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

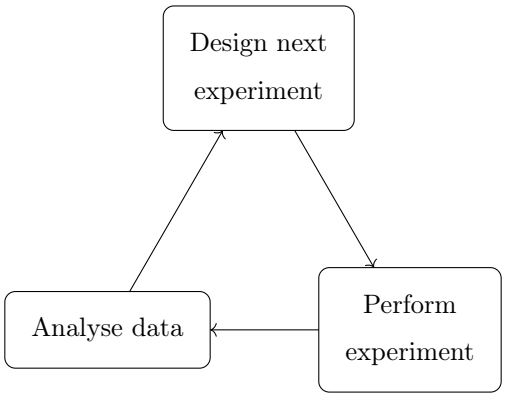
This thesis aims to describe the work I have done so far in my DPhil and discuss the future directions I plan to take in my research. Of primary relevance to my current work and future plans is Chapter 4 on optimal experiment design. Chapter 2 mostly concerns a project I worked on with Benjamin Bloem-Reddy 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 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

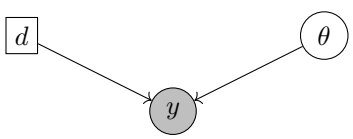


It is only within the context of a proposed data analysis that the optimality or suboptimality of an experiment design can be assessed.

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. In iterated experiment design, we assume that

θ remains fixed between experiments but d is reselected and y resampled at each experiment. The probabilistic data analysis model is represented by $p(y|\theta, d)$.

1.1.2 Criteria for 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

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

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], [Ryan et al., 2015] for further discussion of Bayesian experimental design.

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

$$U(y, \theta, d) = \left(\frac{\partial}{\partial \theta} \log p(y|\theta, d)\right)^2 \tag{1.2}$$

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

$$\mathcal{I}(\theta, d) = \int \left(\frac{\partial}{\partial \theta} \log p(y|\theta, d)\right)^2 p(y|\theta, d)\ dy \tag{1.3}$$

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 $U(d, \theta) = \operatorname{tr} \mathcal{I}(\theta, d)$
- 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

$$U(y, \theta, d) = \log \frac{p(\theta|y, d)}{p(\theta)} = \log \frac{p(y|\theta, d)}{p(y|d)} \tag{1.4}$$

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.

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

$$U(d) = \text{EIG}(d) = \int p(y, \theta|d) \log \frac{p(\theta|y, d)}{p(\theta)} \, dy \, d\theta \tag{1.5}$$

The EIG can be interpreted in a number of ways

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

$$\text{IG}(y, d) = \text{KL}(\ p(\theta|y, d) \ ||\ p(\theta)\ ) \tag{1.6}$$

   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

$$\text{APE}(d) = \int p(y, \theta|d) \log p(\theta|y, d) \, dy \, d\theta \tag{1.7}$$

$$= -\int p(y|d) H(\ p(\theta|y, d)\ ) \tag{1.8}$$

   where H is the differential entropy. Then

$$\text{EIG}(d) = H(\ p(\theta)\ ) - \text{APE}(d) \tag{1.9}$$

   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

$$\text{MI}(x, y) = \text{KL}\,(\ p(x, y)\ ||\ p(x)p(y)\ ) \tag{1.10}$$

   then we have

$$\text{MI}(y, \theta|d) = \text{KL}\,(\ p(y, \theta|d)\ ||\ p(y|d)p(\theta)\ ) \tag{1.11}$$

$$= \int p(y, \theta|d) \log \frac{p(y, \theta|d)}{p(\theta)p(y|d)} \tag{1.12}$$

$$= \text{EIG}(d) \tag{1.13}$$

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

$$H(\ p(y|d)\ ) \tag{1.14}$$

   the aleatoric uncertainty under parameter θ is

$$H(\ p(y|\theta, d)\ ) \tag{1.15}$$

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

$$\mathbb{E}_{\theta \sim p(\theta)} \left[H \left(\, p(y|\theta, d) \, \right) \right] \tag{1.16}$$

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

$$H \left(\, p(y|d) \, \right) - \mathbb{E}_{\theta \sim p(\theta)} \left[H \left(\, p(y|\theta, d) \, \right) \right] \tag{1.17}$$

$$= - \int p(y|d) \log p(y|d) + \int p(y, \theta|d) \log p(y|\theta, d) \tag{1.18}$$

$$= \text{EIG}(d) \tag{1.19}$$

1.1.3 Criteria for 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, 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. See **??** for an explicit representation of experiment design as RL. A typical approximate strategy is to optimise the expected utility at each experiment (the greedy strategy), although 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.

Within

1.1.4 Statistical validity

References for consistency

1.1.5 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

$$U(d) = \mathbb{E}_{(Y,\Theta)\sim p(y,\theta|d)} \left\{ \log \frac{p(\Theta|Y,d)}{p(\Theta)} \right\} \tag{1.20}$$

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

$$R_p = \left(\frac{1}{2} [1 + \tan^{-1} i] - 4c^2 \sin^2 i \right) \frac{\Delta\alpha}{\alpha} \tag{1.21}$$

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 $\partial_d \text{EIG}$? 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.

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).

5.3 Multi-step experiment design as reinforcement learning

5.3.1 Setup

Finite time horizon $t = 1, ..., T$.

- 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. These are the states we will encounter in our trees. Occasionally it is feasible to compute the belief states b_t, encoding the full posterior for θ given that history.
- Actions $a_t = d_{t+1}$. Transitions are fully deterministic.
- 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 $r_t = r(t, \theta, d_t, Y_t)$. We take r to be a non-random function. The rewards are functions of the random observations.

5.3.2 Connection to information-criterion

Horizon 1

Suppose $T = 1$. And choose the following reward function

$$r(t, \theta, d, y) = \log \frac{p(\theta|y, d)}{p(\theta)} = \log \frac{p(y|\theta, d)}{p(y|d)} \tag{5.1}$$

The Q-function of action d_1 is the expected reward

$$Q(\cdot, d_1) = E_{Y \sim p(y|\theta,d)}[r(t, \theta, d_1, Y)] \tag{5.2}$$

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

$$Q(p(\theta), d_1) = E_{\Theta \sim p(\theta)}\{E_{Y \sim p(y|\Theta,d)}[r(t, \Theta, d_1, Y)]\} \tag{5.3}$$

which reduces to the familiar expression

$$Q(p(\theta), d_1) = \int p(y, \theta|d) \log \frac{p(\theta|y, d)}{p(\theta)} d\theta dy \tag{5.4}$$

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

$$r(T, \theta, b_T(\theta)) = \log \frac{b_T(\theta)}{p(\theta)} \tag{5.5}$$

we have updated b according to Bayes Theorem so

$$b_T(\theta) = p(\theta|Y_{1:T}, d_{1:T}) \tag{5.6}$$

Bibliography

[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.

[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.

[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.

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

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

[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*.

[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.

[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.