A Unified Stochastic Gradient Approach to Designing Bayesian-Optimal Experiments

Adam Foster[†] **Martin Jankowiak**[‡] **Matthew O'Meara**[§] **Yee Whye Teh**[†] **Tom Rainforth**[†✠]

[†]Department of Statistics, University of Oxford, Oxford, UK
[‡]Uber AI, San Francisco, CA, USA
[§]University of Michigan, Ann Arbor, MI, USA
[✠]Christ Church, University of Oxford, Oxford, UK
adam.foster@stats.ox.ac.uk

Abstract

We introduce a fully stochastic gradient based approach to Bayesian optimal experimental design (BOED). Our approach utilizes variational lower bounds on the expected information gain (EIG) of an experiment that can be simultaneously optimized with respect to both the variational and design parameters. This allows the design process to be carried out through a single unified stochastic gradient ascent procedure, in contrast to existing approaches that typically construct a pointwise EIG estimator, before passing this estimator to a separate optimizer. We provide a number of different variational objectives including the novel adaptive contrastive estimation (ACE) bound. Finally, we show that our gradient-based approaches are able to provide effective design optimization in substantially higher dimensional settings than existing approaches.

1 INTRODUCTION

The design of experiments is a key problem in almost every scientific discipline. Namely, one wishes to construct an experiment that is most informative about the investigated process, while minimizing its cost. For example, in a psychological trial, we want to ensure questions posed to participants are pertinent and do not have predictable responses. In a pharmaceutical trial, we want to minimize the number of participants needed to test our hypotheses. In an online automated help system, we want to ensure we ask questions that identify the user's problem as quickly as possible.

In all these scenarios, our ultimate high-level aim is to choose designs that maximize the information gathered by the experiment. A powerful and broadly used approach for formalizing this aim is Bayesian optimal experimental design (BOED) (Chaloner and Verdinelli, 1995; Lindley, 1956; Myung et al., 2013). In BOED, we specify a Bayesian model for the experiment and then choose the design that maximizes the expected information gain (EIG) from running it. More specifically, let θ denote the latent variables we wish to learn about from running the experiment and let $\xi \in \Xi$ represent the experimental design. By introducing a prior $p(\theta)$ and a predictive distribution $p(y|\theta, \xi)$ for experiment outcomes y, we can calculate the EIG under this model by taking the expected reduction in posterior entropy

$$I(\xi) \triangleq \mathbb{E}_{p(y|\xi)}\left[H[p(\theta)] - H[p(\theta|y,\xi)]\right], \qquad (1)$$

where $H[\cdot]$ represents the entropy of a distribution and $p(\theta|y,\xi) \propto p(\theta)p(y|\theta,\xi)$. Our experimental design process now becomes that of the finding the design ξ^* that maximizes $I(\xi)$.

Unfortunately, finding ξ^* is typically a very challenging problem in practice. Even evaluating $I(\xi)$ for a single design is computationally difficult because it represents a *nested* expectation and thus has no direct Monte Carlo estimator (Rainforth et al., 2018; Zheng et al., 2018). Though a large variety of approaches for performing this estimation have been suggested (Myung et al., 2013; Watson, 2017; Kleinegesse and Gutmann, 2018; Foster et al., 2019), the resulting BOED strategies share a critical common feature: they estimate $I(\xi)$ on a point-by-point basis and feed this estimator to an outer-level optimizer that selects the design.

This framework can be highly inefficient for a number of reasons. For example, it adds an extra level of nest-

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

ing to the overall computation process: $I(\xi)$ must be separately estimated for each ξ, substantially increasing the overall computational cost. Furthermore, one must typically resort to gradient-free methods to carry out the resulting optimization, which means it is difficult to scale the overall BOED process to high dimensional design settings due to a dearth of optimization schemes which remain effective in such settings.

To alleviate these inefficiencies and open the door to applying BOED in high-dimensional settings, we introduce an alternative to this two-stage framework by introducing unified objectives that can be directly maximized to simultaneously estimate $I(\xi)$ and optimize ξ. Specifically, by building on the work of Foster et al. (2019), we construct variational lower bounds to $I(\xi)$ that can be simultaneously optimized with respect to both the variational and design parameters. Optimizing the former ensures that we achieve a tight bound that in turn gives accurate estimates of $I(\xi)$, while simultaneously optimizing the latter circumvents the need for an expensive outer optimization process. Critically, this approach allows the optimization to be performed using stochastic gradient ascent (SGA) (Robbins and Monro, 1951) and therefore scaled to substantially higher dimensional design problems than existing approaches.

To account for the varying needs of different problem settings, we introduce several classes of suitable variational lower bounds. Most notably, we introduce the adaptive contrastive estimation (ACE) bound: an EIG variational lower bound that can be made arbitrarily tight, while remaining amenable to simultaneous SGA on both the variational parameters and designs.

We demonstrate[1] the applicability of our unified gradient approach using a wide range of experimental design problems, including a real-world high-dimensional example from the pharmacology literature (Lyu et al., 2019). We find that our approaches are able to effectively optimize the EIG, consistently outperforming baseline two-stage approaches, with particularly large gains achieved for high-dimensional problems. These gains lead, in turn, to improved designs and more informative experiments.

2 BACKGROUND

2.1 Bayesian optimal experimental design

When experimentation is costly, time consuming, or dangerous, it is essential to design experiments to learn the most from them. To choose between potential designs, we require a metric of the quality of a candidate design. In the BOED framework dating back to Lindley (1956), this metric represents how much more certain we will become in our knowledge of the world after doing the experiment and analyzing the data. We prefer designs that will lead to strong conclusions even if we are not yet sure what those conclusions will be.

Specifically, we consider an experiment with design ξ, latent variable θ and outcome y. For example, ξ may represent the question posed to a participant in a psychology trial, y their answer, and θ their underlying psychological characteristic which is being studied. The BOED framework begins with a Bayesian model of the experimental process. This model consists of a likelihood $p(y|\theta, \xi)$ that predicts the experimental outcome under design ξ and latent variable θ and a prior $p(\theta)$ which incorporates initial beliefs about the unknown θ. After conducting the experiment, our beliefs about θ are updated to the posterior $p(\theta|y, \xi)$. The information gained about θ from doing the experiment with design ξ and obtaining outcome y is the reduction in entropy from the prior to the posterior

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

As it stands, information gain cannot be evaluated until after the experiment. To define a metric that will let us choose between designs before experimentation, we can use the *expected* information gain (EIG), $I(\xi)$, by taking the expectation of IG over hypothesized outcomes y using the marginal distribution under our model, $p(y|\xi)$, to give

$$I(\xi) \triangleq \mathbb{E}_{p(y|\xi)}\left[H[p(\theta)] - H[p(\theta|y, \xi)]\right] \tag{3}$$

which can be rewritten in the form of a mutual information between θ and y with ξ fixed, namely

$$I(\xi) = \mathrm{MI}_\xi(\theta; y) = \mathbb{E}_{p(\theta)p(y|\theta,\xi)}\left[\log \frac{p(y|\theta, \xi)}{p(y|\xi)}\right]. \tag{4}$$

The Bayesian optimal design, ξ^*, is now the one which maximizes EIG over the set of feasible designs Ξ

$$\xi^* = \arg\max_{\xi \in \Xi} \; I(\xi). \tag{5}$$

In *iterated* experimental design, we design a sequence $\xi_1, ..., \xi_T$ of experiments. At time t, the prior $p(\theta)$ in (4) is replaced by the posterior given the previous experiment designs and observed outcomes, namely

$$p(\theta|\xi_{1:t-1}, y_{1:t-1}) \propto p(\theta) \prod_{\tau=1}^{t-1} p(y_\tau|\theta, \xi_\tau). \tag{6}$$

This now allows us to construct adaptive experiments, wherein we use information gathered from previous iterations to select the designs used at future iterations.

[1] Supporting code is provided at `https://github.com/ae-foster/pyro/tree/sgboed-reproduce`.

2.2 Estimating expected information gain

Making even a single point estimate of EIG when solving (5) can be challenging because we must first estimate the unknown $p(y|\xi)$ or $p(\theta|y,\xi)$, and then take an expectation over $p(\theta)p(y|\theta,\xi)$. Nested Monte Carlo (NMC) estimators (Rainforth et al., 2018), which make a Monte Carlo approximation of both the inner and outer integrals, converge relatively slowly: at a rate $\mathcal{O}(T^{-1/3})$ in the total computational budget T.

Foster et al. (2019) noted that this approach is inefficient because it makes a separate Monte Carlo approximation of the integrand for every sample of the outer integral. To share information between different samples, they proposed a number of variational estimators that used amortization, i.e. they attempted to learn the functional form of the integrand rather than approximating it afresh each time. One of their approaches was based on amortized variational inference and required an *inference network* $q_\phi(\theta|y)$ which takes as input ϕ, y and outputs a distribution over θ. For any $q_\phi(\theta|y)$, we can construct a lower bound on $I(\xi)$. This is the Barber-Agakov (BA), or posterior, lower bound (Barber and Agakov, 2003)

$$I_{BA}(\xi,\phi) \triangleq \mathbb{E}_{p(\theta)p(y|\theta,\xi)}[\log q_\phi(\theta|y)] + H[p(\theta)]\,, \quad (7)$$

which was also used by (Pacheco and Fisher, 2019) and which has found use representation learning (Poole et al., 2019) and maximizing information transmission over noisy channels (Barber and Agakov, 2003).

To make high-quality approximations to $I(\xi)$, and simultaneously learn a good posterior approximation, Foster et al. (2019) maximize this bound with respect to ϕ. This approach is most effective when the bound is tight, i.e. $\max_\phi I_{BA}(\xi,\phi) = I(\xi)$. For $I_{BA}(\xi,\phi)$, this occurs when it is possible to have $q_\phi(\theta|y) = p(y|\theta,\xi)$, i.e. when the inference network is powerful enough to find the true posterior distribution for every y.

To obtain high-quality approximations of $I(\xi)$ even when the inference network cannot capture the true posterior, Foster et al. (2019) also considered another variational estimator: variational nested Monte Carlo (VNMC). This uses the inference network $q_\phi(\theta|y)$ in conjunction with additional samples to improve the estimate of the integrand. They showed that this leads to the following *upper* bound on $I(\xi)$

$$I_{VNMC}(\xi,\phi,L) \triangleq \mathbb{E}\left[\log \frac{p(y|\theta_0,\xi)}{\frac{1}{L}\sum_{\ell=1}^{L}\frac{p(\theta_\ell)p(y|\theta_\ell,\xi)}{q_\phi(\theta_\ell|y)}}\right], \quad (8)$$

where the expectation is over $p(\theta_0)p(y|\theta_0,\xi)q_\phi(\theta_{1:L}|y)$. The inference network in VNMC is trained by minimization, in the same way I_{BA} is trained by maximization. I_{VNMC} has the attractive feature that the bound becomes tight as $L \to \infty$, even if $q_\phi(\theta_\ell|y)$ is not powerful enough to directly represent the true posterior.

2.3 Optimizing the EIG

The experimental design problem is to find the design that maximizes the EIG. Therefore, as well as finding a way to estimate EIG, existing approaches subsequently need to find a way of searching across Ξ to find promising designs. At a high-level, most existing approaches propose a two-stage procedure in which noisy estimates of $I(\xi)$ are made, and a separate optimization procedure selects the candidate design ξ to evaluate next.

Kleinegesse and Gutmann (2018) and Foster et al. (2019) both use Bayesian optimization (BO) for this outer optimization step, a black-box optimization method that is tolerant to noise in the estimates of the objective function (Snoek et al., 2012), in this case $I(\xi)$. Some approaches (Watson, 2017; Lyu et al., 2019) instead select a finite number of candidate designs in Ξ and estimate $I(\xi)$ at each candidate, with some refining this process further by adaptively allocating computational resources between these designs (Vincent and Rainforth, 2017; Rainforth, 2017). Another suggested approach is to use MCMC methods to carry out this outer optimization (Amzal et al., 2006; Müller, 2005).

3 GRADIENT-BASED BOED

Our central proposal is to replace the two-stage procedure outlined above with a single stage that simultaneously estimates $I(\xi)$ and optimizes ξ. This has the critical advantage of allowing SGA to be directly applied to the design optimization. Not only does this provide substantial computational gains over approaches which must construct separate estimates for each design considered, but it also provides the potential to scale to substantially higher dimensional design problems than those which can be effectively tackled with existing approaches. Since we take gradients with respect to ξ, we henceforth assume that Ξ is continuous.

In our approach, we utilize variational *lower bounds* on I. Specifically, suppose we have a bound $\mathcal{L}(\xi,\phi) \leq I(\xi)$ with variational parameters ϕ. For fixed ξ, the estimate of $I(\xi)$ improves as we maximize with respect to ϕ. We propose to maximize $\mathcal{L}$ *jointly* with respect to (ξ,ϕ). As we train ϕ, the variational approximation improves; as we train ξ our design moves to regions where the lower bound on EIG is largest. By tackling this as a single optimization problem over (ξ,ϕ), we obviate the need to have an outer optimizer for ξ. Using a lower bound is important because it allows us to perform a single maximization over (ξ,ϕ), rather than a more complex

optimization such as the max-min optimization that would result if we used an upper bound.

In practice, we do not have lower bounds on I that we can evaluate and differentiate in closed form. Instead, we have bounds that are expectations over $p(\theta)p(y|\theta,\xi)$. Fortunately, we can still maximize these lower bounds with respect to (ξ,ϕ) by using SGA, which is known to remain effective in high dimensions (Bottou, 2010).

3.1 Barber-Agakov (BA)

We now make our first concrete proposal for the lower bound $\mathcal{L}(\xi,\phi)$: the BA bound I_{BA}, as defined in (7). The difference is we will now optimize (ξ,ϕ) jointly whereas previously only ϕ was trained using gradients. To perform SGA, we use the following unbiased estimators for $\partial I_{BA}/\partial\phi$ and $\partial I_{BA}/\partial\xi$

$$\widehat{\frac{\partial I_{BA}}{\partial\phi}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\phi}\log q_\phi(\theta_n|y_n), \tag{9}$$

$$\widehat{\frac{\partial I_{BA}}{\partial\xi}} = \frac{1}{N}\sum_{n=1}^{N}\log q_\phi(\theta_n|y_n)\frac{\partial}{\partial\xi}\log p(y_n|\theta_n,\xi) \tag{10}$$

where $\theta_n, y_n \overset{\text{i.i.d.}}{\sim} p(\theta)p(y|\theta,\xi)$. The estimator of $\partial I_{BA}/\partial\xi$ is a score function estimator, other possibilities are discussed in Section 3.6.

3.2 Adaptive contrastive estimation (ACE)

The BA bound provides one specific case of our one-stage procedure for optimal experimental design. We now introduce a new lower bound that improves upon I_{BA}. The potential issue with the BA bound is that it may not be sufficiently tight, which happens when the inference network cannot represent the true posterior. One possible solution is to introduce additional samples, as in the VNMC estimator (8). However, we cannot use VNMC directly for a one-stage procedure: since it is an upper bound, we must minimize it with respect to ϕ, but we still wish to maximize with respect to ξ.

Looking more closely at the VNMC bound, we see that its main failure case is when the denominator strongly *under-estimates* $p(y|\xi)$, which can happen when all the inner samples $\theta_1,...,\theta_L$ miss regions where the joint $p(\theta_\ell)p(y|\theta_\ell,\xi)$ is large. In addition to the samples $\theta_{1:L}$, we also have the original sample θ_0 from which y was sampled. One way to avoid the under-estimation in the denominator would be to include this sample, giving

$$I_{ACE}(\xi,\phi,L) = \mathbb{E}\left[\log\frac{p(y|\theta_0,\xi)}{\frac{1}{L+1}\sum_{\ell=0}^{L}\frac{p(\theta_\ell)p(y|\theta_\ell,\xi)}{q_\phi(\theta_\ell|y)}}\right] \tag{11}$$

where the expectation is with respect to $p(\theta_0)p(y|\theta_0,\xi)q(\theta_{1:L}|y)$. In fact, by including θ_0 we cause the denominator to now *over-estimate* $p(y|\xi)$ which results in a new **lower bound** on $I(\xi)$ which can be jointly maximized with respect to (ξ,ϕ). The samples $\theta_{1:L}$ can now be seen as contrasts to the original sample θ_0. For this reason, we call $\theta_{1:L}$ *contrastive samples* and we call (11) the **adaptive contrastive estimate (ACE)** of EIG. The following theorem establishes that I_{ACE} is a valid lower bound on the EIG which becomes tight as $L\to\infty$.

Theorem 1. *For any model $p(\theta)p(y|\theta,\xi)$ and inference network $q_\phi(\theta|y)$, we have the following:*

1. *I_{ACE} is a lower bound on $I(\xi)$ and we can characterize the error term as an expected KL divergence:*

$$I(\xi) - I_{ACE}(\xi,\phi,L) = \mathbb{E}_{p(y|\xi)}\left[KL\left(P(\theta_{0:L}|y)\middle\|\prod_\ell q_\phi(\theta_\ell|y)\right)\right] \geq 0,$$

$$P(\theta_{0:L}|y) = \frac{1}{L+1}\sum_{\ell=0}^{L}p(\theta_\ell|y,\xi)\prod_{k\neq\ell}q_\phi(\theta_k|y).$$

2. *As $L\to\infty$, we recover the true EIG: $\lim_{L\to\infty} I_{ACE}(\xi,\phi,L) = I(\xi)$.*
3. *The ACE bound is monotonically increasing in L: $I_{ACE}(\xi,\phi,L_2) \geq I_{ACE}(\xi,\phi,L_1)$ for $L_2 \geq L_1 \geq 0$.*
4. *If the inference network equals the true posterior $q_\phi(\theta|y) = p(\theta|y,\xi)$, then $I_{ACE}(\xi,\phi,L) = I(\xi), \forall L$.*

See Appendix A for the proof and additional results. Gradient estimation for ACE is discussed in Section 3.6. We note that, to the best of our knowledge, I_{ACE} has not previously appeared in the BOED literature.[2]

3.3 Prior contrastive estimation (PCE)

Theorem 1 tells us that I_{ACE} can become close to $I(\xi)$ if either: 1) the inference network becomes close to the true posterior $p(\theta|y,\xi)$, 2) we increase the number of contrastive samples L. The BA bound only becomes tight in case 1). A special case of ACE is to replace the inference network $q_\phi(\theta|y)$ with a fixed distribution and rely on the contrastive samples to make good estimates of $I(\xi)$, only becoming tight in case 2), i.e. as $L\to\infty$. This simplification can speed up training, since we no longer need to learn additional parameters ϕ.

To explore this, we propose the **prior contrastive estimation (PCE)** bound, in which the prior $p(\theta)$ is used to generate contrastive samples:

$$I_{PCE}(\xi,L) \triangleq \mathbb{E}\left[\log\frac{p(y|\theta_0,\xi)}{\frac{1}{L+1}\sum_{\ell=0}^{L}p(y|\theta_\ell,\xi)}\right], \tag{12}$$

[2] Aside from a recent blog post (Sobolev, 2019) we believe this bound has not previously been suggested in any context.

where the expectation is over $p(\theta_0)p(y|\theta_0,\xi)p(\theta_{1:L})$. Whilst inherently less powerful than ACE, PCE can be effective when the prior and posterior are similar, such that $p(\theta)$ is a suitable proposal to estimate $p(y|\xi)$.

Though, to the best of our knowledge, this bound has not been applied to BOED before, we note that it shares a connection to the information noise contrastive estimation (InfoNCE) bound on mutual information used in representation learning (van den Oord et al., 2018). Given K data samples x_k, corresponding representations z_k, and a critic $f_\psi(x,z) \geq 0$, we have

$$\mathrm{MI}(x;z) \geq \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \log \frac{f_\psi(x_k,z_k)}{\frac{1}{K}\sum_{\ell=1}^{K} f_\psi(x_\ell,z_k)}\right] \quad (13)$$

where the expectation is over $p(x)p(z|x)$, $p(x)$ is the data distribution, and $p(z|x)$ is the encoder. Poole et al. (2019) showed that the encoder density $p(z|x)$ is the optimal critic, although it is rarely known in closed form in the representation learning context. Writing θ for x and y for z, we note the mathematical connection between this optimal case and I_{PCE}.

3.4 Likelihood-free ACE

In some models such as random effects models, the likelihood $p(y|\theta,\xi)$ is not known in closed form but can be sampled from. This presents a problem when computing I_{ACE} or its derivatives because the likelihood appears in (11). To allow ACE to be used for these kinds of models, we now show that using a unnormalized approximation to the likelihood still results in a valid lower bound on the EIG. In fact, if using a parametrized likelihood approximation f_ψ, it is then possible to train ψ jointly with (ξ,ϕ) to approximate the likelihood, learn an inference network, and find the optimal design through the solution to a single optimization problem. The following theorem, whose proof is presented in Appendix A, shows that replacing the likelihood with an unnormalized approximation does result in a valid lower bound on EIG.

Theorem 2. *Consider a model $p(\theta)p(y|\theta,\xi)$ and inference network $q_\phi(\theta|y)$. Let $f_\psi(\theta,y) \geq 0$ be an unnormalized likelihood approximation. Then,*

$$I(\xi) \geq \mathbb{E}\left[\log \frac{f_\psi(\theta_0,y)}{\frac{1}{L+1}\sum_{\ell=0}^{L} \frac{p(\theta_\ell)f_\psi(\theta_\ell,y)}{q_\phi(\theta_\ell|y)}}\right] \quad (14)$$

where the expectation is over $p(\theta_0)p(y|\theta_0,\xi)q_\phi(\theta_{1:L}|y)$.

3.5 Iterated experimental design with ACE

In iterated experimental design, we replace $p(\theta)$ by $p(\theta|y_{1:t-1},\xi_{1:t-1})$ as per (6). We can sample $p(\theta|y_{1:t-1},\xi_{1:t-1})$ by performing inference. Whilst variational inference also provides a closed form estimate of the posterior density, some other inference methods do not. This is problematic because the prior density appears in (11). Fortunately, it is sufficient to know the density *up to proportionality* (Foster et al., 2019). Indeed if $p(\theta) = A \cdot \gamma(\theta)$ where A does not depend on (ξ,ϕ,y) and γ is an unnormalized density, then

$$I(\xi) \geq \mathbb{E}\left[\log \frac{p(y|\theta_0,\xi)}{\frac{1}{L+1}\sum_{\ell=0}^{L} \frac{\gamma(\theta_\ell)p(y|\theta_\ell,\xi)}{q_\phi(\theta_\ell|y)}}\right] - \log A \quad (15)$$

and the derivatives of $\log A$ are simply zero.

3.6 Gradient estimation for ACE

To optimize the ACE bound with respect to (ξ,ϕ) we need unbiased gradient estimators of $\partial I_{ACE}/\partial\xi$ and $\partial I_{ACE}/\partial\phi$. The simplest form of the ξ-gradient is

$$\frac{\partial I_{ACE}}{\partial \xi} = \mathbb{E}\left[\frac{\partial g}{\partial \xi} + g \cdot \frac{\partial}{\partial \xi}\log p(y|\theta_0,\xi)\right] \quad (16)$$

where the expectation is with respect to $p(\theta_0)p(y|\theta,\xi)q(\theta_{1:L}|y)$, and

$$g(y,\theta_{0:L},\phi,\xi) = \log \frac{p(y|\theta_0,\xi)}{\frac{1}{L+1}\sum_{\ell=0}^{L} \frac{p(\theta_\ell)p(y|\theta_\ell,\xi)}{q_\phi(\theta_\ell|y)}}. \quad (17)$$

Estimating the expectation (16) directly using Monte Carlo gives the score function, or REINFORCE, estimator. Unfortunately, this is often high variance, and reducing gradient estimator variance is often important in solving challenging experimental design problems.

One variance reduction method is reparameterization. For this, we introduce random variables $\epsilon, \epsilon'_{1:L}$ which do not depend on (ξ,ϕ) along with representations of y and θ as deterministic functions of these variables: $y = y(\theta_0,\xi,\epsilon)$ and $\theta_\ell = \theta(y,\phi,\epsilon'_\ell)$. This now permits the reparameterized gradient

$$\frac{\partial I_{ACE}}{\partial \xi} = \mathbb{E}\left[\frac{\partial g}{\partial \xi} + \frac{\partial g}{\partial y}\frac{\partial y}{\partial \xi} + \sum_{\ell=1}^{L} \frac{\partial g}{\partial \theta_\ell}\frac{\partial \theta_\ell}{\partial y}\frac{\partial y}{\partial \xi}\right] \quad (18)$$

where the expectation is over $p(\theta_0)p(\epsilon)p(\epsilon'_{1:L})$. A Monte Carlo approximation of this expectation is typically a much lower variance estimator for the true ξ-gradient.

Alternatively, if y is a discrete random variable we can sum over the possible values $\mathcal{Y}$. This approach is known as Rao-Blackwellization and gives

$$\frac{\partial I_{ACE}}{\partial \xi} = \sum_{y\in\mathcal{Y}} \mathbb{E}\left[\frac{\partial g}{\partial \xi}\, p(y|\theta_0,\xi) + g\,\frac{\partial}{\partial \xi} p(y|\theta_0,\xi)\right] \quad (19)$$

where the expectation is now over $p(\theta_0)\prod_{\ell=1}^{L} q_\phi(\theta_\ell|y)$.

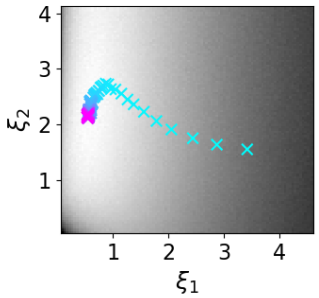


Figure 1: A sample trajectory for the death process. The grayscale shows the EIG surface (white is maximal), whilst crosses show the optimization trajectory of ξ using ACE with pink representing later steps. See Sec. 4.2 for details.

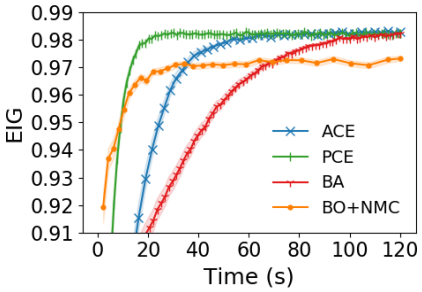


Figure 2: Optimization of EIG for the death process as a function of wall clock time. We depict the mean and ± 1 standard error (s.e.) from 100 runs. The final EIG values (rightmost points) are as follows: [ACE] $\mathbf{0.9830 \pm 0.0001}$, [PCE] 0.9822 ± 0.0001, [BA] 0.9822 ± 0.0002, [BO] 0.9732 ± 0.0009. See Sec. 4.2 for details.

Turning to $\partial I_{ACE}/\partial\phi$, we note that if $\theta_{1:L}$ are reparameterizable (i.e. can be expressed $\theta_\ell = \theta(y, \phi, \epsilon'_\ell)$), then we can utilize the double reparameterization of Tucker et al. (2018); for full details see Appendix A.1.

4 EXPERIMENTS

We now learn optimal experimental designs in five scenarios: the **death process**, a well known two-dimensional design problem from epidemiology; a non-conjugate **regression** model with a 400-dimensional design; an ablation study in the setting of **advertising**; a real-world **biomolecular docking** problem from pharmacology in 100 dimensions; and a **constant elasticity of substitution** iterated design problem in behavioural economics with 6 dimensional designs.

4.1 Evaluating experimental designs

We first discuss which metrics we will use to judge the quality of the designs we obtain. Our primary metric on designs is, of course, the EIG. We prefer designs with high EIGs. In some cases, we can evaluate the EIG analytically. In other cases, we can use a sufficiently large number of samples in a NMC (Rainforth et al., 2018) estimator to be sure that we have estimates that are sufficiently accurate to compare designs.

To explore the limits of our methods, we will also consider scenarios where neither of these approaches is suitable. In these cases, we pair the ACE lower bound (with ξ fixed for evaluation) with the VNMC upper bound (Foster et al., 2019) to trap the true EIG value—if the lower bound of one design is higher than the upper bound for another, we can be sure that the first design is superior (noting that the bounds themselves can be tractably estimated to a very high accuracy).

In some settings, when we know the true optimal design ξ^*, we will also consider the *design error* $\|\xi^* - \xi\|$, i.e. how close our design is to the optimal design.

In iterated experiment design, as well as designing experiments, we must also perform inference on the latent variable θ after each iteration. Here, we also investigate the quality of the final posterior. Specifically, if $p(\theta|y_{1:t}, \xi_{1:t})$ is the posterior after t experiments, we use the posterior entropy, and the posterior RMSE $\mathbb{E}_{\theta\sim p(\theta|y_{1:t},\xi_{1:t})}\left[(\theta - \theta^*)^2\right]^{1/2}$. We prefer low entropies and low RMSE values.

4.2 Death process

We consider an example from epidemiology, the death process (Cook et al., 2008; Kleinegesse and Gutmann, 2018), in which a population of $N = 10$ individuals transitions from healthy to infected states at a constant but unknown rate θ. We can measure the number of infected individuals at two different times ξ_1 and $\xi_1 + \xi_2$ where $\xi_1, \xi_2 \geq 0$. Our aim is to infer the infection rate θ from these observations. For full details of the prior and likelihood used, see Appendix B.2.

On this problem, we apply gradient methods with Rao-Blackwellization over the 66 possible outcomes. Figure 1 shows a sample optimization trajectory with the approximate EIG surface. We compare against BO using the Rao-Blackwellized NMC estimator of Vincent and Rainforth (2017). Figure 2 shows that, for the allowed time budget, all gradient methods perform better than BO even on this two-dimensional problem.

4.3 Regression

We now compare our one-stage gradient approaches to experimental design against a two-stage baseline on a high-dimensional design problem. We choose a general purpose Bayesian linear regression model with n observations and p features. The design ξ is an $n \times p$ matrix; the latent variables are $\theta = (\mathbf{w}, \sigma)$, where $\mathbf{w}$ is the p dimensional regression coefficient and σ^2 is the scalar variance. The n outcomes are generated using a Normal likelihood $y_i \sim N(\boldsymbol{\xi}_i \cdot \mathbf{w}, \sigma)$ for $i = 1, ..., n$. Here $\boldsymbol{\xi}_i$ is the ith row of ξ. To avoid trivial solutions,

Table 1: Regression results. We estimate lower and upper bounds on the final EIG and present the mean and ± 1 s.e. from 10 runs. See Sec. 4.3 for details.

Method	EIG l.b.	EIG u.b.
ACE	16.1 ± 0.1	20.7 ± 0.2
PCE	16.6 ± 0.1	21.5 ± 0.2
BA	16.4 ± 0.2	21.1 ± 0.2
BO + VNMC	7.3 ± 0.1	9.6 ± 0.1
Random Search + VNMC	7.1 ± 0.1	9.4 ± 0.1

we enforce the constraint $\|\boldsymbol{\xi}_i\|_1 = 1$ for all i. We use independent priors $w_j \sim \text{Laplace}(1)$ for $j = 1, ..., p$ and $\sigma \sim \text{Exp}(1)$. See Appendix B.3 for complete details.

We set $n = p = 20$ and applied five methods to this 400 dimensional design problem: BA, ACE and PCE, as well as the VNMC estimator of Foster et al. (2019), with both BO and random search to optimize over Ξ. The results are presented in Table 1. We note that the gradient methods strongly outperform the gradient-free baselines, with about double the final EIG.

4.4 Advertising

We now conduct a detailed ablation study on the effects of dimension on the quality of experimental designs produced using our gradient approaches and BO. To further isolate the distinction between one-stage and two-stage approaches to BOED, we choose a setting in which we can compute $I(\xi)$ analytically. We give BO, but not the gradient methods, access to a EIG oracle when making point evaluations of $I(\xi)$, i.e. our two-stage baseline is spared the need to estimate $I(\xi)$. Thus we put BO in the best possible position and ensure any gains are due to improvements from using gradient-based optimization.

Suppose that we are given an advertising budget of B dollars that we need to allocate among D regions, i.e. we choose $\boldsymbol{\xi} \geq \mathbf{0}$ with $\sum_{i=1}^{D} \xi_i = B$. After conducting an ad campaign, we observe a vector of sales $\boldsymbol{y}$. We use this data to make inferences about the underlying market opportunities $\boldsymbol{\theta}$ in each region. Our prior incorporates the knowledge that neighbouring regions are more correlated than distant ones—this leads to an interesting experimental design problem because information can be pooled between regions. We can also compute the true EIG and optimal design ξ^* analytically. For full details, see Appendix B.4.

We compare the performance of four estimation and optimization methods on this problem, see Fig. 3 for the results. The three gradient-based methods (ACE, PCE, BA) perform best, with the BO baseline struggling in dimensions $D \geq 6$, even though the latter has access to an EIG oracle. PCE performed well in low dimensions, but degraded as the dimension increases

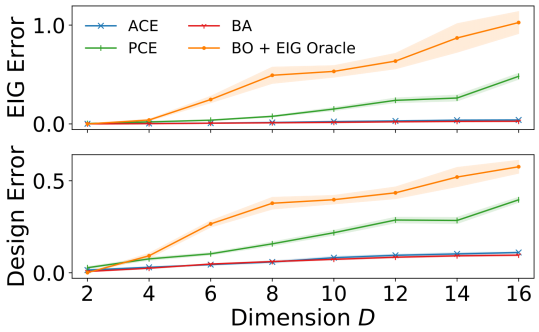


Figure 3: Mean absolute EIG and design errors for the marketing model in Sec. 4.4 averaged over 10 runs. The EIG is normalized such that an EIG error of unity corresponds to doing no better than a uniform budget, i.e. $\xi_i = B/D$ for $i = 1, ..., D$.

and sampling from the prior becomes increasingly inefficient, ACE and BA avoid this by learning adaptive proposal distributions. We note that since in this case the family of variational distributions used in ACE and BA include the true posterior, both methods yield similar performance.

4.5 Biomolecular docking

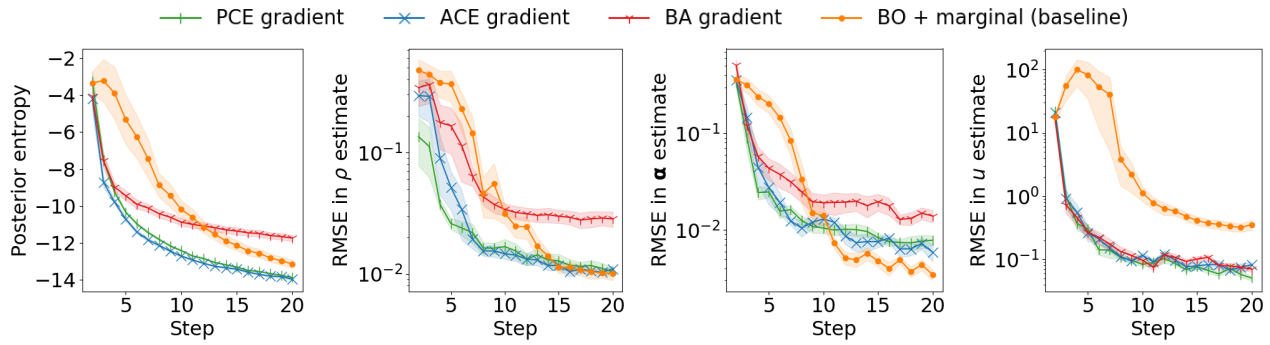
We now consider an experimental design problem of interest to the pharmacology community. Having demonstrated that our one-stage gradient methods compare favourably with two stage approaches, we now compare against designs crafted by domain experts.

In molecular docking, computational techniques are used to predict the binding affinity between a compound and a receptor. When synthesized in the lab, the two may bind—this is called a *hit*. Learning a well-calibrated hit-rate model can guide how many compounds to evaluate for additional objectives, such as drug-likeness or toxicity, before experimental testing. Lyu et al. (2019) modelled the probability of outcome y_i being a hit, given the predicted binding affinity, or docking score $\xi_i \in [-75, 0]$, as

$$p(y_i = 1|\theta, \xi) = \text{bottom} + \frac{\text{top} - \text{bottom}}{1 + e^{-(\xi_i - \text{ee50}) \times \text{slope}}} \quad (20)$$

where $\theta = (\text{top}, \text{bottom}, \text{ee50}, \text{slope})$ with priors given in Appendix B.5.

Of 150 million compounds, Lyu et al. (2019) selected a batch of compounds to experimentally test to best fit the sigmoid hit-rate model. They considered 6 candidate designs and selected one that maximized the EIG estimated by NMC. Here, we instead apply gradient-based BOED to search across candidate designs which consist of 100 docking scores $\xi_1, ..., \xi_{100}$. To evaluate our final designs, we present upper and lower bounds on the final EIG: see Table 2. We see that all gradient methods are able to outperform experts in terms of



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

Figure 4: Improvement in the posterior in the sequential CES experiment. Each step took 120 seconds for each method. We present the mean and ± 1 standard error from 10 runs. See Sec. 4.6 for details.

Table 2: Biomolecular docking results showing the mean and ± 1 s.e. from 10 runs. For the expert, we took the best design of Lyu et al. (2019) appropriately rescaled to consist of 100 docking scores for comparison.

Method	EIG lower bound	EIG upper bound
ACE	$\mathbf{1.0835 \pm 0.0003}$	$\mathbf{1.0852 \pm 0.0001}$
PCE	1.0825 ± 0.0002	1.0839 ± 0.0002
BA	1.0780 ± 0.0003	1.0794 ± 0.0003
Expert	1.0191	1.0227

EIG, and that ACE appears the best of the gradient methods. Figure 5 shows our designs are qualitatively different to those produced by experts.

4.6 Constant elasticity of substitution

We finally turn to *iterated* experimental design in which we produce designs, generate data and make inference repeatedly. This problem therefore captures the end-to-end-process of experimentation and inference.

We consider an experiment in behavioural economics that was previously also considered by Foster et al. (2019). In this experiment, a participant is asked to compare baskets $\mathbf{x}, \mathbf{x}'$ of goods. The model assumes that their response (on a slider) is based on the difference in utility of the baskets, and the constant elasticity of substitution (CES) model (Arrow et al., 1961) governed by latent variables $(\rho, \boldsymbol{\alpha}, u)$ is then used for this utility. The aim is to learn $(\rho, \boldsymbol{\alpha}, u)$ characterizing the participant's utility. In the experiment, there are 20 sequential steps of experimentation with the same participant. We compare our gradient-based approach against the most successful approach of Foster et al. (2019) that approximates the marginal density to form an upper bound on EIG, and BO to optimize ξ. For full details, see Appendix B.6.

Figure 4 shows that gradient-based methods are effective on this problem; both ACE and PCE decrease the posterior entropy and RMSEs on the latent variables faster and further than the baseline, whereas BA does not do so well. We suggest that the similar performance of ACE and PCE is due to the smaller changes in the posterior at middle and late steps, after much data has been accumulated: when the posterior does not change much at each step, $p(\theta|y_{1:t-1}, \xi_{1:t-1})$ forms an effective proposal for estimating $p(y_t|\xi_t)$.

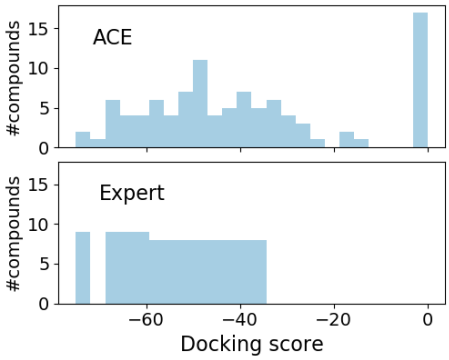


Figure 5: Designs for the biomolecular docking problem obtained by ACE and by Lyu et al. (2019). Designs consist of 100 docking scores at which to test compounds.

5 CONCLUSIONS

We have introduced a new approach for Bayesian experimental design that does away with the two stages of estimating EIG and separately optimizing over Ξ. We use stochastic gradients to maximize a lower bound on $I(\xi)$ and so find optimal designs by solving a single optimization problem. This unification leads to substantially improved performance, especially on high-dimensional design problems.

Of the three lower bounds, I_{BA}, I_{ACE} and I_{PCE}, we note that in all five experiments ACE generally did as well as the better of BA and PCE: we therefore recommend it as the default choice. BA performed well when the inference network could closely approximate the true posterior; PCE performed well when the prior was an adequate proposal for estimating $p(y|\xi)$ and does not require the training of variational parameters.

Acknowledgements

AF gratefully acknowledges funding from EPSRC grant no. EP/N509711/1. YWT'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. TR gratefully acknowledges funding from Tencent AI Labs and a junior research fellowship supported by Christ Church, Oxford.

References

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.

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.

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

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. *Journal of Machine Learning Research*, 2018.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

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

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

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.

Adam Foster, Martin Jankowiak, Elias Bingham, Paul Horsfall, Yee Whye Teh, Thomas Rainforth, and Noah Goodman. Variational Bayesian Optimal Experimental Design. In *Advances in Neural Information Processing Systems 32*, pages 14036–14047. Curran Associates, Inc., 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Steven Kleinegesse and Michael Gutmann. Efficient Bayesian experimental design for implicit models. *arXiv preprint arXiv:1810.09912*, 2018.

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

Jiankun Lyu, Sheng Wang, Trent E Balius, Isha Singh, Anat Levit, Yurii S Moroz, Matthew J O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, et al. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224, 2019.

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

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.

Jason Pacheco and John Fisher. Variational information planning for sequential decision making. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2028–2036, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180, 2019.

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

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.

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

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.

Artem Sobolev. Thoughts on mutual information: More estimators, 2019. URL http://artem.sobolev.name/posts/2019-08-10-thoughts-on-mutual-information-more-estimators.html.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J Maddison. Doubly reparameterized gradient estimators for monte carlo objectives. *arXiv preprint arXiv:1810.04152*, 2018.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

Andrew B Watson. Quest+: A general multidimensional bayesian adaptive psychometric method. *Journal of Vision*, 17(3):10–10, 2017.

Sue Zheng, Jason Pacheco, and John Fisher. A robust approach to sequential information theoretic planning. In *International Conference on Machine Learning*, pages 5941–5949, 2018.

A GRADIENT-BASED BOED

We begin with the proof of Theorem 1, which we restate for convenience.

Theorem 1. *For any model $p(\theta)p(y|\theta,\xi)$ and inference network $q_\phi(\theta|y)$, we have the following:*

1. *I_{ACE} is a lower bound on $I(\xi)$ and we can characterize the error term as an expected KL divergence:*

$$I(\xi) - I_{ACE}(\xi,\phi,L) = \mathbb{E}_{p(y|\xi)}\left[KL\left(P(\theta_{0:L}|y)\middle\|\prod_\ell q_\phi(\theta_\ell|y)\right)\right] \geq 0,$$

$$P(\theta_{0:L}|y) = \frac{1}{L+1}\sum_{\ell=0}^{L} p(\theta_\ell|y,\xi)\prod_{k\neq\ell} q_\phi(\theta_k|y).$$

2. *As $L\to\infty$, we recover the true EIG: $\lim_{L\to\infty} I_{ACE}(\xi,\phi,L) = I(\xi)$.*

3. *The ACE bound is monotonically increasing in L: $I_{ACE}(\xi,\phi,L_2) \geq I_{ACE}(\xi,\phi,L_1)$ for $L_2 \geq L_1 \geq 0$.*

4. *If the inference network equals the true posterior $q_\phi(\theta|y) = p(\theta|y,\xi)$, then $I_{ACE}(\xi,\phi,L) = I(\xi), \forall L$.*

We add the further technical assumption that $p(\theta)p(y|\theta,\xi)/q_\phi(\theta|y)$ is bounded.

Proof. To begin with 1., we have the error term $\delta = I(\xi) - I_{ACE}(\xi,\phi,L)$ which can be written

$$\delta = \mathbb{E}\left[\log \frac{\frac{1}{L+1}\sum_{\ell=0}^{L}\frac{p(\theta_\ell)p(y|\theta_\ell,\xi)}{q_\phi(\theta_\ell|y)}}{p(y|\xi)}\right] \tag{21}$$

$$= \mathbb{E}\left[\log \frac{\frac{1}{L+1}\sum_{\ell=0}^{L} p(\theta_\ell|y)\prod_{k\neq\ell} q_\phi(\theta_k|y)}{\prod_{\ell=0}^{L} q_\phi(\theta_\ell|y)}\right] \tag{22}$$

$$= \mathbb{E}\left[\log \frac{P(\theta_{0:\ell}|y)}{\prod_{\ell=0}^{L} q_\phi(\theta_\ell|y)}\right] \tag{23}$$

where the expectation is over $p(y|\xi)p(\theta_0|y,\xi)\prod_{\ell=1}^{L} q_\phi(\theta_\ell|y)$. Note that the integrand is symmetric under a permutation of the labels $0,...,L$, so its expectation will be the same over the distribution $p(y|\xi)p(\theta_\ell|y,\xi)\prod_{k\neq\ell} q_\phi(\theta_k|y)$. Since $P(\theta_{0:L})$ is a mixture of distributions of this form, this then implies that the expectation will be the same if it is taken over the distribution $p(y|\xi)P(\theta_{0:L})$, yielding

$$\delta = \mathbb{E}_{p(y|\xi)P(\theta_{0:L}|y)}\left[\log \frac{P(\theta_{0:L}|y)}{\prod_{\ell=0}^{L} q_\phi(\theta_\ell|y)}\right] \tag{24}$$

which is the expected KL divergence required. We therefore have $\delta \geq 0$.

For 2., we use that $p(\theta)p(y|\theta,\xi)/q_\phi(\theta|y)$ is bounded. The ACE denominator is a consistent estimator of the marginal likelihood. Indeed,

$$\frac{1}{L+1}\frac{p(\theta_0)p(y|\theta_0,\xi)}{q_\phi(\theta_0|y)} \to 0 \tag{25}$$

and

$$\frac{1}{L+1}\sum_{\ell=1}^{L}\frac{p(\theta_\ell)p(y|\theta_\ell,\xi)}{q_\phi(\theta_\ell|y)} \to p(y|\xi) \text{ a.s.} \tag{26}$$

as $L\to\infty$ by the Strong Law of Large Numbers, since

$$\mathbb{E}_{q_\phi(\theta|y)}\left[\frac{p(\theta)p(y|\theta,\xi)}{q_\phi(\theta|y)}\right] = p(y|\xi). \tag{27}$$

This establishes the a.s. pointwise convergence of the ACE integrand to $\log p(y|\theta_0,\xi)/p(y|\xi)$. Hence by Bounded Convergence Theorem,

$$\hat{I}_{ACE}(\xi,\phi,L) \to I(\xi) \tag{28}$$

as $L\to\infty$.

To establish 3., we use a similar approach to 1. We let $\varepsilon = I_{ACE}(\xi,\phi,L_2) - I_{ACE}(\xi,\phi,L_1)$. Then

$$\varepsilon = \mathbb{E}\left[\log \frac{\frac{1}{L_1+1}\sum_{\ell=0}^{L_1}\frac{p(\theta_\ell)p(y|\theta_\ell,\xi)}{q(\theta_\ell|y)}}{\frac{1}{L_2+1}\sum_{\ell=0}^{L_2}\frac{p(\theta_\ell)p(y|\theta_\ell,\xi)}{q(\theta_\ell|y)}}\right] \tag{29}$$

$$= \mathbb{E}\left[\log \frac{Q(\theta_{0:L_2}|y)}{\frac{1}{L_2+1}\sum_{\ell=0}^{L_2} p(\theta_\ell|y)\prod_{k\neq\ell} q(\theta_k|y)}\right] \tag{30}$$

where the expectation is over $p(y|\xi)p(\theta_0|y,\xi)\prod_{\ell=1}^{L_2} q(\theta_\ell|y)$ and

$$Q(\theta_{0:L_2}|y) = \frac{1}{L_1+1}\sum_{\ell=0}^{L_1} p(\theta_\ell|y)\prod_{k\neq\ell}^{L_2} q(\theta_k|y). \tag{31}$$

As in 1., the integrand is unchanged if we permute the labels $0,...,L_1$. By this symmetry, the expectation is the same when taken over the distribution $p(y|\xi)Q(\theta_{0:L_2}|y)$. We therefore recognise ε as the expectation of a KL divergence. Hence $\varepsilon \geq 0$ as required.

4. follows by Bayes Theorem, i.e.

$$\frac{p(\theta)p(y|\theta,\xi)}{p(\theta|y,\xi)} = p(y|\xi). \tag{32}$$

which completes the proof. □

We also present the proof of Theorem 2.

Theorem 2. *Consider a model $p(\theta)p(y|\theta,\xi)$ and inference network $q_\phi(\theta|y)$. Let $f_\psi(\theta,y) \geq 0$ be an unnormalized likelihood approximation. Then,*

$$I(\xi) \geq \mathbb{E}\left[\log \frac{f_\psi(\theta_0,y)}{\frac{1}{L+1}\sum_{\ell=0}^{L}\frac{p(\theta_\ell)f_\psi(\theta_\ell,y)}{q_\phi(\theta_\ell|y)}}\right] \tag{14}$$

where the expectation is over $p(\theta_0)p(y|\theta_0,\xi)q_\phi(\theta_{1:L}|y)$.

Proof. Initially, we note that the contrastive samples $\theta_1, ..., \theta_L$ do not carry additional information about θ_0. Formally, we consider the mutual information between θ_0 and the random variable $(y, \theta_1, ..., \theta_L)$. Using the Chain Rule for mutual information we have

$$\begin{aligned} &\mathrm{MI}(\theta_0; (y, \theta_1, ..., \theta_L)) \\ &= \mathrm{MI}(\theta_0; y) + \mathrm{MI}(\theta_0; (\theta_1, ..., \theta_L)|y) \end{aligned} \tag{33}$$

Now $\mathrm{MI}(\theta_0; (\theta_1, ..., \theta_L)|y) = 0$ since θ_ℓ $(\ell > 0)$ are conditionally independent of θ_0 given y. Therefore

$$\mathrm{MI}(\theta_0; (y, \theta_1, ..., \theta_L)) = \mathrm{MI}(\theta_0; y) = I(\xi). \tag{34}$$

We now use the Donsker-Varadhan representation of mutual information (Donsker and Varadhan, 1975). Specifically, for random variables A, B with joint distribution $p(a, b)$ and any measurable function $T(a, b)$ we have

$$\begin{aligned} &\mathrm{MI}(A; B) \\ &\geq \mathbb{E}_{p(a,b)}[T(a, b)] - \log \mathbb{E}_{p(a)p(b)}\left[e^{T(a,b)}\right]. \end{aligned} \tag{35}$$

We now use this representation with $a = \theta_0, b = (y, \theta_1, ..., \theta_L)$ and $T(a, b)$ the integrand

$$T(\theta_0, (y, \theta_{1:L})) = \log \frac{f_\psi(\theta_0, y)}{\frac{1}{L+1}\sum_{\ell=0}^{L} \frac{p(\theta_\ell) f_\psi(\theta_\ell, y)}{q_\phi(\theta_\ell|y)}}. \tag{36}$$

We compute the second term in (35), $Z = \mathbb{E}_{p(a)p(b)}\left[e^{T(a,b)}\right]$.

$$Z = \mathbb{E}_{p(\theta_0)p(y|\xi)q_\phi(\theta_{1:L}|y)}\left[\frac{f_\psi(\theta_0, y)}{\frac{1}{L+1}\sum_{\ell=0}^{L} \frac{p(\theta_\ell) f_\psi(\theta_\ell, y)}{q_\phi(\theta_\ell|y)}}\right] \tag{37}$$

$$= \mathbb{E}_{p(y|\xi)q_\phi(\theta_{0:L}|y)}\left[\frac{\frac{p(\theta_0) f_\psi(\theta_0, y)}{q_\phi(\theta_0|y)}}{\frac{1}{L+1}\sum_{\ell=0}^{L} \frac{p(\theta_\ell) f_\psi(\theta_\ell, y)}{q_\phi(\theta_\ell|y)}}\right] \tag{38}$$

$$= \mathbb{E}_{p(y|\xi)q_\phi(\theta_{0:L}|y)}\left[\frac{\frac{1}{L+1}\sum_{\ell=0}^{L} \frac{p(\theta_\ell) f_\psi(\theta_\ell, y)}{q_\phi(\theta_\ell|y)}}{\frac{1}{L+1}\sum_{\ell=0}^{L} \frac{p(\theta_\ell) f_\psi(\theta_\ell, y)}{q_\phi(\theta_\ell|y)}}\right] \tag{39}$$

$$= 1 \tag{40}$$

where the second to last line follows by symmetry. This establishes that $\log Z = 0$, and so (14) constitutes a valid lower bound on $I(\xi)$. That is

$$I(\xi) \geq \mathbb{E}\left[\log \frac{f_\psi(y, \theta_0)}{\frac{1}{L+1}\sum_{\ell=0}^{L} \frac{p(\theta) f_\psi(y, \theta_\ell)}{q_\phi(\theta_\ell, y)}}\right] \tag{41}$$

which completes the proof. □

The following theorem establishes a condition under which the maximum of the ACE objective converges to the maximum of the EIG as $L \to \infty$.

Theorem 3. *Consider a model $p(\theta)p(y|\theta, \xi)$ such that*

$$C \triangleq \sup_{\xi \in \Xi} \inf_{\phi \in \Phi} \mathbb{E}_{p(\theta)p(y|\theta,\xi)}\left[\frac{p(\theta|y, \xi)}{q_\phi(\theta|y, \xi)}\right] < \infty. \tag{42}$$

and $I^* \triangleq \sup_{\xi \in \Xi} I(\xi) < \infty$. Let $q_\phi(\theta|y)$ be an inference network and let

$$I_L = \sup_{\xi \in \Xi, \phi \in \Phi} I_{ACE}(\xi, \phi, L). \tag{43}$$

Then,

$$0 \leq I^* - I_L \leq \frac{C - 1}{L + 1} \tag{44}$$

and in particular $I_L \to I^*$ as $L \to \infty$.

Proof. We have $0 \leq I^* - I_L$ since I_{ACE} is a lower bound on $I(\xi)$ by Theorem 1.

Next, we consider $\Delta(\xi, \phi, L) = I(\xi) - I_{ACE}(\xi, \phi, L)$. We have

$$\Delta = \mathbb{E}_{p(\theta_0)p(y|\theta_0,\xi)q_\phi(\theta_{1:L}|y)}\left[\log \frac{Y_L}{p(y|\xi)}\right] \tag{45}$$

where

$$Y_L = \frac{1}{L+1}\sum_{\ell=0}^{L} w_\ell \quad \text{and} \quad w_\ell = \frac{p(\theta_\ell)p(y|\theta_\ell, \xi)}{q_\phi(\theta_\ell|y)}; \tag{46}$$

we write (45) as

$$\Delta = \mathbb{E}\left[\log\left(1 + \frac{Y_L - p(y|\xi)}{p(y|\xi)}\right)\right] \tag{47}$$

and we apply the inequality $\log(1 + x) \leq x$ to give

$$\Delta \leq \mathbb{E}\left[\frac{Y_L - p(y|\xi)}{p(y|\xi)}\right]. \tag{48}$$

We now observe that for $\ell > 0$, $\mathbb{E}_{q_\phi(\theta_\ell|y)}[w_\ell] = p(y|\xi)$ and hence, taking a partial expectation over $\theta_{1:L}$ we have

$$\Delta \leq \mathbb{E}_{p(\theta_0)p(y|\theta_0,\xi)}\left[\frac{w_0 - p(y|\xi)}{(L+1)p(y|\xi)}\right] \tag{49}$$

$$\leq \frac{1}{L+1}\left(\mathbb{E}_{p(\theta_0)p(y|\theta_0,\xi)}\left[\frac{p(\theta_0|y, \xi)}{q_\phi(\theta_0|y)}\right] - 1\right) \tag{50}$$

Hence

$$I^* - I_L = \sup_{\xi \in \Xi} I(\xi) - \sup_{\xi \in \Xi, \phi \in \Phi} I_{ACE}(\xi, \phi, L)] \tag{51}$$

$$\leq \sup_{\xi \in \Xi}[I(\xi) - \sup_{\phi \in \Phi} I_{ACE}(\xi, \phi, L)] \tag{52}$$

$$\leq \sup_{\xi \in \Xi} \inf_{\phi \in \Phi}[\Delta(\xi, \phi, L)] \tag{53}$$

$$\leq \frac{C - 1}{L + 1} \tag{54}$$

as required. □

A.1 Double reparametrization

We have the ϕ-gradient of the ACE objective

$$\frac{\partial I_{ACE}}{\partial \phi} = \mathbb{E}_{p(\theta_0)p(y|\theta_0,\xi)}\left[-\frac{\partial \mathcal{L}}{\partial \phi}\bigg|_{\theta_0,y}\right] \tag{55}$$

where $\mathcal{L}$ is our estimate of the marginal likelihood with gradient

$$\frac{\partial \mathcal{L}}{\partial \phi}\bigg|_{\theta_0,y} = \frac{\partial}{\partial \phi}\,\mathbb{E}_{q_\phi(\theta_{1:L}|y)}\left[\log\left(\sum_{\ell=0}^{L} w_\ell\right)\middle|\,\theta_0, y\right] \tag{56}$$

where

$$w_\ell = \frac{p(\theta_\ell)p(y|\theta_\ell,\xi)}{q_\phi(\theta_\ell|y)}. \tag{57}$$

If $q_\phi(\theta|y)$ is reparameterizable as a function of ϕ, then we can apply *double* reparameterization to this gradient. Indeed, were it not for the w_0 term, this would be exactly the IWAE of Burda et al. (2015). We exploit the double reparameterization of Tucker et al. (2018) with a minor variation to account for w_0 to obtain a low variance gradient estimator.

The doubly reparametrized gradient for ACE takes the form

$$\frac{\partial I_{ACE}}{\partial \phi} = \mathbb{E}_{p(\theta_0)p(y|\theta_0,\xi)q_\phi(\theta_{1:L}|y)}\left[\sum_{\ell=0}^{L} v_\ell\right] \tag{58}$$

where

$$v_0 = \frac{w_0}{\sum_{m=0}^{L} w_m}\frac{\partial}{\partial \phi}\log q_\phi(\theta_0|y) \tag{59}$$

and for $\ell > 0$

$$v_\ell = -\left(\frac{w_\ell}{\sum_{m=0}^{L} w_m}\right)^2 \frac{\partial \log w_\ell}{\partial \theta_\ell}\frac{\partial \theta_\ell}{\partial \phi}. \tag{60}$$

A.2 Alternative gradient

We begin with an observation: the true integrand when computing the EIG as an expectation over $p(\theta)p(y|\theta,\xi)$ is given by

$$g_*(y,\theta,\xi) = \log\frac{p(y|\theta,\xi)}{p(y|\xi)}. \tag{61}$$

Recall the score function identity

$$\mathbb{E}_{p(x|\xi)}\left[\frac{\partial}{\partial \xi}\log p(x|\xi)\right] = 0. \tag{62}$$

We have

$$\begin{aligned}
&\mathbb{E}_{p(\theta)p(y|\theta,\xi)}\left[\frac{\partial g_*}{\partial \xi}\right] && (63)\\
&= \mathbb{E}_{p(\theta)p(y|\theta,\xi)}\left[\frac{\partial}{\partial \xi}\log\frac{p(y|\theta,\xi)}{p(y|\xi}\right] && (64)\\
&= \mathbb{E}_{p(\theta)}\left(\mathbb{E}_{p(y|\theta,\xi)}\left[\frac{\partial}{\partial \xi}p(y|\theta,\xi)\right]\right) && (65)\\
&\quad - \mathbb{E}_{p(y|\xi)}\left[\frac{\partial}{\partial \xi}\log p(y|\xi)\right] \\
&= 0 && (66)
\end{aligned}$$

by two applications of the score function identity. This suggests that, as g becomes close to g_*, the $\partial g/\partial \xi$ term in (16) has expectation close to zero, and primarily contributes variance to the gradient estimator.

Theorem 2 shows that if we remove the $\partial g/\partial \xi$ term, the resulting algorithm still optimizes a valid lower bound on $I(\xi)$. Specifically, removing this term is equivalent to the following gradient-coordinate algorithm. First, we choose the family $f_\psi(\theta, y)$ to be $p(y|\theta,\psi)$. Then at time step t we do the following

1. Set $\psi_t = \xi_t$
2. Take a gradient step with respect to (ξ, ϕ) to update ξ_t, ϕ_t

Importantly, the new gradient does not include a $\partial g/\partial \xi$ term, but is the gradient of a valid lower bound on EIG. In practice, this alternative gradient did not yield substantially different performance from the standard approach of including the $\partial g/\partial \xi$ term. All our experiments used the standard approach for simplicity.

B EXPERIMENTS

B.1 Implementation

All experiments were implemented in PyTorch 1.4.0 (Paszke et al., 2019) and Pyro 0.3.4 (Bingham et al., 2018). Supporting code can be found at `https://github.com/ae-foster/pyro/tree/sgboed-reproduce`, see 'README.md' for details on how to run the experiments.

B.2 Death process

We place the prior $\theta \sim \text{LogNormal}(0,1)$ on the infection rate and have the likelihood

$$\begin{aligned} I_1 &\sim \text{Binomial}(N, e^{-\theta\xi_1}) \\ I_2 &\sim \text{Binomial}(N - I_1, e^{-\theta\xi_2}). \end{aligned} \tag{67}$$

We also have the constraint $\xi_1, \xi_2 \geq 0$.

Table 3: Death process. We present the final EIG for each method (computed using NMC with 200000 samples).

Method	EIG mean ±1 s.e.
ACE	$\mathbf{0.9830 \pm 0.0001}$
PCE	0.9822 ± 0.0001
BA	0.9822 ± 0.0002
ACE without RB	0.9789 ± 0.0006
PCE without RB	0.9710 ± 0.0025
BA without RB	0.9322 ± 0.0045
BO with NMC	0.9732 ± 0.0009

For each method, we fixed a computational budget of 120 seconds, and did 100 independent runs. For gradient methods, we used the Adam optimizer (Kingma and Ba, 2014) with learning rate 10^{-3} and the default momentum parameters. The inference network made a separate Gaussian approximation to the posterior for each of the 66 outcomes. To evaluate $I(\xi)$ for comparison we used NMC with a large number of samples: 20000 for Figure 2 and 200000 for the final values in the caption and in Table 3. For the BO, we used a Matern52 kernel with variance 1 and lengthscale 0.25, and the GP-UCB1 algorithm (Srinivas et al., 2009) for acquisition.

We used the following number of samples for our Rao-Blackwellized estimators

Method	Number of samples
ACE	10 + 660
PCE	10
BA	10
NMC	2000

B.3 Regression

We consider the following prior on $\theta = (\mathbf{w}, \sigma)$

$$w_j \overset{\text{i.i.d.}}{\sim} \text{Laplace}(1) \text{ for } j = 1, ..., p \tag{68}$$
$$\sigma \sim \text{Exponential}(1) \tag{69}$$

with the likelihood

$$y_i \sim N\left(\sum_{j=1}^{p} \xi_{ij} w_j, \sigma\right) \text{ for } i = 1, ...n. \tag{70}$$

This represents a standard regression model, although with non-Gaussian prior distributions we cannot compute the posterior or true EIG analytically. To ensure the EIG has a finite maximum, we impose the following constraint

$$\sum_j |\xi_{ij}| = 1 \text{ for } i = 1, ..., n. \tag{71}$$

In practice, we set $n = p = 20$.

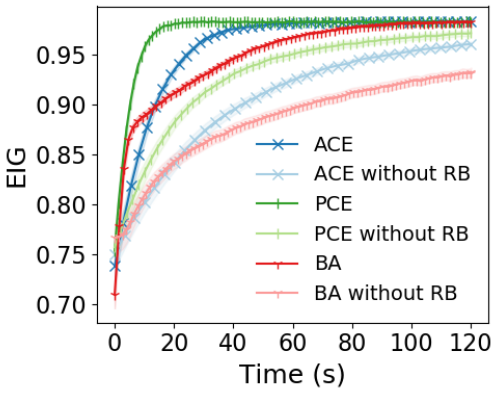


Figure 6: The EIG against time for the death process: comparing Rao-Blackwellization against no Rao-Blackwellization. Each method had a 120 second time budget.

For each of our five methods, we fixed the computational budget to 15 minutes and did 10 independent runs. For gradient methods, we used a learning rate of 10^{-3} and the Adam optimizer with default momentum parameters. The inference network used the following variational family

$$\mathbf{w} \sim N(\boldsymbol{\mu}, s\Sigma_0) \tag{72}$$
$$\sigma \sim \Gamma(\alpha, \beta) \tag{73}$$

and we used a neural network with the following architecture

Operation	Size	Activation
Input $\to$ H1	64	ReLU
H1 $\to$ H2	64	ReLU
H2 $\to \boldsymbol{\mu}$	20	-
H2 $\to (\alpha, \beta)$	2	Softplus
H2 $\to s$	1	Softplus
Σ_0	20×20	-

For BO and random search, point evaluations of $I(\xi)$ were made using VNMC. Each VNMC evaluation took 1000 steps, with the optimization as above (but with ξ fixed). We used a GP with Matern52 kernel with lengthscale 5, variance 10. We used a GP-UCB1 acquisition rule, and terminated once 15 minutes had passed. For random search, we sampled designs using a standard unit Gaussian.

We used the following number of samples

Method	Inner samples L	Outer samples N
ACE	10	10
PCE	10	10
BA	n/a	100
VNMC	10	10

To evaluate designs, we used ACE/VNMC. We first trained ACE using the same procedure as above, for

20000 steps. Then we made the final ACE/VNMC evaluations using the fixed inference network and $L = 2.5 \times 10^3$ inner samples, $N = 10^5$ outer samples.

B.4 Advertising

We introduce a LogNormal likelihood and a D-dimensional latent variable $\boldsymbol{\theta}$ governed by a Normal prior, the joint density of our model is

$$p(\boldsymbol{y}, \boldsymbol{\theta}|\boldsymbol{\xi}) = \mathcal{LN}(\boldsymbol{y}|\boldsymbol{\theta} \odot \boldsymbol{\xi}, \sigma^2 \boldsymbol{\xi}) \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \boldsymbol{\Lambda}_0) \tag{74}$$

where σ controls the observation noise, $\boldsymbol{\Lambda}_0$ is a non-diagonal precision matrix and $\odot$ denotes the Hadamard product. Since there are correlations among the D regions, the optimal advertising budget (w.r.t. gaining information about $\boldsymbol{\theta}$) allocates more money to the regions that are tightly correlated.

Throughout we assume that the number of regions D is even. We set the budget to scale with the number of dimensions, $B = \frac{D}{2}$, set $\sigma = 1$ and choose the prior precision matrix to be

$$\boldsymbol{\Lambda}_0 = (1 + \tfrac{1}{D})\mathbb{I}_D - \tfrac{1}{D}\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}} \qquad \boldsymbol{u}^{\mathrm{T}} \equiv (\alpha, ..., \alpha, 1, ..., 1)$$

where the first $\frac{D}{2}$ components of $\boldsymbol{u}$ equal α and the last $\frac{D}{2}$ components equal 1. We shall see that $\alpha = 0.1$ controls the degree of asymmetry in the optimal design. Discarding an irrelevant constant, we can compute the exact EIG using the formula:

$$I(\boldsymbol{\xi}) = \frac{1}{2} \log\det \boldsymbol{\Lambda}_{\text{post}} \qquad \boldsymbol{\Lambda}_{\text{post}} = \boldsymbol{\Lambda}_0 + \frac{1}{\sigma^2}\text{diag}(\boldsymbol{\xi})$$

Using the matrix determinant lemma for rank-1 matrix updates we can then compute

$$\log\det \boldsymbol{\Lambda}_{\text{post}} = \sum_{i=1}^{D} \log(1 + \tfrac{1}{D} + \xi_i) + \log\left(1 - \sum_{i=1}^{\frac{D}{2}} \left\{ \frac{\alpha^2}{1 + \frac{1}{D} + \xi_i} \right\} - \sum_{i=1+\frac{D}{2}}^{D} \left\{ \frac{1}{1 + \frac{1}{D} + \xi_i} \right\}\right).$$

By symmetry the optimum (it is easy to check that it is a maximum) of $\text{EIG}(\boldsymbol{\xi})$ will satisfy $\xi_i = \xi_{i+1}$ for $i = 1, ..., \frac{D}{2} - 1, \frac{D}{2} + 1, ..., D$. In other words $\boldsymbol{\xi}$ is entirely specified by ξ_1 and ξ_D, which must satisfy $\xi_1 + \xi_D = 1$ because of the constraint on the budget $B = \frac{D}{2}$. Thus we have reduced the EIG maximization problem to a univariate optimization problem that can easily be solved to machine precision, for example by gradient methods or brute force bisection. This analytic solution gives us the ground truth EIG, used within BO and for evaluation, and the true optimal design, used for evaluation.

For each of the four methods (ACE, PCE, BA and BO) we fix the computational budget to 120 seconds per design optimization. For the gradient-based methods this corresponds to 1×10^4, 2×10^4, and 1.8×10^4 gradient steps for ACE, PCE, and BA, respectively. For the BO baseline, we run 110 steps of a GP-UCB-like algorithm (Srinivas et al., 2009) in batch-mode, resulting in a total budget of 1650 function evaluations of the EIG oracle. Note that for all four methods the runtime dependence on the dimension D is negligible in the regime in which we are operating; consequently we use the same number of gradient or BO steps for all D.

For the gradient-based methods, we use the Adam optimizer with default momentum hyperparameters and an initial learning rate of $\ell_0 = 0.1$ that is exponentially decayed towards a final learning rate ℓ_f that depends on the particular method. In particular we set $\ell_f = 1 \times 10^{-4}$, $\ell_f = 1 \times 10^{-5}$, and $\ell_f = 3 \times 10^{-4}$ for the ACE, PCE, and BA methods, respectively. For the BO baseline, we used a Matérn kernel with a fixed length scale $\ell = 0.2$. These hyperparameters were chosen by running a grid search with $D = 16$ and choosing hyperparameters that minimized the mean absolute EIG error.

Finally we note that in Fig. 3 at each dimension D we normalize the EIG by the factor

$$Z = \text{EIG}(\boldsymbol{\xi}^*) - \text{EIG}(\boldsymbol{\xi}_{\text{uniform}}) \tag{75}$$

where $\boldsymbol{\xi}^*$ and $\boldsymbol{\xi}_{\text{uniform}}$ are the optimal and uniform budget designs, respectively. Consequently after normalization the absolute error for the uniform budget design $\boldsymbol{\xi}_{\text{uniform}}$ is equal to 1.

B.5 Biomolecular docking

For the docking model, we used the following independent priors

$$\text{top} \sim \text{Beta}(25, 75) \tag{76}$$
$$\text{bottom} \sim \text{Beta}(4, 96) \tag{77}$$
$$\text{ee50} \sim N(-50, 15^2) \tag{78}$$
$$\text{slope} \sim N(-0.15, 0.1^2). \tag{79}$$

For the design $\xi = (\xi_1, ..., \xi_{100})$ we had 100 binary responses

$$y_i \sim \text{Bern}\left(\text{bottom} + \frac{\text{top} - \text{bottom}}{1 + e^{-(\xi_i - \text{ee50}) \times \text{slope}}}\right). \tag{80}$$

For gradient methods, we used the Adam optimizer with learning rate 10^{-3} and default momentum parameters. For each method, we took 5×10^5 gradient steps (each method converged within this number of

steps). The inference network was mean-field with the same distributional families as the prior. We used the following neural architecture

Operation	Size	Activation
Input → H1	64	ReLU
H1 → H2	64	ReLU
H2 → top	2	Softplus
H2 → bottom	2	Softplus
H2 → ee50 mean	1	-
H2 → ee50 s.d.	1	Softplus
H2 → slope mean	1	-
H2 → slope s.d.	1	Softplus

We used the following number of samples

Method	Inner samples L	Outer samples N
ACE	10	10
PCE	10	10
BA	n/a	100

For the expert method, the design of Lyu et al. (2019), which comprised 580 compounds, was subsampled to comprise 100 compounds for a fair comparison.

For evaluation, we used ACE/VNMC, first training ACE for 25000 steps using the same learning rate as above. With the fixed inference network, we made ACE and VNMC evaluations using $L = 2 \times 10^3$ inner samples, $N = 4 \times 10^6$ outer samples.

B.6 Constant elasticity of substitution

We used the exact set-up of Foster et al. (2019). Specifically, we take $U(\mathbf{x}) = (\sum_i x_i^\rho \alpha_i)^{1/\rho}$ and place the following priors on $\rho, \boldsymbol{\alpha}, u$

$$\rho \sim \text{Beta}(1,1) \tag{81}$$
$$\boldsymbol{\alpha} \sim \text{Dirichlet}([1,1,1]) \tag{82}$$
$$\log u \sim N(1,3) \tag{83}$$
$$\mu_\eta = u \cdot (U(\mathbf{x}) - U(\mathbf{x}')) \tag{84}$$
$$\sigma_\eta = \tau u \cdot (1 + \|\mathbf{x} - \mathbf{x}'\|) \tag{85}$$
$$\eta \sim N(\mu_\eta, \sigma_\eta^2) \tag{86}$$
$$y = f(\eta) \tag{87}$$

where f is the censored sigmoid function and $\tau = 0.005$. All designs $\xi = (\mathbf{x}, \mathbf{x}')$ were constrained to $[0, 100]^6$.

For gradient methods, we used the Adam optimizer with learning rate 10^{-3} and default momentum parameters. To make the design process 120 seconds per step, we used the following number of gradient steps

Method	Number of steps
ACE	1500
PCE	2500
BA	5000

We found that there was insufficient time to effectively train a neural network guide. Instead we used a mean-field variational family with the same distributional families as the prior, and a linear model using the following features: $\text{logit}(y), \log|\text{logit}(y)|, \mathbf{1}(y > 0.5)$.

We used the following number of samples

Method	Inner samples L	Outer samples N
ACE	10	10
PCE	10	10
BA	n/a	100

For the baseline, we used the marginal upper bound of Foster et al. (2019) with the same variational family used in that paper—an f-transformed Normal with additional point masses at the end-points. We used a GP with a Matérn52 kernel, lengthscale 20, variance set from data, and a GP-UCB1 algorithm to make acquisitions which were done in batches of 8.

At each stage of the sequential experiment, the posterior was fitted using mean-field variational inference using the same distributional families as the prior.

C FUTURE WORK

In this paper, we have focused on continuous design spaces in which gradient methods are applicable. One possible extension of our work would be to facilitate a unified one-stage approach to experimental design over *discrete* design spaces. In this case, the lower bounds I_{BA}, I_{ACE} and I_{PCE} remains valid, and performing a joint maximization over (ξ, ϕ) on any of these objectives may be an attractive choice, although gradient optimization would no longer be appropriate for ξ. We envisage that one could apply existing methods for discrete optimization to the joint optimization problem over design and variational parameters. For instance, a continuous relaxation of the discrete variables, or MCMC-style updates on the discrete variables might be used. Future work might further explore this direction.