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 $n = p = 20$.

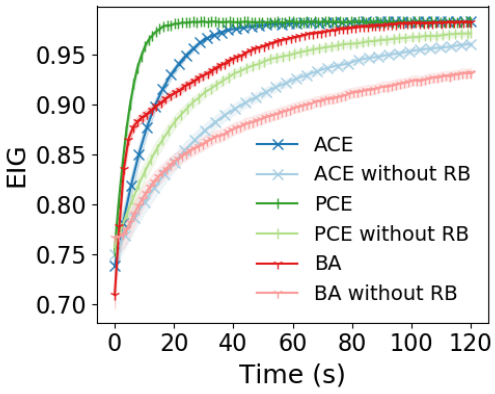

Figure 6: The EIG against time for the death process: comparing Rao-Blackwellization against no Rao-Blackwellization. Each method had a 120 second time budget.

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.