DEPARTMENT OF
STATISTICS

Title TBD

A thesis submitted in partial fulfillment of the requirements for
transfer of status from PRS to DPhil

November 2018

Adam Foster
University College

Department of Statistics
University of Oxford

Contents

Acknowledgements

I have been fortunate, in the first year of my DPhil, to have worked with a number of brilliant people. First and foremost, I would like to acknowledge the support, advice and patience of my supervisor, Yee Whye Teh. Also within Oxford, both Benjamin Bloem-Reddy and Tom Rainforth have been incredibly generous with their time and expertise. Without them, I would have struggled to achieve much in my first year. I would also like to thank Emile Mathieu for his unwavering support and friendship. I am grateful to Noah Goodman for taking a gamble on an unknown DPhil student and supporting my application to Uber Technologies for an internship. Noah's enthusiasm and guidance made my internship a joy, and also set me on a new and exciting research path. I'd like to extend my thanks to Martin Jankowiak and Eli Bingham, as well as the whole pyro team, for their continued help and support.

Preface

TODO change this!!!!!

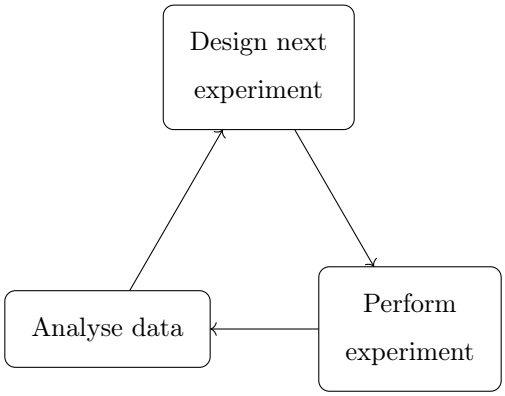
This thesis aims to describe the work I have done so far in my DPhil and discuss the future directions I plan to take in my research. Of primary relevance to my current work and future plans is Chapter **??** on optimal experiment design. Chapter **??** mostly concerns a project I worked on with Benjamin Bloem-Reddy which resulted in a UAI paper and oral presentation [Bloem-Reddy et al., 2018]. Chapter **??** relates to a question posed to me by Yee Whye Teh, namely, 'is probabilistic programming useful for Bayesian nonparametrics?'. Our workshop paper [Bloem-Reddy et al., 2017] and my open source contributions to the language `pyro` informed this chapter. Chapter **??**, the heart of this thesis, constitutes a draft of a paper that I plan to submit to ICML 2019, and represents the culmination of my internship with Uber. As detailed in Chapter 3, it is this project that I have found most exciting in my DPhil so far and that my future work will broadly be an extension of.

Chapter 1

Introduction and literature review

Much of machine learning is concerned with the analysis of given data. By contrast, the field of optimal experiment design (OED) is concerned with the creation of new data by experimentation or query. In many contexts, careful design of the experiment leads to more efficient learning. The gain in efficiency can be dramatic [Myung et al., 2013].

Idealised active learning can be represented as follows

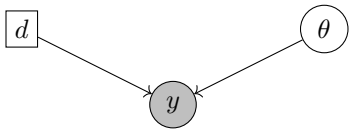


It is only within the context of a proposed data analysis that the optimality or suboptimality of an experiment design can be assessed.

1.1 Foundations

1.1.1 Problem specification

We assume the data analysis model for the experiment takes the form given by the graphical model



in which d represents the (non-random) design of the experiment, θ represents a latent variable and y represents the observed outcome of the experiment. The joint density is

$$p(y, \theta|d) = p(\theta)p(y|\theta, d) \tag{1.1}$$

We can generalize to the sequential design setting by standard Bayesian updating, namely at experiment iteration t, we replace $p(\theta)$ with $p(\theta|d_{1:t-1}, y_{1:t-1})$, where $d_{1:t-1}$ and $y_{1:t-1}$ are the designs and outcomes at previous steps of the experiment. The likelihood $p(y_t|\theta, d_t)$ is assumed unchanged. That is, experimentation does not change the behaviour of the underlying system from time to time, conditional upon θ.

1.1.2 One-step design

We begin with the case of designing one experiment, after which we will collect and analyse our data. In selecting an optimal experiment design, we must first say with respect to what utility a design is considered optimal. We begin by very briefly discussing a range of choices for utility. We then focus on a natural Bayesian choice of utility, the expected information gain.

Suppose $\mathcal{D}$ is the space of admissible designs. We seek a utility $U : \mathcal{D} \to \mathbb{R}$. The optimal design would then be

$$d^* = \text{argmax}_{d \in \mathcal{D}}\, U(d). \tag{1.2}$$

Suppose initially that we can assign a real-valued utility $U(y, \theta, d)$ to the event of observing y under design d when the true latent variable was θ. We can average over y to obtain

$$U(\theta, d) = \int p(y|\theta, d)\; U(y, \theta, d)\; dy \tag{1.3}$$

We can deal with θ in various ways

- The Bayesian approach [Chaloner and Verdinelli, 1995] places a prior $p(\theta)$ on θ and takes $U(d) = \int p(\theta)\; U(\theta, d)\; d\theta$
- The minimax approach [Fedorov, 1972] takes $U(d) = \inf_\theta U(\theta, d)$
- The local approach [Pronzato, 2010] begins with an estimate $\hat{\theta}$ and sets $U(d) = U(\hat{\theta}, d)$

Here, we primarily focus on the Bayesian approach. This is not an arbitrary decision, it can be motivated from decision theoretic considerations [Lindley, 1972]. See [Chaloner and Verdinelli, 1995] for further discussion of Bayesian experimental design.

Alternatively, we could take the (matrix-valued) 'utility'

$$U(y, \theta, d) = \left(\frac{\partial}{\partial \theta} \log p(y|\theta, d) \right)^2 \tag{1.4}$$

which leads to the the Fisher Information Matrix, used in many experiment design criteria [Pronzato, 2010],

defined as

$$\mathcal{I}(\theta, d) = \int \left(\frac{\partial}{\partial \theta} \log p(y|\theta, d) \right)^2 p(y|\theta, d)\ dy \tag{1.5}$$

We can obtain a scalar utility from $\mathcal{I}(\theta, d)$ by choosing from the 'alphabetical' criteria [Box, 1982] which are defined as

- D-optimality $U(d, \theta) = \det\ \mathcal{I}(\theta, d)$
- A-optimality $U(d, \theta) = \text{tr}\ \mathcal{I}(\theta, d)$
- E-optimality $U(d, \theta) = \max_i \lambda_i$ where λ_i are the eigenvalues of $\mathcal{I}(\theta, d)$

Work dating back to [Lindley, 1956] instead uses an information-theoretic utility[1]

$$U(y, \theta, d) = \log \frac{p(\theta|y, d)}{p(\theta)} = \log \frac{p(y|\theta, d)}{p(y|d)} \tag{1.6}$$

and Lindley established that this is the only form that satisfies certain intuitive properties of an informative experiment. For this reason, we will focus on this utility. See [Chaloner and Verdinelli, 1995, Ryan et al., 2015] for a fuller discussion of utility functions used in experiment design.

With the information-theoretic utility and Bayesian averaging, we arrive at the following form for $U(d)$, called the **expected information gain (EIG)**

$$U(d) = \text{EIG}(d) = \iint p(y, \theta|d) \log \frac{p(\theta|y, d)}{p(\theta)}\ dy\ d\theta \tag{1.7}$$

The EIG can be interpreted in a number of ways

1. As the expectation of information gain. If we define

$$\text{IG}(y, d) = \text{KL}(\ p(\theta|y, d)\ ||\ p(\theta)\ ) \tag{1.8}$$

   then $\text{EIG}(d) = \mathbb{E}_{y \sim p(y|d)}[\text{IG}(y, d)]$.

2. From APE. Define the **average posterior entropy** (APE) as

$$\text{APE}(d) = \iint p(y, \theta|d) \log p(\theta|y, d)\ dy\ d\theta \tag{1.9}$$

$$= -\int p(y|d) H(\ p(\theta|y, d)\ )\, dy \tag{1.10}$$

   where H is the differential entropy. Then

$$\text{EIG}(d) = H(\ p(\theta)\ ) - \text{APE}(d) \tag{1.11}$$

   and the prior entropy is a constant w.r.t. d. Thus EIG maximisation corresponds to APE minimisation.

[1] It can also be shown [Chaloner and Verdinelli, 1995] that this utility leads to a (modified form) of D-optimality for linear models.

3. Mutual information. Recall the mutual information is defined as

$$\mathrm{MI}(x, y) = \mathrm{KL}\left(p(x, y) \,||\, p(x)p(y) \right) \tag{1.12}$$

then we have

$$\begin{aligned}
\mathrm{MI}(y, \theta|d) &= \mathrm{KL}\left(p(y, \theta|d) \,||\, p(y|d)p(\theta) \right) && (1.13)\\
&= \iint p(y, \theta|d) \log \frac{p(y, \theta|d)}{p(\theta)p(y|d)} dy\, d\theta && (1.14)\\
&= \mathrm{EIG}(d). && (1.15)
\end{aligned}$$

4. Epistemic uncertainty. The total entropy or uncertainty in response y is

$$H_{tot} = H\left(p(y|d) \right) \tag{1.16}$$

the aleatoric uncertainty under parameter θ is

$$H_{alea}(\theta) = H\left(p(y|\theta, d) \right) \tag{1.17}$$

Under prior $p(\theta)$, the expected aleatoric uncertainty is

$$H_{alea} = \mathbb{E}_{\theta\sim p(\theta)}\left[H\left(p(y|\theta, d) \right) \right] \tag{1.18}$$

The epistemic uncertainty, under $p(\theta)$, is

$$\begin{aligned}
H_{epist} =& H_{tot} - H_{alea} && (1.19)\\
=& H\left(p(y|d) \right) - \mathbb{E}_{\theta\sim p(\theta)}\left[H\left(p(y|\theta, d) \right) \right] && (1.20)\\
=& -\int p(y|d) \log p(y|d) dy + \iint p(y, \theta|d) \log p(y|\theta, d) dy\, d\theta && (1.21)\\
=& \mathrm{EIG}(d) && (1.22)
\end{aligned}$$

We shall see that the connection to mutual information in particular is handy for estimating EIG, because modern techniques for estimating MI have been developed in the recent past.

1.1.3 Multi-step design

We turn to the case of multi-step, or adaptive, experiment design. In this context, we are able to base the next experiment on previous experimental designs and outcomes. Designing a sequence of multiple experiments, with a view to maximise expected utility can be viewed as a Partially Observable Markov Decision Process (POMDP) [Marchant et al., 2014], and falls within the scope of reinforcement learning [Pang et al., 2018]. The problem is also referred to as backward induction or stochastic dynamic programming. The formal reframing of OED as a POMDP is laid out below.

Setup

Suppose we have a determinic, finite time horizon $t = 1, ..., T$. We specify as Partially Observable Markov Decision Process (POMDP) as follows.

- States $s_t = (\theta, h_t)$, where $h_t = d_{1:t}, y_{1:t}$ the history of designs d and outcomes y up to the current time. Here $y_t \sim p(y|\theta, d)$ is the outcome of performing the experiment using design d_t. The practical state s'_t consists of the sufficient statistics for θ obtained from h_t. These can be used to compute the belief states b_t, encoding the full posterior for θ given that history h_t.
- Actions $a_t = d_{t+1}$. Transitions correspond to running the experiment and producing the outcome y_{t+1}.
- Observations $o_t = h_t$. Thus the only unobserved part of the state (θ, h_t) is the latent θ.
- Rewards $r_t = r(t, \theta, h_t)$. We take r to be a non-random function. Note that in many OED settings, we take $r_t = 0$ for $t < T$. Intuitively, this means we only care about our final understanding of or action upon the system, not the path taken to it. This is the choice made by [González et al., 2016] among others.

Under this set-up, the *optimal experiment design policy* is a map π from histories h_t to actions a_t which maximises the total reward

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^t r_t \middle| \pi\right] \tag{1.23}$$

where $\gamma \in [0, 1]$ is the discount factor, often set to 1 in finite horizons.

Connection to EIG

Horizon 1 Suppose $T = 1$. Choose the following reward function

$$r(1, \theta, h) = \log \frac{p(\theta|y, d)}{p(\theta)} = \log \frac{p(y|\theta, d)}{p(y|d)} \tag{1.24}$$

The Q-function of action d_1 is the expected reward

$$Q(s_0, d_1) = E_{y \sim p(y|\theta,d)}[r(t, \theta, d_1, y)] \tag{1.25}$$

Since we have no observation of θ, the belief Q-function of the belief state $p(\theta)$ and action d_1 is

$$Q(p(\theta), d_1) = E_{\theta \sim p(\theta)}\{E_{y \sim p(y|\theta,d)}[r(t, \theta, d_1, y)]\} \tag{1.26}$$

which reduces to the familiar expression

$$Q(p(\theta), d_1) = \iint p(y, \theta|d) \log \frac{p(\theta|y, d)}{p(\theta)} \, d\theta \, dy \tag{1.27}$$

Horizon T This formalism provides a convenient way to avoid the greedy approach to sequential design that is compatible with the information-theoretic objective of [Lindley, 1956].

Suppose the belief at time t is $b_t(\theta)$. This can be computed from the sufficient stats s'_t. We take the reward to be 0 at $t < T$ and

$$r(T, \theta, h_T) = r(T, \theta, b_T(\theta)) = \log \frac{b_T(\theta)}{p(\theta)} \tag{1.28}$$

we have updated b according to Bayes Theorem so

$$b_T(\theta) = p(\theta | y_{1:T}, d_{1:T}) \tag{1.29}$$

This reward structure represents the total information gained about θ from all experiments.

In fact, we can rewrite this reward to take non-zero values at earlier times, setting

$$r(t, \theta, h_t) = r(t, \theta, b_t(\theta)) = \log \frac{b_t(\theta)}{b_{t-1}(\theta)} \tag{1.30}$$

and this is equivalent to the previous formulation. To see this, consider the belief Q-function

$$\begin{aligned}
Q(b_t(\theta), d_{t+1}) &= \int b_t(\theta) \int p(y_{t+1}|\theta, d_{t+1}) \log \frac{b_{t+1}(\theta)}{b_t(\theta)} && (1.31)\\
&\quad + \int p(y_{t+2}|\theta, d_{t+1}, y_{t+1}) \log \frac{b_{t+2}(\theta)}{b_{t+1}(\theta)} + \dots \, dy_{t+2} \, dy_{t+1} \, d\theta && (1.32)\\
&= \int b_t(\theta) \int p(y_{t+1:t+2}|\theta, d_{t+1}) \log \frac{b_{t+2}(\theta)}{b_t(\theta)} + \dots \, dy_{t+1:t+2} \, d\theta && (1.33)\\
&= \dots && (1.34)\\
&= \int b_t(\theta) \int p(y_{t+1:T}|\theta, d_{t+1}) \log \frac{b_T(\theta)}{b_t(\theta)} \, dy_{t+1:T} \, d\theta && (1.35)
\end{aligned}$$

where $p(y_{t+1:T}|\theta, d_{t+1})$ assumes an optimal strategy after step $t+1$, under reward (1.30). We can now see by induction that the optimal strategy and Q-functions are the same for either choice of reward structure.

The greedy approach

In reinforcement learning, greediness refers to maximising the one-step-ahead reward, namely

$$a_t = \text{argmax}_{a_t \in \mathcal{A}} \left(\mathbb{E}[r_{t+1}|a_t] \right) \tag{1.36}$$

which, with the reward of (1.30), corresponds to one-step EIG maximisation at each step. We primarily focus on this form of multi-step optimisation because it removes all aspects of future planning beyond a single step from an already difficult problem.

Non-greedy approaches

Some have considered non-greedy strategies [González et al., 2016] [Pang et al., 2018]. See [Ryan et al., 2018, sec 6.1] for a summary of 'backwards induction' approaches.

Optional stopping

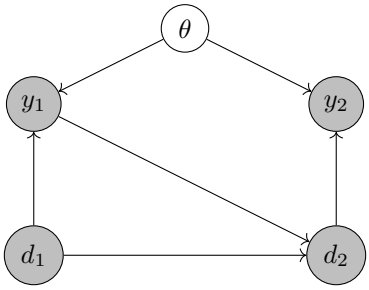
We finally mention another possible complication. Rather than a fixed and finite time horizon T, we may we allowed to continue experimentation indefinitely, choosing when to stop. One natural choice to stopping criterion is to terminate when the posterior entropy reaches a threshold value.

Optional stopping in Bayesian data analysis is a hotly debated topic
[Rouder, 2014, de Heide and Grünwald, 2017], but can be rigorously justified using martingale theory with certain stopping rules [Shafer et al., 2011] in the context of hypothesis testing. Of particular interest are stopping criteria based on non-negative supermartingales. We note in passing that the posterior entropy is a supermartingale (see (1.11)) and is non-negative if θ is discrete.

1.1.4 Theoretical considerations

The proposed experimentation strategy in which we design future experiments on the basis of previous observations may, at first sight, cause some consternation to the theoretical statistician. The first question we seek to answer is 'in what sense is the posterior obtained from multi-step OED the same as that obtained by a pre-ordained experimentation strategy?' The second line of questioning concerns the asymptotics of multi-step OED. 'Is multi-step OED statistically consistent and does it provide a faster convergence rate than other methods?'

To the first question, the answer is simply that the posterior is the same as if the experimentation strategy had been pre-ordained. Indeed, let us regard the designs as random variables and consider a 2-step experiment with the following graphical model



and the following conditional density for θ

$$p(\theta|y_1,d_1,y_2,d_2) = \frac{p(\theta,y_1,d_1,y_2,d_2)}{p(y_1,d_1,y_2,d_2)} \tag{1.37}$$

$$= \frac{p(\theta)p(d_1)p(y_1|\theta,d_1)p(d_2|y_1,d_1)p(y_2|\theta,d_2)}{\int p(\theta)p(d_1)p(y_1|\theta,d_1)p(d_2|y_1,d_1)p(y_2|\theta,d_2)d\theta} \tag{1.38}$$

$$= \frac{p(d_1)p(d_2|y_1,d_1)\ p(\theta)p(y_1|\theta,d_1)p(y_2|\theta,d_2)}{p(d_1)p(d_2|y_1,d_1)\ \int p(\theta)p(y_1|\theta,d_1)p(y_2|\theta,d_2)d\theta} \tag{1.39}$$

$$= \frac{p(\theta)p(y_1|\theta,d_1)p(y_2|\theta,d_2)}{\int p(\theta)p(y_1|\theta,d_1)p(y_2|\theta,d_2)d\theta} \tag{1.40}$$

showing that the dependence between d_2 and d_1, y_1 need not bother us. A related question from [Berry and Fristedt, 1985] was whether the map that takes the observed data to the posterior is a measurable function. This was addressed in [Berry and Fristedt, 1985, pp. 18-20] in the restricted setting of the multi-armed bandit.

A powerful answer to the second question was given by [Paninski, 2005] and is a form of Bernstein–von Mises Theorem for EIG maximisation OED. Under relatively mild conditions, for any neighbourhood $\mathcal{U}$ of the true parameter value θ_0 we have

$$p(\mathcal{U}|y_{1:T}, d_{1:T}) \to 1 \text{ as } T \to \infty \text{ in probability} \tag{1.41}$$

Under further conditions, we can show that the posteriors $p(\theta|y_{1:T}, d_{1:T})$ are asymptotically Normal with covariance matrix Σ_{info}. The same result holds for i.i.d. sampling of designs, giving rise to Σ_{iid}. We have that $\det \Sigma_{\text{info}} \leq \det \Sigma_{\text{iid}}$. This led [Paninski, 2005] to say "Thus, information maximization is in a rigorous sense asymptotically more efficient than the i.i.d. sampling strategy." Related results were obtained by [Pronzato, 2010] and [Hu, 1998].

1.2 Estimation of EIG

The estimation of EIG, and quantities mathematically equivalent to it, has received attention from a diverse group of researchers.

1.2.1 Challenges in EIG estimation

Recall

$$\text{EIG}(d) = \iint p(y, \theta|d) \log \frac{p(\theta|y, d)}{p(\theta)} \, dy \, d\theta = \iint p(y, \theta|d) \log \frac{p(y|\theta, d)}{p(y|d)} \, dy \, d\theta. \tag{1.42}$$

The computation of this integral is challenging since neither $p(\theta|y, d)$ nor $p(y|d)$ nor the outer integral can, in general, be found in closed form.

A further complication arises when the likelihood $p(y|\theta, d)$ cannot be computed pointwise. For example, this is the case in the presence of nuisance variables, also known as random effects. These are additional latent variables, ψ, that we do not consider variables of interest and so we do not want to waste resources reducing our uncertainty for them. Such models arise frequently in scientific applications, for instance accounting for individual variation between participants in a survey. With random effects ψ we have

$$p(y|\theta, d) = \int p(y|\theta, \psi, d) p(\psi|\theta) d\psi \tag{1.43}$$

which is typically intractable.

We survey existing approaches to EIG estimation.

1.2.2 Nested Monte Carlo

The estimator of [Vincent and Rainforth, 2017] and [Myung et al., 2013], among others, is a nested Monte Carlo (NMC) estimator

$$\mathrm{EIG}(d) \approx \frac{1}{N}\sum_{n=1}^{N}\left[\log p(y_n|\theta_n,d) - \log\left(\frac{1}{M}\sum_{m=1}^{M} p(y_n|\theta_m,d)\right)\right] \tag{1.44}$$

where

$$y_n, \theta_n \overset{\text{iid}}{\sim} p(y,\theta|d) \tag{1.45}$$

$$\theta_m \overset{\text{iid}}{\sim} p(\theta) \tag{1.46}$$

are all independent.

The drawbacks of such an estimator were noted in [Rainforth et al., 2018]. Most notably, while simple Monte Carlo estimators converge with a mean squared error rate $\mathcal{O}(N^{-1})$ in the total number of samples, NMC estimators converge at a much slower $\mathcal{O}(N^{-2/3})$ rate and are biased, though consistent [Rainforth et al., 2018].

A saving grace of the MC approach is a speed-up due to [Vincent and Rainforth, 2017] in the case that $\mathcal{Y}$, the sample space of y, is finite. Then

$$\mathrm{EIG}(d) \approx \sum_{y}\left[\frac{1}{N}\sum_{n=1}^{N} p(y|\theta_n,d)\log p(y|\theta_n,d) - \left(\frac{1}{N}\sum_{n=1}^{N} p(y|\theta_n,d)\right)\log\left(\frac{1}{N}\sum_{n=1}^{N} p(y|\theta_n,d)\right)\right] \tag{1.47}$$

where

$$\theta_n \overset{\text{iid}}{\sim} p(\theta). \tag{1.48}$$

Since this is a continuous function of vanilla Monte Carlo estimators, it converges at a rate $\mathcal{O}(N^{-1})$.

We finally mention that, whilst not present in the literature, NMC can be readily extended to the case of random effects

$$\mathrm{EIG}(d) \approx \frac{1}{N}\sum_{n=1}^{N}\left[\log\left(\frac{1}{M'}\sum_{m'=1}^{M'} p(y_n|\theta_n,\psi_{nm'},d)\right) - \log\left(\frac{1}{M}\sum_{m=1}^{M} p(y_n|\theta_m,\psi_m,d)\right)\right] \tag{1.49}$$

where

$$y_n, \theta_n \overset{\text{iid}}{\sim} p(y,\theta|d) \tag{1.50}$$

$$\psi_{nm'}|\theta_n \sim p(\psi|\theta_n) \tag{1.51}$$

$$\theta_m, \psi_m \overset{\text{iid}}{\sim} p(\theta,\psi) \tag{1.52}$$

and we note that we can replace $\psi_{nm'}$ with ψ_m when θ and ψ are independent under the prior $p(\theta,\psi)$. Again, when $\mathcal{Y}$ is finite a speed-up is possible, although we can no longer avoid nesting Monte Carlo estimators.

1.2.3 Inference based approaches

A number of authors [Long et al., 2013, Ryan et al., 2015] use some form of Laplace approximation to $p(\theta|y,d)$ to estimate expected information gain. In [Ouyang et al., 2016], probabilistic programming is used to completely solve the inference problem (in finite spaces) en route to estimating EIG.

1.2.4 Mutual information estimation

As mentioned previously, EIG estimation is mathematically equivalent to mutual information estimation, a topic that has received recent attention in part due to the connection with Generative Adversarial Networks (GANs) [Chen et al., 2016] and disentanglement [Chen et al., 2018]. In the following section, we drop d from our graphical model, and consider a joint density $p(y,\theta) = p(\theta)p(y|\theta)$.

Since $p(\theta|y)$ is typically an intractable, we might use an approximation $q(\theta|y)$. An idea used by [Barber and Agakov, 2004] and [Chen et al., 2016] is to bound the mutual information in terms of an amortised posterior $q(\theta|y)$ as

$$\mathrm{MI}(y,\theta) = \int p(\theta) \int p(y|\theta) \log \frac{p(\theta|y)}{p(\theta)} \, dy \, d\theta \tag{1.53}$$

$$\geq \int p(\theta) \int p(y|\theta) \log \frac{q(\theta|y)}{p(\theta)} \, dy \, d\theta. \tag{1.54}$$

For a fuller derivation and discussion of this and related bounds, see Chapter 2 on variational optimal experiment design.

A more recent idea is to use the Donsker-Varadhan Representation of the KL divergence to estimate mutual information [Belghazi et al., 2018]. We have

$$\mathrm{MI}(y,\theta) = \mathrm{KL}\left(\, p(y,\theta) \, || \, p(y)p(\theta) \, \right) = \sup_T \left\{ \mathbb{E}_{p(y,\theta)}[T(y,\theta)] - \log \left(\mathbb{E}_{p(\theta)p(y)}[e^{T(y,\theta)}] \right) \right\} \tag{1.55}$$

where the supremum is taken over measurable T. Note that the optimising T is given by

$$T^*(y,\theta) = \log \frac{p(y,\theta)}{p(y)p(\theta)} + C \text{ where } C \text{ is any constant} \tag{1.56}$$

The importance of (1.55) is that we no longer need access to any densities to estimate the mutual information.

Practical implementations arising from both these objective functions follow from choosing a suitable parametric family for T (or, in the former case, for $q(\theta|y)$). One then optimises the bound w.r.t. the parameters of the family using finite sample approximations to the expectations. We note that such an idea is intimately connected with the GAN [Nowozin et al., 2016].

1.3 Optimisation of EIG

So far, little attention has been paid to the design d. We suppose now that $d \in \mathcal{D}$ and we seek

$$d^* = \text{argmax}_{d \in \mathcal{D}} \quad \text{EIG}(d) \tag{1.57}$$

the optimal design. As outlined in the previous section, we can have only approximate estimates of $\text{EIG}(d)$. Thus we seek a global optimum of a unknown function which we have only noisy evaluations of. This puts us squarely in the domain of Bayesian optimisation [Shahriari et al., 2016]. When $\mathcal{D}$ is finite, we might instead call it a bandit problem.

Bayesian optimisation in its simplest form requires

1. A model of the unknown function
2. An acquisition rule to decide which design(s) should be queried at the next iteration

It is a fascinating fact that we are now back in the setting of OED. The variable of interest is the location of d^*, the maximiser of the unknown function. Other features of the function can be regarded as random effects. See [Hernández-Lobato et al., 2014] for further discussion on the connection between Bayesian optimisation and OED, in particular, the connection of EIG to Bayesian optimisation.

A popular approach to Bayesian optimisation is to choose a Gaussian process (GP) model of the unknown function, and an upper confidence bound (UCB) acquisition rule [Srinivas et al., 2009].

One aspect that sets EIG maximisation apart from conventional Bayesian optimisation is the ability for us to obtain more accurate estimates of the unknown function EIG by varying the amount of computational resources assigned to estimation. This was explored by [Vincent and Rainforth, 2017] in a finite $\mathcal{D}$ setting. The number of NMC samples was increased for the most promising designs. [McLeod et al., 2017] tackled a more general problem of variable cost objectives, taking a GP based approach.

1.4 Applications

Optimal experiment design is a broadly applicable subject and has been re-discovered a number of times in various fields. The purpose of this section is twofold: firstly, to note specific problems tackled by OED and to identify the model $p(y, \theta|d)$ and design space $\mathcal{D}$ in use; secondly, to more broadly suggest which types of models and experiment design problems are relevant to a certain field.

1.4.1 Bayesian optimisation

We begin with a very brief section on Bayesian optimisation, as we discussed this area in 1.3.

In Bayesian optimisation, models are regression models and typically nonlinear in nature. Gaussian process models are popular. The parameter of interest is the location of the optimum
[Hernández-Lobato et al., 2014]. The design space $\mathcal{D}$ consists of either single points in the input space

$\mathcal{X}$ (single acquisition), or vectors with entries in $\mathcal{X}$ (multiple acquisition). A number of experiment design strategies, here referred to as acquisition rules, have been proposed. See [Ginsbourger et al., 2008, Azimi et al., 2012] for a discussion of single and multiple acquisition rules.

1.4.2 Active learning

Whilst active learning is formally as broad as OED, we can identify certain models of particular interest in this field.

Beginning with [Cohn et al., 1996], the authors look at models for regression such as the neural network, mixture of Gaussians and locally weighted regression. The entire function (or its governing parameters) was the variable of interest. The design space was simply the input space (single acquisition).

We note in passing that an OED treatment of neural networks and deep regression models requires a credible Bayesian approach to deep learning, something that remains an area of research [Gal and Ghahramani, 2016].

The connection between active learning and Bayesian optimal design was explored in [Golovin et al., 2010]. Here, the parameter space, termed the hypothesis space $\mathcal{H}$, as well as the test space (design space) and outcome space are all finite. The authors begin with a problem in which the outcome of an experiment is deterministic and the goal is to determine exactly which hypothesis is correct. In the noiseless setting, the sequential design can be encoded as a decision tree and the problem is called the Optimal Decision Tree problem. This problem is known to be NP-hard. The extension to probabilistic modelling comes when the outcome of a test is random. See [Nowak, 2009] for a related example.

1.4.3 Psychology

An important use of OED in psychology is to distinguish between competing theories. Here, the parameter space is a finite set of theories and all other variables in the model are regarded as random effects. For instance, [Ouyang et al., 2016] apply OED to the 5-4 experiment for category learning [Medin and Schaffer, 1978]. The experiment aimed to distinguish two competing models of category learning: the *exemplar model* and the *prototype model*.

In [Vincent and Rainforth, 2017], OED for delayed and risky choice (DARC) experiments was studied. In such experiments, we want to model how humans discount future rewards relative to present ones. A single experiment takes the following form: 'Would you prefer £A_1 at time t_1 or £A_2 at time t_2'? The parameter of interest is the discount factor and the model was a (non-linear) variant on probit regression.

Models employed in psychology (and other social sciences) may not be Bayesian. Those doing Bayesian data analysis typically use linear models and their extensions – generalized linear mixed models (GLMMs) [Kruschke, 2014, Gelman et al., 2013]. In many cases, the parameters of interest will be certain coefficients in these models.

1.4.4 Big data analysis

In most applications, we assume that experiments will be carried out after design. In [Drovandi et al., 2017], the authors instead assume that a large dataset already exists. The design therefore consists of points, or sets of points, in the existing dataset. The examples focused on logistic regression models. In general, such settings have a well-defined design space, the models to be fitted are regression models.

1.4.5 Natural science

Natural sciences typically have complex and powerful predictive models. In [Vanlier et al., 2012], a complex biochemical network was modelled with a differential equation. Highly non-linear equations for amplitude versus offset experiments are just one example from physics [van Den Berg et al., 2003]. On the other hand, with these well-developed theories the statistical models used in natural sciences can be very simple – little more than Gaussian measurement error was incorporated in both these examples. The design spaces were the times at which to measure a response function, and the physical set-up of the experiment respectively.

Chapter 2

Estimating EIG

As outlined in Sections 1.2 and 1.3, the main technical barriers to OED using EIG maximisation are the estimation and optimisation of EIG. Here we address methods for estimation.

2.1 Variational optimal experiment design

Note: This section is broadly based on a paper recently submitted to the NIPS BDL workshop; we intend to submit a longer version to ICML 2019.

The core contribution of this section is to introduce efficient variational methods for EIG estimation that are applicable to a wide variety of models. The first method, which is related to amortised variational inference [Dayan et al., 1995, Kingma and Welling, 2014, Paige and Wood, 2016, Rezende et al., 2014, Stuhlmüller et al., 2013], employs an approximate posterior distribution, parameterised by the design and the experimental outcome. In a similar manner the second method employs a variational distribution for the marginal density over experimental outcomes for a given design. Both methods can benefit from recent advances in defining flexible families of amortised variational distributions using neural networks (e.g. normalising flows [Rezende and Mohamed, 2015, Tabak and Turner, 2013]). For this reason we developed our system in Pyro [Bingham et al., 2018], a deep probabilistic programming language that provides first class support for neural networks and variational methods.

The Nested Monte Carlo (NMC) approach (see Section 1.2.2) is inefficient because it constructs an independent estimate of $p(\theta|y,d)$ or $p(y|d)$ for each outcome y. Our key insight is that by taking a variational approach, we can instead learn an *amortized* approximation for either $p(\theta|y,d)$ or $p(y|d)$, and then use this approximation to efficiently estimate the EIG. In essence, the estimate of $p(y_1|d)$ provides information about $p(y_2|d)$ for similar y_1 and y_2 (presuming the density is smooth) and so it is more efficient to learn the functional form for $p(y|d)$ (or $p(\theta|y,d)$), than to treat separate values of y as distinct inference problems.

2.1.1 Bounding EIG

We construct a variational bound, $\mathcal{L}_p(d)$, using the amortized posterior $q_p(\theta|y,d)$:

$$\begin{aligned}
\text{EIG}(d) &= \iint p(y,\theta|d)\log\frac{p(\theta|y,d)q_p(\theta|y,d)}{q_p(\theta|y,d)}\,dy\,d\theta + H\left(\ p(\theta)\ \right) && (2.1)\\
&= \iint p(y,\theta|d)\log q_p(\theta|y,d)\,dy\,d\theta + H\left(\ p(\theta)\ \right) + \mathbb{E}_{p(y|d)}\left[\text{KL}\left(p(\theta|y,d)||q_p(\theta|y,d)\right)\right] && (2.2)\\
&\geq \iint p(y,\theta|d)\log q_p(\theta|y,d)\,dy\,d\theta + H\left(\ p(\theta)\ \right) \triangleq \mathcal{L}_p(d). && (2.3)
\end{aligned}$$

In analogy with variational inference, this bound is tight when $q_p(\theta|y,d) = p(\theta|y,d)$. Alternatively, we can instead introduce a marginal density $q_m(y|d)$, which results in an upper bound $\mathcal{U}_m(d)$:

$$\begin{aligned}
\text{EIG}(d) &= \iint p(y,\theta|d)\log p(y|\theta,d)\,dy\,d\theta - \int p(y|d)\log\frac{p(y|d)q_m(y|d)}{q_m(y|d)}dy && (2.4)\\
&= \iint p(y,\theta|d)\log p(y|\theta,d)\,dy\,d\theta - \int p(y|d)\log q_m(y|d)\,dy - \text{KL}\left(p(y|d)||q_m(y|d)\right) && (2.5)\\
&\leq \iint p(y,\theta|d)\log p(y|\theta,d)\,dy\,d\theta - \int p(y|d)\log q_m(y|d)\,dy \triangleq \mathcal{U}_m(d), && (2.6)
\end{aligned}$$

where the bound becomes tight for $q_m(y|d) = p(y|d)$.

2.1.2 Estimation

Just as in variational inference, the bounds in the previous section can be maximised with stochastic gradient methods [Robbins and Monro, 1951]. Concretely, suppose $\mathcal{Q}$ is a family of amortised variational approximations $q_p(\theta|y,d;\phi)$ indexed by ϕ. We can estimate EIG by maximizing the lower bound $\mathcal{L}_p(d;\phi)$:

$$\text{EIG}(d) \approx \max_\phi \mathcal{L}_p(d;\phi) = \max_\phi \left\{\iint p(y,\theta|d)\log q_p(\theta|y,d;\phi)\,dy\,d\theta\right\} + H\left(\ p(\theta)\ \right) \tag{2.7}$$

To do so only requires that we can generate samples from the model, $y_i,\theta_i \sim p(y,\theta|d)$; in a probabilistic programming context this corresponds to running the model forwards with no conditioning. We can then construct the required Monte Carlo estimates for the gradient as

$$\nabla_\phi \mathcal{L}_p(d;\phi) \approx \nabla_\phi \left\{\frac{1}{N}\sum_{i=1}^{N}\log q_p(\theta_i|y_i,d;\phi)\right\} \quad \text{where} \quad y_i,\theta_i \overset{\text{i.i.d.}}{\sim} p(y,\theta|d), \tag{2.8}$$

noting that no re-parameterization is required as $p(y,\theta|d)$ is independent of ϕ. An analogous scheme can be constructed for the upper bound $\mathcal{U}_m(d;\phi)$, except that we now perform a minimization.

2.1.3 Consistency

Suppose that ϕ_n are the values obtained from the procedure outlined above. Disregarding Monte Carlo estimation error for a moment, we can see immediately from (2.2) that provided $\forall y:\ \text{KL}\left(\ p(\theta|y,d)\ ||\ q_p(\theta|y,d;\phi_n)\ \right)\downarrow 0$ as $n\to\infty$ then by Monotone Convergence, the EIG estimates converge to $\text{EIG}(d)$ as $n\to\infty$. The condition rarely applies because it requires the approximating family to include the true posterior. If

$\mathrm{KL}\left(p(\theta|y,d) \;||\; q_p(\theta|y,d;\phi_n) \right)$ is simply monotonic decreasing, then we have convergence to (2.7), the best estimate possible within the given variational family.

2.1.4 Accounting for random effects

Note that the lower bound $\mathcal{L}_p(d)$ can be computed whether or not the model contains random effects (see Section 1.2.1 for a discussion of random effects). On the other hand, the definition of $\mathcal{U}_m(d)$ involves $p(y|\theta,d)$ which is typically intractable in the case of random effects.

Fortunately, we can still make progress. Starting from

$$\mathrm{EIG}(d) = \iint p(y,\theta|d) \log p(y|\theta,d) dy\, d\theta - \int p(y|d) \log p(y|d) dy \tag{2.9}$$

and we can bound each term separately in terms of two approximate densities: $q_m(y|d)$ for the marginal and $q_\ell(y|\theta,d)$ for the likelihood. Specifically, we have from Gibbs' inequality

$$-\int p(y|d) \log p(y|d) dy \leq -\int p(y|d) \log q_m(y|d) dy \tag{2.10}$$

$$\iint p(y,\theta|d) \log p(y|\theta,d) dy\, d\theta \geq \iint p(y,\theta,|d) \log q_\ell(y|\theta,d) dy\, d\theta\,. \tag{2.11}$$

Here we can no longer derive a direct bound on the EIG, but we can still use these inequalities to train to amortized densities, which will yield the true EIG if they match the true densities. Namely, suppose $\mathcal{Q}_1$ is a family of variational distributions $q_m(y|d;\phi_1)$ indexed by ϕ_1 and $\mathcal{Q}_2$ is a family of variational distributions $q_\ell(y|\theta,d;\phi_2)$ indexed by ϕ_2. Then a suitable objective for learning ϕ_1, ϕ_2 is

$$\mathcal{D}_{\phi_1,\phi_2}(d) \triangleq -\iint p(y,\theta,|d) \log q_\ell(y|\theta,d;\phi_2) dy\, d\theta - \int p(y|d) \log q_m(y|d;\phi_1) dy \tag{2.12}$$

$$\{\phi_1^*, \phi_2^*\} = \mathrm{argmin}_{\phi_1,\phi_2} \mathcal{D}_{\phi_1,\phi_2}(d) \tag{2.13}$$

where the optimization can be performed using stochastic gradient methods, as in the main paper. Once these approximations have been learned, we can plug them back into (2.9) to give

$$\mathrm{EIG}(d) \approx \iint p(y,\theta,|d) \log q_\ell(y|\theta,d;\phi_2^*) dy\, d\theta - \int p(y|d) \log q_m(y|d;\phi_1^*) dy \tag{2.14}$$

which can then itself be approximated by conventional Monte Carlo sampling.

2.1.5 Choice of variational family

The success of variational OED relies heavily on a good choice of variational family (guide, in the language of Pyro). Here we briefly outline techniques that aid writing guides for Generalised Linear Mixed Models

(GLMMs), which we define hierarchically as

$$\theta, \psi \sim N(\mu_0, \Sigma_0) \tag{2.15}$$

$$\zeta = X_d \theta \tag{2.16}$$

$$\eta = \zeta + \tilde{X}_d \psi \tag{2.17}$$

$$y \sim p(y|\eta) \tag{2.18}$$

which is a slight relaxation of typical GLMMs.

We first consider posterior guides, q_p and propose a multivariate Gaussian approximation. It can be shown that the bound (2.3) is maximised when the mean and covariance of q_p match that of the true posterior. We have

$$\mathbb{E}(\theta|y,d) = \mathbb{E}(\mathbb{E}[\theta|y,\zeta],d]|y,d) \tag{2.19}$$

$$\mathbb{E}[\theta|\zeta,d] = (X_d^T X_d)^{-1} X_d^T \zeta \tag{2.20}$$

$$\mathbb{E}(\theta|y,d) = (X_d^T X_d)^{-1} X_d^T \mathbb{E}(\zeta|y) \tag{2.21}$$

so we have done the amortisation over X_d analytically.

Similarly, the covariance $\mathbb{C}\text{ov}(\theta|y,d)$ is **not quite– need correction for those random effects**

$$(X_d^T X_d)^{-1} X_d^T \mathbb{C}\text{ov}(\zeta|y) X_d (X_d^T X_d)^{-1} \tag{2.22}$$

For marginal guides, simply note that q_ℓ depends on θ only through $X_d\theta$ –probably a shift in the mean...

2.1.6 Experiments

We validate our EIG estimators on a selection of generalized linear models. These serve as useful benchmarks, since they are workhorse models in many different scientific disciplines. Our results are summarized in Table 2.1 and Fig. 2.1-2.4. In all four cases, both estimators (i.e. the posterior method based on q_p and the marginal method[1] based on q_m) gave significantly lower variance than the NMC baseline, and in all but one case a significantly lower bias as well. We note that NMC especially struggles with random effects (LinReg + RE). More worryingly still, the bias of the NMC estimator can exhibit strong systematic variation as a function of the design, see Fig. 2.1-2.4. This is problematic because it can lead to the choice of a significantly suboptimal design. It is also worth emphasizing the utility of having multiple variational methods at our disposal: while the marginal method yields poor EIG estimates for the model with a large output dimension, the posterior method delivers high quality estimates. Finally, we consider an example (NΓ^{-1}Reg) that is not purely Gaussian. Here our method still performs well, despite the variational families not containing the true posterior or marginal.

[1] correcting for random effects as necessary

	LinReg		LinReg + RE		LinReg large dim(y)		NΓ^{-1}Reg	
	bias	2std	bias	2std	bias	2std	bias	2std
NMC	1.37	1.93	5.33	3.84	3.13	2.97	3.39	3.20
Posterior	-0.23	0.25	-0.55	0.41	-0.29	0.31	-0.50	0.51
Marginal	0.34	0.15	0.36	0.20	4.57	0.29	1.59	0.64

Table 2.1: Bias and variance (we report 2σ) of EIG estimation averaged over 10 runs and 11 designs. Each method was run for 10 seconds. For more details on the models and experimental setup see Appendix 4.1. Note that the directions of the bias for the posterior and marginal match the fact that they are lower and upper bounds, as would be expected.

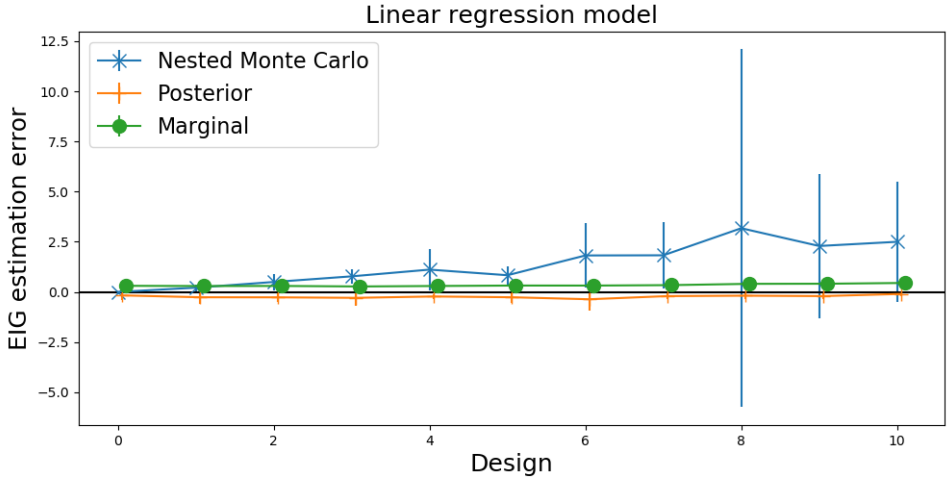


Figure 2.1: LinReg: EIG estimates for a linear regression model over 11 designs. We plot the mean and twice the standard deviation from 10 runs. Computational time was set to 10 seconds for comparison.

2.2 Extensions

2.2.1 Donsker-Varadhan

The bounds in Section 2.1.1 are partly inspired by the first bound described in Section 1.2.4. We also considered using bounds based upon the Donsker-Varadhan representation, mentioned in the same section. The Donsker-Varadhan representation gives

$$\text{EIG}(d) \leq \mathbb{E}_{p(y,\theta|d)}[T_d(y,\theta)] - \log\left(\mathbb{E}_{p(\theta)p(y|d)}[e^{T_d(y,\theta)}]\right) \tag{2.23}$$

Recall that the optimising T_d is given by

$$T_d^*(y,\theta) = \log \frac{p(y,\theta|d)}{p(y|d)p(\theta)} + C \text{ where } C \text{ is any constant.} \tag{2.24}$$

Whilst direction comparison between Donsker-Varadhan our previously proposed methods is difficult if T_d is a generic neural network, we can instead try either of

$$T_{p,d} = \log \frac{q_p(\theta|y,d)}{p(\theta)} \tag{2.25}$$

$$T_{m,d} = \log \frac{p(y|\theta,d)}{q_m(y|d)} \tag{2.26}$$

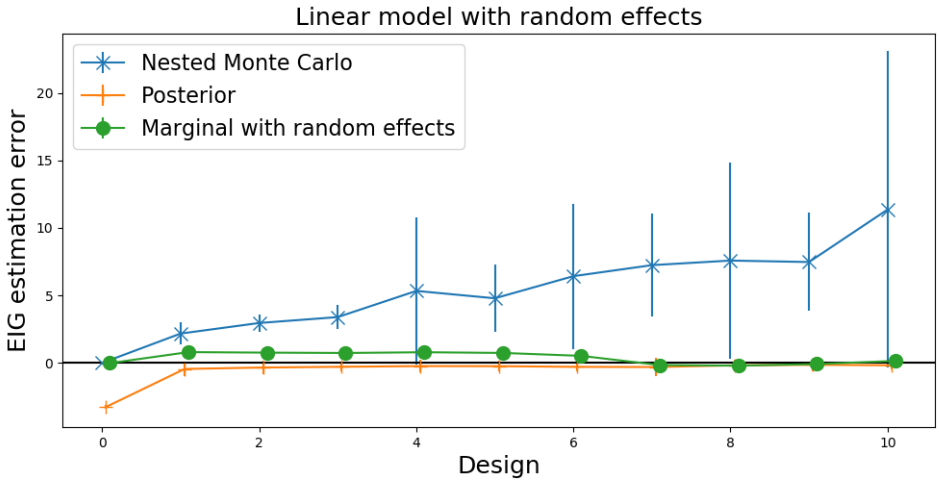


Figure 2.2: LinReg + RE: EIG estimates for a linear regression model with random effects. Settings as in Figure 2.1.

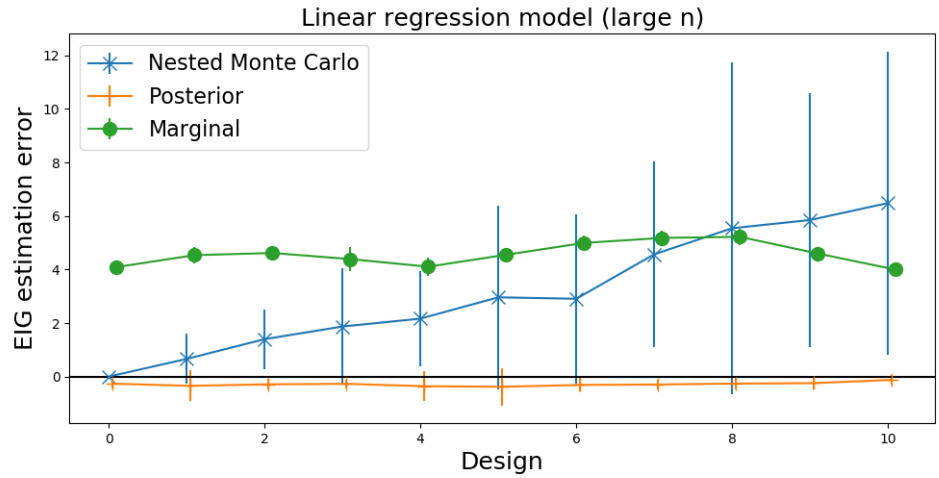


Figure 2.3: LinReg large dim(y): with settings as in Figure 2.1.

where q_p and q_m are the variational densities of Section 2.1.1. The form of T_d^* guarantees that the Donsker-Varadhan bound will be tight under the same conditions that give tightness in the bounds of the previous section.

When compared, the Donsker-Varadhan bound fared poorly against the simpler bounds in experiments. A single evaluation of the bound required many more samples to be numerically stable.

2.2.2 Analytic entropy

Our proposed method is based on taking Monte Carlo sums of log densities to approximate integrals such as

$$\mathrm{APE}(d) \approx -\iint p(y,\theta|d)\log q_p(\theta|y,d;\phi)\,dy\,d\theta, \tag{2.27}$$

which can be rewritten as

$$\mathrm{APE}(d) \approx -\int p(y|d)\int p(\theta|y,d)\log q_p(\theta|y,d;\phi)\,d\theta\,dy \tag{2.28}$$

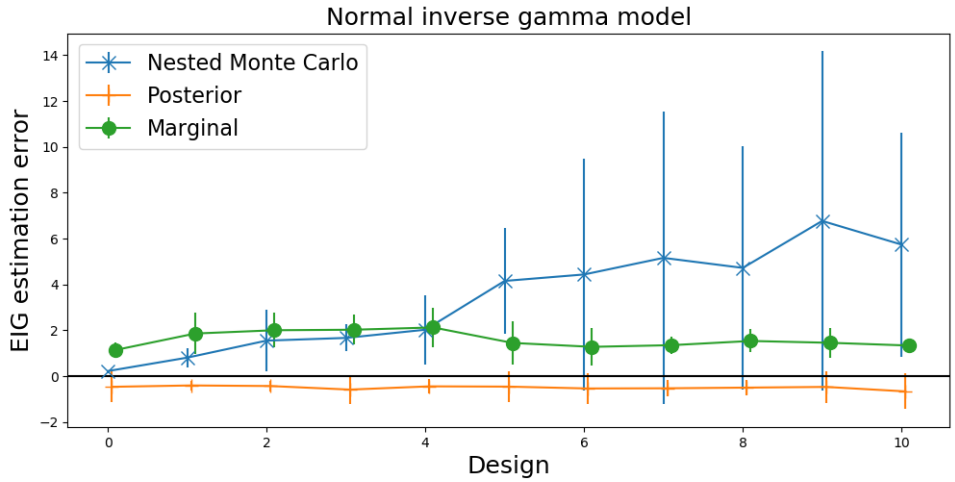


Figure 2.4: NΓ^{-1}Reg: EIG estimates for a Normal inverse-Gamma model. Settings as in Figure 2.1.

we could instead have used the following

$$\text{APE}(d) \approx -\int p(y|d) \int q_p(\theta|y,d;\phi) \log q_p(\theta|y,d;\phi)\, d\theta\, dy = -\int p(y|d) H\left(\, q_p(\theta|y,d;\phi)\,\right) dy. \tag{2.29}$$

If the variational family $\mathcal{Q}$ for q_p had been chosen so that $H\left(\, q_p(\theta|y,d;\phi)\,\right)$ is easily calculable, the new approximation may have significantly lower Monte Carlo estimation error.

On the other hand, the new approximation does not bound APE(d) and so an entirely different mechanism to *learn* q_p must be employed. A correct learning approach is suggested by [**?**], which showed that if $\text{KL}\left(\, q_n \,||\, p\,\right) \to 0$ then $H\left(\, q_n\,\right) \to H\left(\, p\,\right)$. In experiments, we trained q_p using (2.7) just as in the original method. We used the analytic entropy in the final step.

In practice, most computation time was spent learning q_p rather than making the final Monte Carlo estimate and so the gains from this alternative were negligible.

2.2.3 Experiments

We compare the two suggested extensions against the posterior method of Section 2.1. We used the same settings as the LinReg example. The biases and variances can be seen in Table 2.2 and Figure 2.5. We see poorer performance from Donsker-Varadhan and little to tell between posterior, and posterior using analytic entropy.

	LinReg2	
	bias	2std
Posterior	-0.24	0.32
Posterior, with analytic entropy	-0.33	0.53
Donsker-Varadhan	-0.46	1.01

Table 2.2: Bias and variance (we report 2σ) of EIG estimation averaged over 10 runs and 11 designs. Each method was run for 10 seconds. For more details on the models and experimental setup see Appendix 4.1.

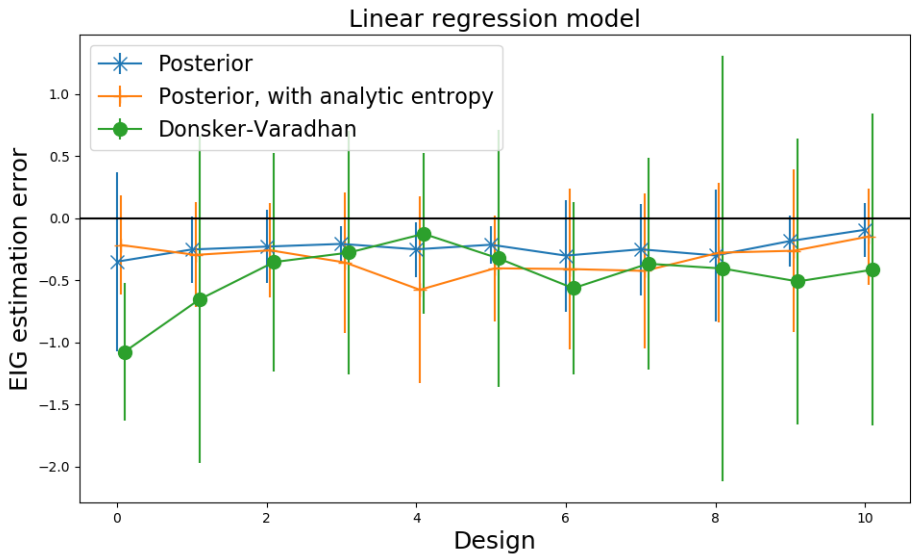


Figure 2.5: LinReg2: EIG estimates for a linear model, with extension methods. Settings as in Figure 2.1.

Chapter 3

Future directions

3.1 EIG

3.1.1 EIG estimation on simpler models

Current project. Focus is on linear-type models (see Kruschke) that are used in applied stats. We can implement some semi-whitebox methods here. The aim is to use these models and EIG estimators in active learning loops. So we want sub-second estimation. This leads to methods based on relatively structued guides.

3.1.2 EIG estimation on complex models

Very related, but taking a more black box approach. Assume that the model is too complex to build a structured guide, but that the experiment is very expensive. So we can spend more time on EIG estimation. Deep learning approaches, like Donsker-Varadhan, might look more attractive.

3.1.3 Theory of EIG estimators

Are estimators statistically consistent? Can we estimate, bound or approximate the error, or the relative error across different d?

3.1.4 EIG gradients

How best to estimate the gradient $\partial_d \text{EIG}$? Can we obtain bounds? What would Rainforth gradient estimation look like? Can we optimise EIG in a GAN-like fashion – iterative updates of q and d.

3.1.5 EIG optimisation

Are there special features of EIG that we can exploit when using Bayes opt, or something else, to do EIG optimisation?

3.1.6 Model misspecification

How best to deal with model misspecification in experiment design. A uniform increase in y entropy does not change design... what would be the right paradigm for this?

3.1.7 Sequential design and active learning

Further considerations for using EIG estimation/optimisation in a live active learning loop.

3.1.8 Optional stopping

Suppose we use posterior entropy as an optional stopping criterion, and use EIG for sequential experiment design. How would this impact final conclusions that we are able to make about data?

3.1.9 Dynamic models

Experiment design for systems that change as a result of the experimentation. Things like the atmosphere or a pond.

3.2 Beyond EIG

3.2.1 Causal inference

What if we design an experiment for causal strucuture learning? *And* information? How do these fields intersect? Speak to Robin Evans.

3.2.2 Power

This is a theoretical question. How does the Bayesian notion of EIG intersect with frequentist notions of experiment design, in particular, statistical power?

3.2.3 Cost

Designing experiments for information, but with a cost assicatiated with each experiment. Sequential case may be more interesting that one shot (which seems simple).

3.2.4 Non-greedy

Related to above. Solving the non-greedy experimental design problem brings in elements from POMDPs and RL. Should we use EIG here? Should we use RL reward functions? Are they in some sense (approximately) the same? Could greedy EIG optimisation arise as a good approximation to the RL task?

3.2.5 Other criteria

In active learning, they have criteria about the expected misclassification, and some other criteria. Can we connect these? In classical experiment design they have all these mysterious criteria like D-optimality and so on.

3.2.6 Experiment design for model criticism

Rather than assuming the model to be true and looking to gain information within the model, suppose instead that we have an empirical distribution and seek a new experiment to best expose flaws in the whole model. For instance, when comparing the posterior predictive and empirical distributions (possibly conditional on an input).

Chapter 4

Appendix

4.1 Experiment details

4.1.1 LinReg

A classical Bayesian linear regression model has the following form

$$\theta \sim N(\mu_0, \Sigma_0) \tag{4.1}$$

$$y|\theta, d \sim N(X_d\theta, \sigma^2 I) \tag{4.2}$$

where X_d is the design matrix.

In our LinReg example, we took:

$$\mu_0 = 0 \tag{4.3}$$

$$\Sigma_0 = \begin{pmatrix} 10^2 & 0 \\ 0 & 0.1^2 \end{pmatrix} \tag{4.4}$$

$$\sigma^2 = 1 \tag{4.5}$$

$$X_d = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \text{ a } (10 \times 2) \text{ matrix} \tag{4.6}$$

with all 11 possible designs considered.

We chose families of variational distributions that include the true posterior (or true marginal). For

the amortised posterior, we set $\phi = (\Lambda, \delta, \Sigma_{\mathrm{p}})$ and let

$$q_p(\theta|y, d; \phi) \sim N(\mu_{\mathrm{p}}, \Sigma_{\mathrm{p}}) \tag{4.7}$$

$$\text{where } \mu_{\mathrm{p}} = (X_d^T X_d + \Lambda)^{-1}(X_d^T X_d + \Lambda)(y + \delta) \tag{4.8}$$

and Λ is a diagonal matrix and Σ_{p} is positive definite. For the marginal, we simple take $\phi = (\mu_{\mathrm{m}}, \Sigma_{\mathrm{m}})$ and

$$q_m(y|d; \phi) \sim N(\mu_{\mathrm{m}}, \Sigma_{\mathrm{m}}) \tag{4.9}$$

Finally, for each of our variational methods we used the Adam optimizer [**?**] with a learning rate specified below. Each iteration used N_t samples, with T iterations in total. We used N samples for the final evaluation. NMC settings are N, M [Vincent and Rainforth, 2017] and we took the advice of the authors to set $N = M^2$.

The exact parameter settings, to get about 10 seconds of computation for each method, were

NMC		Posterior				Marginal			
N	M	N_t	T	lr	N	N_t	T	lr	N
110^2	110	10	1200	0.05	500	10	1200	0.05	500

4.1.2 LinReg + RE

In this experiment, we extended the model to include random effects. Specifically,

$$\theta \sim N(\mu_0, \Sigma_0) \tag{4.10}$$

$$\tilde{\theta} \sim N(\tilde{\mu}_0, \tilde{\Sigma}_0) \tag{4.11}$$

$$y|\theta, d \sim N(X_d\theta + \tilde{X}\tilde{\theta}, \sigma^2 I) \tag{4.12}$$

where

$$\tilde{\mu}_0 = 0 \tag{4.13}$$

$$\tilde{\Sigma}_0 = I_{10} \tag{4.14}$$

$$\tilde{X} = I_{10} \tag{4.15}$$

Here θ is the random variable of interest, while $\tilde{\theta}$ is a nuisance variable that needs to be integrated out. The variational distribution for the likelihood, q_ℓ, was the same as q_m, except that the mean was shifted by $X_d\theta$.

The exact parameter settings, to get about 10 seconds of computation for each method, were

NMC		Posterior				Marginal			
N	M	N_t	T	lr	N	N_t	T	lr	N
52^2	52	10	150	0.05	500	10	600	0.05	500

4.1.3 LinReg large $\dim(y)$

This experiment was identical to LinReg, except that we took X_d to have dimensions 20×2, with 11 designs as before. We also altered the marginal variational distribution to reflect the new dimension of y. Other than that, the specification of all variational distributions was identical.

The exact parameter settings, to get about 10 seconds of computation for each method, were

NMC		Posterior				Marginal			
N	M	N_t	T	lr	N	N_t	T	lr	N
90^2	90	10	1000	0.05	500	10	700	0.05	500

4.1.4 NΓ^{-1}Reg

We changed the model to

$$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta) \tag{4.16}$$

$$\theta \sim N(\mu_0, \Sigma_0) \tag{4.17}$$

$$y|\theta, \sigma^2, d \sim N(X_d\theta, \sigma^2 I) \tag{4.18}$$

where $\alpha = 3$ and $\beta = 2$.

We used a mean-field posterior variational distribution. For θ, we used the same variational distribution as for LinReg. For σ^2 we used an inverse Gamma variational distribution. We augmented the parameters ϕ with α_p, b_0 and took $\beta_\mathrm{p} = b_0 + \frac{1}{2}(y^T y - y^T X_d \mu_\mathrm{p})$. Then

$$q_p(\sigma^2|y, d; \phi) \sim \Gamma^{-1}(\alpha_\mathrm{p}, \beta_\mathrm{p}) \tag{4.19}$$

The marginal variational distribution was as in LinReg (a Gaussian).

The exact parameter settings, to get about 10 seconds of computation for each method, were

NMC		Posterior				Marginal			
N	M	N_t	T	lr	N	N_t	T	lr	N
110^2	110	10	800	0.05	500	10	1200	0.05	500

4.1.5 LinReg2

We used the model of LinReg, and the following parameter settings

The exact parameter settings, to get about 10 seconds of computation for each method, were

Posterior				Posterior, with analytic entropy				Donsker-Varadhan			
N_t	T	lr	N	N_t	T	lr	N	N_t	T	lr	N
10	1200	0.05	500	10	1200	0.05	50	100	100	0.05	500

Bibliography

[Azimi et al., 2012] Azimi, J., Jalali, A., and Fern, X. (2012). Hybrid batch bayesian optimization. *arXiv preprint arXiv:1202.5597*.

[Barber and Agakov, 2004] Barber, D. and Agakov, F. (2004). The im algorithm: a variational approach to information maximization. *Advances in Neural Information Processing Systems*, 16:201.

[Belghazi et al., 2018] Belghazi, I., Rajeswar, S., Baratin, A., Hjelm, R. D., and Courville, A. (2018). Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

[Berry and Fristedt, 1985] Berry, D. A. and Fristedt, B. (1985). Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, 5.

[Bingham et al., 2018] Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. (2018). Pyro: Deep Universal Probabilistic Programming. *arXiv preprint arXiv:1810.09538*.

[Bloem-Reddy et al., 2018] Bloem-Reddy, B., Foster, A., Mathieu, E., and Teh, Y. W. (2018). Sampling and inference for beta neutral-to-the-left models of sparse networks. In *Uncertainty in Artifical Intelligence*.

[Bloem-Reddy et al., 2017] Bloem-Reddy, B., Mathieu, E., Foster, A., Rainforth, T., Teh, Y. W., Ge, H., Lomelí, M., and Ghahramani, Z. (2017). Sampling and inference for discrete random probability measures in probabilistic programs. In *NIPS Workshop on Advances in Approximate Bayesian Inference*.

[Box, 1982] Box, G. E. (1982). Choice of response surface design and alphabetic optimality. Technical report, WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER.

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

[Chen et al., 2018] Chen, T. Q., Li, X., Grosse, R., and Duvenaud, D. (2018). Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*.

[Chen et al., 2016] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180.

[Cohn et al., 1996] Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145.

[Dayan et al., 1995] Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The helmholtz machine. *Neural computation*, 7(5):889–904.

[de Heide and Grünwald, 2017] de Heide, R. and Grünwald, P. D. (2017). Why optional stopping is a problem for bayesians. *arXiv preprint arXiv:1708.08278*.

[Drovandi et al., 2017] Drovandi, C. C., Holmes, C., McGree, J. M., Mengersen, K., Richardson, S., and Ryan, E. G. (2017). Principles of experimental design for big data analysis. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 32(3):385.

[Fedorov, 1972] Fedorov, V. (1972). *Theory of optimal experiments*. Academic Press, New York.

[Gal and Ghahramani, 2016] Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

[Gelman et al., 2013] Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

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

[Golovin et al., 2010] Golovin, D., Krause, A., and Ray, D. (2010). Near-optimal bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems*, pages 766–774.

[González et al., 2016] González, J., Osborne, M., and Lawrence, N. (2016). Glasses: Relieving the myopia of bayesian optimisation. In *Artificial Intelligence and Statistics*, pages 790–799.

[Hernández-Lobato et al., 2014] Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, pages 918–926.

[Hu, 1998] Hu, I. (1998). On sequential designs in nonlinear problems. *Biometrika*, 85(2):496–503.

[Kingma and Welling, 2014] Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *ICLR*.

[Kruschke, 2014] Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

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

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

[Long et al., 2013] 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.

[Marchant et al., 2014] Marchant, R., Ramos, F., Sanner, S., et al. (2014). Sequential bayesian optimisation for spatial-temporal monitoring. In *UAI*, pages 553–562.

[McLeod et al., 2017] McLeod, M., Osborne, M. A., and Roberts, S. J. (2017). Practical bayesian optimization for variable cost objectives. *arXiv preprint arXiv:1703.04335*.

[Medin and Schaffer, 1978] Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, 85(3):207.

[Myung et al., 2013] Myung, J. I., Cavagnaro, D. R., and Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of mathematical psychology*, 57(3-4):53–67.

[Nowak, 2009] Nowak, R. (2009). Noisy generalized binary search. In *Advances in neural information processing systems*, pages 1366–1374.

[Nowozin et al., 2016] Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279.

[Ouyang et al., 2016] Ouyang, L., Tessler, M. H., Ly, D., and Goodman, N. (2016). Practical optimal experiment design with probabilistic programs. *arXiv preprint arXiv:1608.05046*.

[Paige and Wood, 2016] Paige, B. and Wood, F. (2016). Inference networks for sequential monte carlo in graphical models. In *International Conference on Machine Learning*, pages 3040–3049.

[Pang et al., 2018] Pang, K., Dong, M., Wu, Y., and Hospedales, T. (2018). Meta-learning transferable active learning policies by deep reinforcement learning. *arXiv preprint arXiv:1806.04798*.

[Paninski, 2005] Paninski, L. (2005). Asymptotic theory of information-theoretic experimental design. *Neural Computation*, 17(7):1480–1507.

[Pronzato, 2010] Pronzato, L. (2010). One-step ahead adaptive d-optimal design on a finite design space is asymptotically optimal. *Metrika*, 71(2):219–238.

[Rainforth et al., 2018] Rainforth, T., Cornish, R., Yang, H., Warrington, A., and Wood, F. (2018). On nesting monte carlo estimators. In *International Conference on Machine Learning*, pages 4264–4273.

[Rezende and Mohamed, 2015] Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*.

[Rezende et al., 2014] Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *ICML*.

[Robbins and Monro, 1951] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

[Rouder, 2014] Rouder, J. N. (2014). Optional stopping: No problem for bayesians. *Psychonomic Bulletin & Review*, 21(2):301–308.

[Ryan et al., 2018] Ryan, E., Drovandi, C., McGree, J., and Pettitt, A. (2018). Fully bayesian optimal experimental design: A review. *Preprint eprints.qut.edu.au/75000/1/75000.pdf*.

[Ryan et al., 2015] Ryan, E. G., Drovandi, C. C., and Pettitt, A. N. (2015). Fully bayesian experimental design for pharmacokinetic studies. *Entropy*, 17(3):1063–1089.

[Shafer et al., 2011] Shafer, G., Shen, A., Vereshchagin, N., Vovk, V., et al. (2011). Test martingales, bayes factors and p-values. *Statistical Science*, 26(1):84–101.

[Shahriari et al., 2016] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175.

[Srinivas et al., 2009] Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

[Stuhlmüller et al., 2013] Stuhlmüller, A., Taylor, J., and Goodman, N. (2013). Learning stochastic inverses. In *Advances in neural information processing systems*, pages 3048–3056.

[Tabak and Turner, 2013] Tabak, E. and Turner, C. V. (2013). A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164.

[van Den Berg et al., 2003] van Den Berg, J., Curtis, A., and Trampert, J. (2003). Optimal nonlinear bayesian experimental design: an application to amplitude versus offset experiments. *Geophysical Journal International*, 155(2):411–421.

[Vanlier et al., 2012] Vanlier, J., Tiemann, C. A., Hilbers, P. A., and van Riel, N. A. (2012). A bayesian approach to targeted experiment design. *Bioinformatics*, 28(8):1136–1142.

[Vincent and Rainforth, 2017] Vincent, B. T. and Rainforth, T. (2017). The darc toolbox: automated, flexible, and efficient delayed and risky choice experiments using bayesian adaptive design.