Parametric Model Uncertainty in Bayesian Optimal Experimental Design

Candidate Number: 1035161

University of Oxford

A thesis submitted for the degree of

MMath Mathematics and Statistics

Trinity 2022

Contents

Chapter 1

Introduction

1.1 A Motivating Example: Optimal Experimental Design in Drug Trials

We start with a short example, taken from the field of toxicology [1], which aims to illustrate some of the main motivations at the core of Bayesian Optimal Experimental Design (BOED).

Suppose that a lab is investigating the effect of a drug's dosage on a chronic condition. Only 40 rats affected by such condition are available in the lab.
If a rat is administered a dosage ξ, then the probability that it will continue to be sick can be modelled as:

$$p(\xi;\ \theta) = \frac{1}{1 + exp(\gamma - \rho\,(\xi - \theta)^2)} \tag{1.1}$$

$\gamma = 3.1$ and $\rho = 2.9$ are known parameters, while $\theta \sim N(3.5,\, 0.75)$ is an unknown parameter which represents the optimal dosage. Intuitively, if the dosage is too low, the rat will continue to show symptoms of the chronic condition, while if it is too high, the rat will develop some other serious complications.

The team divides the rats into 4 groups of fixed size 10 and is now facing the question of which dosage $\{\xi_i\}_{i=1}^4$ to administer to each group, given the aim of gaining as much information as possible on the optimal dosage θ.

Formally, given θ and dosage ξ_i, the number of rats y_i which continue to be sick after drug administration is distributed as:

$$y_i|\,\xi_i,\,\theta\ \sim Binomial\,(10,\,p(\xi_i;\,\theta)) \tag{1.2}$$

Since, according to the prior, $\theta \in [2.0, 5.0]$ with an approximate probability of 95%, the toxicologists choose to start by administering a dose $\xi_1 = 2.0$ to the first group, and then to sequentially increase the dose by one unit for each of the following groups, leading to experimental designs $(\xi_1, \xi_2, \xi_3, \xi_4) = (2.0, 3.0, 4.0, 5.0)$.

On the other hand, at each time step i, the BOED framework leverages the data from previous experiments $\{\xi_j, y_j\}_{j=1}^{i-1}$ to make a more informative choice of ξ_i.
Details on how ξ_i is chosen and what it means to be 'more informative' are being intentionally omitted here (they are covered in detail in Chapter 2); nevertheless Table 1.1 is sufficient to give us an intuitive understanding of what is going on:
the algorithm observes a low number of sick mice at $\xi_1 = 5.0$, so it learns that θ must be around that value, it therefore proceeds to test dosages in that range with $\xi_2 = 5.2$ and $\xi_3 = 5.4$; when it observes a high number of sick mice $y_3 = 8$, it learns that θ is likely to be lower than 5.0 and picks $\xi_4 = 3.3$ to gain a data point in that area.

Approach	$\xi_1 \to y_1$	$\xi_2 \to y_2$	$\xi_3 \to y_3$	$\xi_4 \to y_4$
Equally spaced ξ_is	$\mathbf{2.0} \to 10$	$\mathbf{3.0} \to 8$	$\mathbf{4.0} \to 0$	$\mathbf{5.0} \to 3$
BOED	$\mathbf{5.0} \to 3$	$\mathbf{5.2} \to 2$	$\mathbf{5.4} \to 8$	$\mathbf{3.3} \to 4$

Table 1.1: Experimental designs and corresponding observed values for both the equally spaced dosages and BOED approaches.

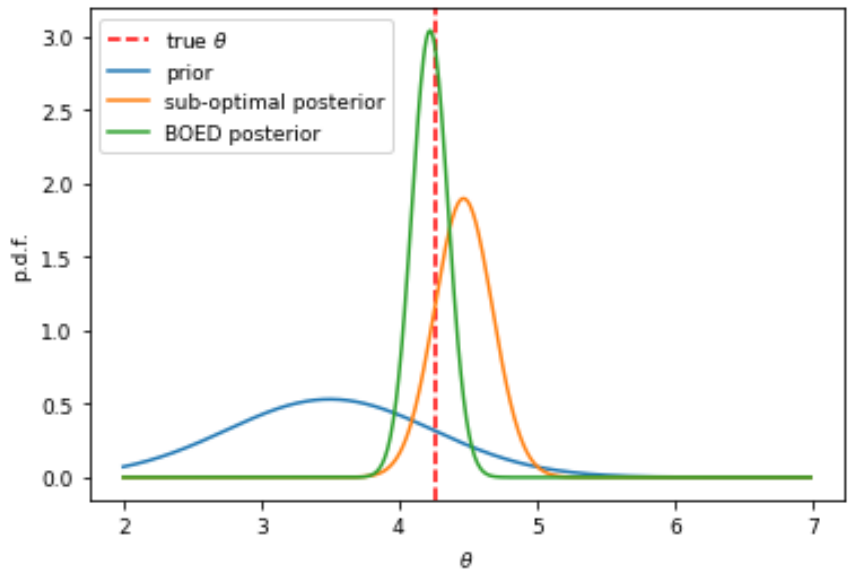


Figure 1.1: Approximate posteriors for the 'equally spaced ξ_is' and BOED approaches.

Once the full experimental history $h \coloneqq \{(\xi_i, y_i)\}_{i=1}^4$ is observed, we proceed to calculate an approximate Gaussian posterior $p(\theta\,|h)$ through variational inference.
BOED gives us a much narrower posterior with $p(\theta\,|h) \sim N(4.22, 0.13)$, whose mean is very close to the true optimal dosage $\theta = 4.27$.
As can be seen in Figure 1.1, the cost for sub-optimal choices of ξ_i is being paid by the other approach with a higher-variance posterior $p(\theta\,|h) \sim N(4.47, 0.21)$.

1.2 Aims of the Thesis

Now that this motivating example is concluded and some of the main ideas in BOED have been introduced, we can briefly highlight some of the questions we aim to explore in this thesis[1].

In Equation (1.1) it can be seen that parameters γ and ρ are assumed to be known. We can certainly imagine scenarios in which this is not the case and, analogously to the prior we placed on θ, we have a prior on these parameters.
This scenario is often tackled by treating such parameters - generally called *nuisance variables* - as part of the target variable, leading us to optimise for the information we gain on $\Theta \coloneqq (\theta, \gamma, \rho)$. It is clear to see how this approach may lead to sub-optimal experimental designs by choosing ξ that give us a lot of information on (γ, ρ) - which we are not interested in - but little information on θ.
We propose alternatives to this technique in the context of Sequential BOED and explore their performance via a number of computational experiments.
Moreover, we analyse two recent advanced approaches based on Deep Learning and prove how one of them can be effortlessly extended to deal with experimental settings where nuisance variables are present.

[1]The word count for the thesis is 6888 words (7359 with the Appendix). Such numbers were produced using the LaTeX editor Kile.

Chapter 2

Formalism of Bayesian Optimal Experimental Design

2.1 Theoretical set-up

The theoretical framework of Bayesian Optimal Experimental Design (BOED) is relatively straight-forward.

We have a latent variable θ we wish to learn about, on which we have a prior $p(\theta)$.
We have a set of experiments Ξ available and we aim to choose the experiment $\xi \in \Xi$ which is the most informative on θ.
Each experiment ξ will produce an outcome y of which we know the conditional distribution $p(y \,|\, \xi, \theta)$ - which we call *model*.

The notion of how "informative" an experiment ξ is, can be formalised by the expected information gain in θ due to the experiment-outcome pair (ξ, y):

$$EIG(\xi) \;\coloneqq\; \mathbb{E}_{p(y \,|\, \xi)} \left[H\left(p(\theta)\right) - H\left(p(\theta \,|\, y, \xi)\right) \right] \tag{2.1}$$

where $H(\cdot)$ is the Shannon entropy [2].

Equation (2.1) can be further simplified to:

$$\begin{aligned} EIG(\xi) &= \mathbb{E}_{p(y,\,\theta\,|\,\xi)}\left[\, log\left(\frac{p(y\,|\,\theta,\xi)}{p(y\,|\,\xi)}\right)\right] && (i) \\ &= \mathbb{E}_{p(y,\,\theta\,|\,\xi)}\left[\, log\left(\frac{p(y,\theta\,|\,\xi)}{p(\theta)\,p(y\,|\,\xi)}\right)\right] && (ii) \\ &= \mathbb{E}_{p(y,\,\theta\,|\,\xi)}\left[\, log\left(\frac{p(\theta\,|\,y,\,\xi)}{p(\theta)}\right)\right] && (iii) \end{aligned} \tag{2.2}$$

For completeness, the proof for (2.1) $\implies$ (i) is given in the Appendix.
$(i) \Rightarrow (ii) \Rightarrow (iii)$ follows immediately from Bayes rule and the fact that $p(\theta) = p(\theta|\xi)$.

2.2 Estimators for the expected information gain

The different expressions in (2.2) give rise to different approaches to estimate $EIG(\xi)$.

Looking at (i), we can see that it requires the closed form of $p(y\,|\,\xi)$ which we can only compute by integrating over θ to get $p(y\,|\,\xi) = \mathbb{E}_{p(\theta)}\left[p(y\,|\,\theta,\xi)\right]$.
A very natural approach to approximate $EIG(\xi)$ is therefore the following Nested Montecarlo Estimator:

$$EIG(\xi) \approx \hat{\mu}_{NMC} \coloneqq \frac{1}{N}\sum_{n=1}^{N} log\left(\frac{p(y_n\,|\,\theta_{n,0},\,\xi)}{\frac{1}{M}\sum_{m=1}^{M} p(y_n\,|\,\theta_{n,m},\,\xi)}\right) \tag{2.3}$$

$$\text{where } \theta_{n,m} \overset{i.i.d.}{\sim} p(\theta)\,,\; y_n \sim p(y\,|\,\theta = \theta_{n,0},\,\xi)$$

Rainforth et al. [3] showed that, given total number of samples $T = O(N \cdot M)$, if N and M are chosen such that $M \propto \sqrt{N}$, then $\hat{\mu}_{NMC}$ converges to $EIG(\xi)$ with a rate of $O(T^{-2/3})$ in the mean squared error.

In order to get around this slow convergence rate, Foster et al. [4] drew from the field of variational inference [5] [6] to propose a series of estimators which are all based on the following key idea:
Rather than estimating the inner expectation on a point-by-point basis (for example in (2.3) you have to estimate $p(y_n\,|\,\xi)$ afresh for each outer sample), the aim is to learn a functional approximation of the density of interest. Once this is learnt, standard Montecarlo can be applied, generally leading to faster overall convergence.

We now briefly review three of the variational estimators proposed which will also be used in Chapter 4 for our computational experiments.

2.2.1 Variational Posterior and Variational Marginal Estimators

In (2.2) (iii) the density $p(\theta \,|\, y, \xi)$ is intractable. We can use variational inference with a family of densities $q_\phi(\theta \,|\, y, \xi)$ parametrised by ϕ to approximate it. Consequently, we can obtain the following expression for the EIG:

$$EIG(\xi) \approx \mathcal{L}_{post}(\xi) \coloneqq \mathbb{E}_{p(y,\theta \,|\, \xi)} \left[log \left(\frac{q_\phi(\theta \,|\, y, \xi)}{p(\theta)} \right) \right] \tag{2.4}$$

A key property of this approximation is that it is actually a lower bound for $EIG(\xi)$ (the proof consists of expressing $EIG(\xi) - \mathcal{L}_{post}(\xi)$ as a Kullback-Leibler divergence). This feature allows us to use Stochastic Gradient Ascent-based methods [7] to find the optimal posterior $q_{\phi^*}(\cdot \,|\, \cdot, \xi)$ within the variational family of distributions. We can then use standard Montecarlo to produce the following estimator:

$$EIG(\xi) \approx \hat{\mu}_{post}(\xi) \coloneqq \frac{1}{N} \sum_{n=1}^{N} log \left(\frac{q_{\phi^*}(\theta_n \,|\, y_n, \xi)}{p(\theta_n)} \right) \tag{2.5}$$

$$\text{where } (\theta_n,\, y_n) \overset{i.i.d.}{\sim} p(y,\, \theta \,|\, \xi)$$

Two main drawbacks of the Variational Posterior are that the bound $EIG(\xi) \geq \mathcal{L}_{post}(\xi)$ is tight if and only if $q_\phi(\theta \,|\, y, \xi) = p(\theta \,|\, y, \xi)$, therefore if the variational family fails to approximate well the true shape of the posterior, our estimates for the expected information gain will be considerably downward biased; secondly θ is often higher dimensional than y, making variational inference significantly harder.

Motivated by this second concern, Foster et al. [4] also proposed a variational bound where q_ϕ is a function of y alone, called Variational Marginal.
This estimator follows from doing variational inference on $p(y \,|\, \xi)$ in expression (2.2) (i) resulting in:

$$EIG(\xi) \approx \mathcal{U}_{marg}(\xi) \coloneqq \mathbb{E}_{p(y,\theta \,|\, \xi)} \left[log \left(\frac{p(y \,|\, \theta, \xi)}{q_\phi(y \,|\, \xi)} \right) \right] \tag{2.6}$$

Analogously to what done with the Variational Posterior Estimator, it can be proven that $\mathcal{U}_{marg}(\xi) \geq EIG(\xi)$, leading to the following Montecarlo estimator with ϕ^* found by Stochastic Gradient Descent:

$$EIG(\xi) \approx \hat{\mu}_{marg}(\xi) \coloneqq \frac{1}{N} \sum_{n=1}^{N} log \left(\frac{p(y_n \,|\, \theta_n,\, \xi)}{q_{\phi^*}(y_n \,|\, \xi)} \right) \tag{2.7}$$

$$\text{where } (\theta_n,\, y_n) \overset{i.i.d.}{\sim} p(y,\, \theta \,|\, \xi)$$

2.2.2 Variational Nested Montecarlo

Despite the computational benefits of $\hat{\mu}_{post}$ and $\hat{\mu}_{marg}$, and the fact that they can provide approximate lower and upper bounds for the true EIG, the biasedness of the two estimators is still a considerable issue.
A trade-off between the speed of variational approaches and the unbiasedness of the Nested Montecarlo estimator is Variational Nested Montecarlo.

The underlying idea is to learn a proposal $q_\phi(\theta \,|\, y, \xi)$ and then use it to do importance sampling to estimate the intractable expectation $p(y \,|\, \xi) = \mathbb{E}_{p(\theta)} \left[p(y\, \theta,\, |\, \xi) \right]$.

$$\begin{aligned} p(y \,|\, \xi) &= \int_\theta p(y\,, \theta \,|\, \xi)\, d\theta \\ &= \int_\theta q_\phi(\theta \,|\, y,\, \xi) \frac{p(y\,, \theta \,|\, \xi)}{q_\phi(\theta \,|\, y,\, \xi)}\, d\theta \\ &= \mathbb{E}_{q_\phi(\theta \,|\, y, \xi)} \left[\frac{p(y\,, \theta \,|\, \xi)}{q_\phi(\theta \,|\, y,\, \xi)} \right] \end{aligned} \tag{2.8}$$

We approximate (2.2) (i) using L samples from (2.8) to get the following upper bound:

$$EIG(\xi) \leq \mathcal{U}_{VNMC}(\xi,\, L) \coloneqq \mathbb{E} \left[log \left(p(y \,|\, \theta_0, \xi) \right) - log \left(\frac{1}{L} \sum_{l=1}^{L} \frac{p(y,\, \theta_l \,|\, \xi)}{q_\phi(\theta_l \,|\, y, \xi)} \right) \right] \tag{2.9}$$

where the expectation is taken w.r.t. $(y,\, \theta_{0:L}) \sim p(y,\, \theta_0 \,|\, \xi) \prod_{l=1}^{L} q_\phi(\theta_l \,|\, y,\, \xi)$.

There are two key improvements that Variational Nested Montecarlo brings to the table:

$$lim_{L \to +\infty} \mathcal{U}_{VNMC}(\xi,\, L) = EIG(\xi) \qquad \forall\, distribution\ q_\phi(\theta \,|\, y,\, \xi) \tag{2.10}$$

This means that even when the variational family $q_\phi(\theta|\, y\,,\, \xi)$ does not contain the true distribution $p(\theta\,|\,y,\,\xi)$, the upper-bound can be made arbitrarily tight by increasing the number of inner samples L.

$$\mathcal{U}_{VNMC}(\xi,\, L) = EIG(\xi) \qquad if\ \ q_\phi(\theta|\, y\,,\, \xi) = p(\theta\,|\,y,\,\xi) \tag{2.11}$$

In practice, this second property means that if $q_\phi(\theta\,|\,y\,,\,\xi)$ approximates well $p(\theta\,|\,y,\,\xi)$, then there is no need to have the inner number of samples $M \propto \sqrt{N}$, (where N is the outer number of samples) as in the standard Nested Montecarlo case. This makes it possible to have remarkably tight upperbounds $\mathcal{U}_{VNMC}(\xi,\, L)$ even for moderate values of L.
$EIG(\xi)$ can therefore be estimated as:

$$\hat{\mu}_{VNMC}(\xi) \coloneqq \frac{1}{N}\sum_{n=1}^{N}\left[log\left(p(y_n\,|\,\theta_{n,0},\,\xi) - log\left(\frac{1}{M}\sum_{m=1}^{M}\frac{p(y_n,\,\theta_{n,m},\,|\,\xi)}{q_\phi(\theta_{n,m}\,|\,y_n,\,\xi)}\right)\right]\right. \tag{2.12}$$

$$\text{where } \theta_{n,0} \overset{i.i.d.}{\sim} p(\theta)\,,\ y_n \sim p(y\,|\,\theta=\theta_{n,0},\,\xi)\,,\ \theta_{n,m} \sim q_\phi(\theta\,|\,y=y_n,\,\xi)$$

A further advance was made in a second paper by Foster et al. [8] where they proved that if in (2.9) we add the outer sample θ_0 in the inner approximation for $p(y\,|\,\xi)$, then $\mathcal{U}_{VNMC}(\xi,\, L)$ becomes a lower bound:

$$EIG(\xi) \geq \mathcal{L}_{VNMC}(\xi,\, L) \coloneqq \mathbb{E}\left[log\left(\,p(y\,|\,\theta_0,\xi)\,\right) - log\,\left(\frac{1}{L+1}\sum_{l=0}^{L}\frac{p(y,\,\theta_l\,|\,\xi)}{q_\phi(\theta_l\,|\,y,\xi)}\right)\right] \tag{2.13}$$

where the expectation is taken w.r.t. $(y,\,\theta_{0:L}) \sim p(y,\,\theta_0\,|\,\xi)\prod_{l=1}^{L} q_\phi(\theta_l\,|\,y,\,\xi)$.

The main idea of the proof is again to express $EIG(\xi) - \mathcal{L}_{VNMC}(\xi,\, L)$ as a KL divergence.

2.3 High-dimensional continuous design spaces

So far, we have been trying to estimate $EIG(\xi)$ separately for each $\xi \in \Xi$, to then choose the experimental design with the highest expected information gain $\xi_{opt} \coloneqq arg\,max_{\,\xi\in\Xi}\, EIG(\xi)$.

This becomes problematic when the design space Ξ is continuous and high-dimensional.

If we take a closer look at (2.12) for example, we can see that the variational family used has functions of the type $q_\phi(\,\cdot\,|\,\cdot\,,\,\xi) \,:\, \Theta \,\times\, \mathcal{Y} \to \mathbb{R}^+$. This means that we have to learn a separate functional approximation of $p(\theta\,|\,y,\,\xi)$ for each ξ.
We can instead work with a variational family that is shared across all experimental designs $q_\phi(\,\cdot\,|\,\cdot\,,\,\cdot) \,:\, \Theta \,\times\, \mathcal{Y} \,\times\, \Xi \to \mathbb{R}^+$.
If this is used together with a lower-bound on the information gain such as $\mathcal{L}_{VNMC}$ in (2.23), it allows us to take care of the outer optimisation problem (i.e. the choice of ξ_{opt}) by working with the joint objective $\mathcal{L}_{VNMC}(\phi,\,\xi)$ and use gradient ascent to find at the same time ϕ^* and ξ_{opt}, avoiding the need for a two-step process [8].

2.4 Iterated Experimental Design

The motivating example in Chapter 1 briefly highlighted the power of BOED when we are dealing with a sequence of experiments.
The main aims of this section are to formalise the assumptions we are making on the experimental process, to illustrate how the estimators for the EIG introduced in Section 2.2 can be extended to settings with sequential experiments, and finally to briefly outline some of the limitations of this approach.

Suppose we have resources to carry out T experiments. We work under the assumption that experimental outcomes are independent conditioned on the choice of experimental design and the latent variable:

$$\{\, y_t \,|\, \xi_t, \theta \,\}_{t=1}^{T} \text{ is a set of independent random variables} \qquad (***)$$

Let h_t represent the experimental history up to time t, that is $h_t \coloneqq \{\, (\xi_i, y_i) \,\}_{i=1}^{t}$.
At each time step t we aim to choose the experimental design ξ_t^{opt} that maximises the expected information gain given the experimental history h_{t-1} already available:

$$\xi_t^{opt} \coloneqq arg\,max\; I_{h_{t-1}}(\xi_t) \tag{2.14}$$

$$\text{where } I_{h_{t-1}}(\xi_t) \;\coloneqq\; \mathbb{E}_{p(y_t\,|\,h_{t-1},\xi_t)}\left[H\left(p(\theta\,|\,h_{t-1})\right) - H\left(p(\theta\,|\,h_t)\right)\right] \tag{2.15}$$

Analogous estimators to those from Section 2.2 can be derived in the case of sequential experimental design. However, we now have to both sample from an intractable

density, and evaluate its closed form.
Let's look for example at the Variational Posterior lower bound, which becomes:

$$\mathcal{L}_{post}(\xi_t) = \mathbb{E}_{p(y_t,\, \theta \,|\, \xi_t,\, h_{t-1})} \left[log \left(\frac{q_{\phi^*}(\theta \,|\, y_t,\, \xi_t)}{p(\theta \,|\, h_{t-1})} \right) \right] \tag{2.16}$$

Both $p(y_t,\, \theta_t \,|\, \xi,\, h_{t-1})$ and $p(\theta \,|\, h_{t-1})$ are not known.
Using Bayes' theorem and then conditional independence of the experimental outcomes $(***)$ we can rewrite $p(\theta \,|\, h_{t-1})$ as:

$$\begin{aligned} p(\theta \,|\, h_{t-1}) &\coloneqq p(\theta \,|\, y_{1:t-1},\, \xi_{1:t-1}) \\ &= \frac{p(\theta)\, p(y_{1:t-1} \,|\, \xi_{1:t-1},\, \theta)}{p(y_{1:t-1} \,|\, \xi_{1:t-1})} \\ &= p(\theta) \prod_{i=1}^{t-1} p(y_i \,|\, \theta,\, \xi_i) \,/\, p(y_{1:t-1} \,|\, \xi_{1:t-1}) \\ &\propto p(\theta) \prod_{i=1}^{t-1} p(y_i \,|\, \theta,\, \xi_i) \end{aligned} \tag{2.17}$$

Similarly, using again conditional independence:

$$p(y_t,\, \theta \,|\, \xi_t,\, h_{t-1}) = p(\theta \,|\, h_{t-1})\, p(y_t \,|\, \theta,\, \xi_t) \tag{2.18}$$

By plugging (2.17)-(2.18) into (2.16), we obtain:

$$\mathcal{L}_{post}(\xi_t) = \mathbb{E}_{p(\theta \,|\, h_{t-1})\, p(y_t \,|\, \theta, \xi_t)} \left[log \left(\frac{q_{\phi^*}(\theta \,|\, y_t,\, \xi_t)}{p(\theta) \prod_{i=1}^{t-1} p(y_i \,|\, \theta,\, \xi_i)} \right) \right] + C \tag{2.19}$$

Crucially, $C = log(\, p(\, y_{1:t-1} \,|\, \xi_{1:t-1}))$ is independent of (ξ_t, ϕ), therefore it does not affect neither the SGA-based procedure that produces the estimate ϕ^*, nor the choice of ξ_t^{opt}.

However, the expectation is taken with respect to the distribution $p(\theta \,|\, h_{t-1})\, p(y_t \,|\, \theta,\, \xi_t)$ where $p(\theta \,|\, h_{t-1})$ is still unknown.
This is generally tackled by variational inference, leading to the following two-step iterative procedure:
For each time t we first find a variational approximation for $p(\theta \,|\, h_{t-1}) \approx q_\psi(\theta)$, we

then proceed to estimate ξ_t^{opt} using some Montecarlo estimator:

$$\mathcal{L}_{post}(\xi_t) \approx \mathbb{E}_{q_{\psi^*}(\theta \mid h_{t-1})\, p(y_t \mid \theta, \xi_t)} \left[log \left(\frac{q_{\phi^*}(\theta \mid y_t,\, \xi_t)}{p(\theta) \prod_{i=1}^{t-1} p(y_i \mid \theta,\, \xi_i)} \right) \right] + C \tag{2.20}$$

The limitation of Iterated Experimental Design is that it myopically picks the design ξ_t^{opt} which maximises the intermediate information gain $I_{h_{t-1}}(\xi_t)$ without taking into account the number of remaining experiments.
This greedy strategy can as a consequence lead to sub-optimal design choices; an alternative is presented in Chapter 5.

Chapter 3

Parametric Model Uncertainty

So far we have worked under the assumption that we have an exact closed form of the *model* $p(y \mid \xi, \theta)$, from which we can sample and which we can evaluate inside our estimators.
In real-world experimental settings this is not always the case. For example, one might only be able to sample from $p(y \mid \xi, \theta)$ but not to evaluate its density, or there might be uncertainty regarding the type of function that models the experiment-outcome relation.

The current BOED framework would exacerbate the latter issue by forcing the experimentalist to choose a specific model for $p(y \mid \xi, \theta)$ beforehand and thus leaving the experimental design choices exposed to the risk of being optimised with respect to a wrong model.
Model misspecification is indeed a non-trivial hole within the Bayesian Optimal Experimental Design literature.

We aim to take a more narrow, but still somewhat flexible view on model misspecification, by focussing on cases where the uncertainty on $p(y \mid \xi, \theta)$ can be expressed in parametric form.

For example, let's take a look back at the toxicology example from Chapter 1, which had the following model:

$$y \mid \xi, \theta \sim Binomial\,(10, p(\xi, \theta)) \tag{3.1}$$

$$\text{with} \quad p(\xi, \theta) = \frac{1}{1 + exp(\gamma - \rho\,(\xi - \theta)^2)} \tag{3.2}$$

In Chapter 1, γ and ρ were assumed to be known fixed parameters. We can see that, in this case, the model depends exclusively on θ and ξ and hence falls within the class analysed in Chapter 2.

However, we could very naturally have some uncertainty about the parameters $(\gamma,\, \rho)$ which leads to the true model belonging to a wider family of distributions $p(y \,|\, \xi,\, \theta) \in \{\, p(y \,|\, \xi,\, \theta,\, \alpha) \,:\, \alpha \in A \,\}$ where in this case $\alpha = (\gamma,\, \rho)$ and A is the set of admissible parameters α. This is exactly the case which we aim to investigate.

3.1 Theoretical framework

Similarly to Section 2.4, we assume we have resources to carry out T experiments. Instead of having an initial prior $p(\theta)$, we have an initial joint prior on both the latent variable of interest θ and the nuisance parameter α.
In this case, the distribution of $y \,|\, \xi,\, \theta$ is not known, but if we also condition on the nuisance variable α, we assume we can both sample from and evaluate this density:

$$\text{Known distributions:} \qquad p(\theta,\, \alpha) \qquad p(y \,|\, \xi,\, \theta,\, \alpha) \tag{3.3}$$

Analogously to the Iterated Experimental Design framework of Section 2.4, at each time step t we aim to maximise the expected information gain given the available experimental history.

$$\xi_t^{opt} \coloneqq arg\,max\; I_{h_{t-1}}\,(\xi_t) \tag{3.4}$$

$$\text{where } I_{h_{t-1}}\,(\xi_t) \;\coloneqq\; \mathbb{E}_{p(y \,|\, h_{t-1},\, \xi_t)}\left[H\left(p(\theta \,|\, h_{t-1})\right) - H\left(p(\theta \,|\, h_t)\right)\right] \tag{3.5}$$

However, writing the difference of entropies inside expectation (3.5) directly in an amenable form such as in (2.2) - which allows us to either do a 2-level nested Monte-carlo or variational inference with a 1-level Montecarlo - is not feasible.
Nevertheless, we can still manage to derive similar types of estimators and to obtain the same asymptotic complexities by taking the following approach:

$$I_{h_{t-1}}\,(\xi_t) = \int_{\alpha} p(\alpha \,|\, h_{t-1})\, I_{h_{t-1}}(\xi_t \,|\, \alpha)\, d\alpha \tag{3.6}$$

$$\text{where } I_{h_{t-1}}(\xi_t \,|\, \alpha) \;\coloneqq\; \mathbb{E}_{p(y \,|\, h_{t-1},\, \xi_t,\, \alpha)}\left[H(p(\theta \,|\, h_{t-1},\, \alpha)) - H\left(p(\theta \,|\, h_t,\, \alpha)\right)\right] \tag{3.7}$$

Crucially, with α and h_{t-1} fixed, simpler expressions for $I_{h_{t-1}}(\xi_t \,|\, \alpha)$ can be derived and a lot of the techniques introduced in Section 2 can be recycled to give useful bounds and estimators for $I_{h_{t-1}}(\xi_t)$.

3.2 Estimators

Let's start for example by trying to obtain a bound on $I_{h_{t-1}}(\xi_t)$ analogous to the Variational Posterior from 2.2.1.

Following the steps of Proof A.1 (see Appendix), we can obtain expressions for $I_{h_{t-1}}(\xi_t \,|\, \alpha)$ that remind those for $EIG(\xi)$ in (2.2):

$$
\begin{aligned}
I_{h_{t-1}}(\xi_t \,|\, \alpha) &= \mathbb{E}_{p(y_t, \theta \,|\, \alpha, \xi_t, h_{t-1})} \left[log \left(\frac{p(y_t \,|\, \theta,\, \alpha,\, \xi_t)}{p(y_t \,|\, \alpha,\, \xi_t,\, h_{t-1})} \right) \right] && (i) \\
&= \mathbb{E}_{p(y_t, \theta \,|\, \alpha, \xi_t, h_{t-1})} \left[log \left(\frac{p(y_t,\, \theta \,|\, \alpha,\, \xi_t,\, h_{t-1})}{p(\theta \,|\, \alpha,\, h_{t-1})\, p(y_t \,|\, \alpha,\, \xi_t,\, h_{t-1})} \right) \right] && (ii) \qquad (3.8) \\
&= \mathbb{E}_{p(y_t, \theta \,|\, \alpha, \xi_t, h_{t-1})} \left[log \left(\frac{p(\theta \,|\, \alpha,\, y_t,\, \xi_t,\, h_{t-1})}{p(\theta \,|\, \alpha,\, h_{t-1})} \right) \right] && (iii)
\end{aligned}
$$

Plugging (3.8) (iii) into (3.6) we have that the intermediate expected information gain can be rewritten as:

$$
I_{h_{t-1}}(\xi_t) = \mathbb{E}_{p(\alpha \,|\, h_{t-1})} \left[\mathbb{E}_{p(y_t, \theta \,|\, \alpha, \xi_t, h_{t-1})} \left[log \left(\frac{p(\theta \,|\, \alpha,\, y_t,\, \xi_t,\, h_{t-1})}{p(\theta \,|\, \alpha,\, h_{t-1})} \right) \right] \right] \qquad (3.9)
$$

Using the fact that $p(\alpha \,|\, h_{t-1}) = p(\alpha \,|\, h_{t-1},\, \xi_t)$, we can express (3.9) as a single expectation taken with respect to $p(y_t,\, \theta,\, \alpha \,|\, \xi_t,\, h_{t-1}) = p(\theta,\, \alpha \,|\, h_{t-1})\, p(y_t \,|\, \theta,\, \alpha,\, \xi_t)$, bringing us closer to the aim of finding a 1-level Montecarlo estimator:

$$
I_{h_{t-1}}(\xi_t) = \mathbb{E}_{p(\theta, \alpha \,|\, h_{t-1})\, p(y_t \,|\, \theta, \alpha, \xi_t)} \left[log \left(\frac{p(\theta \,|\, \alpha,\, y_t,\, \xi_t,\, h_{t-1})}{p(\theta \,|\, \alpha,\, h_{t-1})} \right) \right] \qquad (3.10)
$$

There are now three pieces left: (a) we need to be able to sample from the distribution with respect to which expectation (3.10) is taken, (b) we need to use variational inference to approximate the nominator $p(\theta \,|\, \alpha,\, y_t,\, \xi_t,\, h_{t-1})$, (c) we want to be able

to evaluate the denominator $p(\theta \,|\, \alpha,\, h_{t-1})$ at most up to an additive constant.

(a) As done in Section 2.4 for θ, at each time-step t we have to update our prior on $(\theta,\, \alpha)$. This is done using variational inference to approximate $p(\theta,\, \alpha \,|\, h_{t-1}) \approx q_\psi(\theta,\, \alpha)$.

(b) Analogously to equation (2.4) we must again use variational inference to get $p(\theta \,|\, \alpha,\, y_t,\, \xi_t,\, h_{t-1}) \approx q_\phi(\theta \,|\, \alpha,\, y_t,\, \xi_t)$.

So far, the process seems to be proceeding smoothly, with the machinery from the "non-uncertain" BOED translating to equivalent techniques for the "uncertain" BOED. However, (c) will show that this is not always possible and that we have to pay a price - in terms of the quality of our estimators - for the extra flexibility that the Parametric Model Uncertainty (PMU) framework allows.

(c) In 2.4 we were able to express the denominator $p(\theta \,|\, y_{1:t-1},\, \xi_{1:t-1})$ as a product of known distributions and distributions independent of the parameters of interest. The latter could be safely discarded as they did not affect the choice of $(\phi^*,\, \xi_t^{opt})$ and hence led to an estimator where the denominator could be evaluated exactly (2.20). In PMU, the denominator $p(\theta \,|\, \alpha,\, h_{t-1})$ can be rewritten as:

$$
\begin{aligned}
p(\theta \,|\, \alpha,\, h_{t-1}) &\coloneqq p(\theta \,|\, \alpha,\, y_{1:t-1},\, \xi_{1:t-1}) \\
&= \frac{p(\theta,\, \alpha,\, y_{1:t-1} \,|\, \xi_{1:t-1})}{p(\alpha,\, y_{1:t-1} \,|\, \xi_{1:t-1})} \\
&= \frac{p(\theta,\, \alpha) \prod_{i=1}^{t-1} p(y_i \,|\, \theta,\, \alpha,\, \xi_i)}{p(\alpha \,|\, y_{1:t-1},\, \xi_{1:t-1})\, p(y_{1:t-1} \,|\, \xi_{1:t-1})} \\
&\propto p(\theta,\, \alpha) \prod_{i=1}^{t-1} p(y_i \,|\, \theta,\, \alpha,\, \xi_i) \,/\, p(\alpha \,|\, y_{1:t-1},\, \xi_{1:t-1})
\end{aligned}
\tag{3.11}
$$

The extra term $p(\alpha \,|\, y_{1:t-1},\, \xi_{1:t-1})$ cannot be evaluated exactly and requires us to use the variational approximation $p(\theta,\, \alpha \,|\, h_{t-1}) \approx q_\psi(\theta,\, \alpha)$ from (a).
Generally, $q_\psi(\theta,\, \alpha)$ will allow for easy marginalisation (for ex. in the case of a multivariate Gaussian) giving $p(\alpha \,|\, y_{1:t-1},\, \xi_{1:t-1}) \approx q_\psi(\alpha)$.

This means that Variational Posterior estimator will now have two variational approximations inside the expectation:

$$\mathcal{L}_{post}(\xi_t) \approx \mathbb{E}_{q_{\psi^*}(\theta)\, p(y_t \mid \theta, \xi_t)} \left[log \left(\frac{q_{\phi^*}(\theta \mid y_t,\, \xi_t)}{p(\theta) \prod_{i=1}^{t-1} p(y_i \mid \theta,\, \xi_i)} \right) \right] + C \tag{3.12}$$

Variational Posterior approximation for the standard Iterated BOED framework (2.20).

$$\mathcal{L}^{PMU}_{post}(\xi_t) \approx \mathbb{E}_{q_{\psi^*}(\theta,\, \alpha)\, p(y_t \mid \alpha, \theta, \xi_t)} \left[log \left(\frac{q_{\phi^*}(\theta \mid \alpha,\, y_t,\, \xi_t)\, q_{\psi^*}(\alpha)}{p(\theta,\, \alpha) \prod_{i=1}^{t-1} p(y_i \mid \theta,\, \alpha,\, \xi_i)} \right) \right] + C \tag{3.13}$$

Variational Posterior approximation for the Parametric Model Uncertainty BOED framework.

Using a similar approach, the Variational Marginal and Variational Nested Monte-carlo estimators can be adapted to the Model Uncertainty framework.
In this section, with the purpose of maintaining the narrative flow, we only provide a table outlining the salient characteristics of each estimator; their full expressions and derivations are provided in Appendix A.2.

The key takeaway from Table 3.1 is that the convergence rates obtained for standard BOED remain unchanged when we allow for Parametric Model Uncertainty (PMU), even though we have to work on bigger spaces when doing variational inference.

Estimator	Nested MC	Variational Posterior	Variational Marginal	Variational Nested MC
Number of variational approximations inside $\mathbb{E}$ (standard BOED)	0	1	1	1
Number of variational approximations inside $\mathbb{E}$ (PMU BOED)	0	2	1	2
MSE asymptotic convergence rate (standard BOED)	$O(T^{-2/3})$	$O(T^{-1})$	$O(T^{-1})$	$O(T^{-1})$
MSE asymptotic convergence rate (PMU BOED)	$O(T^{-2/3})$	$O(T^{-1})$	$O(T^{-1})$	$O(T^{-1})$
Total dimension of the input space(s) of the var. approx. used inside $\mathbb{E}$ (standard BOED)	0	$d(\Theta) + d(\mathcal{Y})$	$d(\mathcal{Y})$	$d(\Theta)+d(\mathcal{Y})+$ $(d(\Xi))^*$
Total dimension of the input space(s) of the var. approx. used inside $\mathbb{E}$ (PMU BOED)	0	$2 \cdot d(A) +$ $d(\Theta) + d(\mathcal{Y})$	$d(A) + d(\mathcal{Y})$	$2 \cdot d(A) +$ $d(\Theta)+d(\mathcal{Y})+$ $(d(\Xi))^*$
Total dimension of the input space(s) of the var. approx. used for sampling and inside $\mathbb{E}$ (standard BOED)	$d(\theta)$	$2 \cdot d(\Theta)+d(\mathcal{Y})$	$d(\Theta) + d(\mathcal{Y})$	$2 \cdot d(\Theta) +$ $d(\mathcal{Y}) +$ $(d(\Xi))^*$
Total dimension of the input space(s) of the var. approx. used for sampling and inside $\mathbb{E}$ (PMU BOED)	$d(A) + d(\theta)$	$2 \cdot d(A) + 2 \cdot$ $d(\Theta) + d(\mathcal{Y})$	$2 \cdot d(A) +$ $d(\Theta) + d(\mathcal{Y})$	$2 \cdot d(A) + 2 \cdot$ $d(\Theta)+d(\mathcal{Y})+$ $(d(\Xi))^*$

Table 3.1: Comparison of estimators for $I_{h_{t-1}}(\xi_t)$ both in standard BOED and in PMU BOED.
* $d(\Xi)$ appears only when we use the unified objective for (ξ^{opt}, ϕ^*) from section 2.3.

Chapter 4

Computational Experiments

We now aim to test the theoretical framework developed in Section 3 against the most common approach when dealing with nuisance variables, which is to treat them as part of the target variable $\Theta \coloneqq (\theta, \alpha)$ and maximising the overall EIG.
We compare these two frameworks to the standard BOED baseline, where we assume the true value of α is known, and also highlight how problematic model misspecification can be, strengthening the case for experimentalists to incorporate model uncertainty whenever this is present.

4.1 Experiment 1: Optimal Drug Dosage

Let's go back to the motivating example in Chapter 1.

At the time we placed a prior on our latent variable of interest $\theta \sim p(\theta)$, but assumed the remaining parameters were known constants.
We now also have uncertainty on $\gamma \sim p(\gamma)$.

$$p(\theta, \gamma) \sim \mathcal{N}\left(\begin{pmatrix}3.5\\3.1\end{pmatrix}, \begin{pmatrix}0.75^2 & 0\\0 & 0.5^2\end{pmatrix}\right) \qquad \rho = 2.9 \tag{4.1}$$

$$y|\,\xi, \theta \;\sim \mathit{Binomial}\,(10,\, f(\xi, \theta)) \tag{4.2}$$

$$\text{with} \quad f(\xi,\, \theta) = \frac{1}{1 + exp(\gamma - \rho\,(\xi - \theta)^2)} \tag{4.3}$$

We analyse and compare four approaches in a setting where the number of sequential experiments is $T = 20$:

(a) *Fully specified model*: In this case we sample an initial $\gamma \sim \mathcal{N}(3.1,\, 0.5^2)$ and feed it to the model, making it possible use standard iterated BOED along the variational posterior estimator from (2.5).

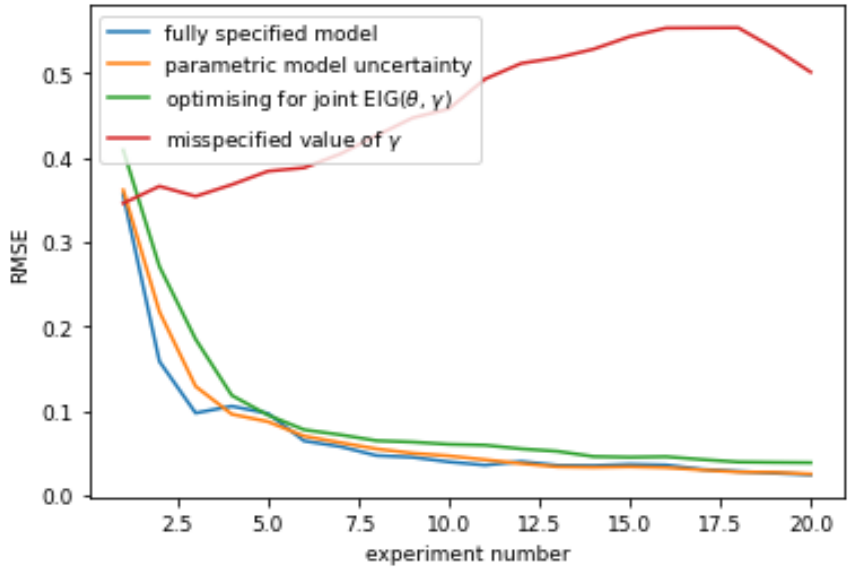
(b) *Parametric model uncertainty framework* used with the variational posterior $\mathcal{L}^{PMU}_{post}(\xi_t)$ from (3.13).

(c) *Optimising for EIG on* $\Theta \coloneqq (\theta,\, \gamma)$: Here we use again the variational posterior estimator (2.5) and pick designs which maximise the joint EIG.

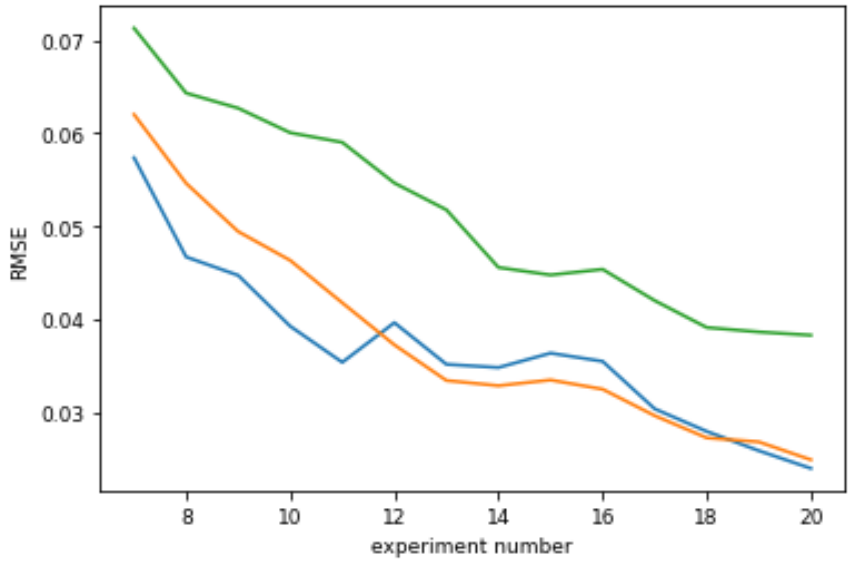
(d) *Misspecified model*: We feed the model a value of γ which is one standard deviation away from the true value ($\gamma_{real} \sim p(\gamma) \to \gamma_{miss} = \gamma_{real} + 0.5$) and then treat it as a fully specified model.

First of all we should decide on a metric to use to compare the four approaches.
As a matter of fact, comparing the cumulative EIGs would not make sense because (c) is measuring the EIG on $(\theta,\, \gamma)$, making it a completely different quantity.
Even measuring the variance or entropy of the posterior approximation $q_{\phi^*}(\theta \,|\, h_t)$ after each experiment would not be a sensible choice, because the misspecified model tends to converge to a distribution concentrated away from the true value of θ (so a low entropy of the posterior would not necessarily imply a better model).
As a consequence, we decide to focus on a relatively natural metric which is the squared distance between the true value of θ and the approximate MAP estimate for θ after each experiment.
Since we use a Gaussian variational approximation $p(\theta \,|\, h_t) \approx q_{\phi^*_t}(\theta \,|\, h_t)$ with $\phi^* = (\mu^*_t,\, \sigma^*_t)$, this translate to measuring $(\theta - \mu^*_t)^2$ after each experiment.
We simulate 10 full experimental runs and then calculate the root-mean-squared error for each time step t.

Figure 4.1 shows that Parametric Model Uncertainty (PMU) outperforms the joint EIG approach throughout the whole experimental run.
It is also worth noticing that within a few experiments PMU manages to 'bridge the gap' with the fully specified BOED, leading to roughly equivalent performances



(a) Full experimental run



(b) Zoom-in

Figure 4.1: Root-mean-squared error for the θ estimates of the 4 different BOED approaches

in the long-run for the two approaches. As the fully specified BOED represents our benchmark (we obviously cannot do better than it using PMU), this shows that PMU can be particularly effective.
Moreover, we can see that if BOED is used with a misspecified model, it performs extremely poorly. This suggests that, unless nuisance variables are known with absolute certainty, either PMU or optimisation for the joint EIG should be used, since if the model is correctly specified our estimates would be only marginally better, while if the model is misspecified we run the risk of our estimates being completely off.

4.2 Experiment 2: Location Finding

In this experiment we assume that there is a hidden source of noise $\theta \in \mathbb{R}^2$ and that we can sequentially measure noise intensity at T locations $(\xi_t)_{t=1}^T$, with noise intensity y_t decreasing roughly quadratically as the Euclidean distance between ξ_t and θ grows.

$$p(\theta) = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0.2^2 & 0\\0 & 0.2^2\end{pmatrix}\right) \tag{4.4}$$

$$p(\alpha) \sim LogNormal(0.68,\, 0.08^2) \qquad b = 0.25 \qquad m = 0.5 \tag{4.5}$$

$$y\,|\theta,\, \xi \sim \mathcal{N}(log\,\mu(\theta,\, \xi),\, 0.05^2) \tag{4.6}$$

$$\text{with} \quad \mu(\theta, \xi) = b + \frac{\alpha}{m + ||\theta - \xi||^2} \tag{4.7}$$

As before θ and α are picked with similar standard deviations (0.2 and 0.16 respectively).

The experiment follows the same lines of Experiment 1, with the RMSE being taken over 10 experimental runs, but with the exception that Nested Montecarlo rather than Variational Posterior is being used to approximate EIGs.

We can see that in this case the 20 experiments are not sufficient for PMU to converge to the baseline. However, it must be noted that, compared to Experiment 1, incorporating stochasticity of α causes a lot more of uncertainty at the start, with the RMSE after 1 experiment being approximately twice in the case of model uncertainty

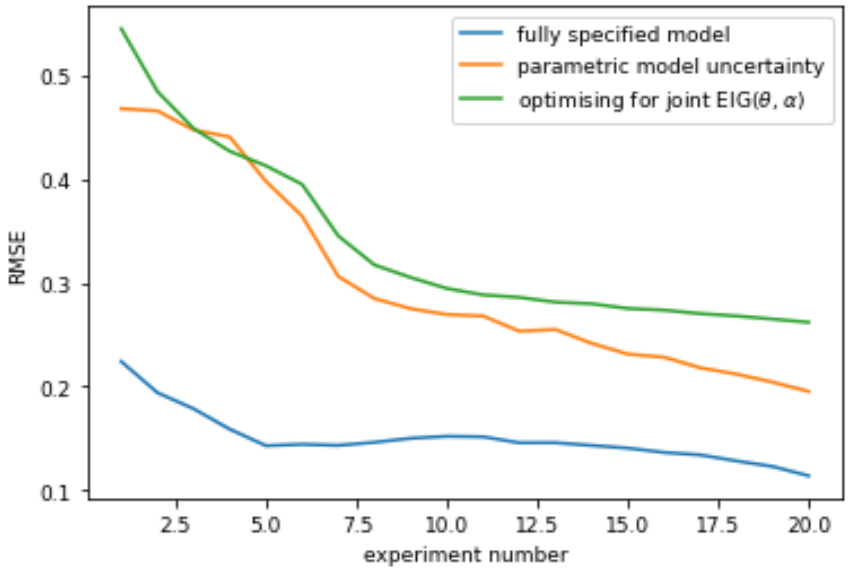


Figure 4.2: Root-mean-squared errors for the θ estimates

than in the fully specified case.
Nevertheless, PMU still outperforms the standard approach of optimising for the joint EIG.

Chapter 5

Policy-based Bayesian Optimal Experimental Design

In the last couple of years innovative approaches to BOED that leverage the power of Deep Learning have been proposed. They have made incredible improvements in the field possible by successfully tackling two key issues:

(a) Instead of greedily choosing at each time-step the experiment ξ_t which maximises the intermediate expected information gain $I_{h_{t-1}}(\xi_t)$, they learn a function π_ϕ - that maps the available experimental history h_{t-1} to the next experimental choice ξ_t - which is trained using the total expected information gain as the objective.

(b) The mapping π_ϕ can be trained ahead of time, circumventing the need for expensive computations during an experimental run, thus leading to almost instantaneous deployment times and significantly expanding the range of real-world applications of BOED.

In this section we aim to give an initial brief overview of the Deep Adaptive Design framework [9] and outline two fundamental theoretical properties that are at its core. We then proceed to fully prove that such properties continue to hold in the Parametric Model Uncertainty framework, laying the theoretical foundations for an approach to BOED that is both able to exploit the power of Deep Learning and to incorporate model uncertainty in its experimental design choices.

5.1 Deep Adaptive Design (DAD)

Let's start with the high-level assumption that we have a function π, called *policy*, that given an experimental history h_{t-1} returns a deterministic choice of design ξ_t. For a fixed policy π we can easily simulate full experimental histories according to the following algorithm:

Sample $\theta \sim p(\theta)$
for $t \leq T$ **do**
  $\xi_t = \pi(h_{t-1})$
  sample $y_t \sim p(y \,|\, \theta,\, \xi_t)$
end for
return $h_T = \{(\xi_t,\, y_t)\}_{t=1}^T$

This implies that the density of the generative process is:

$$p(\theta)\, p(h_T \,|\, \theta,\, \pi) = p(\theta) \prod_{t=1}^{T} p(y_t \,|\, \theta,\, \xi_t) \tag{5.1}$$

Our aim is now to learn a policy π that maximises the total expected information gain:

$$\mathcal{I}_T(\pi) \coloneqq \mathbb{E}_{p(\theta)\, p(h_t \,|\, \theta,\, \pi)} \left[\sum_{t=1}^{T} I_{h_{t-1}}(\xi_t) \right] \tag{5.2}$$

where the intermediate EIG is defined, exactly like in (2.15), as:

$$I_{h_{t-1}}(\xi_t) \coloneqq \mathbb{E}_{p(y \,|\, h_{t-1},\, \xi_t)} \left[H\left(p(\theta \,|\, h_{t-1})\right) - H\left(p(\theta \,|\, h_t)\right) \right] \tag{5.3}$$

Crucially, it was proven that the sum of intermediate EIGs inside expectation (5.2) can be expressed in a much more compact form:

$$\mathcal{I}_T(\pi) = \mathbb{E}_{p(\theta)\, p(h_t \,|\, \theta,\, \pi)} \left[log\, p(h_T \,|\, \theta,\, \pi) - log\, p(h_T \,|\, \pi) \right] \tag{5.4}$$

Despite the fact that $log\, p(h_T \,|\, \pi)$ is intractable, using the same types of contrastive bounds from Section 2.3, we can derive an intuitive lower bound on $\mathcal{I}_T(\pi)$:

$$\mathcal{I}_T(\pi) \geq \mathcal{L}_T(\pi,\, L) \coloneqq \mathbb{E}_{p(\theta_0,\, h_T \,|\, \pi)\, p(\theta_{1:L})} \left[log\, p(h_T \,|\, \theta_0,\, \pi) - log \left(\frac{1}{L+1} \sum_{l=0}^{L} p(h_T \,|\, \theta_l,\, \pi) \right) \right] \tag{5.5}$$

where we are essentially approximating $p(h_T \mid \pi) = E_{p(\theta)}[p(h_T \mid \theta, \pi)]$ with L samples, and adding $p(h_T \mid \theta_0, \pi)$ ensures $\mathcal{L}_T$ is actually a lower bound.

A second important property is the invariance of the total EIG under permutations of the same history:

Statement. *Let $\sigma \in S_k$ be a permutation acting on an experimental history $h_k^1 = (\xi_1, y_1), ..., (\xi_k, y_k)$ giving $h_k^2 = (\xi_{\sigma(1)}, y_{\sigma(1)}), ..., (\xi_{\sigma(k)}, y_{\sigma(k)})$. Then we have that:*

$$\mathbb{E}\left[\sum_{t=1}^{T} I_{h_{t-1}}(\xi_t) \middle| h_k = h_k^1\right] = \mathbb{E}\left[\sum_{t=1}^{T} I_{h_{t-1}}(\xi_t) \middle| h_k = h_k^2\right]$$

Intuitively this means that if we have observed k pairs $(\xi_t, y_t)_{t=1}^K$, their order does not affect the total EIG and the optimal choice of designs for the remaining $T - k$ steps will be the same for all permutations.

So far we have introduced π as some abstract function that given h_{t-1} returns the next design to choose ξ_t.
We can now use Permutation Invariance and the lower bound $\mathcal{L}_T$ to properly formalise the form of π and how it can be trained.
We start by encoding each design-outcome pair $(\xi_i,\, y_i)$ with a neural network E_{ϕ_1}. For an experimental history $h_t = (\xi_i,\, y_i)_{i=1}^t$, we then sum the t representations together to obtain:

$$R(h_t) \coloneqq \sum_{i=1}^{t} E_{\phi_1}(\xi_i,\, y_i) \tag{5.6}$$

$R(h_t)$ is now fed into a second neural network F_{ϕ_2} which outputs the design chosen by π:

$$\pi_\phi(h_t) = F_{\phi_2}(R(h_t)) \qquad \text{where } \phi = (\phi_1,\, \phi_2) \tag{5.7}$$

$R(h_t)$ is clearly permutation invariant, this means that the symmetry will not have to be learnt by the network during training, significantly improving performance [10].

We can now pair π_ϕ with the lower-bound $\mathcal{L}_T(\pi,\, L)$ and use SGA-based methods to find the optimal parameters ϕ^*.

5.2 Extending DAD to deal with Parametric Model Uncertainty

A key idea in the last section was to express the total EIG as a single quantity (rather than as a sum with T terms), which made the training of the neural nets underlying π_ϕ possible.
A very natural question to ask ourselves, is whether we can manage to obtain a simple form of the total EIG also under the Parametric Model Uncertainty assumptions of Section 3.
We show that this is indeed possible. The proof is relatively long and dry, but it represents, along with Section 3, one of the two main chunks of original work proposed in this dissertation.

Statement. *Suppose that the experimental setting satisfies the assumptions of the Parametric Model Uncertainty framework. Then for a given policy π the total EIG takes the following form:*

$$\begin{aligned} \mathcal{I}_T(\pi) &:= \mathbb{E}_{p(\theta,\alpha)\,p(h_T|,\theta,\alpha)}\left[\sum_{t=1}^{T} I_{h_{t-1}}(\xi_t)\right] \\ &= \mathbb{E}_{p(\theta,\alpha,h_T|\pi)}\left[log\, p(h_T|\theta,\alpha,\pi) - log\, p(h_T|\alpha,\pi)\right] \end{aligned} \tag{5.8}$$

Proof. Using expression (3.8)(i) we have:

$$I_{h_{t-1}}(\xi_t) = \mathbb{E}_{p(\theta,\alpha|h_{t-1})p(y_t|\theta,\alpha,\xi_t)}\left[log\left(\frac{p(y_t|\theta,\alpha,\xi_t)}{p(y_t|\alpha,\xi_t,h_{t-1})}\right)\right] \tag{5.9}$$

The inner fraction can be rewritten as:

$$\begin{aligned} \frac{p(y_t|\theta,\alpha,\xi_t)}{p(y_t|\alpha,\xi_t,h_{t-1})} &= \frac{p(\theta|h_{t-1},\alpha)\,p(y_t|\theta,\alpha,\xi_t)}{p(\theta|h_{t-1},\alpha)\,p(y_t|\alpha,\xi_t,h_{t-1})} \\ &= \frac{p(\theta|h_{t-1},\alpha)\,p(y_t|\theta,\alpha,\xi_t,h_{t-1})}{p(\theta|h_{t-1},\alpha)\,p(y_t|\alpha,\xi_t,h_{t-1})} && \text{as } y_t|\theta,\alpha,\xi_t \perp\!\!\!\perp h_{t-1} \\ &= \frac{p(\theta|h_{t-1},\alpha,\xi_t)\,p(y_t|\theta,\alpha,\xi_t,h_{t-1})}{p(\theta|h_{t-1},\alpha)\,p(y_t|\alpha,\xi_t,h_{t-1})} && \text{as } \xi_t = \pi(h_{t-1}) \\ &= \frac{p(\theta|y_t,\alpha,\xi_t,h_{t-1})}{p(\theta|h_{t-1},\alpha)} && \text{by Bayes' theorem} \end{aligned}$$

(5.9) can now be expressed as:

$$\begin{aligned} \mathrm{I}_{h_{t-1}}(\xi_t) &= \mathbb{E}_{p(\theta,\alpha|h_{t-1})p(y_t|\theta,\alpha,\xi_t)}\left[log\left(\frac{p(\theta|y_t,\alpha,\xi_t,h_{t-1})}{p(\theta|h_{t-1},\alpha)}\right)\right] \\ &= \mathbb{E}_{p(\theta,\alpha|h_{t-1})}\left[-log\, p(\theta|h_{t-1},\alpha)\right] + \mathbb{E}_{p(y_t,\theta,\alpha|\xi_t,h_{t-1})}\left[log\, p(\theta|h_{t-1},\xi_t,y_t,\alpha)\right] \end{aligned} \tag{5.10}$$

where the first expectation has been simplified since the integrand is independent of $p(y_t|\theta,\alpha,\xi_t)$.
If we now use on the second expectation the fact that:

$$\begin{aligned} p(y_t,\theta,\alpha|\xi_t,h_{t-1}) &= p(\alpha|\xi_t,h_{t-1})\, p(y_t|\xi_t,h_{t-1},\alpha)\, p(\theta|h_{t-1},\xi_t,\alpha,y_t) \\ &= p(\alpha|h_{t-1})\, p(y_t|\xi_t,h_{t-1},\alpha)\, p(\theta|h_{t-1},\xi_t,\alpha,y_t) \end{aligned}$$

(5.10) can be rewritten as the difference of two expected entropies:

$$\begin{aligned} \mathrm{I}_{h_{t-1}}(\xi_t) &= \mathbb{E}_{p(\alpha|h_{t-1})}\left[H(p(\theta|h_{t-1},\alpha))\right] - \mathbb{E}_{p(\alpha|h_{t-1})\,p(y_t|\xi_t,h_{t-1},\alpha)}\left[H(p(\theta|h_{t-1},\xi_t,y_t,\alpha))\right] \\ &= \mathbb{E}_{p(\alpha|h_{t-1})\,p(y_t|\xi_t,h_{t-1},\alpha)}\left[H(p(\theta|h_{t-1},\alpha)) - H(p(\theta|h_{t-1},\xi_t,y_t,\alpha))\right] \end{aligned} \tag{5.11}$$

where in the last line we used the fact that the first entropy is independent of $p(y_t|\xi_t,h_{t-1},\alpha)$.

$I_{h_{t-1}}(\xi_t)$ can be seen in (5.11) to depend only on h_{t-1} and $\xi_t = \pi(h_{t-1})$.

$$\begin{aligned} \mathcal{I}_T(\pi) &\coloneqq \mathbb{E}_{p(\theta,\alpha)\,p(h_T|\pi,\theta,\alpha)}\left[\sum_{t=1}^{T} I_{h_{t-1}}(\xi_t)\right] \\ &= \mathbb{E}_{p(\theta,\alpha,h_T|\pi)}\left[\sum_{t=1}^{T} I_{h_{t-1}}(\xi_t)\right] \\ &= \mathbb{E}_{p(h_T|\pi)}\left[\sum_{t=1}^{T} I_{h_{t-1}}(\xi_t)\right] && \text{by integrating over } (\theta,\alpha) \\ &= \sum_{t=1}^{T} \mathbb{E}_{p(h_{t-1}|\pi)}\left[I_{h_{t-1}}(\xi_t)\right] && \text{as } I_{h_{t-1}}(\xi_t) \perp\!\!\!\perp h_T \setminus h_{t-1} \end{aligned} \tag{5.12}$$

Plugging (5.11) into (5.12) we have:

$$\mathcal{I}_T(\pi) = \sum_{t=1}^{T} \mathbb{E}_{p(h_{t-1}|\pi)} \left[\mathbb{E}_{p(\alpha|h_{t-1})\, p(y_t|\xi_t, h_{t-1}, \alpha)} \left[H(p(\theta|h_{t-1}, \alpha)) - H(p(\theta|h_{t-1}, \xi_t, y_t, \alpha)) \right] \right] \tag{5.13}$$

Since:

$$\begin{aligned}
p(h_{t-1}|\pi)\, p(\alpha|h_{t-1})\, p(y_t|\xi_t, h_{t-1}, \alpha) &= \\
p(h_{t-1}|\pi)\, p(\alpha|h_{t-1}, \pi)\, p(y_t|\pi, h_{t-1}, \alpha) &= && \text{as } \alpha \perp\!\!\!\perp \pi \text{ and } \xi_t = \pi(h_{t-1}) \\
p(h_{t-1}|\pi)\, p(\alpha, y_t|h_{t-1}, \pi) &= \\
p(h_{t-1}, y_t, \alpha|\pi) &= \\
p(h_t, \alpha|\pi) && \text{because } h_t = h_{t-1} \cup \{\pi(h_{t-1}), y_t\}
\end{aligned}$$

we can rewrite (5.13) as:

$$\begin{aligned}
\mathcal{I}_T(\pi) &= \sum_{t=1}^{T} \mathbb{E}_{p(h_t,\alpha|\pi)} \left[H(p(\theta|h_{t-1}, \alpha) - H(p(\theta|h_{t-1}, \xi_t, y_t, \alpha)) \right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{p(h_t,\alpha|\pi)} \left[H(p(\theta|h_{t-1}, \alpha) - H(p(\theta|h_t, \alpha)) \right] && \text{as } h_t = h_{t-1} \cup \{(\xi_t, y_t)\} \\
&= \sum_{t=1}^{T} \mathbb{E}_{p(h_T,\alpha|\pi)} \left[H(p(\theta|h_{t-1}, \alpha) - H(p(\theta|h_t, \alpha)) \right] && \text{using indep. w.r.t. } h_T \setminus h_t
\end{aligned} \tag{5.14}$$

Crucially, this is a telescoping sum which gives:

$$\mathcal{I}_T(\pi) = \mathbb{E}_{p(h_T,\alpha|\pi)} \left[H(p(\theta|\alpha)) - H(p(\theta|h_T, \alpha)) \right] \tag{5.15}$$

We now aim to further simplify each of the two terms in (5.15) separately.

$$\begin{aligned}
\mathbb{E}_{p(h_T,\alpha|\pi)}\left[H(p(\theta|\alpha))\right] &= \mathbb{E}_{p(h_T,\alpha|\pi)}\left[\mathbb{E}_{p(\theta|\alpha)}\left(-log\, p(\theta|\alpha)\right)\right] \\
&= \mathbb{E}_{p(\alpha|\pi)\,p(h_T|\alpha,\pi)}\left[\mathbb{E}_{p(\theta|\alpha)}\left(-log\, p(\theta|\alpha)\right)\right] \\
&= \mathbb{E}_{p(\alpha|\pi)}\left[\mathbb{E}_{p(\theta|\alpha)}\left(-log\, p(\theta|\alpha)\right)\right] \qquad \text{as } p(h_T|\alpha,\pi) \perp\!\!\!\perp p(\theta|\alpha) \\
&= \mathbb{E}_{p(\alpha,\theta|\pi)}\left[-log\, p(\theta|\alpha)\right] \\
&= \mathbb{E}_{p(\alpha,\theta|\pi)\,p(h_T|\alpha,\theta,\pi)}\left[-log\, p(\theta|\alpha)\right] \qquad \text{as } p(h_T|\alpha,\theta,\pi) \perp\!\!\!\perp p(\theta|\alpha) \\
&= \mathbb{E}_{p(\theta,\alpha,h_T|\pi)}\left[-log\, p(\theta|\alpha)\right]
\end{aligned} \tag{5.16}$$

$$\begin{aligned}
\mathbb{E}_{p(h_T,\alpha|\pi)}\left[-H(p(\theta|h_T,\alpha))\right] &= \mathbb{E}_{p(h_T,\alpha|\pi)\,p(\theta|h_T,\alpha)}[log\, p(\theta|h_T,\alpha)] \\
&= \mathbb{E}_{p(\theta,h_T,\alpha|\pi)}\left[log\, p(\theta|h_T,\alpha)\right]
\end{aligned} \tag{5.17}$$

Plugging (5.16)-(5.17) into (5.15) give:

$$\begin{aligned}
\mathcal{I}_T(\pi) &= \mathbb{E}_{p(\theta,\alpha,h_T|\pi)}\left[-log\, p(\theta|\alpha) + log\, p(\theta|h_T,\alpha)\right] \\
&= \mathbb{E}_{p(\theta,\alpha,h_T|\pi)}\left[log\, p(h_T|\theta,\alpha,\pi) - log\, p(h_T|\alpha,\pi)\right]
\end{aligned} \tag{5.18}$$

since $p(\theta|h_T,\alpha) = p(\theta|h_T,\alpha,\pi) = p(h_T|\theta,\alpha,\pi)\,p(\theta|\alpha)\,/\,p(h_T|\alpha,\pi)$ □

Now that we have derived a closed-form expression for $\mathcal{I}_T(\pi)$, we can exploit it to effortlessly extend DAD to the setting of Parametric Model Uncertainty (PMU).
It can be easily proven that, using (3.8), Permutation Invariance is preserved under the PMU assumptions.
This allows us to use the same architecture from Section 5.1 (neural network encoder, pooled representation, and then emitter network) $\pi_\phi(h_t) = F_{\phi_2}(R(h_t))$.
Moreover, (5.18) can also be leveraged to produce a contrastive lower bound on $\mathcal{I}_T(\pi)$:

$$\mathcal{L}_T(\pi,L) \coloneqq \mathbb{E}_{p(\theta_0,\alpha,h_T|\pi)\,p(\theta_{1:L}|\alpha)}\left[log\, p(h_T|\theta,\alpha,\pi) - log\left(\frac{1}{L+1}\sum_{l=0}^{L} p(h_T|\theta_l,\alpha,\pi)\right)\right] \tag{5.19}$$

$\mathcal{L}_T(\pi,L)$ can be used as an objective to train the parameters of π_ϕ by stochastic gradient ascent, completing the picture of this extension of DAD, which makes it possible to have model uncertainty and to use Deep Learning at the same time.

5.3 Limitations of iDAD with model uncertainty

A second Deep Learning-based approach to BOED, proposed as recently as 2021, also deserves a brief mention.
Implicit Deep Adaptive Design (iDAD) [11] extends the DAD framework to implicit models (settings where we can only sample from the model, but not evaluate its density) and to instances where the experiments are not conditionally independent.
Formally, the experimental outcomes are distributed according to:

$$y_t = f(\epsilon_t;\, \xi_t(h_{t-1}),\, \theta,\, h_{t-1}) \tag{5.20}$$

where $\theta \sim p(\theta)$, $\epsilon_t \sim p(\epsilon)$, f is a deterministic function and we assume we can compute $\partial f/\partial \xi$ and $\partial f/\partial h$.

The problem with trying to incorporate PMU into this framework, is that the noise variables ϵ_t are assumed to be independent of everything else, making it impossible to reframe the advances on implicit models in a way that can be used for PMU.
Moreover, in PMU we are able to evaluate the explicit form of the model; this allows us to have tighter bounds and simpler architectures of π_ϕ, strengthening the case for our extension of DAD.

Chapter 6

Conclusion and Future Work

In summary, we have proposed a novel approach (PMU) for including uncertainty on model parameters in Bayesian Optimal Experimental Design.
We have derived bounds and estimators for the expected information gain (EIG) in such setting, showing via a series of computational experiments that PMU slightly outperforms the standard approach of optimising for the joint EIG.
Moreover, we laid the theoretical foundations for an extension of the Deep Adaptive Design (DAD) framework which can incorporate parametric model uncertainty in its design choices.
A very natural continuation of this line of work, perhaps too ambitious for the purpose of a Part C dissertation, would be to implement the DAD extension on top of the existing code, and to assess its performance on a number of experiments, analogously to what we did in the non-policy-based setting.

Appendix A

Appendix

A.1 Proof of (2.2)

Statement. $EIG(\xi) \coloneqq \mathbb{E}_{p(y\,|\,\xi)}\left[H\left(p(\theta)\right) - H\left(p(\theta\,|\,y,\,\xi)\right)\right] = \mathbb{E}_{p(y,\,\theta\,|\,\xi)}\left[\,log\left(\frac{p(y\,|\,\theta,\xi)}{p(y\,|\,\xi)}\right)\right]$

Proof. Using the definition of Shannon entropy $H(p(x)) \coloneqq -\int_x p(x)\; log(p(x))\; dx$ we can write the expected information gain as:

$$EIG(\xi) = \int_y p(y\,|\,\xi)\,\left[-\int_\theta p(\theta)\,log\left[p(\theta)\right]\,d\theta + \int_\theta p(\theta\,|\,y,\,\xi)\,log\left[p(\theta\,|\,y,\,\xi)\right]\,d\theta\right] dy \tag{A.1}$$

From Bayes' rule we have that $p(\theta\,|\,y,\,\xi) = \frac{p(y\,|\,\theta,\xi)\;p(\theta\,|\,\xi)}{p(y\,|\,\xi)} = \frac{p(y\,|\,\theta,\xi)\;p(\theta)}{p(y\,|\,\xi)}$ since $p(\theta\,|\,\xi) = p(\theta)$ - intuitively if we know the choice of experimental design ξ but not the experimental outcome y, we do not learn anything on θ.

$$\begin{aligned}
&= \int_y p(y\,|\,\xi)\,\left[-\int_\theta p(\theta)\,log\left[p(\theta)\right]\,d\theta + \int_\theta \frac{p(\theta)\,p(y\,|\theta,\,\xi)}{p(y\,|\,\xi)}\left(log\left[\frac{p(y\,|\,\theta,\,\xi)}{p(y\,|\,\xi)}\right] + log\left[p(\theta)\right]\right)\,d\theta\right]\,dy \\
&= \int_y p(y\,|\,\xi)\,\left[\int_\theta \left(-p(\theta) + p(\theta)\,\frac{p(y\,|\theta,\xi)}{p(y\,|\,\xi)}\right)\;log\left[p(\theta)\right]\,d\theta\right]\,dy\;+ \qquad (*) \\
&+ \int_y p(y\,|\,\xi)\,\left[\int_\theta \frac{p(\theta)\,p(y\,|\,\theta,\,\xi)}{p(y\,|\,\xi)}\,log\left[\frac{p(y\,|\,\theta,\,\xi)}{p(y\,|\,\xi)}\right]\,d\theta\right]\,dy \qquad (**)
\end{aligned} \tag{A.2}$$

The second term of the sum $(**)$ can be rewritten as:

$$\begin{aligned}
(**) &= \int_y \int_\theta p(\theta)\,p(y\,|\,\theta,\,\xi)\,log\left[\frac{p(y\,|\,\theta,\,\xi)}{p(y\,|\,\xi)}\right]\,d\theta\,dy \\
&= \int_y \int_\theta p(y,\,\theta\,|\,\xi)\,log\left[\frac{p(y\,|\,\theta,\,\xi)}{p(y\,|\,\xi)}\right]\,d\theta\,dy \\
&= EIG(\xi)
\end{aligned} \tag{A.3}$$

We are therefore left to show that $(*) = 0$

$$
\begin{aligned}
(*) &= \int_y p(y \,|\, \xi) \left[\int_\theta \left(-p(\theta) + p(\theta) \frac{p(y \,|\theta, \xi)}{p(y \,|\, \xi)} \right) \, log \left[p(\theta) \right] \, d\theta \right] dy \\
&= \int_y \int_\theta \left(p(\theta) \, p(y \,|\, \xi) - p(y, \, \theta \,|\, \xi) \right) \, log(p(\theta)) \, d\theta \\
&= \int_y \int_\theta \left(p(y, \, \theta, \, |\, \xi) - p(y, \, \theta \,|\, \xi) \right) \, log(p(\theta)) \, d\theta \\
&= 0
\end{aligned}
\tag{A.4}
$$

where the last implication follows again from the fact that $p(\theta) = p(\theta \,|\, \xi)$. □

A.2 PMU Variational Bounds

Following much of the same machinery developed in Section 3, expressions for the Variational Marginal and Variational Nested Montecarlo estimators in the context of Parameteric Model Uncertainty can be derived.

$$
I_{h_{t-q}}(\xi_t) = \mathbb{E}_{p(\theta, \alpha | h_{t-1}) \, p(y_t | \theta, \alpha, \xi_t)} \left[log \left(\frac{p(y_t | \theta, \alpha, \xi_t)}{p(y_t | \alpha, \xi_t, h_{t-1})} \right) \right] \tag{A.5}
$$

$$
\implies \mathcal{L}^{PMU}_{post}(\xi_t) = \mathbb{E}_{q_{\phi^*}(\theta, \alpha) \, p(y_t | \theta, \alpha, \xi_t)} \left[log \left(\frac{p(y_t | \theta, \alpha, \xi_t)}{q_{\phi^*}(y_t | \alpha, \xi_t)} \right) \right] \tag{A.6}
$$

$$
\mathcal{L}^{PMU}_{VNMC}(\xi_t) = \mathbb{E} \left[log \left(\frac{p(y_t | \theta_0, \alpha, \xi_t)}{\frac{1}{L+1} \sum_{l=0}^{L} \frac{p(\theta_l | h_{t-1}, \alpha) \, p(y_t | \theta_l, \alpha, \xi_t)}{q_{\phi^*}(\theta_l | y_t, \xi_t, \alpha)}} \right) \right] \tag{A.7}
$$

$$
\text{where } \mathbb{E} \text{ is taken w.r.t. } q_{\psi^*}(\theta_0, \alpha) \, p(y_t | \theta_0, \alpha, \xi_t) \prod_{l=1}^{L} q_{\phi^*}(\theta_l | y_t, \xi_t, \alpha) \tag{A.8}
$$

Using $p(\theta_l | h_{t-1}, \alpha) \approx \frac{p(\theta_l, \alpha) \prod_{i=1}^{t-1} p(y_i | \theta_l, \alpha, \xi_i)}{q_{\psi^*}(\alpha) p(y_{1:t-1} | \xi_{1:t-1})}$ the Variational Nested MC lower-bound becomes:

$$
\mathcal{L}^{PMU}_{VNMC}(\xi_t) = \mathbb{E} \left[log \left(\frac{p(y_t | \theta_0, \alpha, \xi_t)}{\frac{1}{L+1} \sum_{l=0}^{L} \frac{p(\theta_l, \alpha) \prod_{i=1}^{t} p(y_i | \theta_l, \alpha, \xi_i)}{q_{\psi^*}(\alpha) q_{\phi^*}(\theta_l | y_t, \xi_t, \alpha)}} \right) \right] + C \tag{A.9}
$$

A.3 Code

The computational side of the project was conducted using a machine with the following specifications:

GPU: NVIDIA GeForce GTX 1050 (6GB)
CPU: 2.8 GHz Intel(R) Core(TM) i7-7700HQ

Experiments were implemented in PyTorch [12] along with the probabilitistic programming language Pyro [13].
All the commented code can be found in the following anonymised repository:
`https://github.com/1035161/Parametric-Model-Uncertainty-in-BOED`

Bibliography

[1] Kathryn Chaloner and Isabella Verdinelli. "Bayesian Experimental Design: A Review". In: *Statistical Science* 10 (1995), pp. 273–304.

[2] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[3] Tom Rainforth et al. "On Nesting Monte Carlo Estimators". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 4267–4276. URL: https://proceedings.mlr.press/v80/rainforth18a.html.

[4] Adam Foster et al. "Variational Bayesian Optimal Experimental Design". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/d55cbf210f175f4a37916eafe6c04f0d-Paper.pdf.

[5] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773. eprint: https://doi.org/10.1080/01621459.2017.1285773. URL: https://doi.org/10.1080/01621459.2017.1285773.

[6] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[7] Léon Bottou. "Stochastic Gradient Descent Tricks". In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 421–436. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_25. URL: https://doi.org/10.1007/978-3-642-35289-8_25.

[8] Adam Foster et al. "A Unified Stochastic Gradient Approach to Designing Bayesian-Optimal Experiments". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 26–28 Aug 2020, pp. 2959–2969. URL: https://proceedings.mlr.press/v108/foster20a.html.

[9] Adam Foster et al. "Deep Adaptive Design: Amortizing Sequential Bayesian Experimental Design". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 3384–3395. URL: https://proceedings.mlr.press/v139/foster21a.html.

[10] Manzil Zaheer et al. "Deep Sets". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/f22e4747da1aa27e363d86d40ff442Paper.pdf.

[11] Desi R Ivanova et al. "Implicit Deep Adaptive Design: Policy-Based Experimental Design without Likelihoods". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 25785–25798. URL: https://proceedings.neurips.cc/paper/2021/file/d811406316b669ad3d370d78b51b1d2e-Paper.pdf.

[12] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[13] Eli Bingham et al. "Pyro: Deep Universal Probabilistic Programming". In: *Journal of Machine Learning Research* 20.28 (2019), pp. 1–6. URL: http://jmlr.org/papers/v20/18-403.html.