Variational Bayesian Optimal Experimental Design

Adam Foster[†*] **Martin Jankowiak**[‡] **Eli Bingham**[‡] **Paul Horsfall**[‡]
Yee Whye Teh[†] **Tom Rainforth**[†] **Noah Goodman**[‡§]
[†]Department of Statistics, University of Oxford, Oxford, UK
[‡]Uber AI Labs, Uber Technologies Inc., San Francisco, CA, USA
[§]Stanford University, Stanford, CA, USA
adam.foster@stats.ox.ac.uk

Abstract

Bayesian optimal experimental design (BOED) is a principled framework for making efficient use of limited experimental resources. Unfortunately, its applicability is hampered by the difficulty of obtaining accurate estimates of the expected information gain (EIG) of an experiment. To address this, we introduce several classes of fast EIG estimators by building on ideas from amortized variational inference. We show theoretically and empirically that these estimators can provide significant gains in speed and accuracy over previous approaches. We further demonstrate the practicality of our approach on a number of end-to-end experiments.

1 Introduction

Tasks as seemingly diverse as designing a study to elucidate human cognition, selecting the next query point in an active learning loop, and designing online feedback surveys all constitute the same underlying problem: designing an experiment to maximize the information gathered. Bayesian optimal experimental design (BOED) forms a powerful mathematical abstraction for tackling such problems [8, 23, 37, 43] and has been successfully applied in numerous settings, including psychology [30], Bayesian optimization [16], active learning [15], bioinformatics [42], and neuroscience [38].

In the BOED framework, we construct a predictive model $p(y|\theta, d)$ for possible experimental outcomes y, given a design d and a particular value of the parameters of interest θ. We then choose the design that optimizes the expected information gain (EIG) in θ from running the experiment,

$$\text{EIG}(d) \triangleq \mathbb{E}_{p(y|d)}\big[H[p(\theta)] - H[p(\theta|y,d)]\big], \tag{1}$$

where $H[\cdot]$ represents the entropy and $p(\theta|y,d) \propto p(\theta)p(y|\theta,d)$ is the posterior resulting from running the experiment with design d and observing outcome y. In other words, we seek the design that, in expectation over possible experimental outcomes, most reduces the entropy of the posterior over our target latent variables. If the predictive model is correct, this forms a design strategy that is (one-step) optimal from an information-theoretic viewpoint [24, 37].

The BOED framework is particularly powerful in sequential contexts, where it allows the results of previous experiments to be used in guiding the designs for future experiments. For example, as we ask a participant a series of questions in a psychology trial, we can use the information gathered from previous responses to ask more pertinent questions in the future, that will, in turn, return more information. This ability to design experiments that are self-adaptive can substantially increase their efficiency: fewer iterations are required to uncover the same level of information.

In practice, however, the BOED approach is often hampered by the difficulty of obtaining fast and high-quality estimates of the EIG: due to the intractability of the posterior $p(\theta|y,d)$, it constitutes

[*] Part of this work was completed by AF during an internship with Uber AI Labs.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

a nested expectation problem and so conventional Monte Carlo (MC) estimation methods cannot be applied [33]. Moreover, existing methods for tackling nested expectations have, in general, far inferior convergence rates than those for conventional expectations [22, 30, 32]. For example, nested MC (NMC) can only achieve, at best, a rate of $\mathcal{O}(T^{-1/3})$ in the total computational cost T [33], compared with $\mathcal{O}(T^{-1/2})$ for conventional MC.

To address this, we propose a variational BOED approach that sidesteps the double intractability of the EIG in a principled manner and yields estimators with convergence rates in line with those for conventional estimation problems. To this end, we introduce four efficient and widely applicable variational estimators for the EIG. The different methods each present distinct advantages. For example, two allow training with implicit likelihood models, while one allows for asymptotic consistency even when the variational family does not contain the target distribution.

We theoretically confirm the advantages of our estimators, showing that they all have a convergence rate of $\mathcal{O}(T^{-1/2})$ when the variational family contains the target distribution. We further verify their practical utility using a number of experiment design problems inspired by applications from science and industry, showing that they provide significant empirical gains in EIG estimation over previous methods and that these gains lead, in turn, to improved end-to-end performance.

To maximize the space of potential applications and users for our estimators, we provide[2] a general-purpose implementation of them in the probabilistic programming system Pyro [5], exploiting Pyro's first-class support for neural networks and variational methods.

2 Background

The BOED framework is a model-based approach for choosing an experiment design d in a manner that optimizes the information gained about some parameters of interest θ from the outcome y of the experiment. For instance, we may wish to choose the question d in a psychology trial to maximize the information gained about an underlying psychological property of the participant θ from their answer y to the question. In general, we adopt a Bayesian modelling framework with a prior $p(\theta)$ and a predictive model $p(y|\theta, d)$. The information gained about θ from running experiment d and observing y is the reduction in entropy from the prior to the posterior:

$$\mathrm{IG}(y, d) = H[p(\theta)] - H[p(\theta|y, d)]\,. \tag{2}$$

At the point of choosing d, however, we are uncertain about the outcome. Thus, in order to define a metric to assess the utility of the design d we take the expectation of $\mathrm{IG}(y, d)$ under the marginal distribution over outcomes $p(y|d) = \mathbb{E}_{p(\theta)}[p(y|\theta, d)]$ as per (1). We can further rearrange this as

$$\mathrm{EIG}(d) = \mathbb{E}_{p(y,\theta|d)}\left[\log \frac{p(\theta|y,d)}{p(\theta)}\right] = \mathbb{E}_{p(y,\theta|d)}\left[\log \frac{p(y,\theta|d)}{p(\theta)p(y|d)}\right] = \mathbb{E}_{p(y,\theta|d)}\left[\log \frac{p(y|\theta,d)}{p(y|d)}\right] \tag{3}$$

with the result that the EIG can also be interpreted as the mutual information between θ and y given d, or the epistemic uncertainty in y averaged over the prior $p(\theta)$. The Bayesian optimal design is defined as $d^* \triangleq \arg\max_{d\in\mathcal{D}} \mathrm{EIG}(d)$, where $\mathcal{D}$ is the set of permissible designs.

Computing the EIG is challenging since neither $p(\theta|y, d)$ or $p(y|d)$ can, in general, be found in closed form. Consequently, the integrand is intractable and conventional MC methods are not applicable. One common way of getting around this is to employ a nested MC (NMC) estimator [30, 43]

$$\hat{\mu}_{\mathrm{NMC}}(d) \triangleq \frac{1}{N}\sum_{n=1}^{N} \log \frac{p(y_n|\theta_{n,0}, d)}{\frac{1}{M}\sum_{m=1}^{M} p(y_n|\theta_{n,m}, d)} \quad \text{where} \quad \theta_{n,m} \overset{\text{i.i.d.}}{\sim} p(\theta),\ y_n \sim p(y|\theta = \theta_{n,0}, d). \tag{4}$$

Rainforth et al. [33] showed that this estimator, which has a total computational cost $T = \mathcal{O}(NM)$, is consistent in the limit $N, M \to \infty$ with RMSE convergence rate $\mathcal{O}(N^{-1/2} + M^{-1})$, and that it is asymptotically optimal to set $M \propto \sqrt{N}$, yielding an overall rate of $\mathcal{O}(T^{-1/3})$.

Given a base EIG estimator, a variety of different methods can be used for the subsequent optimization over designs, including some specifically developed for BOED [1, 29, 32]. In our experiments, we

[2] Implementations of our methods are available at `http://docs.pyro.ai/en/stable/contrib.oed.html`. To reproduce the results in this paper, see `https://github.com/ae-foster/pyro/tree/vboed-reproduce`.

will adopt Bayesian optimization [39], due to its sample efficiency, robustness to multi-modality, and ability to deal naturally with noisy objective evaluations. However, we emphasize that our focus is on the base EIG estimator and that our estimators can be used more generally with different optimizers.

The static design setting we have implicitly assumed thus far in our discussion can be generalized to sequential contexts, in which we design T experiments $d_1, ..., d_T$ with outcomes $y_1, ..., y_T$. We assume experiment outcomes are conditionally independent given the latent variables and designs, i.e.

$$p(y_{1:T}, \theta | d_{1:T}) = p(\theta) \prod_{t=1}^{T} p(y_t | \theta, d_t). \tag{5}$$

Having conducted experiments $1, ..., t-1$, we can design d_t by incorporating data in the standard Bayesian fashion: at experiment iteration t, we replace the prior $p(\theta)$ in (3) with $p(\theta | d_{1:t-1}, y_{1:t-1})$, the posterior conditional on the first $t-1$ designs and outcomes. We can thus conduct an adaptive sequential experiment in which we optimize the choice of the design d_t at each iteration.

3 Variational Estimators

Though consistent, the convergence rate of the NMC estimator is prohibitively slow for many practical problems. As such, EIG estimation often becomes the bottleneck for BOED, particularly in sequential experiments where the BOED calculations must be fast enough to operate in real-time.

In this section we show how ideas from amortized variational inference [10, 17, 34, 40] can be used to sidestep the double intractability of the EIG, yielding estimators with much faster convergence rates thereby alleviating the EIG bottleneck. A key insight for realizing why such fundamental gains can be made is that the NMC estimator is inefficient because a *separate* estimate of the integrand in (3) is made for each y_n. The variational approaches we introduce instead look to directly learn a *functional approximation*—for example, an approximation of $y \mapsto p(y|d)$—and then evaluate this approximation at multiple points to estimate the integral, thereby allowing information to be shared across different values of y. If M evaluations are made in learning the approximation, the total computational cost is now $T = \mathcal{O}(N + M)$, yielding substantially improved convergence rates.

Variational posterior $\hat{\mu}_{\text{post}}$ Our first approach, which we refer to as the variational posterior estimator $\hat{\mu}_{\text{post}}$, is based on learning an amortized approximation $q_p(\theta|y, d)$ to the posterior $p(\theta|y, d)$ and then using this to estimate the EIG:

$$\text{EIG}(d) \approx \mathcal{L}_{\text{post}}(d) \triangleq \mathbb{E}_{p(y,\theta|d)} \left[\log \frac{q_p(\theta|y, d)}{p(\theta)} \right] \approx \hat{\mu}_{\text{post}}(d) \triangleq \frac{1}{N} \sum_{n=1}^{N} \log \frac{q_p(\theta_n|y_n, d)}{p(\theta_n)}, \tag{6}$$

where $y_n, \theta_n \overset{\text{i.i.d.}}{\sim} p(y, \theta|d)$ and $\hat{\mu}_{\text{post}}(d)$ is a MC estimator of $\mathcal{L}_{\text{post}}(d)$. We draw samples of $p(y, \theta|d)$ by sampling $\theta \sim p(\theta)$ and then $y|\theta \sim p(y|\theta, d)$. We can think of this approach as amortizing the cost of the inner expectation, instead of running inference separately for each y.

To learn a suitable $q_p(\theta|y, d)$, we show in Appendix A that $\mathcal{L}_{\text{post}}(d)$ forms a variational lower bound $\text{EIG}(d) \geq \mathcal{L}_{\text{post}}(d)$ that is tight if and only if $q_p(\theta|y, d) = p(\theta|y, d)$. Barber and Agakov [3] used this bound to estimate mutual information in the context of transmission over noisy channels, but the connection to experiment design has not previously been made.

This result means we can learn $q_p(\theta|y, d)$ by introducing a family of variational distributions $q_p(\theta|y, d, \phi)$ parameterized by ϕ and then maximizing the bound with respect to ϕ:

$$\phi^* = \arg\max_{\phi} \mathbb{E}_{p(y,\theta|d)} \left[\log \frac{q_p(\theta|y, d, \phi)}{p(\theta)} \right], \qquad \text{EIG}(d) \approx \mathcal{L}_{\text{post}}(d; \phi^*). \tag{7}$$

Provided that we can generate samples from the model, this maximization can be performed using stochastic gradient methods [35] and the unbiased gradient estimator

$$\nabla_\phi \mathcal{L}_{\text{post}}(d; \phi) \approx \tfrac{1}{S} \sum_{i=1}^{S} \nabla_\phi \log q_p(\theta_i|y_i, d, \phi) \quad \text{where} \quad y_i, \theta_i \overset{\text{i.i.d.}}{\sim} p(y, \theta|d), \tag{8}$$

and we note that no reparameterization is required as $p(y, \theta|d)$ is independent of ϕ. After K gradient steps we obtain variational parameters ϕ_K that approximate ϕ^*, which we use to compute

a corresponding EIG estimator by constructing a MC estimator for $\mathcal{L}_{\text{post}}(d;\phi)$ as per (6) with $q_p(\theta_n|y_n,d) = q_p(\theta_n|y_n,d,\phi_K)$. Interestingly, the tightness of $\mathcal{L}_{\text{post}}(d)$ turns out to be equal to the expected *forward* KL divergence[3] $\mathbb{E}_{p(y|d)}\left[\text{KL}\left(p(\theta|y,d)||q_p(\theta|y,d,\phi)\right)\right]$ so we can view this approach as learning an amortized proposal by minimizing this expected KL divergence.

Variational marginal $\hat{\mu}_{\text{marg}}$ In some scenarios, θ may be high-dimensional, making it difficult to train a good variational posterior approximation. An alternative approach that can be attractive in such cases is to instead learn an approximation $q_m(y|d)$ to the marginal density $p(y|d)$ and substitute this into the final form of the EIG in (3). As shown in Appendix A, this yields an *upper bound*

$$\text{EIG}(d) \leq \mathcal{U}_{\text{marg}}(d) \triangleq \mathbb{E}_{p(y,\theta|d)}\left[\log \frac{p(y|\theta,d)}{q_m(y|d)}\right] \approx \hat{\mu}_{\text{marg}}(d) \triangleq \frac{1}{N}\sum_{n=1}^{N} \log \frac{p(y_n|\theta_n,d)}{q_m(y_n|d)}, \quad (9)$$

where again $y_n, \theta_n \overset{\text{i.i.d.}}{\sim} p(y,\theta|d)$ and the bound is tight when $q_m(y|d) = p(y|d)$. Analogously to $\hat{\mu}_{\text{post}}$, we can learn $q_m(y|d)$ by introducing a variational family $q_m(y|d,\phi)$ and then performing stochastic gradient descent to *minimize* $\mathcal{U}_{\text{marg}}(d,\phi)$. As with $\hat{\mu}_{\text{post}}$, this bound was studied in a mutual information context [31], but it has not been utilized for BOED before.

Variational NMC $\hat{\mu}_{\text{VNMC}}$ As we will show in Section 4, $\hat{\mu}_{\text{post}}$ and $\hat{\mu}_{\text{marg}}$ can provide substantially faster convergence rates than NMC. However, this comes at the cost of converging towards a biased estimate if the variational family does not contain the target distribution. To address this, we propose another EIG estimator, $\hat{\mu}_{\text{VNMC}}$, which allows one to trade-off resources between the fast learning of a biased estimator permitted by variational approaches, and the ability of NMC to eliminate this bias.[4]

We can think of the NMC estimator as approximating $p(y|d)$ using M samples from the prior. At a high-level, $\hat{\mu}_{\text{VNMC}}$ is based around learning a proposal $q_v(\theta|y,d)$ and then using samples from this proposal to make an importance sampling estimate of $p(y|d)$, potentially requiring far fewer samples than NMC. Formally, it is based around a bound that can be arbitrarily tightened, namely

$$\text{EIG}(d) \leq \mathbb{E}\left[\log p(y|\theta_0,d) - \log \frac{1}{L}\sum_{\ell=1}^{L} \frac{p(y,\theta_\ell|d)}{q_v(\theta_\ell|y,d)}\right] \triangleq \mathcal{U}_{\text{VNMC}}(d,L) \quad (10)$$

where the expectation is taken over $y, \theta_{0:L} \sim p(y,\theta_0|d)\prod_{\ell=1}^{L} q_v(\theta_\ell|y,d)$, which corresponds to one sample y,θ_0 from the model and L samples from the approximate posterior conditioned on y. To the best of our knowledge, this bound has not previously been studied in the literature. As with $\hat{\mu}_{\text{post}}$ and $\hat{\mu}_{\text{marg}}$, we can minimize this bound to train a variational approximation $q_v(\theta|y,d,\phi)$. Important features of $\mathcal{U}_{\text{VNMC}}(d,L)$ are summarized in the following lemma; see Appendix A for the proof.

Lemma 1. *For any given model $p(\theta)p(y|\theta,d)$ and valid $q_v(\theta|y,d)$,*

1. $\text{EIG}(d) = \lim_{L\to\infty} \mathcal{U}_{VNMC}(d,L) \leq \mathcal{U}_{VNMC}(d,L_2) \leq \mathcal{U}_{VNMC}(d,L_1) \quad \forall L_2 \geq L_1 \geq 1$,
2. $\mathcal{U}_{VNMC}(d,L) = \text{EIG}(d) \;\; \forall L \geq 1 \quad$ *if* $\quad q_v(\theta|y,d) = p(\theta|y,d) \;\; \forall y,\theta$,
3. $\mathcal{U}_{VNMC}(d,L) - \text{EIG}(d) = \mathbb{E}_{p(y|d)}\left[\text{KL}\left(\prod_{\ell=1}^{L} q_v(\theta_\ell|y,d)\,\middle|\middle|\,\frac{1}{L}\sum_{\ell=1}^{L} p(\theta_\ell|y,d)\prod_{k\neq\ell} q_v(\theta_k|y,d)\right)\right]$

Like the previous bounds, the VNMC bound is tight when $q_v(\theta|y,d) = p(\theta|y,d)$. Importantly, the bound is also tight as $L \to \infty$, even for imperfect q_v. This means we can obtain asymptotically unbiased EIG estimates even when the true posterior is not contained in the variational family.

Specifically, we first train ϕ using K steps of stochastic gradient on $\mathcal{U}_{\text{VNMC}}(d,L)$ with some fixed L. To form a final EIG estimator, however, we use a MC estimator of $\mathcal{U}_{\text{VNMC}}(d,M)$ where typically $M \gg L$. This final estimator is a NMC estimator that is consistent as $N, M \to \infty$ with ϕ_K fixed

$$\hat{\mu}_{\text{VNMC}}(d) \triangleq \frac{1}{N}\sum_{n=1}^{N}\left(\log p(y_n|\theta_{n,0},d) - \log \frac{1}{M}\sum_{m=1}^{M} \frac{p(y_n,\theta_{n,m}|d)}{q_v(\theta_{n,m}|y_n,d,\phi_K)}\right) \quad (11)$$

where $\theta_{n,0} \overset{\text{i.i.d.}}{\sim} p(\theta)$, $y_n \sim p(y|\theta=\theta_{n,0},d)$ and $\theta_{n,m} \sim q_v(\theta|y=y_n,d,\phi_K)$. In practice, performance is greatly enhanced when the proposal q_v is a good, if inexact, approximation to the posterior. This significantly improves upon traditional $\hat{\mu}_{\text{NMC}}$, which sets $q_v(\theta|y,d) = p(\theta)$ in (11).

[3] See Appendix A for a proof. A comparison with the reverse KL divergence can be found in Appendix G.

[4] In Appendix F, we describe a method using $q_m(y|d)$ as a control variate that can also eliminate this bias and lower the variance of NMC, requiring additional assumptions about the model and variational family.

Implicit likelihood and $\hat{\mu}_{\mathbf{m}+\ell}$ So far we have assumed that we can evaluate $p(y|\theta, d)$ pointwise. However, many models of interest have *implicit likelihoods* from which we can draw samples, but not evaluate directly. For example, models with nuisance latent variables ψ (such as a random effect models) are implicit likelihood models because $p(y|\theta, d) = \mathbb{E}_{p(\psi|\theta)}[p(y|\theta, \psi, d)]$ is intractable, but can still be straightforwardly sampled from.

In this setting, $\hat{\mu}_{\text{post}}$ is applicable without modification because it only requires samples from $p(y|\theta, d)$ and *not* evaluations of this density. Although $\hat{\mu}_{\text{marg}}$ is not directly applicable in this setting, it can be modified to accommodate implicit likelihoods. Specifically, we can utilize *two* approximate densities: $q_m(y|d)$ for the marginal and $q_\ell(y|\theta, d)$ for the likelihood. We then form the approximation

$$\text{EIG}(d) \approx \mathcal{I}_{\text{m}+\ell}(d) \triangleq \mathbb{E}_{p(y,\theta|d)}\left[\log \frac{q_\ell(y|\theta, d)}{q_m(y|d)}\right] \approx \hat{\mu}_{\text{m}+\ell}(d) \triangleq \frac{1}{N}\sum_{n=1}^{N} \log \frac{q_\ell(y_n|\theta_n, d)}{q_m(y_n|d)}. \quad (12)$$

Unlike the previous three cases, $\mathcal{I}_{\text{m}+\ell}(d)$ is not a bound on $\text{EIG}(d)$, meaning it is not immediately clear how to train $q_m(y|d)$ and $q_\ell(y|\theta, d)$ to achieve an accurate EIG estimator. The following lemma shows that we can bound the EIG estimation *error* of $\mathcal{I}_{\text{m}+\ell}$. The proof is in Appendix A.

Lemma 2. *For any given model $p(\theta)p(y|\theta, d)$ and valid $q_m(y|d)$ and $q_\ell(y|\theta, d)$, we have*

$$|\mathcal{I}_{m+\ell}(d) - \text{EIG}(d)| \leq -\mathbb{E}_{p(y,\theta|d)}[\log q_m(y|d) + \log q_\ell(y|\theta, d)] + C, \quad (13)$$

where $C = -H[p(y|d)] - \mathbb{E}_{p(\theta)}[H(p(y|\theta, d)]$ does not depend on q_m or q_ℓ. Further, the RHS of (13) is 0 if and only if $q_m(y|d) = p(y|d)$ and $q_\ell(y|\theta, d) = p(y|\theta, d)$ for almost all y, θ.

This lemma implies that we can learn $q_m(y|d)$ and $q_\ell(y|\theta, d)$ by maximizing $\mathbb{E}_{p(y,\theta|d)}[\log q_m(y|d) + \log q_\ell(y|\theta, d)]$ using stochastic gradient ascent, and substituting these learned approximations into (12) for the final EIG estimator. To the best of our knowledge, this approach has not previously been considered in the literature. We note that, in general, q_m and q_ℓ are learned separately and there need not be any weight sharing between them. See Appendix A.4 for a discussion of the case when we couple q_m and q_ℓ so that $q_m(y|d) = \mathbb{E}_{p(\theta)}[q_\ell(y|\theta, d)]$.

Using estimators for sequential BOED In sequential settings, we also need to consider the implications of replacing $p(\theta)$ in the EIG with $p(\theta|d_{1:t-1}, y_{1:t-1})$. At first sight, it appears that, while $\hat{\mu}_{\text{marg}}$ and $\hat{\mu}_{\text{m}+\ell}$ only require samples from $p(\theta|d_{1:t-1}, y_{1:t-1})$, $\hat{\mu}_{\text{post}}$ and $\hat{\mu}_{\text{VNMC}}$ also require its density to be evaluated, a potentially severe limitation. Fortunately, we can, in fact, avoid evaluating this posterior density. We note that, from (5), we have $p(\theta|y_{1:t-1}, d_{1:t-1}) = p(\theta)\prod_{i=1}^{t-1} p(y_i|\theta, d_i)/p(y_{1:t-1}|d_{1:t-1})$. Substituting this into the integrand of (6) gives

$$\mathcal{L}_{\text{post}}(d_t) = \mathbb{E}_{p(\theta|y_{1:t-1},d_{1:t-1})p(y_t|\theta,d_t)}\left[\log \frac{q_p(\theta|y_t, d_t)}{p(\theta)\prod_{i=1}^{t-1} p(y_i|\theta, d_i)}\right] + \log p(y_{1:t-1}|d_{1:t-1}) \quad (14)$$

where $p(\theta)\prod_{i=1}^{t-1} p(y_i|\theta, d_i)$ can be evaluated exactly and the additive constant $\log p(y_{1:t-1}|d_{1:t-1})$ does not depend on the new design d_t, θ, or any of the variational parameters, and so can be safely ignored. Making the same substitution in (11) shows that we can also estimate $\mathcal{U}_{\text{VNMC}}(d_t, L)$ up to a constant, which can then be similarly ignored. As such, any inference scheme for sampling $p(\theta|d_{1:t-1}, y_{1:t-1})$, approximate or exact, is compatible with all our approaches.

Selecting an estimator Having proposed four estimators, we briefly discuss how to choose between them in practice. For reference, a summary of our estimators is given in Table 1, along with several baseline approaches. First, $\hat{\mu}_{\text{marg}}$ and $\hat{\mu}_{\text{m}+\ell}$ rely on approximating a distribution over y; $\hat{\mu}_{\text{post}}$ and $\hat{\mu}_{\text{VNMC}}$ approximate distributions over θ. We may prefer the former two estimators if $\dim(y) \ll \dim(\theta)$ as it leaves us with a simpler density estimation problem, and vice versa. Second, $\hat{\mu}_{\text{marg}}$ and $\hat{\mu}_{\text{VNMC}}$ require an

Table 1: Summary of EIG estimators. Baseline methods are explained in Section 5.

		Implicit	Bound	Consistent	Eq.
Ours	$\hat{\mu}_{\text{post}}$	✓	Lower	✗	(6)
	$\hat{\mu}_{\text{marg}}$	✗	Upper	✗	(9)
	$\hat{\mu}_{\text{VNMC}}$	✗	Upper	✓	(11)
	$\hat{\mu}_{\text{m}+\ell}$	✓	✗	✗	(12)
Baseline	$\hat{\mu}_{\text{NMC}}$	✗	Upper	✓	(4)
	$\hat{\mu}_{\text{laplace}}$	✗	✗	✗	(75)
	$\hat{\mu}_{\text{LFIRE}}$	✓	✗	✗	(76)
	$\hat{\mu}_{\text{DV}}$	✓	Lower	✗	(77)

explicit likelihood whereas $\hat{\mu}_{\text{post}}$ and $\hat{\mu}_{\text{m}+\ell}$ do not. If an explicit likelihood is available, it typically makes sense to use it—one would never use $\hat{\mu}_{\text{m}+\ell}$ over $\hat{\mu}_{\text{marg}}$ for example. Finally, if the variational families do not contain the target densities, $\hat{\mu}_{\text{VNMC}}$ is the only method guaranteed to converge to the true EIG(d) in the limit as the computational budget increases. So we might prefer $\hat{\mu}_{\text{VNMC}}$ when computation time and cost are not constrained.

4 Convergence rates

We now investigate the convergence of our estimators. We start by breaking the overall error down into three terms: I) variance in MC estimation of the bound; II) the gap between the bound and the tightest bound possible given the variational family; and III) the gap between the tightest possible bound and EIG(d). With variational EIG approximation $\mathcal{B}(d) \in \{\mathcal{L}_{\text{post}}(d), \mathcal{U}_{\text{marg}}(d), \mathcal{U}_{\text{VNMC}}(d, L), \mathcal{I}_{\text{m}+\ell}(d)\}$, optimal variational parameters ϕ^*, learned variational parameters ϕ_K after K stochastic gradient iterations, and MC estimator $\hat{\mu}(d, \phi_K)$ we have, by the triangle inequality,

$$\|\hat{\mu}(d,\phi_K) - \text{EIG}(d)\|_2 \leq \underbrace{\|\hat{\mu}(d,\phi_K) - \mathcal{B}(d,\phi_K)\|_2}_{\text{I}} + \underbrace{\|\mathcal{B}(d,\phi_K) - \mathcal{B}(d,\phi^*)\|_2}_{\text{II}} + \underbrace{|\mathcal{B}(d,\phi^*) - \text{EIG}(d)|}_{\text{III}}$$

where we have used the notation $\|X\|_2 \triangleq \sqrt{\mathbb{E}[X^2]}$ to denote the L^2 norm of a random variable.

By the weak law of large numbers, term I scales as $N^{-1/2}$ and can thus be arbitrarily reduced by taking more MC samples. Provided that our stochastic gradient scheme converges, term II can be reduced by increasing the number of stochastic gradient steps K. Term III, however, is a constant that can only be reduced by expanding the variational family (or increasing L for $\hat{\mu}_{\text{VNMC}}$). Each approximation $\mathcal{B}(d)$ thus converges to a biased estimate of the EIG(d), namely $\mathcal{B}(d, \phi^*)$. As established by the following Theorem, if we set $N \propto K$, the rate of convergence to this biased estimate is $\mathcal{O}(T^{-1/2})$, where T represents the total computational cost, with $T = \mathcal{O}(N + K)$.

Theorem 1. *Let $\mathcal{X}$ be a measurable space and Φ be a convex subset of a finite dimensional inner product space. Let $X_1, X_2, ...$ be i.i.d. random variables taking values in $\mathcal{X}$ and $f : \mathcal{X} \times \Phi \to \mathbb{R}$ be a measurable function. Let*

$$\mu(\phi) \triangleq \mathbb{E}[f(X_1, \phi)] \approx \hat{\mu}_N(\phi) \triangleq \frac{1}{N} \sum_{n=1}^{N} f(X_n, \phi)$$

and suppose that $\sup_{\phi \in \Phi} \|f(X_1, \phi)\|_2 < \infty$. Then $\sup_{\phi \in \Phi} \|\hat{\mu}_N(\phi) - \mu(\phi)\|_2 = \mathcal{O}(N^{-1/2})$. Suppose further that Assumption 1 in Appendix B holds and that ϕ^* is the unique minimizer of μ. After K iterations of the Polyak-Ruppert averaged stochastic gradient descent algorithm of [28] with gradient estimator $\nabla_\phi f(X_t, \phi)$, we have $\|\mu(\phi_K) - \mu(\phi^*)\|_2 = \mathcal{O}(K^{-1/2})$ and, combining with the first result,

$$\|\hat{\mu}_N(\phi_K) - \mu(\phi^*)\|_2 = \mathcal{O}(N^{-1/2} + K^{-1/2}) = \mathcal{O}(T^{-1/2}) \text{ if } N \propto K.$$

The proof relies on standard results from MC and stochastic optimization theory; see Appendix B. We note that the assumptions required for the latter, though standard in the literature, are strong. In practice, ϕ can converge to a local optimum $\phi^\dagger$, rather than the global optimum ϕ^*, introducing an additional asymptotic bias $\left|\mathcal{B}(d, \phi^\dagger) - \mathcal{B}(d, \phi^*)\right|$ into term III.

Theorem 1 can be applied directly to $\hat{\mu}_{\text{marg}}$, $-\hat{\mu}_{\text{post}}$, and $\hat{\mu}_{\text{VNMC}}$ (with fixed $M = L$), showing that they converge respectively to $\mathcal{U}_{\text{marg}}(d, \phi^*)$, $-\mathcal{L}_{\text{post}}(d, \phi^*)$, and $\mathcal{U}_{\text{VNMC}}(d, L, \phi^*)$ at a rate $= \mathcal{O}(T^{-1/2})$ if $N \propto K$ and the assumptions are satisfied. For $\hat{\mu}_{\text{m}+\ell}$, we combine Theorem 1 and Lemma 2 to obtain the same $\mathcal{O}(T^{-1/2})$ convergence rates; see the supplementary material for further details.

The key property of $\hat{\mu}_{\text{VNMC}}$ is that we need not set $M = L$ and can remove the asymptotic bias by increasing M with N. We begin by training ϕ with a fixed value of L, decreasing the error term $\|\mathcal{U}_{\text{VNMC}}(d, L, \phi_K) - \mathcal{U}_{\text{VNMC}}(d, L, \phi^*)\|_2$ at the fast rate $\mathcal{O}(K^{-1/2})$ until $|\mathcal{U}_{\text{VNMC}}(d, L, \phi^*) - \text{EIG}(d)|$ becomes the dominant error term. At this point, we start to increase N, M. Using the NMC convergence results discussed in Sec. 2, if we set $M \propto \sqrt{N}$, then $\hat{\mu}_{\text{VNMC}}$ converges to EIG(d) at a rate $\mathcal{O}((NM)^{-1/3})$. Note that the total cost of the $\hat{\mu}_{\text{VNMC}}$ estimator is $T = \mathcal{O}(KL + NM)$, where typically $M \gg L$. The first stage, costing KL, is fast variational training of an amortized importance sampling proposal for $p(y|d) = \mathbb{E}_{p(\theta)}[p(y|\theta, d)]$. The second stage, costing NM, is slower refinement to remove the asymptotic bias using the learned proposal in an NMC estimator.

Table 2: Bias squared and variance from 5 runs, averaged over designs, of EIG estimators applied to four benchmarks. We use - to denote that a method does not apply and $*$ when it is superseded by other methods. Bold indicates the estimator with the lowest empirical mean squared error.

	A/B test		**Preference**		**Mixed effects**		**Extrapolation**	
	Bias^2	Var	Bias^2	Var	Bias^2	Var	Bias^2	Var
$\hat{\mu}_{\text{post}}$	1.33×10^{-2}	7.15×10^{-3}	4.26×10^{-2}	8.53×10^{-3}	2.34×10^{-3}	2.92×10^{-3}	1.24×10^{-4}	5.16×10^{-5}
$\hat{\mu}_{\text{marg}}$	7.45×10^{-2}	6.41×10^{-3}	$\mathbf{1.10\times10^{-3}}$	$\mathbf{1.99\times10^{-3}}$	-	-	-	-
$\hat{\mu}_{\text{VNMC}}$	3.44×10^{-3}	3.38×10^{-3}	4.17×10^{-3}	9.04×10^{-3}	-	-	-	-
$\hat{\mu}_{\text{m}+\ell}$	$*$	$*$	$*$	$*$	$\mathbf{3.06\times10^{-3}}$	$\mathbf{5.94\times10^{-5}}$	$\mathbf{6.90\times10^{-6}}$	$\mathbf{1.84\times10^{-5}}$
$\hat{\mu}_{\text{NMC}}$	4.70×10^{0}	3.47×10^{-1}	7.60×10^{-2}	8.36×10^{-2}	-	-	-	-
$\hat{\mu}_{\text{laplace}}$	$\mathbf{1.92\times10^{-4}}$	$\mathbf{1.47\times10^{-3}}$	8.42×10^{-2}	9.70×10^{-2}	-	-	-	-
$\hat{\mu}_{\text{LFIRE}}$	2.29×10^{0}	6.20×10^{-1}	1.30×10^{-1}	1.41×10^{-2}	1.41×10^{-1}	6.67×10^{-2}	-	-
$\hat{\mu}_{\text{DV}}$	4.34×10^{0}	8.85×10^{-1}	9.23×10^{-2}	8.07×10^{-3}	9.10×10^{-3}	5.56×10^{-4}	7.84×10^{-6}	4.11×10^{-5}

One can think of the standard NMC approach as a special case of $\hat{\mu}_{\text{VNMC}}$ in which we naively choose $p(\theta)$ as the proposal. That is, standard NMC skips the first stage and hence does not benefit from the improved convergence rate of learning an amortized proposal. It typically requires a much higher total cost to achieve the same accuracy as VNMC.

5 Related work

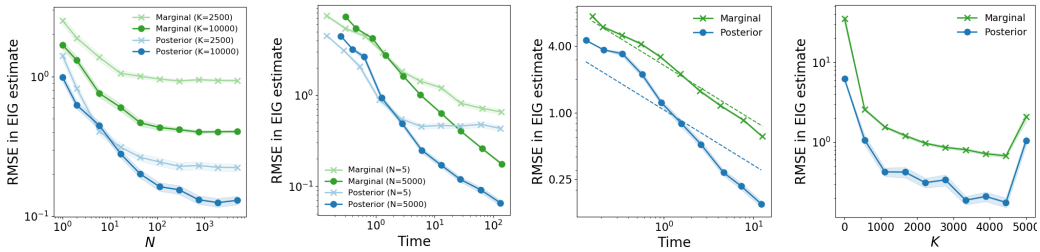
We briefly discuss alternative approaches to EIG estimation for BOED that will form our baselines for empirical comparisons. The **Nested Monte Carlo (NMC)** baseline was introduced in Sec. 2. Another established approach is to use a **Laplace approximation** to the posterior [22, 25]; this approach is fast but is limited to continuous variables and can exhibit large bias. Kleinegesse and Gutmann [18] recently suggested an implicit likelihood approach based on the Likelihood-Free Inference by Ratio Estimation **(LFIRE)** method of Thomas et al. [41]. We also consider a method based on the **Donsker-Varadhan (DV)** representation of the KL divergence [11] as used by Belghazi et al. [4] for mutual information estimation. Though not previously considered in BOED, we include it as a baseline for illustrative purposes. For a full discussion of the DV bound and a number of other variational bounds used in deep learning, we refer to the recent work of Poole et al. [31]. For further discussion of related work, see Appendix C.

6 Experiments

6.1 EIG estimation accuracy

We begin by benchmarking our EIG estimators against the aforementioned baselines. We consider four experiment design scenarios inspired by applications of Bayesian data analysis in science and industry. First, **A/B testing** is used across marketing and design [6, 19] to study population traits. Here, the design is the choice of the A and B group sizes and the Bayesian model is a Gaussian linear model. Second, revealed **preference** [36] is used in economics to understand consumer behaviour. We consider an experiment design setting in which we aim to learn the underlying utility function of an economic agent by presenting them with a proposal (such as offering them a price for a commodity) and observing their revealed preference. Third, fixed effects and random effects (nuisance variables) are combined in **mixed effects** models [14, 20]. We consider an example inspired by item-response theory [13] in psychology. We seek information only about the fixed effects, making this an implicit likelihood problem. Finally, we consider an experiment where labelled data from one region of design space must be used to predict labels in a target region by **extrapolation** [27]. In summary, we have two models with explicit likelihoods (A/B testing, preference) and two that are implicit (mixed effects, extrapolation). Full details of each model are presented in Appendix D.

For each scenario, we estimated the EIG across a grid of designs with a fixed computational budget for each estimator and calculated the true EIG analytically or with brute force computation as appropriate; see Table 2 for the results. Whilst the Laplace method, unsurprisingly, performed best for the Gaussian linear model where its approximation becomes exact, we see that our methods are otherwise more accurate. All our methods outperformed NMC.



(a) Convergence in N (b) Convergence in K (c) Convergence $N = K$ (d) Fixed budget $N + K$

Figure 1: Convergence of RMSE for $\hat{\mu}_{\text{post}}$ and $\hat{\mu}_{\text{marg}}$. (a) Convergence in number of MC samples N with a fixed number K of gradient updates of the variational parameters. (b) Convergence in time when increasing K and with N fixed. (c) Convergence in time when setting $N = K$ and increasing both (dashed lines represent theoretical rates). (d) Final RMSE with $N + K = 5000$ fixed, for different K. Each graph shows the mean with shading representing ± 1 std. err. from 100 trials.

6.2 Convergence rates

We now investigate the empirical convergence characteristics of our estimators. Throughout, we consider a single design point from the A/B test example. We start by examining the convergence of $\hat{\mu}_{\text{post}}$ and $\hat{\mu}_{\text{marg}}$ as we allocate the computational budget in different ways.

We first consider the convergence in N after a fixed number of K updates to the variational parameters. As shown in Figure 1a, the RMSE initially decreases as we increase N, before plateauing due to the bias in the estimator. We also see that $\hat{\mu}_{\text{post}}$ substantially outperforms $\hat{\mu}_{\text{marg}}$. We next consider the convergence as a function of wall-clock time when N is held fixed and we increase K. We see in Figure 1b that, as expected, the errors decrease with time and that when a small value of $N = 5$ is taken, we again see a plateauing effect, with the variance of the final MC estimator now becoming the limiting factor. In Figure 1c we take $N = K$ and increase both, obtaining the predicted convergence rate $\mathcal{O}(T^{-1/2})$ (shown by the dashed lines). We conjecture that the better performance of $\hat{\mu}_{\text{post}}$ is likely due to θ being lower dimensional ($\dim = 2$) than y ($\dim = 10$). In Figure 1d, we instead fix $T = N + K$ to investigate the optimal trade-off between optimization and MC error: it appears the range of K/T between 0.5 and 0.9 gives the lowest RMSE.

Finally, we show how $\hat{\mu}_{\text{VNMC}}$ can improve over NMC by using an improved variational proposal for estimating $p(y|d)$. In Figure 2, we plot the EIG estimates obtained by first running K steps of stochastic gradient with $L = 1$ to learn $q_v(\theta|y, d)$, before increasing M and N. We see that spending some of our time budget training $q_v(\theta|y, d)$ leads to noticeable improvements in the estimation, but also that it is important to increase N and M. Rather than plateauing like $\hat{\mu}_{\text{post}}$ and $\hat{\mu}_{\text{marg}}$, $\hat{\mu}_{\text{VNMC}}$ continues to improve after the initial training period as, albeit at a slower $O(T^{-1/3})$ rate.

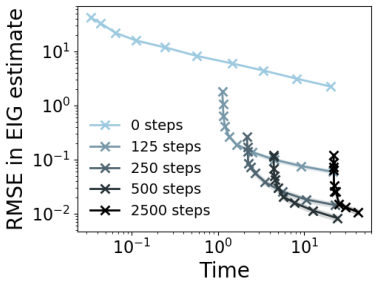
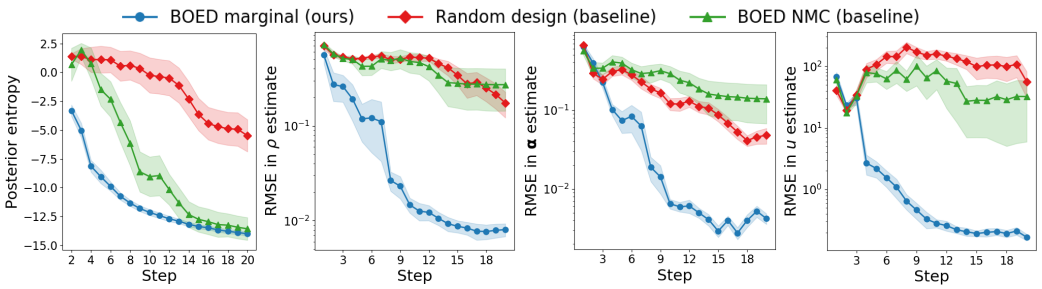


Figure 2: Convergence of $\hat{\mu}_{\text{VNMC}}$ taking $M = \sqrt{N}$. 'Steps' refers to pre-training of the variational posterior (i.e. K), with 0 steps corresponding to $\hat{\mu}_{\text{NMC}}$. Means and confidence intervals as per Fig. 1.

6.3 End-to-end sequential experiments

We now demonstrate the utility of our methods for designing sequential experiments. First, we demonstrate that our variational estimators are sufficiently robust and fast to be used for adaptive experiments with a class of models that are of practical importance in many scientific disciplines. To this end, we run an adaptive psychology experiment with human participants recruited from Amazon Mechanical Turk to study how humans respond to features of stylized faces. To account for fixed effects—those *common* across the population—as well as individual variations that we treat as nuisance variables, we use the mixed effects regression model introduced in Sec. 6.1. See Appendix D for full details of the experiment.

To estimate the EIG for different designs, we use $\hat{\mu}_{\text{m}+\ell}$, since it yields the best performance on our mixed effects model benchmark (see Table 2). Our EIG estimator is integrated into a system that



(a) Entropy (b) Posterior RMSE of ρ (c) Posterior RMSE of $\boldsymbol{\alpha}$ (d) Posterior RMSE of u

Figure 4: Evolution of the posterior in the sequential CES experiment. (a) Total entropy of a mean-field variational approximation of the posterior. (b)(c)(d) The RMSE of the posterior approximations of ρ, $\boldsymbol{\alpha}$ and u as compared to the true values used to simulate agent responses. Note the scale of the vertical axis is logarithmic. All plots show the mean and ± 1 std. err. from 10 independent runs.

presents participants with a stimulus, receives their response, learns an updated model, and designs the next stimulus, all online. Despite the relative simplicity of the design problem (with 36 possible designs) using BOED with $\hat{\mu}_{\mathrm{m}+\ell}$ leads to a more certain (i.e. lower entropy) posterior than random design; see Figure 3.

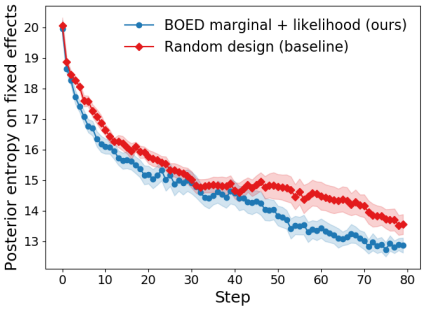


Figure 3: Evolution of the posterior entropy of the fixed effects in the Mechanical Turk experiment in Sec. 6.3. We depict the mean and ± 1 std. err. from 10 experimental trials.

Second, we consider a more challenging scenario in which a random design strategy gleans very little. We compare random design against two BOED strategies: $\hat{\mu}_{\mathrm{marg}}$ and $\hat{\mu}_{\mathrm{NMC}}$. Building on the revealed preference example in Sec. 6.1, we consider an experiment to infer an agent's utility function which we model using the Constant Elasticity of Substitution (CES) model [2] with latent variables $\rho, \boldsymbol{\alpha}, u$. We seek designs for which the agent's response will be informative about $\theta = (\rho, \boldsymbol{\alpha}, u)$. See Appendix D for full details. We estimate the EIG using $\hat{\mu}_{\mathrm{marg}}$ because the dimension of y is smaller than that of θ, and select designs $d \in [0, 100]^6$ using Bayesian optimization. To investigate parameter recovery we simulate agent responses from the model with fixed values of $\rho, \boldsymbol{\alpha}, u$. Figure 4 shows that using BOED with our marginal estimator reduces posterior entropy *and* concentrates more quickly on the true parameter values than both baselines. Random design makes no inroads into the learning problem, while BOED based on NMC particularly struggles at the outset when $p(\theta|d_{1:t-1}, y_{1:t-1})$, the prior at iteration t, is high variance. Our method selects informative designs throughout.

7 Discussion

We have developed efficient EIG estimators that are applicable to a wide range of experimental design problems. By tackling the double intractability of the EIG in a principled manner, they provide substantially improved convergence rates relative to previous approaches, and our experiments show that these theoretical advantages translate into significant practical gains. Our estimators are well-suited to modern deep probabilistic programming languages and we have provided an implementation in Pyro. We note that the interplay between variational and MC methods in EIG estimation is not directly analogous to those in standard inference settings because the NMC EIG estimator is itself inherently biased. Our $\hat{\mu}_{\mathrm{VNMC}}$ estimator allows one to play off the advantages of these approaches, namely the fast learning of variational approaches and asymptotic consistency of NMC.

Acknowledgements

We gratefully acknowledge research funding from Uber AI Labs. MJ would like to thank Paul Szerlip for help generating the sprites used in the Mechanical Turk experiment. AF would like to thank Patrick Rebeschini, Dominic Richards and Emile Mathieu for their help and support. AF gratefully acknowledges funding from EPSRC grant no. EP/N509711/1. YWT's and TR's research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 617071.

References

[1] Billy Amzal, Frédéric Y Bois, Eric Parent, and Christian P Robert. Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical association*, 101(474):773–785, 2006.

[2] Kenneth J Arrow, Hollis B Chenery, Bagicha S Minhas, and Robert M Solow. Capital-labor substitution and economic efficiency. *The review of Economics and Statistics*, pages 225–250, 1961.

[3] David Barber and Felix Agakov. The IM algorithm: a variational approach to information maximization. *Advances in Neural Information Processing Systems*, 16:201–208, 2003.

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Devon Hjelm, and Aaron Courville. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 530–539, 2018.

[5] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1): 973–978, 2019.

[6] George EP Box, J Stuart Hunter, and William G Hunter. Statistics for experimenters. In *Wiley Series in Probability and Statistics*. Wiley Hoboken, NJ, 2005.

[7] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *4th International Conference on Learning Representations, ICLR*, 2016.

[8] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[9] Alex R Cook, Gavin J Gibson, and Christopher A Gilligan. Optimal observation times in experimental epidemic processes. *Biometrics*, 64(3):860–868, 2008.

[10] Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The Helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

[11] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28(1): 1–47, 1975.

[12] Sylvain Ehrenfeld. Some experimental design problems in attribute life testing. *Journal of the American Statistical Association*, 57(299):668–679, 1962.

[13] Susan E Embretson and Steven P Reise. *Item response theory*. Psychology Press, 2013.

[14] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[15] Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems*, pages 766–774, 2010.

[16] José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, pages 918–926, 2014.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

[18] Steven Kleinegesse and Michael U Gutmann. Efficient bayesian experimental design for implicit models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 476–485, 2019.

[19] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1): 140–181, 2009.

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

[21] Tuan Anh Le, Maximilian Igl, Tom Rainforth, Tom Jin, and Frank Wood. Auto-Encoding Sequential Monte Carlo. *International Conference on Learning Representations (ICLR)*, 2018.

[22] Jeremy Lewi, Robert Butera, and Liam Paninski. Sequential optimal design of neurophysiology experiments. *Neural Computation*, 21(3):619–687, 2009.

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

[24] Dennis V Lindley. *Bayesian statistics, a review*, volume 2. SIAM, 1972.

[25] Quan Long, Marco Scavino, Raúl Tempone, and Suojin Wang. Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, 259:24–39, 2013.

[26] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient dynamic discovery of high-value information with partial VAE. *arXiv preprint arXiv:1809.11142*, 2018.

[27] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[28] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[29] Peter Müller. Simulation based optimal design. *Handbook of Statistics*, 25:509–518, 2005.

[30] Jay I Myung, Daniel R Cavagnaro, and Mark A Pitt. A tutorial on adaptive design optimization. *Journal of mathematical psychology*, 57(3-4):53–67, 2013.

[31] Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A Alemi, and George Tucker. On variational lower bounds of mutual information. *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

[32] Tom Rainforth. *Automating Inference, Learning, and Design using Probabilistic Programming*. PhD thesis, University of Oxford, 2017.

[33] Tom Rainforth, Robert Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting Monte Carlo estimators. In *International Conference on Machine Learning*, pages 4264–4273, 2018.

[34] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1278–1286, 2014.

[35] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[36] Paul A Samuelson. Consumption theory in terms of revealed preference. *Economica*, 15(60): 243–253, 1948.

[37] Paola Sebastiani and Henry P Wynn. Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1), 2000.

[38] Ben Shababo, Brooks Paige, Ari Pakman, and Liam Paninski. Bayesian inference and online experimental design for mapping neural microcircuits. In *Advances in Neural Information Processing Systems*, pages 1304–1312, 2013.

[39] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[40] Andreas Stuhlmüller, Jacob Taylor, and Noah Goodman. Learning stochastic inverses. In *Advances in neural information processing systems*, pages 3048–3056, 2013.

[41] Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U Gutmann. Likelihood-free inference by ratio estimation. *arXiv preprint arXiv:1611.10242*, 2016.

[42] Joep Vanlier, Christian A Tiemann, Peter AJ Hilbers, and Natal AW van Riel. A Bayesian approach to targeted experiment design. *Bioinformatics*, 28(8):1136–1142, 2012.

[43] Benjamin T Vincent and Tom Rainforth. The DARC toolbox: automated, flexible, and efficient delayed and risky choice experiments using bayesian adaptive design. 2017.