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

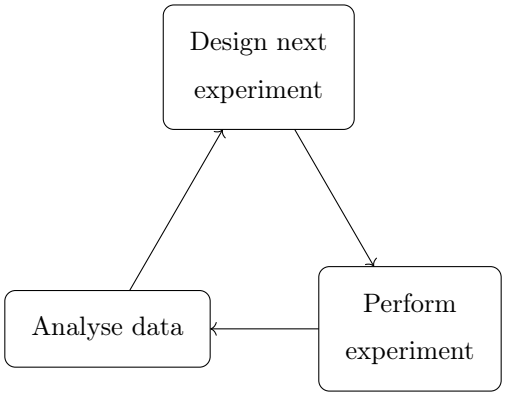
This thesis aims to describe the work I have done so far in my DPhil and discuss the future directions I plan to take in my research. Of primary relevance to my current work and future plans is Chapter 4 on optimal experiment design. Chapter 2 mostly concerns a project I worked on with Benjamin Bloem-Reddy which resulted in a UAI paper and oral presentation [Bloem-Reddy et al., 2018]. Chapter 3 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 4, 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 5, 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 fields of optimal experiment design (OED) and active learning are 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 – rendering previously infeasible experimental programs feasible. (citation needed)

Idealised active learning can be represented as follows

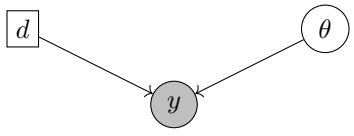


It is only within the context of a proposed data analysis that the optimality or suboptimality of an experiment design can be assessed.

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

Suppose we 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

$$U(\theta, d) = \int p(y|\theta, d)\ U(y, \theta, d)\ dy \tag{1.2}$$

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 $U(d) = U(\hat{\theta}, d)$

The optimal design would then be

$$d^* = \text{argmax}_{d\in\mathcal{D}}\ U(d) \tag{1.3}$$

where $\mathcal{D}$ is the space of admissible designs. 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], [Ryan et al., 2015] 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 [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.

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

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

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

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

$$H\left(\, p(y|d) \, \right) \tag{1.16}$$

the aleatoric uncertainty under parameter θ is

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

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

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

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

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

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

Designing a sequence of multiple experiments, with a view to maximise expected utility can be viewed as a Partially Observable Markov Decision Process (POMDP), 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 π from histories h_t to actions a_t which maximises the total reward

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^t r_t \mid \pi\right] \tag{1.22}$$

where $\gamma \in [0, 1]$ is the discount factor. In a finite horizon setting, we typically set this to 1. When there are only terminal rewards and $\gamma = 1$, this reduces to

$$R_T = \mathbb{E}[r_T \mid \pi] \tag{1.23}$$

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

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

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., 2015, sec 6.1] for a summary of 'backwards induction' approaches.

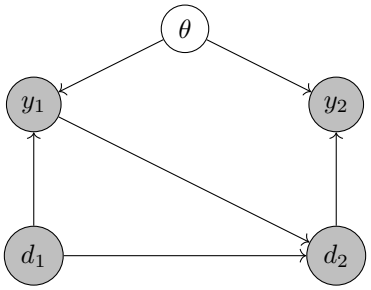
Optional stopping

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.

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

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

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

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

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

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

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

$$\begin{aligned} \mathrm{MI}(y, \theta) &= \int p(\theta) \int p(y|\theta) \log \frac{p(\theta|y)}{p(\theta)} \, dy \, d\theta \qquad (1.53) \\ &\geq \int p(\theta) \int p(y|\theta) \log \frac{q(\theta|y)}{p(\theta)} \, dy \, d\theta. \qquad (1.54) \end{aligned}$$

For a fuller derivation and discussion of this and related bounds, see Chapter **??** 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

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

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

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

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

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

the optimal design. As outlined in the previous section, we can have only approximate estimates of $\mathrm{EIG}(d)$. This puts us squarely in the domain of Bayesian optimisation [**?**]. When $\mathcal{D}$ is finite, we would term it a multi-armed 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 [?] 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 [?].

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. [?] tackled a more general problem of variable cost objectives, taking a GP based approach.

1.4 Applications

TODO: Major revisions needed

Machine learning and statistics

There is long-standing interest in 'classical' statistical models and their design [Youssef,]. Consider a basic linear models with Gaussian noise. Optimal design here can be expressed in terms of the eigenspectrum of XX^T (see [Chaloner, 1984]). For nonlinear models, see the section on Physics. What about GLMs? Likely can solve the problem analytically again. These are great baselines. People in the linear models case are often concerned with proving the equivalence of different kinds of optimality [Youssef,].

In machine learning, experiment design is closely related to two common techniques in, for example, image classification: data augmentation and active learning.

In data augmentation images are rotated, translated, etc to create more training data. We could theoretically optimise the augmentation but this seems wasteful since copying the labels to new images is very easy.

A much more interesting area is *active learning*. In this context, there are a large number of unlabeled images. Labeling is expensive. We select which images to label either up front, or (more typical in active learning) in a sequential manner. The key difference here is that we have a finite pool of unlabeled instances. We may be more interested in reducing uncertainty in the labels of these unlabeled images than in our posterior entropy.

The connection between active learning and Bayesian optimal design was explored in [Golovin et al., 2010]. In this paper, they start from a place where the outcome of a test is deterministic (think of the 12 men on an island problem). 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 OED criterion is introduced later to account for noisy observations and the fact that true parameters need not be known exactly even after all tests have been run.

A particular active learning example can be found in [Nowak, 2009]. We have $\mathcal{H}$ a hypothesis space (read parameter space) and $\mathcal{X}$ a query space (read design space). The goal is to determine the true $h^* \in \mathcal{H}$. Each query outputs a label in $\{-1, 1\}$ corrupted with Bernoulli noise (independent between

queries). The algorithm broadly works by targeting $x \in \mathcal{X}$ where the expected posterior label in near 0 (random guess). The convergence rate of $\mathbb{P}(\hat{h}_i \neq h^*) \to 0$ is studied (shown exponential). The importance of having access to unlabeled data is exploited by [Dasgupta, 2006].

Psychology

For an overview of optimal experiment design in probabilistic programming, [Ouyang et al., 2016] from Noah's group is a good place to start. Experiment design is necessary to distinguish competing theories. We should select models with the highest *expected information gain*, written formally as

$$U(d) = \mathbb{E}_{(Y,\Theta)\sim p(y,\theta|d)} \left\{ \log \frac{p(\Theta|Y,d)}{p(\Theta)} \right\} \tag{1.58}$$

This equation has been studied by mathematical statisticians since the 50's [Lindley, 1956]. We can naively evaluate $U(d)$ in a PPL via *nested inference*.

A canonical experiment discussed in this paper is 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* (learn categories by comparing new items to all previous items) and the *prototype model* (learn categories by remembering a prototypical example). There are two models, so $\Theta = \{m_1, m_2\}$. During the experiment, participants are presented with a sequence of objects. In the training phase, they are also told the correct label after guessing. In testing they have no feedback. The objects varied in four dimensions: colour, shape, size and count; we can consider the space of objects to be $\{0,1\}^4$. Each object has a label A or B. The true labeling mechanism was limited by Medin and Schaffer to be linearly separable. There are 9 inputs in the final testing set and Medin and Schaffer restricted there to be 5 As and 4 Bs. The objects 0000 and 1111 have to be present. Under these restrictions there are 933 possible experiments up to permutation. So $\mathcal{D}$ is a finite set of size 933. $\mathcal{Y}$ is the 'test' responses, ie. the subjects responses when they are not given feedback. Thus $\mathcal{Y} = \{A, B\}^9$. The kind of participant numbers seen were 10-30.

In [Vincent and Rainforth, 2017], the canonical experiment is as follows. We want to model how humans discount future rewards relative to present ones (via utility indifference pricing framework). 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. Formally, $\mathcal{D} = [0,\infty)^2$, $\mathcal{Y} = \{1,2\}$ and $\Theta = [0,\infty)$. These three spaces fit together as follows. We first chose $\mathcal{D}$ the space of possible designs. We subsequently chose $\mathcal{Y}$ the space of possible outcomes. We posited a probabilistic model for Y in terms of parameters θ. Focus on non-nested estimation for finite $\mathcal{Y}$ and sequential design. Sequential design means different participants will be asked different questions based on their previous answers.

Bioinformatics

In [Vanlier et al., 2012], the authors consider experiment design from the perspective that, with little data, many different parameter settings adequately describe the data. Canonical model. Biochemical

network modeled as an ODE.

$$\dot{x} = f(x, u, p)$$
$$\dot{y} = g(x, q) + \xi$$
$$x(0) = x_0$$

u is the input, x, y are time varying (uncontrolled) with x latent and y observed, p, q, x_0 are parameters θ required to simulate the model and do not depend on t, ξ represents measurement noise. We treat ξ as iid Gaussian. The paradigm chosen here is expected variance reduction, as opposed to information gain. (Possibly wrong if people still do that.) The variance is in the posterior predictive density.

Physics

In [van Den Berg et al., 2003], we begin by discussing 'classical' experiment design procedures which assume linear dependence between model and outcome $y = G_{m_0} d$. One can solve this linear equation by least squares, $\hat{d} = G^T(GG^T)^{-1}y$. Define $L = G^T(GG^T)^{-1}$, possibly adding regularization as necessary. Basically, you want to maximize the max eigenvalue of G, which is essentially a gradient. The larger gradient, the more informative the experiment. In a linear setting, the gradient does not depend on the true parameter value.

Now consider linear noise but a nonlinear function between parameters and outcomes. For example, the authors took

$$R_p = \left(\frac{1}{2}[1 + \tan^{-1} i] - 4c^2 \sin^2 i\right) \frac{\Delta\alpha}{\alpha} \tag{1.59}$$

where $\alpha = (\alpha 1 + \alpha 2)/2$, $\Delta\alpha = (\alpha 2 - \alpha 1)$. The parameter we want to optimize is $\alpha 2$.

This is a relatively simple and comprehensible case.

Chapter 2

Probabilistic modelling and inference

Chapter 3

Probabilistic programming

Chapter 4

Optimal experiment design

Chapter 5

Future directions

5.1 EIG

5.1.1 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.

5.1.2 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.

5.1.3 Theory of EIG estimators

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

5.1.4 EIG gradients

How best to estimate the gradient ∂_dEIG? 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.

5.1.5 EIG optimisation

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

5.1.6 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?

5.1.7 Sequential design and active learning

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

5.1.8 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?

5.1.9 Dynamic models

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

5.2 Beyond EIG

5.2.1 Causal inference

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

5.2.2 Power

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

5.2.3 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).

5.2.4 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?

5.2.5 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.

5.2.6 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 6

Appendix

6.1 Multi-step experiment design as reinforcement learning

6.1.1 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. The practical state s'_t consists of the sufficient statistics for θ obtained from h_t. Occasionally it is feasible 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 $o_t = (d_t, Y_t)$. Here $Y_t \sim p(y|\theta, d)$ is the outcome of performing the experiment using design d_t.
- 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 π from histories h_t to actions a_t which maximises the total reward

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^t r_t \mid \pi\right] \tag{6.1}$$

where $\gamma \in [0, 1]$ is the discount factor. In a finite horizon setting, we typically set this to 1.

6.1.2 Connection to information-criterion

Horizon 1

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

$$r(1, \theta, h) = \log \frac{p(\theta|y, d)}{p(\theta)} = \log \frac{p(y|\theta, d)}{p(y|d)} \tag{6.2}$$

The Q-function of action d_1 is the expected reward

$$Q(s_0, d_1) = E_{Y \sim p(y|\theta,d)}[r(t, \theta, d_1, Y)] \tag{6.3}$$

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

$$Q(p(\theta), d_1) = E_{\Theta \sim p(\theta)}\{E_{Y \sim p(y|\Theta,d)}[r(t, \Theta, d_1, Y)]\} \tag{6.4}$$

which reduces to the familiar expression

$$Q(p(\theta), d_1) = \int p(y, \theta|d) \log \frac{p(\theta|y, d)}{p(\theta)} d\theta dy \tag{6.5}$$

Horizon T

This formalism provides a convenient way to avoid the greedy approach to sequential design.

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

$$r(T, \theta, h_T) = r(T, \theta, b_T(\theta)) = \log \frac{b_T(\theta)}{p(\theta)} \tag{6.6}$$

we have updated b according to Bayes Theorem so

$$b_T(\theta) = p(\theta|Y_{1:T}, d_{1:T}) \tag{6.7}$$

Bibliography

[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.

[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, 1984] Chaloner, K. (1984). Optimal bayesian experimental design for linear models. *The Annals of Statistics*, pages 283–300.

[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.

[Dasgupta, 2006] Dasgupta, S. (2006). Coarse sample complexity bounds for active learning. In *Advances in neural information processing systems*, pages 235–242.

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

[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.

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

[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.

[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*.

[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.

[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.

[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.

[Youssef,] Youssef, N. A. A review on optimal experimental design.