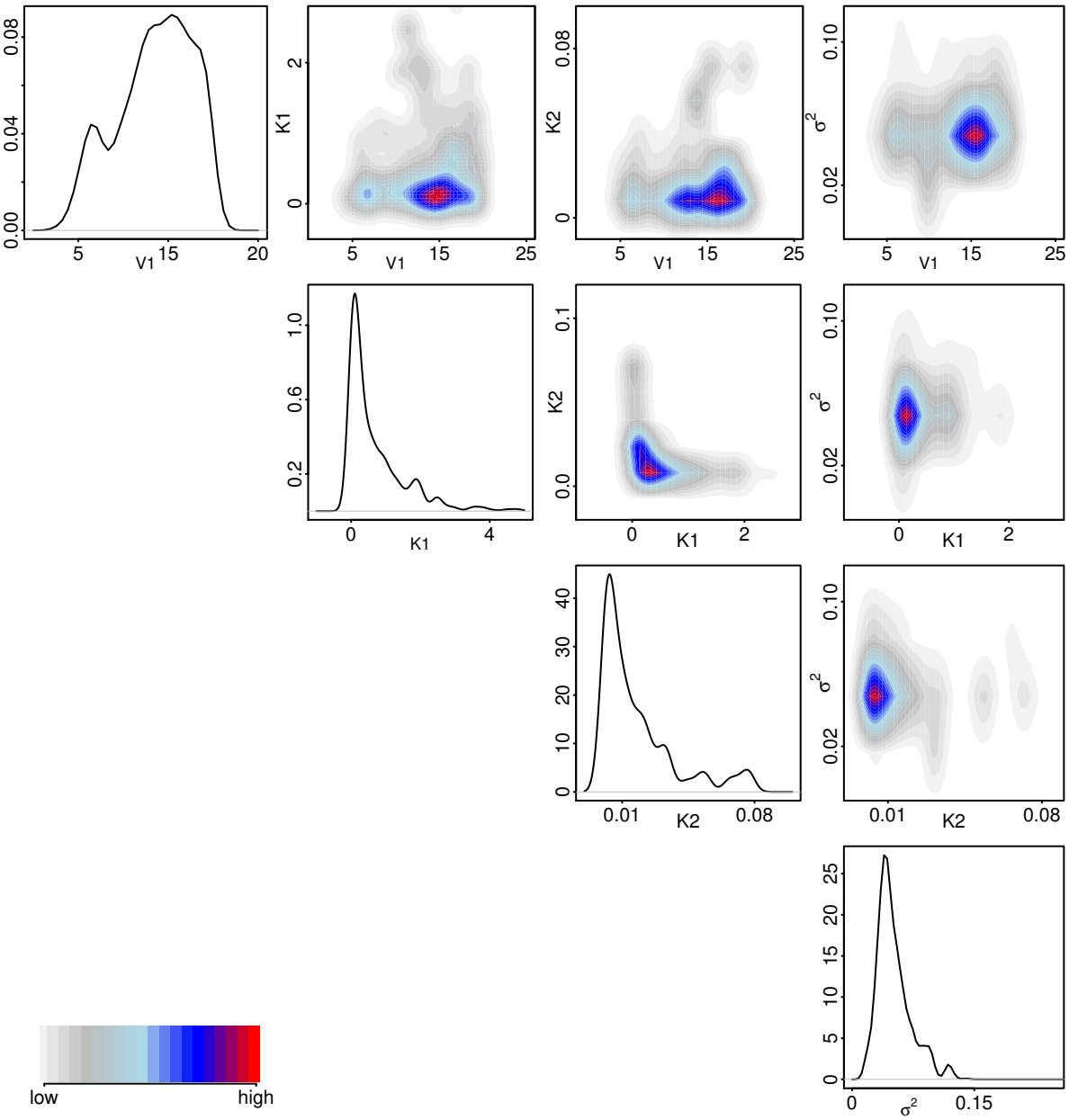


Figure 8.4: Posterior distributions of model parameters of Model 1 given prior information captured by **D**.

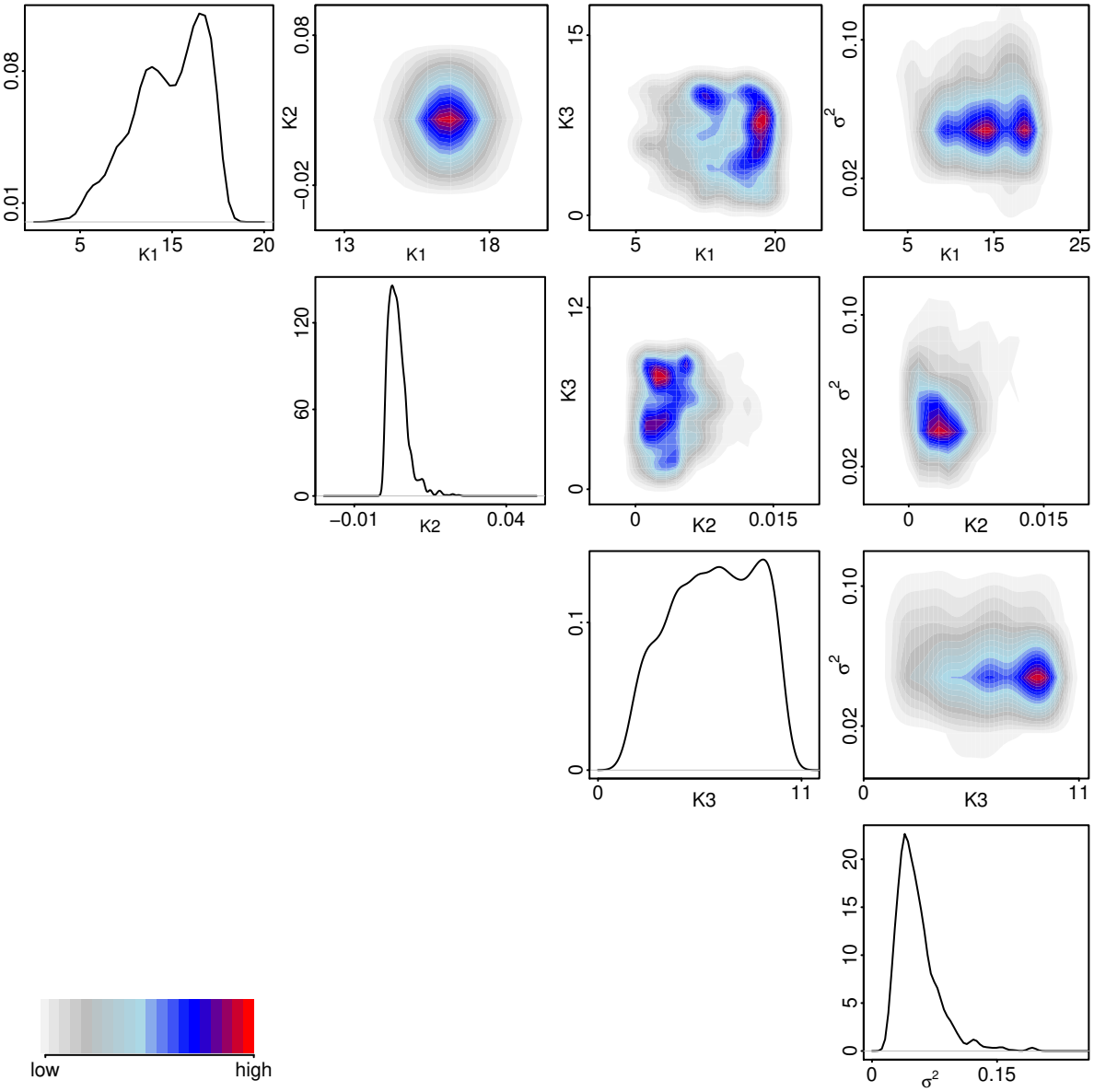


Figure 8.5: Posterior distributions of model parameters of Model 2 given prior information captured by **D**.

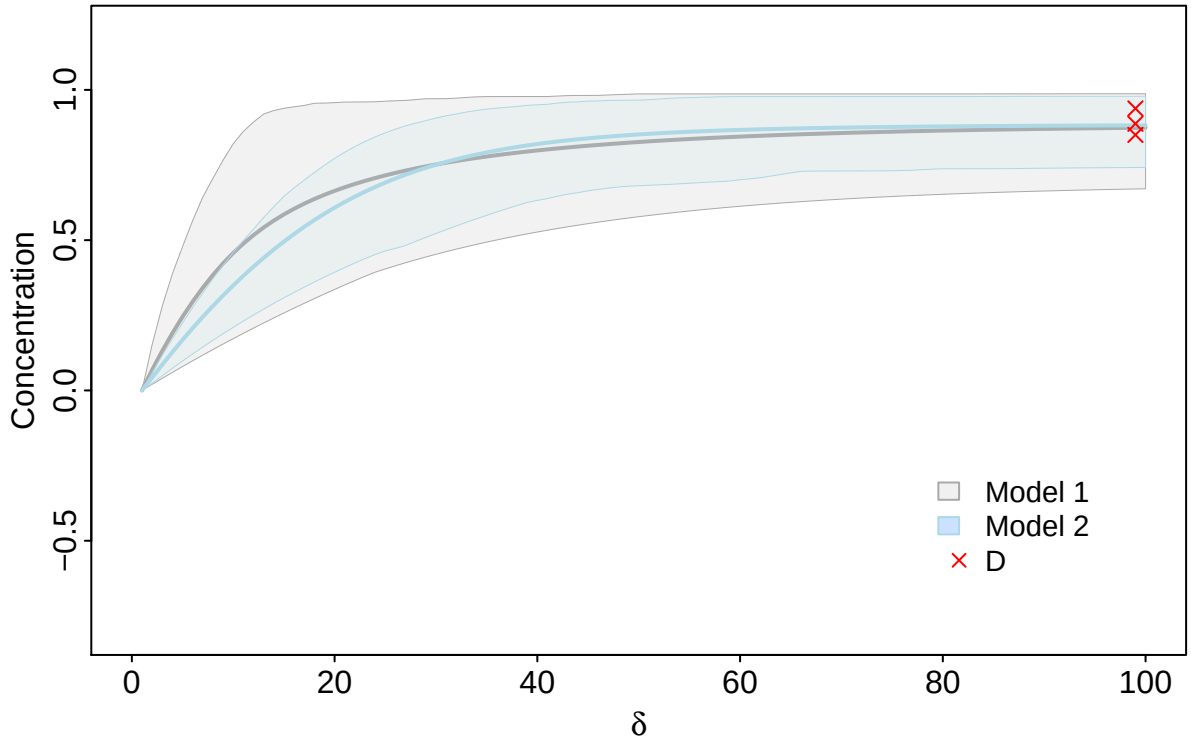


Figure 8.6: Predictions using samples from the posterior distribution of model parameters under the competing hypotheses. Data summarising previous knowledge are marked by red crosses.

Subsequent choice of experimental conditions is attainable through the methodology considered in Chapters 3,5 and 6. Focus is initially placed on efficient evaluation of the expected utility of candidate designs which is explored in Section 8.3. A comparison of the proposed methodology and traditionally adopted Monte-Carlo based approaches is provided in 8.3.1 and 8.3.2 respectively. Exploration of the design space is later considered in Section 8.4.

8.3 Evaluation of the expected utility

This section explores competing methodologies for evaluation of the expected utility in the context of the studied model inference problem. This problem is quite representative of cases when traditionally employed estimation methods become highly inefficient due to the considerable computational demand involved in obtaining predictions from the adopted models. More specifically, solution of the corresponding system of o.d.e.s

requires on average $1.12 \cdot 10^{-4}$ seconds which, as previously shown in Section 3.2.4, results in total computational time of approximately 748 minutes for evaluation of the expected utility of only one potential design using the traditionally adopted Monte-Carlo methods and a population of 100 samples from each of the competing predictive distributions. As a result, full search over the continuous design space $\Delta = [0, 100]$ s is hindered leading to consideration of only a limited selection of experimental conditions, thus inducing potentially sub-optimal designs. On a larger scale, consideration of the sequential setup, described in Section 8.1, would lead to considerable waiting times between the optimisation and experimentation stage, deeming this exchange problematic.

This section employs the variational method introduced in Chapter 5 to tackle these challenges and provides a comparison with empirical results obtained from the Monte-Carlo based approaches. Application of both methodologies relies on samples from the posterior predictive distributions corresponding to the competing hypotheses, depicted in Figure 8.6, given preliminary information obtained at $\boldsymbol{\delta} = \{100\}$ s. Observation of the posterior predictive distributions provides further insight on definition of the class of utility functions $\mathcal{F}$ established in Chapter 5. In summary, $\mathcal{F}$ was shown to represent a class of dissimilarity measures, commonly known as $f-$divergences that in the context of model selection, quantify the discrimination between the competing predictive distributions. A higher utility suggests larger discrimination and thus more confident hypotheses ranking given a new set of predictions. Considering the specific case depicted in Figure 8.6, classification of a newly observed experimental dataset at $\boldsymbol{\delta} = \{15\}$ s to one of the competing hypotheses may potentially be achieved with higher confidence compared to $\boldsymbol{\delta} = \{55\}$ s where there is an almost complete overlap between model predictions and, so, similar observed measurements are expected under any hypothesis. Two members of function class $\mathcal{F}$ are adopted in this study, inducing expressions of the expected utility in terms of the KL divergence and Hellinger distance.

Initial evaluation of the expected utility is restricted to the design points $\{1, 5, 10, 20, 60, 80\}$ s. Exploration of such a limited selection of experimental conditions is likely to provide a sub-optimal solution however interest at this stage focuses solely on the performance of the proposed estimator and its comparison to alternative estimation methods in terms of accuracy and computational time rather than selecting the optimal condition

which is further considered in Section 8.4.

8.3.1 Variational approximation

Estimates of the expected utility are, in this section, obtained through the variational approach, described in Algorithm 3. A summary of the obtained estimates is provided in Figure 8.7 in the form of 95% credibility intervals obtained by replicating each experiment 100 times.

Alternative choices of function φ are considered, more specifically, definitions $\varphi_{KL}(\chi) = -\log\chi$ and $\varphi_H(\chi) = 2\sqrt{(\chi)}$, $\chi > 0$ inducing expressions of the expected utility as the KL divergence and Hellinger distance respectively. Two main properties can be exploited from choice of the latter: 1) symmetricity in the divergence, meaning that $F(m, m') = F(m', m)$, thus reducing the number of partial utility approximations further and 2) restriction of $F(m, m')$ in the $[0, 1]$ interval. As previously observed in Chapter 7, the proposed estimator performs particularly well when estimating smaller divergences and hence, choice of φ_H ensures that only such values are being targeted.

As illustrated in Figure 8.7, both choices ot utility function produce identical shapes of the expected utility surface but with different corresponding value ranges. Interpretation of the expected utility varies among the two examined choices of utility functions. The expression of weighted KL divergence is always non-negative and can be understood along the lines of Jeffreys (1961)'s guidelines for the Bayes' factor, due to the previously established correspondence shown in Chapter 2. The alternative representation in terms of Hellinger distances, attains values only within the interval $[0, 1]$, deeming the obtained results also intuitively interpretable.

Summaries were generated for sample sizes of 250, 500 and 1000 from the predictive distributions under study to demonstrate the improvement in the estimation accuracy as larger samples are considered. In general, populations of 500 or 1000 should be preferred as less samples produce considerably noisy estimates and, as a result, significant overlap between intervals corresponding to the alternative experimental conditions which hinders their comparison. Conversely, a population of 1000 samples achieves sufficiently distinctive discrimination between the compared decisions and so consideration of more predictions does not necessarily justify the additional computational cost incurred by it.

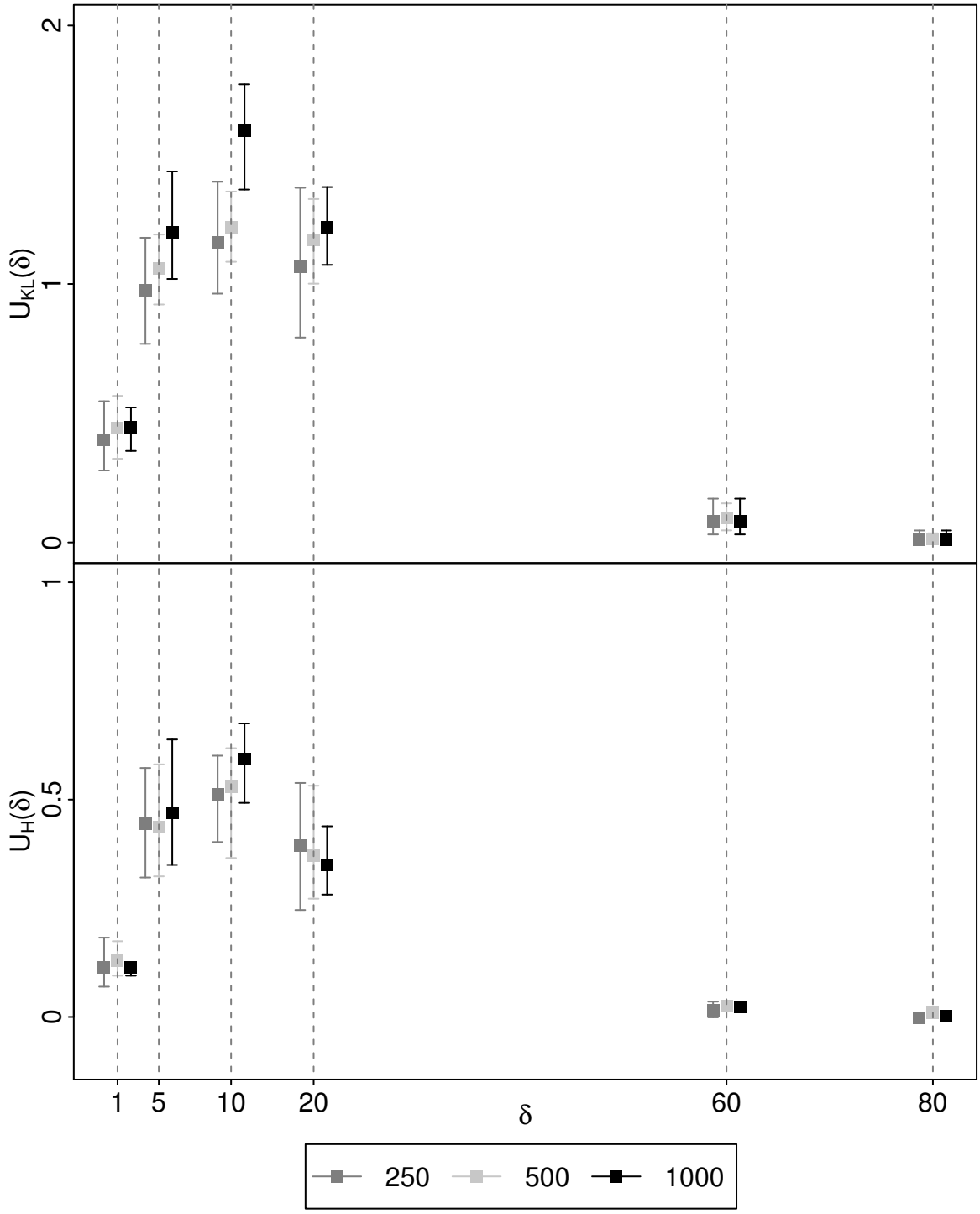


Figure 8.7: Credible intervals (95%) of the expected utility corresponding to φ_{KL} (top) and φ_{H} (bottom). Each set of intervals represents summaries corresponding to populations of samples from the competing predictive distributions of sizes 250, 500 and 1000. Although each set corresponds to exactly the same experimental condition (marked by dashed grey lines), a jitter term has been added to values of the x-axis for clarity.

The RKHS class structure induced by the Gaussian kernel was imposed on function class $\mathcal{G}$ defined by (7.6) with $\|\cdot\|$ expressing the Euclidean metric in $\mathbb{R}$. A penalty term inversely analogous of the considered population N was adopted, setting $\rho = \frac{1}{N}$ for populations of less than 1000 samples and a slower rate of $\rho = \frac{0.1}{N}$, otherwise. A more comprehensive discussion regarding these choices is provided in Chapter 7.

Similarly to findings from estimation of the expected utility in the model discrimination problem of Chapter 7, the estimator was found to perform better, both in terms of accuracy and computational time, at experimental conditions with relatively smaller expected utilities, such as time points 60 or 80 s. Estimation at these points ranged from 3 to 7 minutes while at the remaining conditions of $5, 10$ and 20 s, computational times of up to 30 minutes were observed for populations of 1000 samples from the predictive distributions. The difference in the observed estimation times can be attributed to the complexity of the corresponding optimisation problem. Intuitively, learning from samples with relatively low variability which is subsequently reflected in their low expected utility, as discussed in more detail in Chapter 2, poses an easier problem than comparing two noisy samples where, potentially, more sophisticated models (more basis functions in terms of RKHS) are required to adequately capture such structures.

Given the subset considered in this section, findings suggest that observation of the system close to time point $\boldsymbol{\delta} = \{10\}$ s incurs the highest benefit as can be observed in Figure 8.7. Overall, the obtained results confirm the initial impression with points associated with seemingly larger discrimination between the competing models incurring higher expected utilities and, conversely, points with significant overlap between predictive distributions being assigned a relatively low expected utility. A comparison with estimates obtained through the commonly adopted Monte-Carlo based methods as well as their overall performance is provided in the subsequent section.

8.3.2 Monte-Carlo based methods

Application of Monte-Carlo estimation methods, introduced in Chapter 3, is considered in this section for evaluation of the expected utility. The adopted approach relies on Monte-Carlo integration for estimation of the integral over Y involved in evaluation of the expected utility and subsequently employs thermodynamic integration for estimation

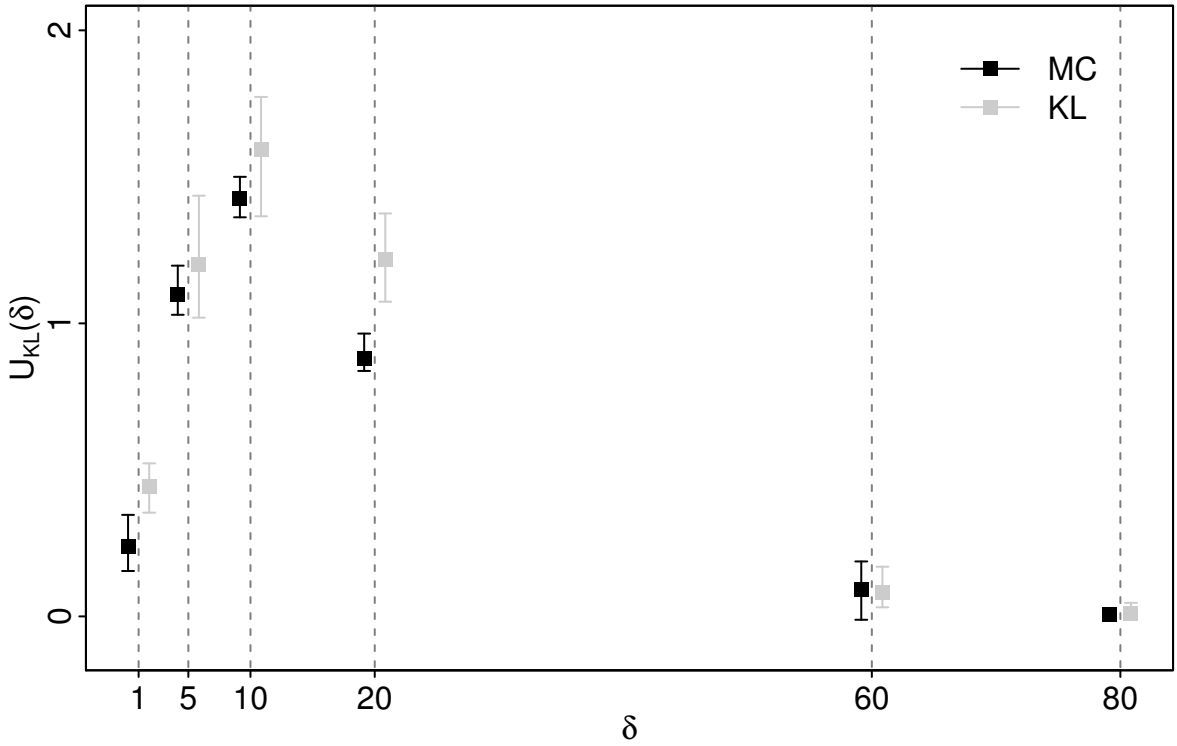


Figure 8.8: Credible intervals (95%) of estimates obtained through Monte-Carlo and $f-$divergence estimation given 500 predictions. Each set corresponds to the same experimental condition, however a jitter term is added to values of the x-axis for clarity.

of the marginal likelihood for each $\boldsymbol{y} \in Y$, as described in Section 3.2.1.

Estimates generated using the studied approach are summarised in Figure 8.8 along with the previously considered results obtained through $f-$divergence estimation. The presented intervals were produced by replicating each experiment 100 times, based on a collection of 500 samples from each of the considered predictive distributions. The quality of the obtained estimates appears to be comparable, however the required computational time differs substantially between the two examined methods. More specifically, estimation of the expected utility of one experimental condition using the proposed variational approach, requires up to 30 minutes while the alternative Monte-Carlo estimator incurs a computational time of approximately 2.5 days for the same population of 500 samples. The considerable difference is attributed to alternative processes involved in the two estimation methods, a detailed discussion of which has been previously considered in Chapter 5 and is further examined in Section 8.6 in the context of this particular case study. Evaluation of the expected utility through Monte-Carlo based methods resorted

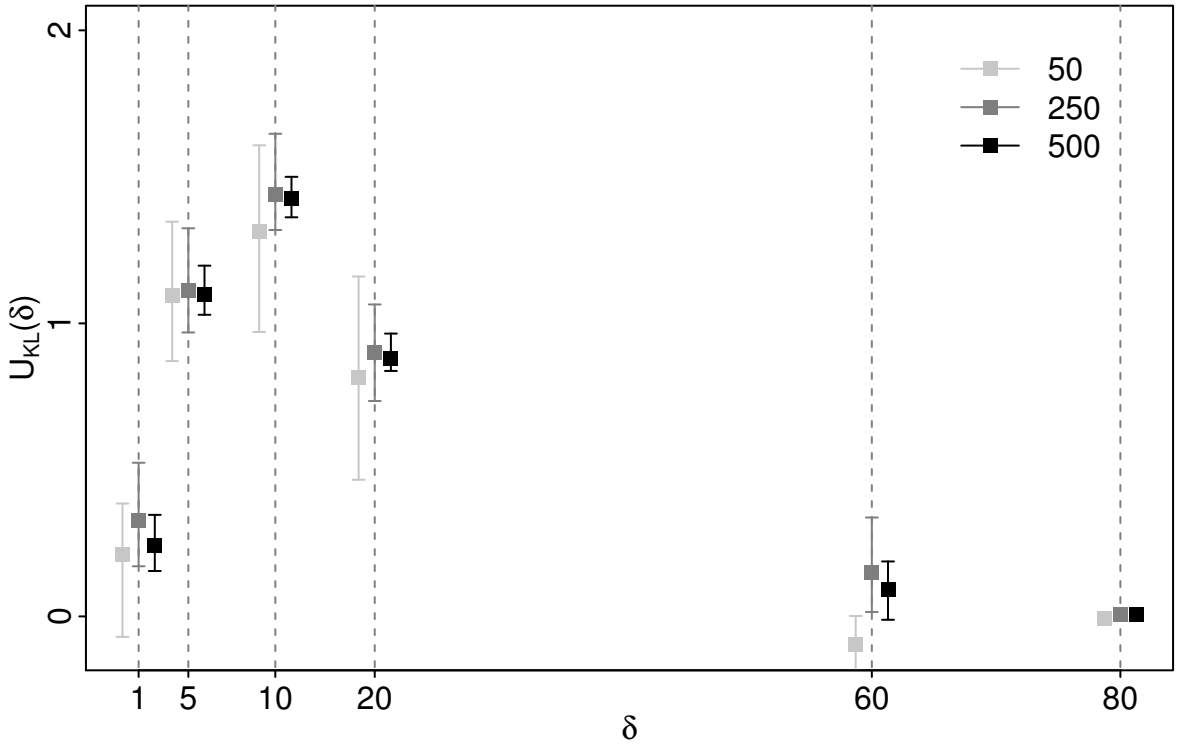


Figure 8.9: Credible intervals (95%) of estimates obtained through Monte-Carlo based methods given populations of differing sizes. Although each set corresponds to exactly the same experimental condition (marked by dashed grey lines), a jitter term is added to values of the x-axis for clarity.

to use of a High Performance Distributed Computing framework, however, it is worth noting that, access to such resources is not always possible. In this case, consideration of smaller sample sizes would be necessary for evaluation of the expected utility within more realistic time frames. Figure 8.9 summarises estimates of the expected utility obtained through the examined Monte-Carlo based methods given sample populations of differing sizes. As illustrated therein, the noise in the obtained estimates increases considerably with the decrease in sample size. In particular, choice of a population of 50 samples from each of the predictive distributions, incurs a significant overlap between intervals corresponding to different candidate experimental conditions, thus hindering the effective discrimination among them while still incurring a considerable computational time of ~ 6.6 hours.

For each choice of sample size, a population of 1000 SMC samples was employed for marginal likelihood estimation, as it was found to be the minimum size required for a

'healthy' population during the sampling phase. The established sequence of intermediate distributions connecting the prior and posterior distributions was based on the tempering schedule described by 7.7.

Overall, results obtained from the two competing methods are in agreement, attributing time point 10 s the highest utility for model comparison. The performance of both considered estimators appears to improve for smaller values of the expected utility, for instance those corresponding to 60 or 80 s, which can be attributed to the lower variability present in the corresponding predictive samples.

Lastly, Figure 8.10 provides a summary of the estimates obtained through Monte-Carlo estimation using the 0-1 utility function u_{0-1}, introduced in Chapter 2, on a sample of 500 predictions. This utility is included as an example of a non-convex φ under which, application of the proposed estimation method is not possible. Overall, the obtained estimates agree with the results of preceding paragraphs, however, significant information loss can be observed comparatively. This is attributed to u_{0-1} capturing simply information on whether the correct model was performed successfully at a particular time point, without any quantification of the corresponding evidence in support of this classification.

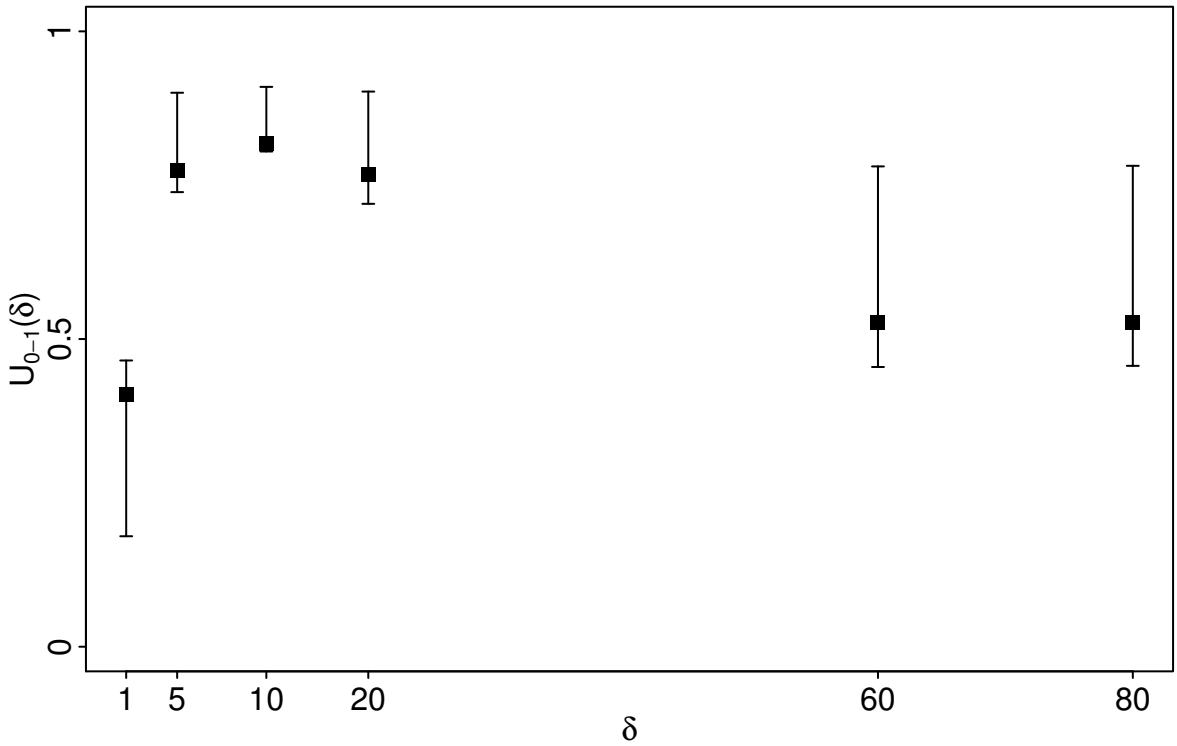


Figure 8.10: Credible intervals (95%) of estimates obtained through nested Monte-Carlo integration using the 0-1 utility function.

8.4 Design optimisation

Sections 8.3.1 and 8.3.2 demonstrate the benefits incurred from use of the proposed estimator of the expected utility. Interest, in this section, focuses on incorporating the variational estimator into a Bayesian Optimisation algorithm for maximisation of the expected utility over the design space, following the methodology that was presented in Chapter 6. Figure 8.11 provides a summary from implementation of the BO algorithm for maximisation of the expected utility following the procedure described in Algorithm **??**. The predictive mean and credible region of the Gaussian process obtained at the latest optimisation step are presented therein, incorporating the information from every observed function evaluation (denoted by red marks). The width of the credible region reflects the estimation error with the greatest effort being placed in inferring the error at the optimum. This was found to be at 12 s and is marked by a dashed grey line.

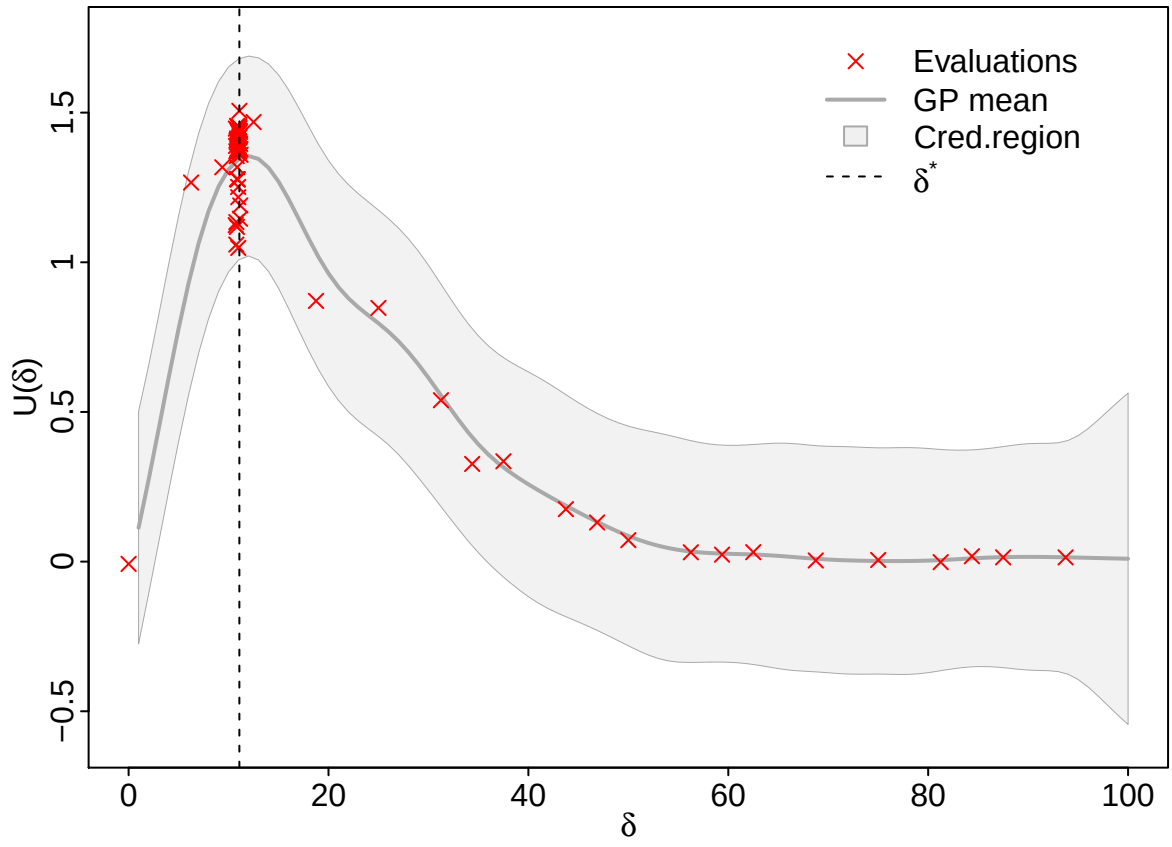


Figure 8.11: Maximisation of the expected utility surface corresponding to choice φ_{KL} using Gaussian process Bayesian optimisation.

The summarised results were generated under definition of the utility function φ_{KL}, previously introduced in Section 8.3.1. A Gaussian process prior with mean vector $\mathbf{0}$ was used for this case study and alternative definitions of the previously considered acquisition functions were explored — the Expected Improvement and Upper Confidence Bound. Both concluded similar results and so only a summary of the former is provided in Figure 8.11. In order to obtain some insight on the shape of the utility surface in absence of pre-existing information, initial evaluations of the expected utility were obtained for 10 selected as a Sobol sequence (Sobol', 1967).

The succeeding section presents an autonomous framework for the study of biochemical systems, incorporating the Bayesian optimisation procedure in combination with the efficient variational estimator implemented in this section.

8.5 Sequential and adaptive design

This section presents an autonomous and efficient framework for the study of biochemical networks using experimental data. Incorporating the individual processes explored in Sections 8.3 and 8.4, a response-adaptive sequence of designs is possible within realistic timescales. The adopted procedure, described in Algorithm 4, establishes a closed-loop setup of data collection, knowledge update and decision making which flexibly integrates new information as it becomes available, leading to efficient and better informed decision making. A summary of each sequential stage composing this procedure is provided in Figure 8.12 with each subfigure depicting the predictive distributions of the competing models given the most recent data at each stage (marked with red crosses) and the optimal experimental condition for observing the system at the subsequent stage (indicated by a dashed grey line). All results presented in this section were obtained under choice φ_{KL} or, commonly, the Shannon entropy, inducing an expression of the expected utility in terms of the KL divergence between the prior predictive distributions obtained at each stage.

At the initial stage (stage 0) estimation of the expected utility relies on predictive samples corresponding to the prior distributions of model parameters without taking into account any experimental data. As previously discussed in Section 8.3, availability of no

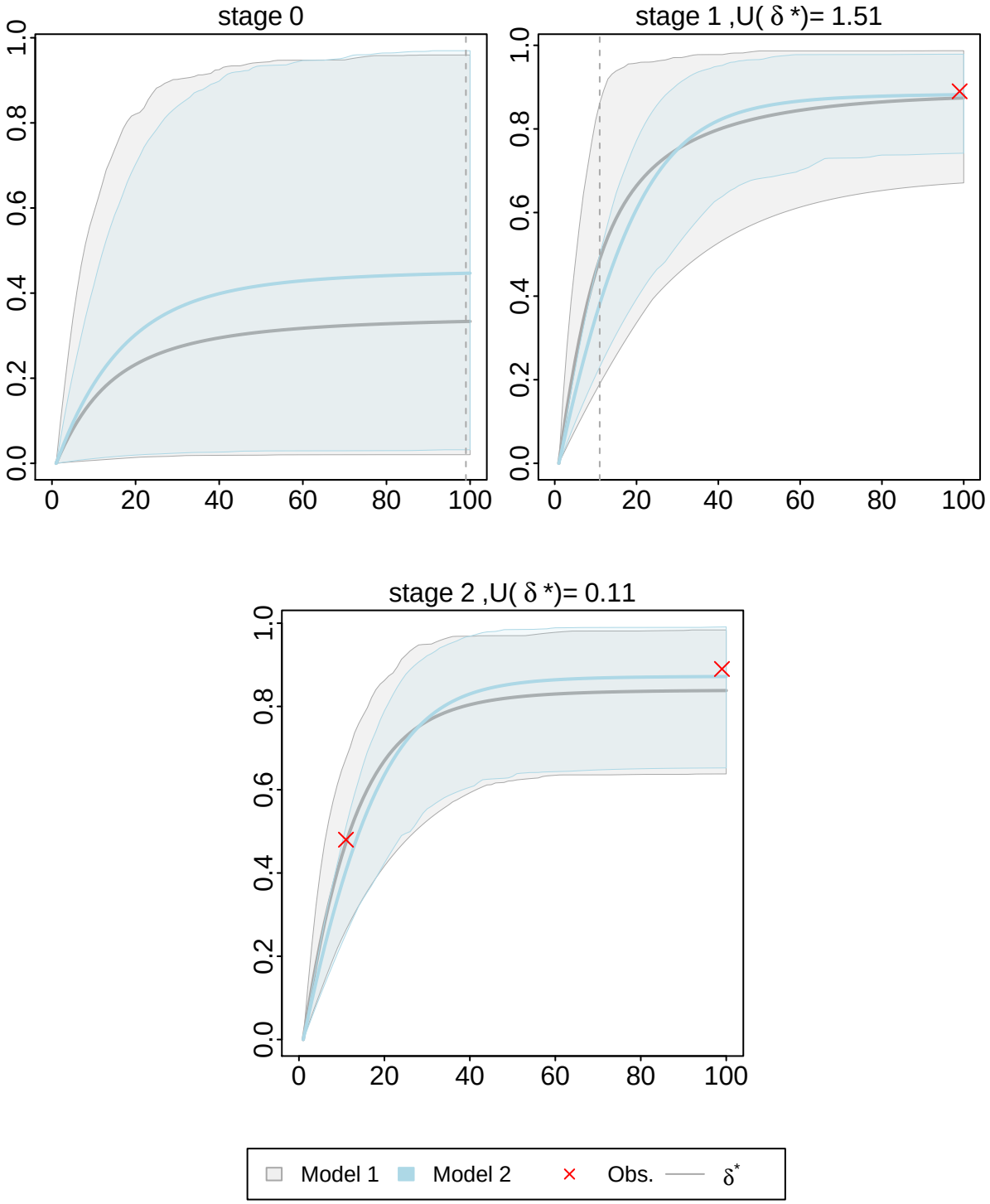


Figure 8.12: A sequential and adaptive experimental design for inferring the structure of a biochemical network given two competing hypotheses. Each subfigure depicts the predictive distribution corresponding to each of the competing models, at each stage of the sequential procedure, given the most recent data (red crosses). The optimal experimental condition δ^*, represented by a dashed grey line, incurs the maximum discrepancy between predictive distributions.

preliminary information at the initial stage of the procedure encourages the assumption of sufficiently broad predictive distributions, accounting for a large number of potential experimental measurements. Due to the significant overlap of predictive distributions following this initial assumption, discrimination between the two models is challenging regardless the time of observation. This was reflected in the obtained expected utility surface characterised by two main traits: 1) a flat surface, indicating that, there is no distinct preference among alternative experimental conditions in Δ and 2) an optimal expected utility induced by choice of φ_{KL} of less than 0.1 which, according to Jeffreys (1961), is not worth considering for experimentation. Since observation of the studied system was essential for overcoming this limitation, at the initial stage an experimental condition was selected randomly as $\boldsymbol{\delta} = \{100\}$ s.

Upon observation of the experimental data, obtained at $\boldsymbol{\delta}$, the existing knowledge was updated to reflect the newly obtained information. This update was incorporated in the posterior distributions of model parameters in a Bayesian manner which were, subsequently, used to provide samples from the corresponding predictive distributions. These served as priors for evaluation of the expected utility at the subsequent stage 1, as depicted in Figure 8.12. The knowledge update incurred a new expected utility surface reflecting the utility of future decisions based on the already observed information. The expected utility surfaces, corresponding to stages 1 and 2 are presented in Figure 8.13. Approximation of the surface and maximisation over the design space was achieved with Bayesian optimisation, as described in Section 8.4.

The information provided therein is used to guide decisions on the experimental design of stages 1 and 2 respectively, following the iterative procedure of Algorithm 4. The results obtained for stage 1 have already been analysed in Section 8.4. The experimental data, obtained at the indicated optimal experimental condition $\boldsymbol{\delta}^* = 12$ s during stage 2 of the sequential process, are shown in Figure 8.12 together with the knowledge update that followed observation of the new information. The updated predictive distributions appear fairly similar after observation of the system at experimental conditions $\{12, 100\}$, deeming discrimination between the candidate models. This initial impression is further reflected in the corresponding expected utility surface — the flatness of the curve suggests that, any experimental condition incurs similar expected utility to its alternatives. More

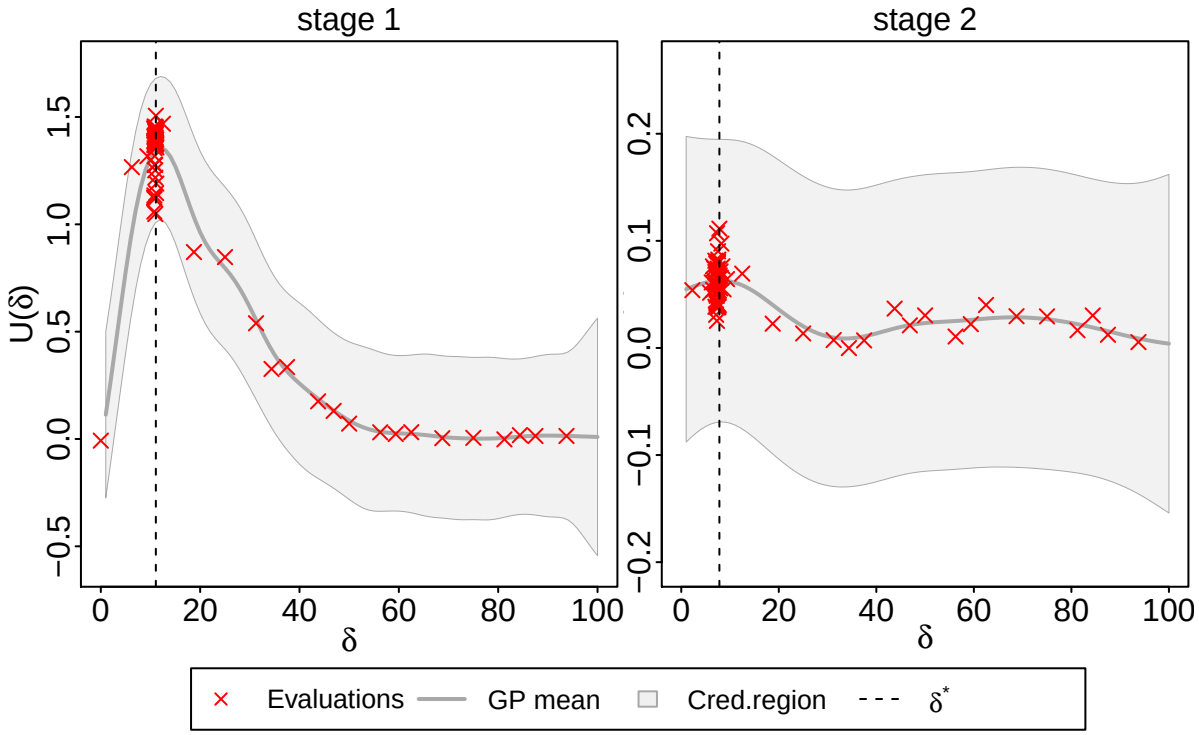


Figure 8.13: Sequentially updated expected utility surface over Δ upon observation of new experimental data at each stage of the adopted design. Each subfigure depicts the expected utility (y-axis) evaluated for all potential experimental conditions (x-axis) at each distinct stage of the process. After observation of the system at stage 0, none of the remaining experimental conditions offers any substantial new information at subsequent stages, resulting in a flat and close to 0 expected utility surface.

importantly, it can be observed that, no elements in Δ benefit the study to such a degree that justifies subsequent experimentation under the proposed condition $\boldsymbol{\delta}^*$ as its corresponding expected utility was found to be only 0.11. Given the considerable cost involved in the data collection stage, the sequential procedure was completed without observation of the system at the optimal condition (8 s) and subsequent knowledge update.

Overall, the most substantial information gain was obtained during the second experimental stage. Having observed the system at times 12 and 100 s, further observation incurred no additional gain worthy of consideration. The optimal sequence of designs is, thus, concluded as $\boldsymbol{\delta}^*_{\text{seq}} = \{12, 100\}$, suggesting observation of the biochemical network early after it has been triggered and later on, once equilibrium has been reached.

8.6 Summary and conclusion

This chapter presented application of a range of methodologies, introduced in Chapters 2, 5 and 6, to a real-life problem in Systems Biology. Section 8.1 outlined the need for experimental design for the optimal study of biochemical systems which was subsequently achieved in Section 8.3.1 through implementation of the proposed estimator. As discussed in Chapter 5, this approach induces a class of information-based utility functions, the properties of which were explored further in this case study. Choice of alternative utility definitions provided by this class demonstrated that, each choice induces a distinct f-divergence between predictive distributions which not only provides a natural representation of the utility as a measure of separation between models but it also allows for exploitation of their corresponding properties.

Estimation using the widely adopted Monte-Carlo integration was also considered in Section 8.3.2 for a comparison. The two methods were found to produce estimates of comparable quality given same-sized populations, however the proposed estimator was shown to be significantly more efficient computationally. The observed improvement can be predominantly attributed to the increased efficiency of the proposed estimation process compared to the currently adopted Monte-Carlo based approaches. More specifically, the variational estimator is extremely data-efficient, employing a vector of potentially infinite basis functions that compose the class of RKHS which introduce an increased expressiveness allowing for more information to be extracted from the predictive samples. Contrarily, traditional Monte-Carlo methods rely on evaluation of the marginal likelihood, acting as a summary statistic for each individual data point of the predictive sample. Marginal likelihood estimation requires sampling from the posterior distribution of model parameters as a closed form expression for it is not available under the models employed in this chapter. As demonstrated in Section 8.3.2, this process is associated with extremely high computational cost which, often, compromises the optimal examination of the studied objective. In addition, although SMC sampling algorithms were shown to be quite well suited to this purpose with the marginal likelihood resulting as a by-product of this procedure, they are known to require a considerable amount of tuning and so their implementation may often be time-consuming. On the contrary, a limited

number of parameters are involved in the proposed estimator, guidelines for choice of which are provided in this thesis.

The proposed estimator was, in Section 8.4, flexibly incorporated into the BO algorithm, achieving complete and efficient exploration of the design space. The two procedures were lastly combined in Section 8.5, establishing a closed loop, automated setup for the study of biochemical systems through optimal experimental design.

Chapter 9

Optimal experimental design for the study of fluorescent kinetics

This chapter presents a real-life application of the proposed optimal experimental design framework to the study of fluorescent kinetics. Observation of such phenomena is typically possible only via specialised instrumentation that allow specification of numerous factors impacting the observed output. Certain settings may be proven to be more beneficial for inference in the studied system and so, optimal experimental design is key to identifying such conditions. The proposed methodology, combined with instrument control software, provides a fully automated solution to the study of such phenomena, establishing a closed-loop sequence of experimentation, knowledge update and optimal decision making, without requiring further human input at any stage. This case study extends on the previously considered problems by showcasing implementation of the proposed methods given a high-dimensional output space and a higher dimensional design space.

9.1 Experimental design in Photophysics

This section provides a brief introduction to terminology and processes from fluorescent spectroscopy that are most relevant to the phenomena examined in this case study.

Fluorescence describes the ability of a substance to absorb and re-emit energy in

the form of light. Absorption of light causes a change in electron distribution, moving the molecule to an excited electronic state which then relaxes back to the ground state via emission of light (Andrews, 2015). Due to this property's dependence on the local environment, fluorescent molecules are often used as sensors for numerous biochemical applications providing information on molecular structures and interactions (Rolinski et al., 2010). A typical experiment consists of exciting a fluorophore molecule and monitoring the average length it spends in this state which is commonly referred to as the **fluorescence decay time**. This measure can reveal useful information about surrounding sites and neighbouring components of the studied system.

A popular instrument for obtaining experimental data of decay measurements adopt time-correlated singe-photon counting (TCSPC) methods (Alghamdi et al., 2018) that output a histogram of fluorescence photon arrival counts from successive excitation cycles. The TCSPC is composed by 1) a light source that is used to excite the system, 2) a black-box environment where molecule interactions occur and 3) a detector of fluorescence. The decay time is subsequently inferred on the basis of the emission and detection states. The system behaviour can be highly sensitive to external factors which can be flexibly regulated by the experimental equipment (Rolinski et al., 2010). In particular, as the decay time is known to often be dependent on the emission and detection wavelengths, optical components allowing the isolation of a desired wavelength are typically placed between stages 1)-2) and 2)-3). The wavelength distribution of an emission corresponding to a fixed excitation wavelength will be referred to as the **emission spectrum** (**?**).

The fluorescence decay, denoted by $I(\boldsymbol{\delta}, t)$ where $\boldsymbol{\delta}$ represents the experimental conditions and t the time of measurement, is defined through the individual mechanisms operating in the system during its excited state (Birch and McLoskey, 2017). Traditionally considered models have so far relied on rather simplifying assumptions that, although well-suited to simpler systems, fail to generalise to more complex structures, for example, in cases when higher sample heterogeneity is present. High sensitivity of these systems on external conditions and intrinsic mechanisms that are yet not entirely understood are some of the challenges currently faced. This chapter is concerned with the application of the proposed optimal experiment design methodology in an attempt to study and establish more realistic models that sufficiently capture the complex behaviour of the examined

systems.

The particular process under study concerns the interaction between an excited and an inactive molecule in a heterogeneous environment. This phenomenon is commonly termed as **fluorescence resonance energy transfer** (FRET;**?**) while the interacting molecules will be referred to as the donor (D) and acceptor (A) molecules respectively. As FRET is distance dependent, a useful application of this phenomenon allows inference on the distance between the donor and acceptor molecules currently exploited in numerous scientific disciplines including medical diagnostics (Knowles et al., 2014), DNA analysis (Chung et al., 2019) and optical imaging. Similarly to the biochemical systems, considered in Chapter 8, FRET is only indirectly observed through fluorescence decay measurements. The role of optimal experimental design to address this issue, enabling the optimal study of such systems is further considered in Section 9.2.

9.2 Case study

An outline of the studied problem is provided in this section. A collection of hypotheses describing alternative potential dynamics is introduced in 9.2.1 while Section 9.2.3 considers the components and setup of the experiments facilitating their study.

9.2.1 Alternative hypotheses

Following the notation adopted in Chapters 2-8, the experimental dataset $\boldsymbol{y}$ is represented by a λ-dimensional vector of fluorescence decay measurements over time, the experimental parameters are summarised by $\boldsymbol{\delta} = (\mu, \nu)$, where μ and ν refer to the excitation and detection wavelengths, measured in nanometres (nm), and lastly $\boldsymbol{\theta}_m$ expresses the remaining uncontrolled parameters associated with model m.

Two hypotheses are formed to describe the rate of FRET at a particular time point. This will be denoted by $k_i(t)$, where $i = 1, 2$ refers to the corresponding hypothesis and finds expression in:

$$k_i(t) = \int_0^\infty \rho_i(r) \frac{\mathrm{d}}{\mathrm{d}t} \left\{ \exp\left[-\left(\frac{R_0}{r} \right)^6 \frac{t}{\tau_D} \right] \right\} \mathrm{d}r \tag{9.1}$$

where r is the distance between the donor and acceptor molecules, τ_D the lifetime of the donor molecule in the absence of energy transfer and R_0 is the Forster distance, determined by the spectral overlap of the donor emission and the acceptor absorption. **Hypothesis 1** assumes a Gaussian distribution for r and so:

$$\rho_1(r) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(r-r_0)^2)}{2\sigma^2}\right] , \tag{9.2}$$

while under **Hypothesis 2**, r is considered to be uniformly distributed within a ring of distances bounded by r_1, r_2, taking the form:

$$\rho_2(r) = \frac{3r^2}{r_2^3 - r_1^3} , \tag{9.3}$$

for $r_1 < r < r_2$ and 0 otherwise. The rate of FRET is incorporated in the expression of the fluorescence decay rate $I(\boldsymbol{\delta}, t)$ as:

$$I(\boldsymbol{\delta}, t) = D_{\boldsymbol{\delta}}(t) + A^1_{\boldsymbol{\delta}}(t) + A^2_{\boldsymbol{\delta}}(t) , \tag{9.4}$$

where:

$$D_{\boldsymbol{\delta}}(t) = \frac{\epsilon_D(\mu)\phi_D F_D(\nu)}{\epsilon_D(\mu) + \epsilon_A(\mu)} \exp\left[-\frac{t}{\tau_D} - \int_0^t k_i(t')\mathrm{d}t'\right] \tag{9.5}$$

$$A^1_{\boldsymbol{\delta}}(t) = \phi_A F_A(\nu) \int_0^t D_{\boldsymbol{\delta}}(t-z) k_i(t-z) \exp\left[-\frac{z}{\tau_A}\right] \mathrm{d}z \tag{9.6}$$

$$A^2_{\boldsymbol{\delta}}(t) = \frac{\epsilon_A(\mu)\phi_A F_A(\nu)}{\epsilon_D(\mu) + \epsilon_A(\mu)} \exp\left[-\frac{t}{\tau_A}\right] . \tag{9.7}$$

The terms ϵ_D, ϵ_A are the extinction coefficients of the donor and acceptor molecules respectively indicating the strength with which they absorb light at a given wavelength, ϕ_D, ϕ_A provide a measure of the efficiency of the fluorescence process and will be referred to as the **quantum yield** of D and A respectively. In essence, the quantum yield expresses the percentage of the photons emitted out of all the photons absorbed. Lastly, quantities F_D and F_A express the emission spectrum associated with a particular wavelength.

To account for the delay between the moment that light is first emitted by the source

and absorbed from the fluorophore as well as the detector's response to the re-emitted light, the light pulse from the source or prompt is convolved with $I(\boldsymbol{\delta}, t)$ resulting in the observed decay:

$$R(\boldsymbol{\delta}, t) = p(t) \otimes I(\boldsymbol{\delta}, t) \ , \tag{9.8}$$

where operation $\otimes$ describes the convolution between the prompt and fluorescence decay.

9.2.2 Statistical models

As the experimental output is expressed in terms of photon arrival counts over time, a Poisson model is a reasonable modelling choice and thus:

$$\boldsymbol{y} \mid \boldsymbol{\delta} \sim \text{Poisson}(R(\boldsymbol{\delta})) \ , \tag{9.9}$$

for $R(\boldsymbol{\delta}) = \{R(\boldsymbol{\delta}, 1), \ldots, R(\boldsymbol{\delta}, \lambda)\}$.

Data for terms $\epsilon_D, \epsilon_A, \phi_D, \phi_A, F_D, F_A, R_0$ were obtained from public databases while τ_D, τ_A and the parameters involved in ρ_1 and ρ_2 were considered as the unknown model parameters. More specifically, $\boldsymbol{\theta}_{m_1} = (\tau_D, \tau_A, r_0, \sigma^2)$ and $\boldsymbol{\theta}_{m_2} = (\tau_D, \tau_A, r_1, r_2)$. Prior distributions were assigned to the unknown model parameters n a Bayesian context:

$$\begin{aligned}
\tau_D &\sim G(3, 2) \\
\tau_A &\sim G(3, 2) \\
r_0 &\sim G(3, 2) \\
\sigma^2 &\sim IG(3, 0.1) \\
r_1, \ r_2 &\sim U(0, 10), \ \text{if } r_1 < r_2 \text{ and 0 otherwise} \ ,
\end{aligned}$$

where the lifetimes τ_D and τ_A were assigned a sufficiently wide *Gamma* (G) distribution while ensuring positivity in their obtained values and similarly for the distance r_0. Based on prior knowledge suggesting that energy transfer is possible when the distance between the donor and acceptor is fairly small, a relatively low variance σ^2 was assumed and, thus assigned the stated inverse Gamma (IG) prior. The bounds r_1 and r_2 were assigned uniform distributions (U) between 0 and 10 under the condition that $r_1 < r_2$, as required

by (9.3). Similarly to Chapters 7 and 8, one of the two competing models was assumed to be the 'true' model and was used to artificially generate experimental data from the observed system. In particular, the model expressing Hypothesis 1 was assumed as the 'true' model.

9.2.3 Experimental setup

This case study is concerned with experiments performed using a TCSPC instrument to observe the behaviour of a studied biomolecular system in an heterogeneous environment. Interest lies in using the obtained output to compare the formed hypotheses attributing alternative distance distributions in FRET that are expressed by models (9.2) and (9.3). The considered FRET system was based on the interaction between a yellow fluorescent protein (YFP) and a cyan fluorescent protein (CFP) as a donor-acceptor pair where excitation of the donor leads to emission from the acceptor molecule assuming that the proteins are close enough for energy transfer to occur. FRET can thus be used for inference of direct protein-protein interaction between YFP and CFP fusion proteins in living or fixed cells.

As previously discussed in Section 9.1, the purpose of this study is to establish models that provide a more realistic representation of the complex nature inherent in photophysical phenomena. Inevitably, such an expression is reliant on specification of highly sophisticated models, evaluation of which is, as a result, computationally demanding.

The case study presented in this chapter represents a problem in which the computational demand of models evaluations stretches to such an extend that deem the traditional optimal experimental design methods intractable, even after resorting to the use of high performance computing resources. Particularly, generating one prediction under expression 9.4, requires 11.6 seconds on average. Thus, evaluation of the expected utility corresponding to one experimental condition under the previously adopted Monte-Carlo approach based on a population of 50 samples and using a high performance parallel computing framework, totals to a computational time of approximately 12 days. Optimal experimental design through Monte-Carlo based methods was, therefore not considered for the study of this model selection problem. The results presented in the following Section 9.3, relied solely on evaluation of the expected utility through the proposed vari-

ational estimator, introduced in Chapter 5. The estimator is further incorporated into a sequential, response-adaptive framework, establishing a fully automated process for the study of fluorescent kinetics in different systems.

9.3 Sequential and adaptive design

The implementation of a fully automated and highly efficient framework is demonstrated in this section for the study of fluorescent dynamics in heterogeneous environments. The response-adaptive algorithm summarised in Algorithm 4 was employed for this purpose, similarly to the previously considered case studies of Chapters 7 and 8.

Under the adopted setup, the observed output $\boldsymbol{y}$ is expressed by a 4096−dimensional vector representing the photon counts recorded by the TCSPC instrument at each of the 4096 time points. An illustration of the competing model predictions for a collection of potential experimental conditions is provided in Figure 9.2 where photon counts are plotted on the logarithmic scale following common practices. As illustrated therein, the experimental output is highly sensitive to choice of particular experimental conditions such as the emission and absorption wavelengths which were, therefore, chosen as the experimental conditions, thus $\boldsymbol{\delta} = (\mu, \nu)$. The work presented in this thesis lays the foundation for the study of additional conditions that may affect the observed output such as the system temperature which can be flexibly incorporated in $\boldsymbol{\delta}$.

The donor molecules typically emit at shorter wavelengths that overlap with the absorption spectrum of the acceptor. The emission and absorption spectra for the YFP -CFP pair are provided in Figure 9.1 in which an overlap between the donor emission and acceptor absorption can be observed at the shaded region corresponding to wavelengths between 450 and 530 nm. Due to this phenomenon, experimental conditions within the shaded region were considered in this case study.

Similarly to the experimental designs established in the studies of Chapters 7 and 8, a sequential and response-adaptive framework was adopted, taking full advantage of the available information at each stage by incorporating it into the study through an update from the prior to the posterior model parameters to reflect the newly obtained knowledge. The initial stage of the sequential procedure considered predictions corresponding

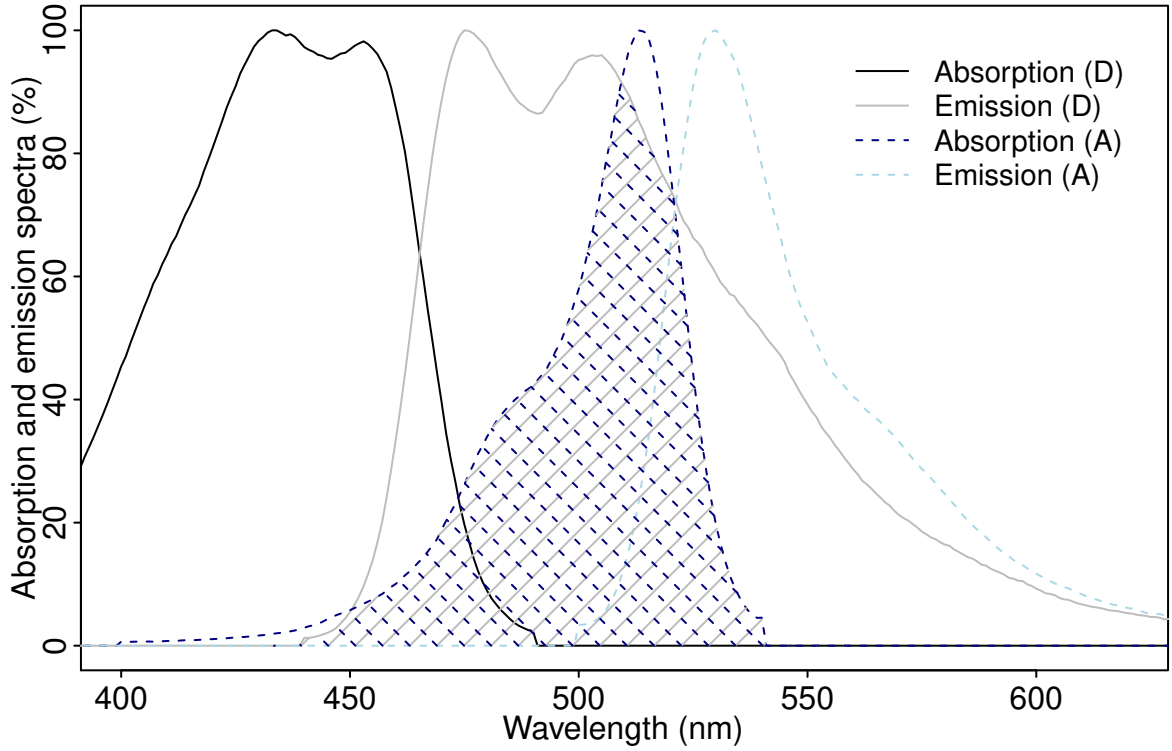


Figure 9.1: Absorption and emission spectra for the YFP (D) - CFP (A) pair.

to the prior distributions of model parameters and the expected utility was employed to quantify the potential benefit of each experimental condition under consideration. Efficient estimation of the expected utility was achieved through the variational estimator proposed in Chapter 5 for the previously considered choices of φ_{KL} and φ_H. Similarly to Chapter 7, the RKHS class structure induced by the Gaussian kernel was imposed on function class $\mathcal{G}$ defined by (7.6) with $\|\cdot\|$ expressing the Euclidean metric in $\mathbb{R}^\lambda$. A sample size of 500 was used for the estimation inducing an inversely analogous penalty term of $\rho = \frac{1}{500}$.

Observation of the system through the TCSPC instrument was performed during the subsequent stage at the proposed condition, enabling the update or formulation of new hypotheses in light of the newly obtained knowledge. Thus, at each stage of the procedure, the considered hypotheses reflected every system measurement observed up to that point allowing more effective learning and efficient use of resources. As there were no particular resource limitations to restrict the number of experiments that could be considered, experimentation was continued until the information provided from newly

observed experimental conditions was no longer beneficial.

Prior predictions obtained at stage 0 of the sequential procedure under the competing hypotheses for a selection of designs is provided in Figure 9.2.

As variables μ and ν are discrete, a grid search over the design space was feasible which concluded designs $\boldsymbol{\delta} = (\mu, \nu) \in [400, 490] \times [440, 490]$ to be the most beneficial for observation of the studied system at the subsequent stage 1. Among them, the optimal design was found to be $\boldsymbol{\delta}^* = (400, 440)$ nm incurring maximum expected utilities of

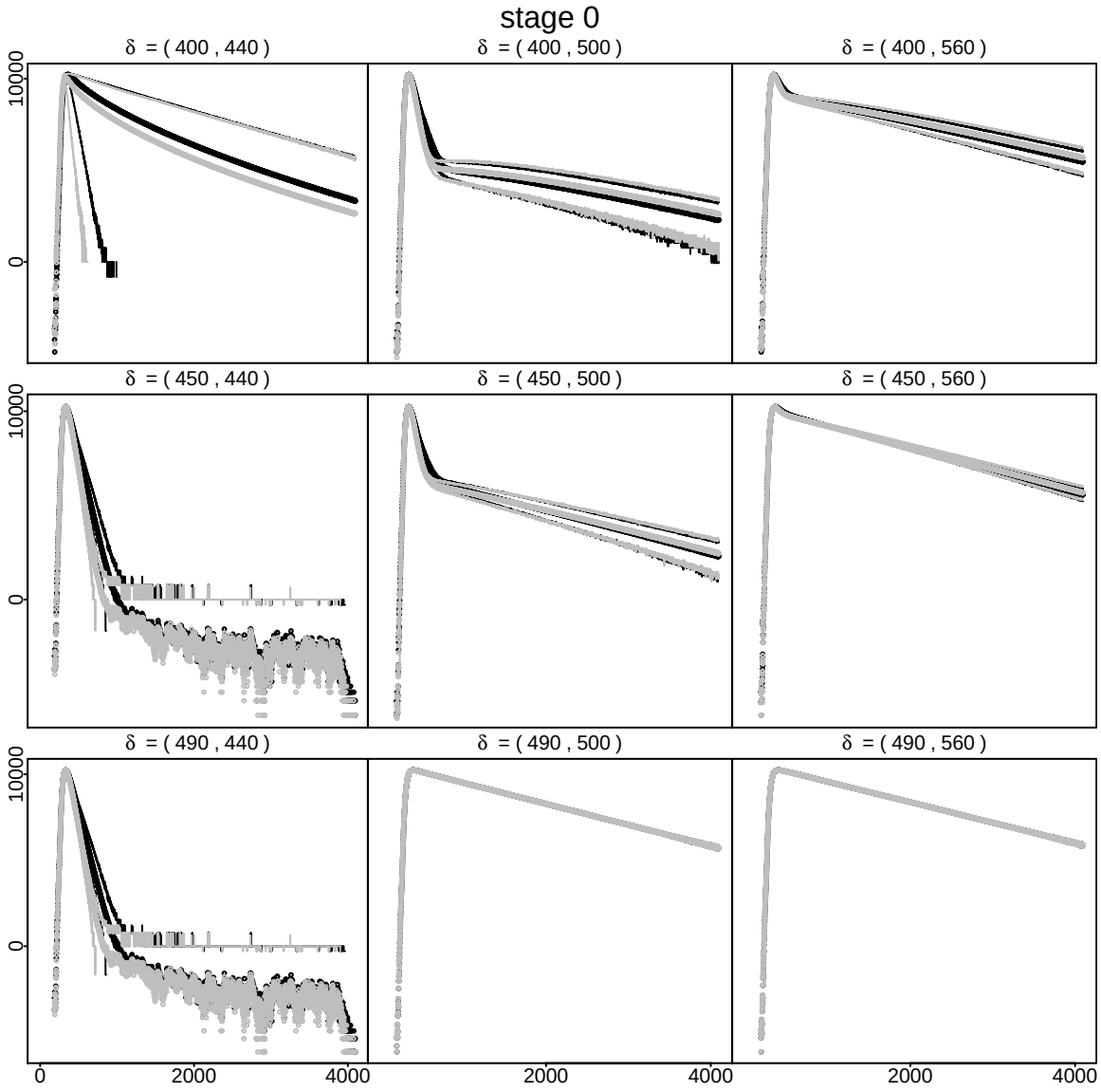


Figure 9.2: Credible intervals (95%) of the prior predictive distributions (y-axis) over the 4096 observed time points (x-axis) corresponding to Hypotheses 1 (black) and 2 (grey) generated at a selection of experimental conditions $\boldsymbol{\delta} = (\mu, \nu)$.

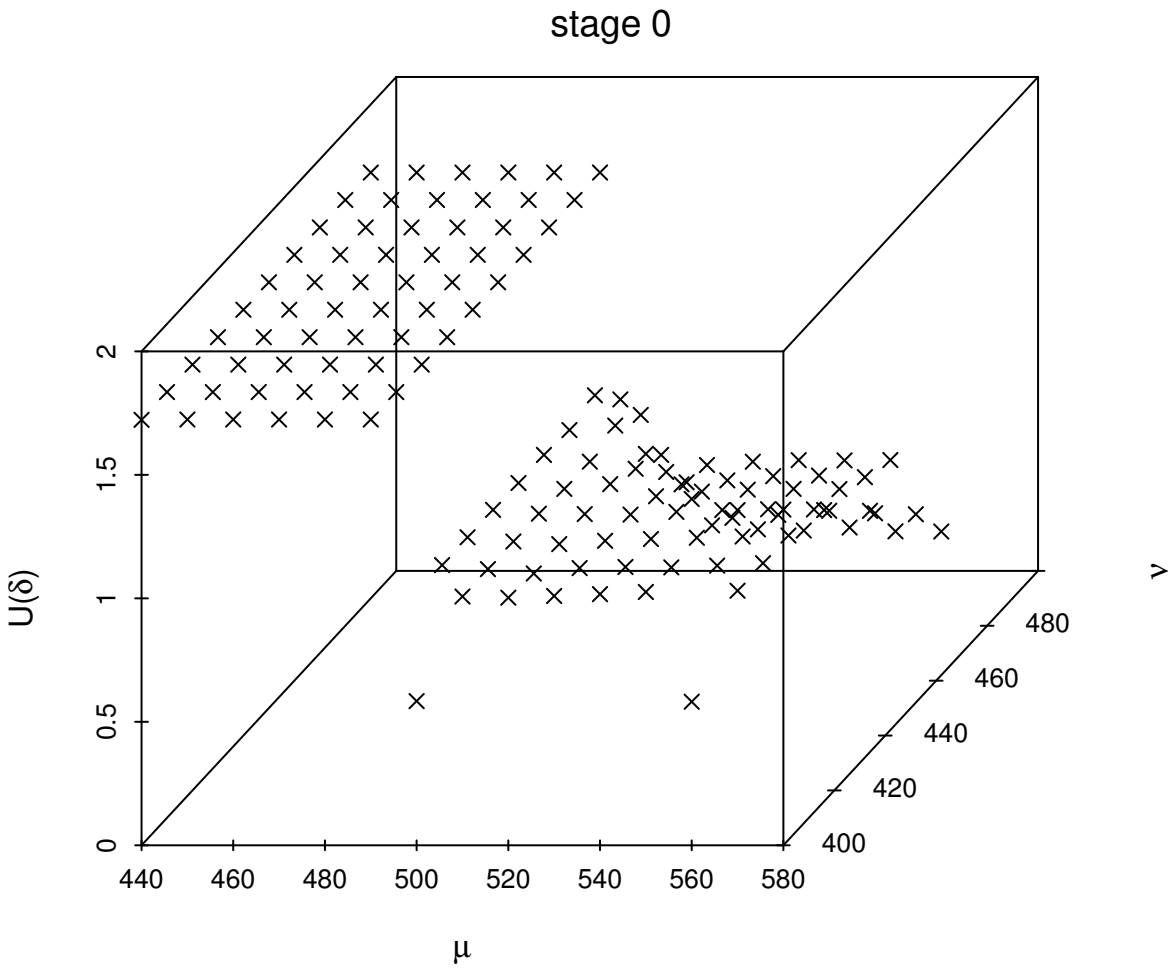


Figure 9.3: Expected utility surface over Δ based on samples from the prior predictive distributions of considered models before the observation of experimental data.

$U_{\mathrm{KL}}(\boldsymbol{\delta}^*) = 1.72$ and $U_H(\boldsymbol{\delta}^*) = 0.8$. The produced expected utility surface induced by choice of φ_{KL} is depicted in Figure 9.3.

Similarly to the previously considered case study of Chapter 9, the shape of the expected utility surface remained the same under the alternative φ_{H} with the estimates being restricted on $[0, 1]$.

Experimental data collected at condition $\boldsymbol{\delta}^* = (400, 440)$ nm during the first stage of the algorithm are depicted in Figure 9.4. The new information was incorporated to update the existing knowledge and was reflected by the posterior distribution of model parameters. The posterior predictive distributions, depicted in Figure 9.5, served as prior predictive distributions at stage 1 of the sequential procedure, providing samples for evaluation of the expected utility corresponding to alternative experimental conditions for consideration at the following stage 2. The corresponding expected utility surface found

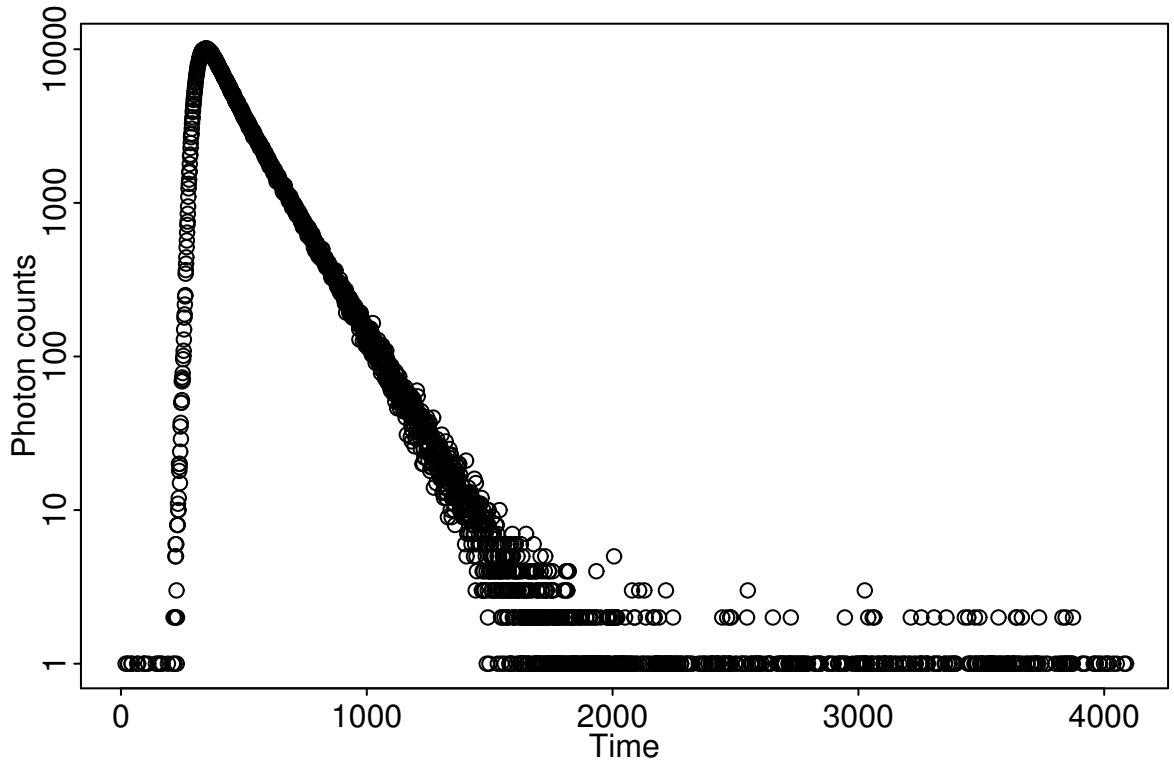


Figure 9.4: Data obtained for experimental condition $\boldsymbol{\delta}^* = (400, 440)$ nm. Photon counts are plotted on the logarithmic scale.

through estimation of the expected utility using the proposed variational estimator under φ_{KL} is provided in Figure 9.6. The most beneficial experimental conditions, again, lay within the subset $\boldsymbol{\delta} = (\mu, \nu) \in [400, 490] \times [440, 490]$ with their corresponding expected utilities being overall higher than the previously produced estimates of stage 0. In particular, the optimal experimental condition $\boldsymbol{\delta}^* = (450, 440)$ nm incurred expected utilities $U_{\mathrm{KL}}(\boldsymbol{\delta}^*) = 2.85$ and $U_H(\boldsymbol{\delta}^*) = 0.84$, suggesting a substantial potential gain from observation of the system at these wavelengths thus encouraging further experimentation.

The sequential procedure continues until the new information provided by the experimental data carry no substantial knowledge for discriminating against the competing hypotheses. Overall, the knowledge update step at each sequential stage was found as the most time-consuming one as sampling from the posterior distribution of model parameters is required in light of the new experimental data. However, under the proposed variational estimator for evaluation of the expected utility, sampling at each stage needs to be performed only once for each model in order to generate the corresponding predictive distributions and thus, the procedure can still be completed under reasonable time

scales. As previously discussed, posterior sampling from the examined models required 15 hours using SMC methods under a high performance distributed computing framework. At each stage, estimation of the expected utility for all conditions within Δ ranged from a few minutes to an hour when considering a collection of 1000, 4096−dimensional predictions from each model under consideration.

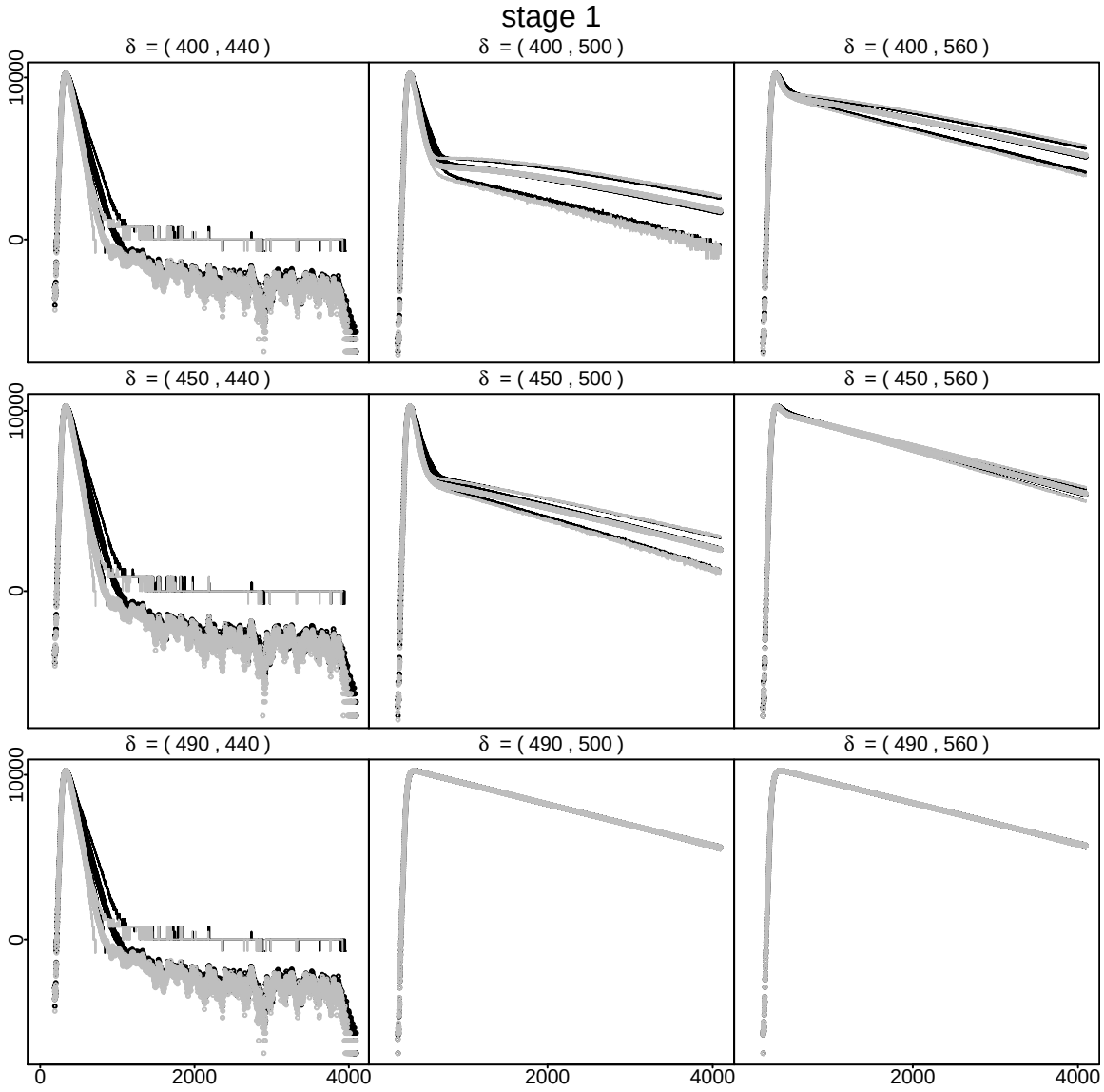


Figure 9.5: Credible intervals (95%) of the posterior predictive distributions (y-axis) over the 4096 observed time points (x-axis) corresponding to Hypotheses 1 (black) and 2 (grey) generated at a selection of experimental conditions $\boldsymbol{\delta} = (\mu, \nu)$.

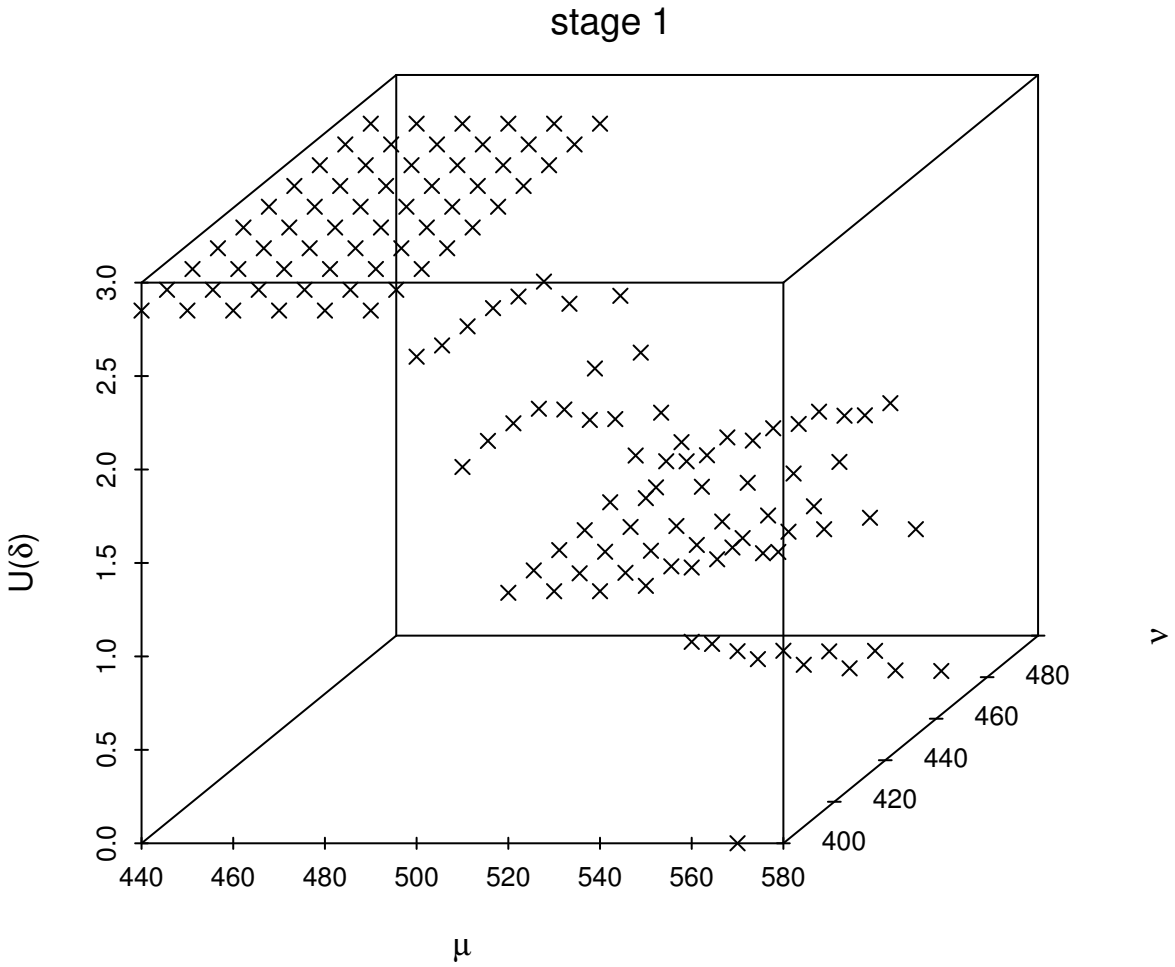


Figure 9.6: Expected utility surface over Δ based on samples from the predictive distributions upon observation of experimental data produced at stage 0.

Summary

This chapter considered a real-life application of the proposed optimal experimental design methodology for the study of fluorescent kinetics in heterogeneous environments. Employment of optimal design within this research area is thought to be key for developing more realistic, and thus more sophisticated, models that adequately capture the complex nature of such phenomena. Due to the high computational cost required in evaluation of the considered models, currently adopted Monte-Carlo based methods are proven to be highly inefficient and thus not suitable for optimal experimental design under such setups. On the contrary as demonstrated in this chapter, implementation of the proposed variational estimator allows consideration of a sequential setup under

which the most recently observed information is incorporated into the study, efficiently informing subsequent decisions. The established framework, combined with instrument control software, provides a fully automated solution for the study of such phenomena within a realistic time frame without relying on any human contribution at any stage. This process serves as an emulator to a hypothetical physicist studying a model inference problem describing a photophysical phenomenon. The study initiates with a set of competing hypotheses which are subsequently tested against obtained experimental data. The physicist proceeds in updating their knowledge given the observed dataset and identifies the optimal decision for the continuation of the study based on the most recent information. This sequential procedure continues until an optimal sequence of experiments has been established.

Chapter 10

Conclusion

This thesis focuses on optimal experimental design methodologies for the study of natural phenomena. The work presented in the preceding chapters can be summarised by the following main objectives:

- development of a methodological framework for efficient estimation of the expected utility addressing ongoing challenges,

- integration of efficient optimisation techniques for maximisation of the expected utility over the design space,

- establishment of a comprehensive and automated framework for the optimal study of natural phenomena.

Chapter 2 initiates with an overview of the existing and most relevant methodology along with a discussion on ongoing challenges. Two main issues are outlined: 1) the need for efficient estimation of the expected utility, particularly under models with computationally demanding likelihoods, and 2) consideration of an optimisation framework that achieves a systematic search of the design space within a realistic time frame without compromising the optimality of the proposed design.

Emphasis is placed on the former as it inevitably appears in optimal experimental design studies, especially under the targeted class of problems. Adopting an information-theoretic approach, a class of utility functions is proposed, including commonly employed

utilities such as the Shannon entropy. Although this choice has been widely considered in the relevant literature, it is often associated with considerably high computational cost under traditionally adopted methodology which hinders the optimality of the induced design. Suboptimal utility functions are therefore often adopted instead which have been previously shown to fail in adequately capturing the targeted expected utility surface (Ryan et al., 2016).

The inefficiency of existing approaches within the studied problem class is attributed to their reliance on two main procedures: 1) *evaluation of the partial utility*, requiring estimation of the marginal likelihood ratio of a prediction $\boldsymbol{y}$ for each pair of competing models and 2) *evaluation of the expected utility* through Monte-Carlo integration, relying on evaluation of the partial utility for a considerably large collection of predictions $\boldsymbol{y}$. Evaluation of the marginal likelihood is, typically, dependent on computationally demanding approximation methods, a problem that is aggravated by the need to perform this calculation an arbitrarily large number of times.

To overcome existing limitations, an efficient estimator for evaluation of the expected utility is proposed in Chapter 5. The examined approach establishes a variational representation of the expected utility. A problem shift is thus achieved — from the initial, computationally demanding evaluation problem to a convex optimisation problem that avoids evaluation of the marginal likelihood altogether. This is accomplished by transforming the initial space induced by the marginal likelihood ratio to its dual function space which is further shown that, under an RKHS structure, provides a practical yet optimal solution to the initial expected utility evaluation problem.

Focus, in this thesis, was restricted on the class of RKHS induced by a Gaussian kernel, however consideration of alternative kernels may, in certain cases, be more suitable to the problem at hand and are, therefore, a promising research topic. Alternative structures of the imposed function class, other than the RKHS, may also constitute an interesting extension of this work. An extensive selection of potential classes is provided in Sugiyama et al. (2012).

A challenge, generally associated with variational approximation methods, is the increase in the estimator variance with the true estimated value (Poole et al., 2019; Song and Ermon, 2019). This issue was, in this thesis, addressed through the consideration

of a bounded optimality criterion, namely choice of a utility function under which the expected utility is expressed in terms of the bounded Hellinger distance instead of the traditional, unbounded from above KL divergence. However, alternative approaches such as the clipped density ratios, proposed by Song and Ermon (2019), or the unnormalised lower bounds, depending on multiple samples, considered in (Poole et al., 2019), are perhaps also worth exploring, when greater flexibility in the choice of utility function is required.

Overall, both the existing and the proposed methodology attempt to accomplish the exact same objective: quantify the discrimination between the predictive distributions of the competing models. The substantial improvement incurred from the proposed algorithm, however, lies on its employment of an extremely efficient procedure as opposed to the traditional Monte-Carlo based methods. In particular, the former operates under a highly flexible, non-parametric framework, allowing direct comparison of one distribution against the other. On the contrary, the alternative approach adopts an indirect comparison of the two distributions under which, information from each predictive sample is extracted through its corresponding marginal likelihood. Subsequent comparison of the predictive samples is achieved via the associated marginal likelihood estimates, acting as summary statistics.

Theoretical properties of the proposed estimator are briefly discussed, demonstrating that this alternative approach to evaluation of the expected utility does not compromise the quality of the produced estimates compared to the traditionally adopted methods. Further empirical results are compared at the subsequent Chapters 7, 8 and 9.

The second challenge of efficient expected utility optimisation over the design space is addressed in Chapter 6. This problem is predominantly associated with studies considering continuous and potentially, high dimensional design spaces, where a deterministic comparison of experimental conditions requires an unrealistic number of operations and alternative methods are suboptimal, failing to adequately explore the optimised surface. The class of Bayesian optimisation algorithms is introduced in an attempt to address these ongoing challenges. Despite their remarkable suitability, Bayesian optimisation algorithms have not posed a particularly popular choice in the optimal experimental design literature.

As further discussed in Chapter 6, Bayesian optimisation algorithms combined with the proposed variational estimator set the foundation to a fully automated framework for the study of modern phenomena under a sequential and response-adaptive setup. The established procedure poses an attempt to simulate the setup of a typical scientific study, automating the process of closed-loop data collection, knowledge update and optimal decision making. Focus in this thesis is restricted to a simpler myopic procedure under which, optimal decisions are proposed one at a time, followed by experimentation and expected utility surface update until a new optimal decision is required again. More sophisticated setups may incur higher benefit, taking into account what lies ahead before concluding on an optimal decision. A non-myopic approach was not deemed particularly beneficial to the case studies considered in this thesis and were thus not explored further in view of the additional complexity incurred by such a setup. This extension may, however, constitute an interesting topic of future research.

Chapters 7, 8 and 9 are devoted to application of the proposed optimal experimental design framework to three case studies. Each case study allows exploration of different aspects of the examined methodology and its comparison against traditionally adopted approaches.

Chapter 7 considers a simple case study of a model discrimination problem on a set of polynomial models. Although the employed class of models lies outside the scope of this thesis, interest in the initial case study was placed on a comparison of the two competing estimation methods when the true estimated value is known. Under this setup, findings showed that both estimators perform particularly well based on an equal number of samples from the predictive distributions under consideration. Although not yet prohibitive towards the Monte-Carlo based method, an apparent difference in the respective computational demand of each method exists.

A real-life problem of experimental design for model inference in Systems Biology is examined in Chapter 8. Alternative hypotheses on the structure of a biochemical network are represented by mathematical models expressed by systems of ordinary differential equations. As a result, considerable computational time is devoted to obtaining each model prediction, building up to an infeasible length required for expected utility evaluation on even a fairly small subset of 5 experimental conditions of a continuous the

design space. Comparison of the competing methodologies is, thus, possible on a small subset of experimental conditions with the resulting estimates being, overall, in agreement and both concluding on the same optimal decision. The proposed variational estimator is subsequently incorporated in a Bayesian optimisation algorithm for an efficient search over the continuous design space. This procedure is further considered within the general, sequential design procedure. Thus, at each cycle of the study, Bayesian optimisation is employed for maximisation of the newly formed expected utility surface, observations from which are acquired through the proposed estimator. Although such a setup is, potentially, more welcoming to consideration of Monte-Carlo based methods due to the reduction in function evaluations achieved by Bayesian optimisation algorithms, it was found that, the number of predictive samples required for completion of the study within a realistic time frame produced particularly noisy estimates, hindering their convergence. It was, therefore, shown that, a trade-off between an acceptable computational time and optimality of the produced design is not always possible in such studies under the traditionally adopted Monte-Carlo based method. Lastly, optimal experimental design under alternative utility functions from the proposed class is explored in that chapter. Particularly, the examined choices result in alternative expressions of the expected utility in terms of $f-$divergences with emphasis being placed on the KL divergence and the Hellinger distance. The induced expected utilities, thus, inherit different properties — such as symmetricity, interpretability, boundedness — which may be exploited depending on the problem under study.

Lastly, application of the examined framework to a model discrimination problem in Spectroscopy is demonstrated in Chapter 9. This case study examines competing models, describing alternative behaviours in fluorescence kinetics. The complexity of such phenomena has previously hindered optimal experimental design efforts for their study, thus limiting the considered options to rather simplistic models that may not adequately capture the system behaviour. The proposed estimator is, thus, particularly well-suited to evaluation of the expected utility under such models as model evaluation is only required to produce samples from their corresponding predictive distributions. Solution of a convex optimisation problem, relying only on computation of the associated Gram matrix, is subsequently employed for estimation of the expected utility. The presented

problem examines the performance of the variational approximation method on a high dimensional output space, proving to scale up well without resulting in unreasonably long waiting times.

Summary

This thesis initiates with a review of the current work and ongoing challenges in optimal experimental design. Lack of methodologies for efficient estimation of the expected utility, particularly under computationally demanding models, poses a predominant concern. In particular, as modern science is progressively concerned with complex phenomena, the need to incorporate sophisticated models in optimal experimental design studies is imperative. A variational estimator allowing efficient and timely optimal decision making under such models is proposed in this thesis, paving the way to a novel experimental design framework. Application of the proposed framework to a range of problems is subsequently explored. To provide a comprehensive coverage of potential challenges the following setups were examined:

- discrimination problems employing more than 2 competing models (Chapter 7),
- continuous design spaces, on which grid search methods for optimisation are prohibitive (Chapter 8),
- increasingly complex and, thus, computationally demanding models (Chapters 8 and 9),
- experimental design problems concerned with optimisation of more than one experimental conditions (Chapter 9),
- high dimensional output space (Chapter 9).

The proposed framework is shown to perform particularly well under each examined setup, successfully addressing challenges that hinder currently considered methodologies. The presented work contributes towards the extension of optimal experimental design to studies of modern phenomena which currently poses an ongoing problem.

Appendix A

Proofs for Properties 1 and 2 of Monte-Carlo estimator

Property 1. *(Unbiased)*

$$\begin{aligned}
\mathbb{E}\left[\hat{I}\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\eta(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\frac{\eta(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\int\left[\frac{\eta(\boldsymbol{x})}{p(\boldsymbol{x})}\right]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{N}\sum_{i=1}^{N}\int\left[\frac{\eta(\boldsymbol{x})}{p(\boldsymbol{x})}\right]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{N}\sum_{i=1}^{N}\int\eta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = I
\end{aligned} \tag{A.1}$$

Property 2. *(Consistent)*

$$\begin{aligned}
\mathrm{Var}\left[\hat{I}\right] &= \mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\eta(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}\left[\frac{\eta(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}\right] \\
&= \frac{1}{N}\mathrm{Var}\left[\frac{\eta(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}\right] \xrightarrow{N\to\infty} 0\ .
\end{aligned} \tag{A.2}$$

Appendix B

Sequential Monte Carlo algorithm

Sequential Monte Carlo methods are sampling tools used to approximate a distribution of interest through a collection of weighted particles. The particles are successively moved and updated using importance sampling steps according to a sequence of intermediate distributions $\pi_1, \ldots, \pi_K$ that connect an initial, relatively easy to sample from, distribution to the possibly more challenging target distribution. The algorithm, therefore, consists of K steps and so at step $k = 1, \ldots, K$, the current population of weighted particles is updated to approximate distribution π_k through a combination of importance sampling and resampling moves. While transitioning towards the distribution of interest, intermediate distributions become increasingly difficult to sample from as they start to look less like the initial flat distribution and more like the target. However, choice of a dense enough sequence allows for smoother transitions between intermediate distributions as little change is observed from one stage to the next one. This discrepancy is accounted for by performing an importance sampling step after each transition. Typically, and throughout this thesis, the initial distribution is chosen to be the prior which sequentially leads to the target posterior distribution of a random vector of interest. In this thesis, the transition between consecutive distributions is regulated by a sequence $\beta = \{\beta_1, \ldots, \beta_K\}$ in the following way:

$$\pi_k = f(\boldsymbol{y} \mid \boldsymbol{\theta})^{\beta_k} p(\boldsymbol{\theta}) \ , \tag{B.1}$$

where $0 = \beta_1 < \beta_1 < \cdots < \beta_{K-1} < \beta_K = 1$. Under this setup, the contribution of the p.d.f. and therefore the information extracted from dataset $\boldsymbol{y}$ is accounted for gradually rather than updating the distribution of $\boldsymbol{\theta}$ to reflect observation of $\boldsymbol{y}$ at once.

The sampler proceeds through the following steps. It is assumed that at step k the current population of L particles $\{\boldsymbol{\theta}_{k-1}^1, \ldots, \boldsymbol{\theta}_{k-1}^L\}$ and corresponding weights $W_{k-1} = \{w_{k-1}^1, \ldots, w_{k-1}^L\}$ converges asymptotically to the intermediate distribution π_{k-1}. Particles are propagated towards distribution π_k through a kernel K and to account for the discrepancy between the updated population and the targeted distribution π_k an importance sampling step is performed, updating the corresponding weights. The Gaussian kernel centered around the current particle is a common choice in the literature. The corresponding weights are in that case updated according to:

$$w_k^i = \frac{\pi_k(\boldsymbol{\theta}_k)}{\sum_{i=1}^{L} w_{k-1}^i \mathcal{N}(\boldsymbol{\theta}_{k-1}, \sigma^2)} \,. \tag{B.2}$$

A potentially more efficient choice of kernel is proposed by Del Moral et al. (2006). In this case, each particle is propagated towards π_k through an MCMC kernel $\mathcal{K}_k(\boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_k)$. As calculation of the updated weights is, under this setup, deemed intractable, the authors introduce a sequence of backward Markov kernels $\mathcal{L}_k$ such that:

$$\pi_k(\boldsymbol{\theta}_{0:k}) = \pi_k(\boldsymbol{\theta}_k) \prod_{l=1}^{K} \mathcal{L}(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k-1}). \tag{B.3}$$

The weights are then updated by:

$$w_n = w_{k-1} \frac{\pi_k(\boldsymbol{\theta}_k)\mathcal{L}_k(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k-1})}{\pi_{k-1}(\boldsymbol{\theta}_{k-1})\mathcal{K}_k(\boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_k)}. \tag{B.4}$$

Guidance on the optimal choice of $\mathcal{L}_n$ is provided by Del Moral et al. (2006). It was found that choice of the suboptimal kernel:

$$\mathcal{L}_n = \frac{\pi_n(\boldsymbol{\theta}_{n-1})\mathcal{K}_n(\boldsymbol{\theta}_{n-1}, \boldsymbol{\theta}_n)}{\pi_n(\boldsymbol{\theta}_n)} \tag{B.5}$$

leads to convenient simplifications and is therefore preferred in this thesis. Under this

choice, (B.4) takes the form:

$$w_n = w_{n-1} \frac{\pi_n(\boldsymbol{\theta}_{n-1})}{\pi_{n-1}(\boldsymbol{\theta}_{n-1})}. \tag{B.6}$$

Equation (B.5) assumes that $\pi_{n-1} \approx \pi_n$. As π_n is often known only up to a normalising constant Z_n where $\pi_n = \frac{q_n}{Z_n}$, the un-normalised weights can be used instead:

$$w_n = w_{n-1} \frac{q_n(s_{n-1})}{q_{n-1}(s_{n-1})} \ . \tag{B.7}$$

For the particular choice of backward kernel, the weights W_n are independent of the obtained population as suggested by equation (B.7).

Degeneracy issues are often encountered after a number of steps in which case only a few particles are assigned very high weights and therefore represent the whole population. When such situations arise, an additional resampling step is performed during which particles are resampled proportionally to their weights and the new draws are assigned equal weights. The effective sample size (ESS; Kong et al. (1994)) is used as an indicator for population degeneracy where values below a predefined threshold (typically $L/2$) suggest that the variability of the population is below the minimum desired level. The SMC sampler is summarised in Algorithm 7.

Algorithm 7 SMC

1: **Initialisation**
Draw sample $(\boldsymbol{\theta}_0^1, \ldots, \boldsymbol{\theta}_0^L)$ from prior $p(\boldsymbol{\theta})$ and assign equal weights $1/L$.
2: **Sampling**
For $k = 1, \ldots, K$:
Update the weights, $w_k^i = w_{k-1}^i f(\boldsymbol{y} \mid \boldsymbol{\theta}_{k-1}^i)^{\beta_k - \beta_{k-1}}$.
Resample according to W_k, if ESS $< (L/2)$, and assign weights $1/L$ to new population.
Propagate $\boldsymbol{\theta}_k^i$ using $\boldsymbol{\theta}_k^i \sim \mathcal{K}_k(\boldsymbol{\theta}_{k-1}^i, \cdot)$.

Appendix C

Additional material for Chapter 5

C.1 Derivation of terms T_1^*, T_2^*, T_3^* in proof of Proposition 1

For $\mathbf{T^*_{1,k}(v_k)}$:

$$T^*_{1,k}(v_k) = \sup_z \{\langle z, u_k\rangle - T_{1,k}(z)\} = \sup_z \left\{\langle z, u_k\rangle + \frac{1}{N}\left\langle z, \Phi(\boldsymbol{y}_k^{P_m})\right\rangle\right\} \tag{C.1}$$

Evaluation of the supremum can be achieved by setting the sub-gradient of each term with respect to z equal to 0 and substitute the output back into the initial expression:

$$\begin{aligned}
&\frac{\partial\left[\langle z, u_k\rangle + \frac{1}{N}\left\langle z, \Phi(\boldsymbol{y}_k^{P_m})\right\rangle\right]}{\partial z} = 0 \\
\Rightarrow &u_k + \frac{1}{N}\Phi(\boldsymbol{y}_k^{P_m}) = 0 \\
\Rightarrow &u_k = -\frac{1}{N}\Phi(\boldsymbol{y}_k^{P_m}).
\end{aligned} \tag{C.2}$$

We can observe that for $u_k = -\frac{1}{N}\Phi(\boldsymbol{y}_k^{P_m})$:

$$T^*_{1,k}(u_k) = \sup_z \left\{-\left\langle z, \frac{1}{N}\Phi(\boldsymbol{y}_k^{P_m})\right\rangle + \frac{1}{N}\left\langle z, \Phi(y_k^{P_m})\right\rangle\right\} = 0$$

and $T^*_{1,k}(u_k) = +\infty$ otherwise.

For $\mathbf{T^*_{2,k}(u_k)}$:

$$T^*_{2,k}(u_k) = \sup_z \{\langle z, v_k\rangle - T_{2,k}(z)\} = \sup_z \left\{\langle z, v_k\rangle - \frac{1}{N}\varphi^*\left(\left\langle z, \Phi(\boldsymbol{y}_k^{P_{m'}})\right\rangle\right)\right\}. \quad \text{(C.3)}$$

Setting the sub-gradient equal to 0:

$$\begin{aligned} & \frac{\partial\left[\langle z, v_k\rangle - \frac{1}{N}\varphi^*\left(\left\langle z, \Phi(\boldsymbol{y}_k^{P_{m'}})\right\rangle\right)\right]}{\partial z} = 0 \\ \Rightarrow \quad & v_k - \frac{1}{N}\varphi^{*\prime}\left(\left\langle z, \Phi(\boldsymbol{y}_k^{P_{m'}})\right\rangle\right)\Phi(\boldsymbol{y}_k^{P_{m'}}) = 0 \\ \Rightarrow \quad & v_k = a_k\Phi(\boldsymbol{y}_k^{P_{m'}}) \ , \end{aligned} \quad \text{(C.4)}$$

where:

$$\begin{aligned} a_k &= \frac{1}{N}\varphi^{*\prime}\left(\left\langle z, \Phi(\boldsymbol{y}_k^{P_{m'}})\right\rangle\right) \\ \Rightarrow \quad \left\langle z, \Phi(\boldsymbol{y}_k^{P_{m'}})\right\rangle &= \Psi(Na_k), \end{aligned} \quad \text{(C.5)}$$

for $\Psi^{-1}(u) = \varphi^{*\prime}(u)$. From properties of the Fenchel transform for real-valued convex functions, it turns out that $\Psi(u) = \varphi'(u)$. Therefore, substituting (C.4) and (C.5) back to (C.3) results in:

$$\begin{aligned} T^*_{2,k}(v_k) &= \langle z, a_k\varphi'(\boldsymbol{y}_k^{P_{m'}})\rangle - \frac{1}{N}\phi^*\left(\varphi'(Na_k)\right) \\ &= a_k\varphi'(Na_k) - \frac{1}{N}\varphi^*\left(\varphi'(Na_k)\right) \ . \end{aligned} \quad \text{(C.6)}$$

If (C.4) does not hold, no maximum is achieved and therefore $T^*_{2,k}(v_k) = +\infty$.

For $\mathbf{T^*_3(u)}$:

$$T^*_3(u) = \sup_z \{\langle z, u\rangle - T_3(z)\} = \sup_z \left\{\langle z, u\rangle - \frac{\rho}{2}\|z\|^2_{\mathcal{H}}\right\} \ . \quad \text{(C.7)}$$

$$\frac{\partial\left[\langle z, u\rangle - \frac{\rho}{2}\|z\|^2_{\mathcal{H}}\right]}{\partial z} = 0$$

$$\Rightarrow \qquad u - \frac{\rho}{z} = 0$$
$$\Rightarrow \qquad z = \frac{u}{\rho}. \tag{C.8}$$

Substituting (C.8) into $F_3(u, z)$ produces the following expression:

$$\begin{aligned} T_3^*(u) &= \left\langle \frac{u}{\rho}, u \right\rangle - \frac{\rho}{2} \left\| \frac{u}{\rho} \right\|_{\mathcal{H}}^2 \\ &= \frac{1}{\rho} \langle u, u \rangle - \frac{1}{2\rho} \|u\|_{\mathcal{H}}^2 \\ &= \frac{1}{2\rho} \|u\|_{\mathcal{H}}^2 , \end{aligned} \tag{C.9}$$

since $\langle u, u \rangle = \|u\|_{\mathcal{H}}^2$.

C.2 Estimation of the Hellinger distance

Similarly to Section 5.4, transformation $\varphi_H(\chi(\boldsymbol{y})) = 2\sqrt{(\chi(\boldsymbol{y}))}$, $\chi > 0$ inducing the Hellinger distance between predictive distributions is considered. Under this definition, expression (5.12) obtains the form:

$$\begin{aligned} \tilde{\boldsymbol{a}} = \arg\inf_{\boldsymbol{a}} \Bigg\{ & - \sum_{k=1}^{N} \frac{2\sqrt{a_k}}{\sqrt{N}} + \frac{1}{2\rho} \sum_{k,l=1}^{N} a_k a_l K\left(\boldsymbol{y}_k^{P_{m'}}, \boldsymbol{y}_l^{P_{m'}}\right) \\ & - \frac{1}{N\rho} \sum_{k,l=1}^{N} a_l K\left(\boldsymbol{y}_k^{P_m}, \boldsymbol{y}_l^{P_{m'}}\right) + \frac{1}{2\rho N^2} \sum_{k,l=1}^{N} K\left(\boldsymbol{y}_k^{P_m}, \boldsymbol{y}_l^{P_m}\right) \Bigg\} . \end{aligned} \tag{C.10}$$

Having obtained vector $\tilde{\boldsymbol{a}}$, an estimate for the Hellinger distance between distributions corresponding to competing models m, m' can be evaluated as:

$$\hat{F}_{\varphi_H;\boldsymbol{\delta}}(m, m') = 1 - \sum_{k=1}^{N} \frac{\sqrt{\hat{a}_k}}{\sqrt{N}} . \tag{C.11}$$

Estimation of the Hellinger distance between predictive distributions for optimal experimental design is considered in Chapters 8 and 9.

References

Alghamdi, A., Vyshemirsky, V., Birch, D. J., and Rolinski, O. J. (2018). Detecting beta-amyloid aggregation from time-resolved emission spectra. *Methods and applications in fluorescence*, 6(2):024002. 119

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142. 15, 48

Amzal, B., Bois, F. Y., Parent, E., and Robert, C. P. (2006). Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical association*, 101(474):773–785. 33

Andrews, D. L. (2015). *Photonics, Volume 4: Biomedical Photonics, Spectroscopy, and Microscopy*. John Wiley & Sons. 119

Atkinson, A. C. (2008). Dt-optimum designs for model discrimination and parameter estimation. *Journal of Statistical planning and Inference*, 138(1):56–64. 19

Atkinson, A. C. and Donev, A. N. (1992). *Optimum experimental designs*. 24

Belghazi, M. I., Rajeswar, S., Baratin, A., Hjelm, D., and Courville, A. (2018). Mine: Mutual information neural estimation. 38, 43, 55

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media. 42

Bingham, D. and Chipman, H. (2002). Optimal designs for model selection. *Technometrics*, pages 1–20. 21

Birch, D. J. and McLoskey, D. (2017). Finding the time for fluorescence. its measurement and applications in life science. *Readout: HORIBA Technical Reports*, (49):13–20. 119

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer. 53

Blackwell, D. (1951). Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102, Berkeley, Calif. University of California Press. 15, 46, 47

Blackwell, D. (1953). Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272. 13, 14, 15

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877. 37, 38

Box, G. E. and Hill, W. J. (1967). Discrimination among mechanistic models. *Technometrics*, 9(1):57–71. 19, 75, 76, 78

Burges, C. J., Smola, A. J., and Scholkopf, B. (1999). Advances in kernel methods. *Support Vector Learning*. 53

Calder, M., Vyshemirsky, V., Gilbert, D., and Orton, R. (2006). Analysis of signalling pathways using continuous time markov chains. In *Transactions on Computational Systems Biology VI*, pages 44–67. Springer. 65

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, pages 273–304. 25

Chung, L. H. C., Birch, D. J., Vyshemirsky, V., Ryadnov, M. G., and Rolinski, O. J. (2019). Tracking insulin glycation in real time by time-resolved emission spectroscopy. *The Journal of Physical Chemistry B*, 123(37):7812–7817. 120

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314. 48

D'Argenio, D. Z. (1990). Incorporating prior parameter uncertainty in the design of sampling schedules for pharmacokinetic parameter estimation experiments. *Mathematical Biosciences*, 99(1):105–118. 25

DeGroot, M. (1970). Optimal statistical decisions. mcgraw-hil book company. *New York*. 14

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436. 28, 33, 140

Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2013). Sequential monte carlo for bayesian sequentially designed experiments for discrete data. *Computational Statistics & Data Analysis*, 57(1):320–335. 29, 84

Drovandi, C. C. and Pettitt, A. N. (2013). Bayesian experimental design for models with intractable likelihoods. *Biometrics*, 69(4):937–948. 1, 29

Dudley, R. (2002). Real analysis and probability , cambridge uni. *Press, Cambridge, UK*. 68

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607. 28, 84

Ghanem, R. G. and Spanos, P. D. (2003). *Stochastic finite elements: a spectral approach*. Courier Corporation. 34

Ginsbourger, D., Le Riche, R., and Carraro, L. (2008). A multi-points criterion for deterministic parallel global optimization based on gaussian processes. 68

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. (2007). A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520. 82

Grünewälder, S., Audibert, J.-Y., Opper, M., and Shawe-Taylor, J. (2010). Regret bounds for gaussian process bandit problems. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 273–280. 68

Huan, X. and Marzouk, Y. M. (2013). Simulation-based optimal bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 232(1):288–317. 34

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press. 11

Jeffreys, H. (1961). *Theory of Probability*. Oxford, Oxford, England, third edition. 20, 87, 99, 105, 114

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233. 37

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795. 20

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):272–304. 23

Kleinegesse, S. and Gutmann, M. U. (2019). Efficient bayesian experimental design for implicit models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 476–485. 35, 65

Klipp, E., Liebermeister, W., Wierling, C., and Kowald, A. (2016). *Systems biology: a textbook*. John Wiley & Sons. 1

Knowles, T. P., Vendruscolo, M., and Dobson, C. M. (2014). The amyloid state and its association with protein misfolding diseases. *Nature reviews Molecular cell biology*, 15(6):384–396. 120

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and bayesian missing data problems. *Journal of the American statistical association*, 89(425):278–288. 141

Kriete, A. and Eils, R. (2013). *Computational systems biology: from molecular mechanisms to disease*. Academic press. 94

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86. 21

Kuo, L., Soyer, R., and Wang, F. (1999). Bayesian statistics vi, chapter optimal design for quantal bioassay via monte carlo methods. 32

Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. 71

Lawrence, N. D., Girolami, M., Rattray, M., and Sanguinetti, G. (2010). *Learning and inference in computational systems biology*. MIT press. 95

Liepe, J., Taylor, H., Barnes, C. P., Huvet, M., Bugeon, L., Thorne, T., Lamb, J. R., Dallman, M. J., and Stumpf, M. P. (2012). Calibrating spatio-temporal models of leukocyte dynamics against in vivo live-imaging data using approximate bayesian computation. *Integrative biology*, 4(3):335–345. 94

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005. 18

Lindley, D. V. (1972). *Bayesian statistics, a review*, volume 2. SIAM. 17

Long, Q., Scavino, M., Tempone, R., and Wang, S. (2013). Fast estimation of expected information gains for bayesian experimental designs based on laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, 259:24–39. 1, 29

McGree, J., Drovandi, C. C., and Pettitt, A. N. (2012). A sequential monte carlo approach to the sequential design for discriminating between rival continuous data models. 28, 29

Mendes, P., Sha, W., and Ye, K. (2003). Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19(suppl_2):ii122–ii129. 98

Metropolis, N. and Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, 44(247):335–341. 25

Močkus, J. (1975). On bayesian methods for seeking the extremum. In *Optimization techniques IFIP technical conference*, pages 400–404. Springer. 35, 65

Močkus, J. (1989). Bayesian approach to global optimization: theory and applications. 65

Moreno, P. J., Ho, P. P., and Vasconcelos, N. (2004). A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Advances in neural information processing systems*, pages 1385–1392. 48

Müller, P. (1999). Simulation-based optimal design. 32, 33

Müller, P. and Parmigiani, G. (1995). Optimal design via curve fitting of monte carlo experiments. *Journal of the American Statistical Association*, 90(432):1322–1330. 31, 32

Neal, R. M. (2001). Annealed importance sampling. *Statistics and computing*, 11(2):125–139. 28

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861. 37, 38, 43, 49, 52, 59

Overstall, A. M., McGree, J. M., and Drovandi, C. C. (2017). An approach for finding fully bayesian optimal designs using normal-based approximations to loss functions. *Statistics and Computing*, pages 1–16. 1

Overstall, A. M. and Woods, D. C. (2016). Multivariate emulation of computer simulators: model selection and diagnostics with application to a humanitarian relief model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(4):483–505. 65

Overstall, A. M. and Woods, D. C. (2017). Bayesian design of experiments using approximate coordinate exchange. *Technometrics*, 59(4):458–470. 32

Overstall, A. M., Woods, D. C., and Martin, K. J. (2019). Bayesian prediction for physical models with application to the optimization of the synthesis of pharmaceutical products using chemical kinetics. *Computational Statistics & Data Analysis*, 132:126–142. 1, 16

Pickup, J. C., Hussain, F., Evans, N. D., Rolinski, O. J., and Birch, D. J. (2005). Fluorescence-based glucose sensors. *Biosensors and Bioelectronics*, 20(12):2555–2565. 1

Poole, B., Ozair, S., Oord, A. v. d., Alemi, A. A., and Tucker, G. (2019). On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*. 133, 134

Pronzato, L. and Walter, É. (1985). Robust experiment design via stochastic approximation. *Mathematical Biosciences*, 75(1):103–120. 25

Pukelsheim, F. (1980). On linear regression designs which maximize information. *Journal of Statistical Planning and Inference*, 4(4):339–364. 23

Rasmussen, C. E. and Williams, C. K. (2006). Gaussian processes for machine learning the mit press. *Cambridge, MA*. 32, 68

Ricklefs, R. E., Schluter, D., et al. (1993). Species diversity in ecological communities: historical and geographical perspectives. 1

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media. 11

Rockafellar, R. T. (1970). *Convex analysis*. Number 28. Princeton university press. 38, 39, 42

Rockafellar, R. T. (1974). *Conjugate duality and optimization*, volume 16. Siam. 41, 42

Rolinski, O. J., Amaro, M., and Birch, D. J. (2010). Early detection of amyloid aggregation using intrinsic fluorescence. *Biosensors and Bioelectronics*, 25(10):2249–2252. 119

Roth, P. M. (1967). Design of experiments for discrimination among rival models. 20

Ruderman, A., Reid, M. D., García-García, D., and Petterson, J. (2012). Tighter variational representations of f-divergences via restriction to probability measures. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1155–1162. 37, 38, 51, 52, 82

Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2016). A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, 84(1):128–154. 2, 26, 29, 30, 33, 133

Ryan, E. G., Drovandi, C. C., Thompson, M. H., and Pettitt, A. N. (2014). Towards bayesian experimental design for nonlinear models that require a large number of sampling times. *Computational Statistics & Data Analysis*, 70:45–60. 1, 20, 27

Ryan, K. J. (2003). 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. 1, 16, 25, 27, 59

Schlaifer, R. and Raiffa, H. (1961). *Applied statistical decision theory*. 17

Silk, D., Kirk, P. D., Barnes, C. P., Toni, T., and Stumpf, M. P. (2014). Model selection in systems biology depends on experimental design. *PLoS Comput Biol*, 10(6):e1003650. 99

Silvey, S. D. (1980). *Optimal Design An Introduction to the Theory for Parameter Estimation*. Springer. 24

Sobol', I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 7(4):784–802. 112

Song, J. and Ermon, S. (2019). Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*. 38, 133, 134

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of*

the 27th International Conference on International Conference on Machine Learning, pages 1015–1022. 72

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press. 47, 133

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574. 38

Vanlier, J., Tiemann, C. A., Hilbers, P. A., and van Riel, N. A. (2014). Optimal experiment design for model selection in biochemical networks. *BMC systems biology*, 8(1):20. 21

von Kügelgen, J., Rubenstein, P. K., Schölkopf, B., and Weller, A. (2019). Optimal experimental design via bayesian optimization: active causal structure learning for gaussian process networks. *arXiv preprint arXiv:1910.03962*. 35, 65

Vyshemirsky, V. and Girolami, M. A. (2008). Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839. 28, 84, 94, 95, 98

Wald, A. (1950). Statistical decision functions. 13, 17, 23

Wilkinson, D. J. (2007). Bayesian methods in bioinformatics and computational systems biology. *Briefings in bioinformatics*, 8(2):109–116. 94