Variational Optimal Experiment Design: Efficient Automation of Adaptive Experiments

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

Abstract

Bayesian optimal experimental design (OED) is a principled framework for making efficient use of limited experimental resources. Unfortunately, the applicability of OED is hampered by the difficulty of obtaining accurate estimates of the expected information gain (EIG) for different experimental designs. We introduce a class of fast EIG estimators that leverage amortised variational inference and show that they provide substantial empirical gains over previous approaches. We integrate our approach into a deep probabilistic programming framework, thus making OED accessible to practitioners at large.

1 Introduction

Tasks as seemingly diverse as designing a study to elucidate an aspect of human cognition, selecting the next query point in an active learning loop, and tuning a microscope all fit tidily within the framework of optimal experiment design (OED). Though numerous approaches to choosing optimal designs exist [5, 30], arguably the most natural is to maximize the expected information gained from the experiment [6, 19, 25, 31, 38]. This information theoretic formulation of experiment design is very general and has been applied in numerous settings, including psychology [22], Bayesian optimisation [14], active learning [12], bioinformatics [37], and neuroscience [32].

Following Bayesian decision theory [20], one begins with a likelihood model and a prior over model parameters, and then chooses the design that maximises the *expected* information gain (EIG) of the parameters of interest. In other words, one seeks the design that, in expectation over possible experimental outcomes, most reduces the entropy of the posterior for the target parameters. This OED framework is particularly powerful in a sequential context, where it allows the results of previous experiments to be used in guiding the designs for future experiments.

For OED to have the broadest possible impact, it should be partially, or even fully, automated. This motivates embedding OED in a probabilistic programming language so that the full experimental pipeline—from model specification and inference to design optimization—can be carried out in a unified system [23, 25]. Designing such a system entails several challenges. Our core contribution is to introduce efficient variational methods for EIG estimation that are applicable to a wide variety of models. The first method, which is related to amortised variational inference [9, 16, 24, 28, 34], employs an approximate posterior distribution, parameterized by the design and experimental outcome. In a similar manner the second method employs a variational distribution for the marginal density over experimental outcomes for a given design. Both methods can benefit from recent advances in defining flexible families of amortised variational distributions using neural networks (e.g. normalising flows [27, 35]). For this reason we developed our system[1] in Pyro [4], a deep probabilistic programming language that provides first class support for neural networks and variational methods.

[1]Our implementation will soon be made available at `https://github.com/uber/pyro`.

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

We note that our methods are also directly applicable to the calculation of mutual informations and Kullback-Leibler divergences of marginal distributions. They thus have a large number of potential applications in estimating objectives for training deep generative models [1, 7, 8, 11, 36].

2 EIG Estimation

Consider a model specified by the joint density $p(y, \theta|d) = p(y|\theta, d)p(\theta)$, where d is the (non-random) design of the experiment, θ is a latent random variable and y is the observed outcome of the experiment. Then the EIG is given by the expected reduction in entropy from the prior to the posterior under the marginal distribution over outcomes $p(y|d) = \mathbb{E}_{p(\theta)}[p(y|\theta, d)]$, that is

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

$$= \iint p(y, \theta|d) \log \frac{p(\theta|y, d)}{p(\theta)} \, dy \, d\theta = \iint p(y, \theta|d) \log \frac{p(y|\theta, d)}{p(y|d)} \, dy \, d\theta, \tag{2}$$

further details on which are given in Appendix B. Computing (2) is extremely challenging, since neither $p(\theta|y, d)$, $p(y|d)$, nor the outer integral can, in general, be found in closed form: it forms a nested estimation. Despite noted drawbacks [26], nested Monte Carlo (NMC) estimation of the EIG remains the go-to approach in the literature [22, 38]. Most notably, while simple Monte Carlo estimators converge with a mean squared error rate $\mathcal{O}(N^{-1})$ in the total number of samples, NMC estimators converge at a much slower $\mathcal{O}(N^{-2/3})$ rate and are biased, though consistent [26].

2.1 Variational Optimal Experimental Design

The NMC approach is inefficient because it constructs an independent estimate of $p(\theta|y, d)$ or $p(y|d)$ for each outcome y. Our key insight is that by taking a variational approach, we can instead learn an *amortized* approximation for either $p(\theta|y, d)$ or $p(y|d)$, and then use this approximation to efficiently estimate the EIG. In essence, the estimate of $p(y_1|d)$ provides information about $p(y_2|d)$ for similar y_1 and y_2 (presuming some smoothness in the density) and so it is more efficient to learn the functional form for $p(y|d)$ (or $p(\theta|y, d)$), than to treat separate values of y as distinct inference problems.

More concretely, we construct a variational bound, $\mathcal{L}_p(d)$, using the amortized posterior approximation $q_p(\theta|y, d)$:

$$\text{EIG}(d) = \iint p(y, \theta|d) \log \frac{p(\theta|y, d)q_p(\theta|y, d)}{q_p(\theta|y, d)} \, dy \, d\theta + H[p(\theta)] \tag{3}$$

$$= \iint p(y, \theta|d) \log q_p(\theta|y, d) \, dy \, d\theta + H[p(\theta)] + \mathbb{E}_{p(y|d)}\left[\text{KL}\left(p(\theta|y, d)||q_p(\theta|y, d)\right)\right] \tag{4}$$

$$\geq \iint p(y, \theta|d) \log q_p(\theta|y, d) \, dy \, d\theta + H[p(\theta)] \triangleq \mathcal{L}_p(d). \tag{5}$$

In analogy with variational inference, this bound is tight when $q_p(\theta|y, d) = p(\theta|y, d)$. Alternatively, we can instead introduce a marginal density approximation $q_m(y|d)$, giving an upper bound $\mathcal{U}_m(d)$:

$$\text{EIG}(d) = \iint p(y, \theta|d) \log p(y|\theta, d) \, dy \, d\theta - \int p(y|d) \log \frac{p(y|d)q_m(y|d)}{q_m(y|d)} dy \tag{6}$$

$$= \iint p(y, \theta|d) \log p(y|\theta, d) \, dy \, d\theta - \int p(y|d) \log q_m(y|d) \, dy - \text{KL}\left(p(y|d)||q_m(y|d)\right) \tag{7}$$

$$\leq \iint p(y, \theta|d) \log p(y|\theta, d) \, dy \, d\theta - \int p(y|d) \log q_m(y|d) \, dy \triangleq \mathcal{U}_m(d), \tag{8}$$

where the bound again becomes tight for $q_m(y|d) = p(y|d)$.

In certain cases, $p(y|\theta, d)$ cannot be computed pointwise. For example, this is the case in the presence of nuisance variables, also known as random effects. These are additional latent variables, ψ, that we do not consider variables of interest, such that we do not want to waste resources reducing our uncertainty for them. Such models arise frequently in scientific applications, for instance accounting for individual variation between participants in a survey. With random effects ψ we have

$$p(y|\theta, d) = \int p(y|\theta, \psi, d)p(\psi|\theta)d\psi \tag{9}$$

which is typically intractable. Fortunately, the posterior bound is unchanged by random effects. Our marginal method can be adapted to this random effects setting by introducing an approximation to $p(y|\theta, d)$ as shown in Appendix C. To estimate the relevant integrals, we can still draw exact samples from $p(y|\theta, d)$ by drawing from the joint $p(y|\theta, \psi, d)p(\psi|\theta)$.

2.2 Estimation

Just as in variational inference, the bounds in the previous section can be maximised with stochastic gradient methods [29]. Concretely, suppose $\mathcal{Q}$ is a family of amortised variational approximations $q_p(\theta|y, d; \phi)$ indexed by ϕ. We can estimate EIG by maximizing the lower bound $\mathcal{L}_p(d; \phi)$:

$$\text{EIG}(d) \approx \max_{\phi} \mathcal{L}_p(d; \phi) = \max_{\phi} \left\{ \iint p(y, \theta|d) \log q_p(\theta|y, d; \phi)\, dy\, d\theta \right\} + H[p(\theta)] \tag{10}$$

To do so only requires that we can generate samples from the model, $y_i, \theta_i \sim p(y, \theta|d)$; in a probabilistic programming context this corresponds to running the model forwards with no conditioning. We can then construct the required Monte Carlo estimates for the gradient as

$$\nabla_\phi \mathcal{L}_p(d; \phi) \approx \nabla_\phi \left\{ \frac{1}{N} \sum_{i=1}^{N} \log q_p(\theta_i|y_i, d; \phi) \right\} \quad \text{where} \quad y_i, \theta_i \overset{\text{i.i.d.}}{\sim} p(y, \theta|d), \tag{11}$$

noting that no re-parameterization is required as $p(y, \theta|d)$ is independent of ϕ. An analogous scheme can be constructed for the upper bound $\mathcal{U}_m(d; \phi)$, expect that we now perform a minimization. Maximizing over the design space can then be done with a variety of optimization methods; in our experiments we make use of Bayesian optimization [33].

3 Experiments

We validate our EIG estimators on a selection of Generalized Linear Mixed Models (GLMMs). These serve as useful benchmarks, since they are workhorse models in many different scientific disciplines. Our results are summarized in Table 1 and Fig. 1-6 in Appendix D.7. In all six cases, both estimators (i.e. the posterior method based on q_p and the marginal method based on q_m) give significantly lower variance than the NMC baseline, and in two of the three cases a significantly lower bias as well. We note that NMC especially struggled with random effects (LinReg + RE). More worryingly still, the bias of the NMC estimator can exhibit strong systematic variation as a function of the design, see Fig. 2 for instance. This is problematic because it can lead to the choice of a significantly suboptimal design. It is also worth emphasizing the utility of having multiple variational methods at our disposal: while the marginal method yields poor EIG estimates for the model with a large output dimension, the posterior method delivers high quality estimates.

Next, we consider examples (NΓ^{-1}Reg and NΓ^{-1}Reg + RE) that are not purely Gaussian. Here our methods still perform well, despite the variational families not containing the true posterior or marginal, with both approaches having lower bias and variance than NMC.

Our final example (SigReg) goes beyond typical GLMMs and introduces a sigmoid non-linearity to the classical mixed effects regression model. Here the marginal approach performed well, but the posterior method struggled. We postulate that this is due to the time taken to train a larger number of parameters for the more complex variational posterior approximation used here.

	LinReg		LinReg + RE		LinReg-HD		NΓ^{-1}Reg		NΓ^{-1}Reg + RE		SigReg	
	Bias	2std	Bias	2std	Bias	2std	Bias	2std	Bias	2std	Bias	2std
NMC	1.37	1.93	5.33	3.84	3.13	2.97	2.66	2.36	5.65	6.26	-0.017	0.13
Posterior	-0.23	0.25	-0.55	0.41	-0.29	0.31	-0.70	0.58	-0.65	0.53	-0.086	0.14
Marginal	0.34	0.15	0.36	0.20	4.57	0.29	1.45	0.58	0.09	0.26	0.0045	0.057

Table 1: Bias and variance (we report 2σ) of EIG estimation. This was averaged over 10 runs and 11 designs, each method being limited to run for 10 seconds total (for SigReg there were 10 runs, 15 designs and 80 seconds of computation). For more details on the models and experimental setup see Appendix D. Note that the directions of the bias for the posterior and marginal match the fact that they are lower and upper bounds, as would be expected.

Acknowledgments

AF gratefully acknowledges funding from EPSRC grant no. EP/N509711/1. AF acknowledges the support of Uber AI Labs. YWT's and TR's research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 617071.

References

[1] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

[2] D. Barber and F. Agakov. The im algorithm: a variational approach to information maximization. *Advances in Neural Information Processing Systems*, 16:201, 2004.

[3] I. Belghazi, S. Rajeswar, A. Baratin, R. D. Hjelm, and A. Courville. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[4] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman. Pyro: Deep Universal Probabilistic Programming. *arXiv preprint arXiv:1810.09538*, 2018.

[5] G. E. Box. Choice of response surface design and alphabetic optimality. Technical report, Wisconsin University-Madison Mathematics Research Center, 1982.

[6] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[7] T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

[8] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[9] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

[10] M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

[11] B. Esmaeili, H. Wu, S. Jain, N. Siddharth, B. Paige, and J.-W. van de Meent. Hierarchical Disentangled Representations. *arXiv.org*, Apr. 2018.

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

[13] P. Hennig and C. J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(Jun):1809–1837, 2012.

[14] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, pages 918–926, 2014.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

[17] S. Kleinegesse and M. Gutmann. Efficient bayesian experimental design for implicit models. *arXiv preprint arXiv:1810.09912*, 2018.

[18] J. Lewi, R. Butera, and L. Paninski. Efficient active learning with generalized linear models. In *Artificial Intelligence and Statistics*, pages 267–274, 2007.

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

[20] D. V. Lindley. *Bayesian statistics, a review*, volume 2. SIAM, 1972.

[21] Q. Long, M. Scavino, R. Tempone, and S. Wang. Fast estimation of expected information gains for bayesian experimental designs based on laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, 259:24–39, 2013.

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

[23] L. Ouyang, M. H. Tessler, D. Ly, and N. Goodman. Practical optimal experiment design with probabilistic programs. *arXiv preprint arXiv:1608.05046*, 2016.

[24] B. Paige and F. Wood. Inference networks for sequential monte carlo in graphical models. In *International Conference on Machine Learning*, pages 3040–3049, 2016.

[25] T. Rainforth. *Automating Inference, Learning, and Design using Probabilistic Programming*. PhD thesis, 2017.

[26] T. Rainforth, R. Cornish, H. Yang, A. Warrington, and F. Wood. On nesting monte carlo estimators. In *International Conference on Machine Learning*, pages 4264–4273, 2018.

[27] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[28] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

[29] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[30] E. G. Ryan, C. C. Drovandi, and A. N. Pettitt. Fully bayesian experimental design for pharmacokinetic studies. *Entropy*, 17(3):1063–1089, 2015.

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

[32] B. Shababo, B. Paige, A. Pakman, and L. Paninski. Bayesian inference and online experimental design for mapping neural microcircuits. In *Advances in Neural Information Processing Systems*, pages 1304–1312, 2013.

[33] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[34] A. Stuhlmüller, J. Taylor, and N. Goodman. Learning stochastic inverses. In *Advances in neural information processing systems*, pages 3048–3056, 2013.

[35] E. Tabak and C. V. Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[36] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[37] J. Vanlier, C. A. Tiemann, P. A. Hilbers, and N. A. van Riel. A bayesian approach to targeted experiment design. *Bioinformatics*, 28(8):1136–1142, 2012.

[38] B. T. Vincent and T. Rainforth. The darc toolbox: automated, flexible, and efficient delayed and risky choice experiments using bayesian adaptive design. 2017.

A Related Work

For completeness, we include a limited discussion of related work that was not referred to in the main text. In the context of information maximization in noisy channels, [2] uses a variational bound on mutual information that is closely related to our 'posterior' bound $\mathcal{L}_p(d)$. In the context of deep learning (e.g. adversarially trained generative models), [3] uses a related bound that is based on the Donsker-Varadhan representation of the KL divergence [10] to estimate mutual information between high dimensional continuous random variables. With particular attention to implicit models, [17] uses density ratio estimation via logistic regression to construct an algorithm for approximate Bayesian OED; this approach is related to our 'marginal' bound $\mathcal{U}_m(d)$ but is not a variational method (no bound is being optimized). The Laplace approximation is used in the context of OED by various authors, see for example [21, 30], with it being used particularly effectively in the specific context of generalized linear models by [18]. Finally, for a recent review of Bayesian OED with a comprehensive set of references (especially from the statistics community) see [30].

B Background on Expected Information Gain Maximization

In this section, we provide a more detailed background on the EIG maximization approach. Under our model, the outcome of the experiment given a design d is distributed according to

$$p(y|d) = \int p(y, \theta|d) d\theta = \int p(y|\theta, d) p(\theta) d\theta, \tag{12}$$

where we have used the fact that $p(\theta) = p(\theta|d)$ because θ is independent of the design. Our aim is to choose the optimal design d under some criterion. We therefore define a utility function, $U(y, d)$, representing the utility of running an experiment according to a design d and obtaining an outcome y. Typically our aim is to maximize information gathered from the experiment, and so we set $U(y, d)$ to be the gain in Shannon information between the prior and the posterior

$$U(y, d) = H[p(\theta)] - H[p(\theta|y, d)] = \int p(\theta|y, d) \log(p(\theta|y, d)) d\theta - \int p(\theta) \log(p(\theta)) d\theta. \tag{13}$$

However, we are still uncertain about the outcome. Thus, we use the expectation of $U(y, d)$ with respect to $p(y|d)$ as our target:

$$\begin{aligned} \text{EIG}(d) &= \int p(y|d) \left(\int p(\theta|y, d) \log(p(\theta|y, d)) d\theta - \int p(\theta) \log(p(\theta)) d\theta \right) dy \\ &= \iint p(y, \theta|d) \log \left(\frac{p(\theta|y, d)}{p(\theta)} \right) d\theta dy \end{aligned} \tag{14}$$

noting that this corresponds to the mutual information between the parameters θ and the observations y. The Bayesian-optimal design is then given by

$$d^* = \arg\max_{d \in \mathcal{D}} \text{EIG}(d). \tag{15}$$

where $\mathcal{D}$ is the permissible set of designs. We can intuitively interpret d^* as being the design that most reduces the uncertainty in θ on average over possible experimental results. If our likelihood model is correct, i.e. if experimental outcomes are truly distributed according to $p(y|\theta, d)$ for a given θ and d, then it is easy to see from the above definition that d^* is the true optimal design, in terms of information gain, given our current information about the parameters $p(\theta)$. In practice, our likelihood model is an approximation of the real world. Nonetheless, EIG maximization remains a very powerful and statistically principled approach that is typically significantly superior to more heuristic alternatives. For example, the state-of-the-art entropy based Bayesian optimization acquisition strategies are particular cases of Bayesian OED [13, 14]. However, a major drawback to the EIG maximization approach is that it is typically difficult and computationally intensive to carry out. Not only does it represent an optimization of an intractable expectation, this expectation is itself nested because the integrand is itself intractable due to the $p(\theta|y, d)$ term.

We note that the EIG admits a number of different interpretations: 1) as the expected gain in Shannon information; 2) as the expected Kullback-Leibler (KL) divergence between posterior and prior; 3) as the expected epistemic uncertainty in the response y; 4) as the negative average posterior entropy (plus a constant); and 5) as the mutual information between y and θ.

B.1 Automating Sequential Design Problems

We have thus far assumed that there is no previous data (i.e. design-outcome pairs). Though this static experimental design setup is of use in its own right, the full potential of EIG maximization is not realized until one considers using it in *sequential* settings. Here EIG provides a framework for adaptively making an optimal series of decisions in an online fashion in the presence of uncertainty. For example, imagine a psychology trial where we ask a participant a series of questions to learn about certain behavior characteristics. If a human is conducting this experiment they are likely to adapt the questions they ask as they learn about the participant to try and maximize the information gathered. Sequential EIG maximization provides a mathematical framework for reasoning about and optimizing such processes, thereby providing a means of developing effective machine learning systems to carry out such tasks.

We can generalize to the sequential design setting by incorporating data in the standard Bayesian fashion such that at experiment iteration t, we replace $p\left(\theta\right)$ with $p\left(\theta|d_{1:t-1}, y_{1:t-1}\right)$, where $d_{1:t-1}$ and $y_{1:t-1}$ are respectively the designs and outcomes at previous iterations. The likelihood $p\left(y_t|\theta, d_t\right)$, on the other hand, is unchanged (presuming it is a parametric distribution) as, conditioned on θ and d, the current outcome is independent of the previous data. Putting this together, we get that the expected information gain criteria for the sequential case is

$$\begin{aligned}\text{EIG}_t(d) = &\iint p\left(\theta|d_{1:t-1}, y_{1:t-1}\right) p(y_t|\theta, d_t) \log(p(y_t|\theta, d_t)) d\theta dy_t \\ &- \int p(y_t|y_{1:t-1}, d_{1:t}) \log(p(y_t|y_{1:t-1}, d_{1:t})) dy_t.\end{aligned} \tag{16}$$

We can now see that these terms are the same as in the non-sequential case, except that expectations are taken with respect to $p\left(\theta|d_{1:t-1}, y_{1:t-1}\right)$ rather than $p(\theta)$.

C Marginal method with random effects

Starting from

$$\text{EIG}(d) = \iint p(y, \theta|d) \log p(y|\theta, d) dy\, d\theta - \int p(y|d) \log p(y|d) dy, \tag{17}$$

we can bound each term separately in terms of two approximate densities: $q_m(y|d)$ for the marginal and $q_\ell(y|\theta, d)$ for the likelihood. Specifically, we have from Gibbs' inequality

$$-\int p(y|d) \log p(y|d) dy \leq -\int p(y|d) \log q_m(y|d) dy \tag{18}$$

$$\iint p(y, \theta|d) \log p(y|\theta, d) dy\, d\theta \geq \iint p(y, \theta, |d) \log q_\ell(y|\theta, d) dy\, d\theta\,. \tag{19}$$

Here we can no longer derive a direct bound on the EIG, but we can still use these inequalities to train an approximate marginal density q_m and an amortized approximate likelihood q_ℓ, which will yield the true EIG if they match the true marginal and likelihood respectively. Namely, suppose $\mathcal{Q}_1$ is a family of variational distributions $q_m(y|d; \phi_1)$ indexed by ϕ_1 and $\mathcal{Q}_2$ is a family of variational distributions $q_\ell(y|\theta, d; \phi_2)$ indexed by ϕ_2. Then a suitable objective for learning ϕ_1, ϕ_2 is

$$\mathcal{D}_{\phi_1,\phi_2}(d) \triangleq -\iint p(y, \theta, |d) \log q_\ell(y|\theta, d; \phi_2) dy\, d\theta - \int p(y|d) \log q_m(y|d; \phi_1) dy \tag{20}$$

$$\{\phi_1^*, \phi_2^*\} = \text{argmin}_{\phi_1,\phi_2} \mathcal{D}_{\phi_1,\phi_2}(d) \tag{21}$$

where the optimization can be performed using stochastic gradient methods, as in the main paper. Once these approximations have been learned, we can plug them back into (17) to give

$$\text{EIG}(d) \approx \iint p(y, \theta, |d) \log q_\ell(y|\theta, d; \phi_2^*) dy\, d\theta - \int p(y|d) \log q_m(y|d; \phi_1^*) dy \tag{22}$$

which can then itself be approximated by conventional Monte Carlo sampling.

D Experiments

D.1 LinReg

A classical Bayesian linear regression model has the following form

$$\theta \sim N(\mu_\theta, \Sigma_{\theta\theta}) \tag{23}$$

$$y|\theta, d \sim N(X_d\theta, \sigma^2 I) \tag{24}$$

where X_d is the design matrix.

In our LinReg example, we took:

$$\mu_\theta = 0 \tag{25}$$

$$\Sigma_{\theta\theta} = \begin{pmatrix} 10^2 & 0 \\ 0 & 0.1^2 \end{pmatrix} \tag{26}$$

$$\sigma^2 = 1 \tag{27}$$

$$X_d = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \text{ a } (10 \times 2) \text{ matrix} \tag{28}$$

with all 11 possible designs considered.

We chose families of variational distributions that include the true posterior (or true marginal). For the amortised posterior, we set $\phi = (\Lambda, \delta, \Sigma_\text{p})$ and let

$$q_p(\theta|y, d; \phi) \sim N(\mu_\text{p}, \Sigma_\text{p}) \tag{29}$$

$$\text{where } \mu_\text{p} = (X_d^T X_d + \Lambda)^{-1} X_d^T (y + \delta) \tag{30}$$

and Λ is a diagonal matrix and Σ_p is positive definite. For the marginal, we simple take $\phi = (\mu_\text{m}, \Sigma_\text{m})$ and

$$q_m(y|d; \phi) \sim N(\mu_\text{m}, \Sigma_\text{m}) \tag{31}$$

Finally, for each of our variational methods we used the Adam optimizer [15] with a learning rate specified below. Each iteration used N_t samples, with T iterations in total. We used N samples for the final evaluation. NMC settings are N, M [38] and we took the advice of the authors to set $N = M^2$.

The exact parameter settings, to get about 10 seconds of computation for each method, were

NMC		Posterior				Marginal			
N	M	N_t	T	lr	N	N_t	T	lr	N
110^2	110	10	1200	0.05	500	10	1200	0.05	500

D.2 LinReg + RE

In this experiment, we extended the model to include random effects. Specifically,

$$\theta \sim N(\mu_\theta, \Sigma_{\theta\theta}) \tag{32}$$

$$\psi \sim N(\mu_\psi, \Sigma_{\psi\psi}) \tag{33}$$

$$y|\theta, d \sim N(X_{d,\theta}\theta + X_{d,\psi}\psi, \sigma^2 I) \tag{34}$$

where

$$\mu_\psi = 0 \tag{35}$$

$$\Sigma_{\psi\psi} = I_{10} \tag{36}$$

$$X_{d,\psi} = I_{10} \tag{37}$$

and $X_{d,\theta}$ was the X_d from the previous experiment. Here θ is the random variable of interest, while ψ is a nuisance variable that needs to be integrated out. The variational distribution for the likelihood, q_ℓ, was the same as q_m, except that the mean was shifted by $X_{d,\theta}\theta$.

The exact parameter settings, to get about 10 seconds of computation for each method, were

NMC		Posterior				Marginal			
N	M	N_t	T	lr	N	N_t	T	lr	N
52^2	52	10	150	0.05	500	10	600	0.05	500

D.3 LinReg-HD

This experiment was identical to LinReg, except that we took X_d to have dimensions 20×2, with 11 designs as before. We also altered the marginal variational distribution to reflect the new dimension of y. Other than that, the specification of all variational distributions was identical.

The exact parameter settings, to get about 10 seconds of computation for each method, were

NMC		Posterior				Marginal			
N	M	N_t	T	lr	N	N_t	T	lr	N
90^2	90	10	1000	0.05	500	10	700	0.05	500

D.4 NΓ^{-1}Reg

We changed the model to

$$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta) \tag{38}$$

$$\theta \sim N(\mu_\theta, \Sigma_{\theta\theta}) \tag{39}$$

$$y|\theta, \sigma^2, d \sim N(X_d\theta, \sigma^2 I) \tag{40}$$

where $\alpha = 3$ and $\beta = 2$.

We used a mean-field posterior variational distribution. For θ, we used the same variational distribution as for LinReg. For σ^2 we used an inverse Gamma variational distribution. We augmented the parameters ϕ with α_p, b_0 and took $\beta_\text{p} = b_0 + \frac{1}{2}(y^T y - y^T X_d \mu_\text{p})$. Then

$$q_p(\sigma^2|y, d; \phi) \sim \Gamma^{-1}(\alpha_\text{p}, \beta_\text{p}) \tag{41}$$

The marginal variational distribution was as in LinReg (a Gaussian).

The exact parameter settings, to get about 10 seconds of computation for each method, were

NMC		Posterior				Marginal			
N	M	N_t	T	lr	N	N_t	T	lr	N
110^2	110	10	800	0.05	500	10	1200	0.05	500

D.5 NΓ^{-1}Reg + RE

This model was identical to the previous one. However, we now consider σ^2 to be a random effect rather than a parameter of interest.

The exact parameter settings, to get about 10 seconds of computation for each method, were

NMC		Posterior				Marginal			
N	M	N_t	T	lr	N	N_t	T	lr	N
60^2	60	10	900	0.05	500	10	600	0.05	500

D.6 SigReg

We first considered a mixed effects regression model, but with different parameters

$$\mu_\theta = 10 \tag{42}$$
$$\Sigma_{\theta\theta} = \left(8^2\right) \tag{43}$$
$$\mu_\psi = 1 \tag{44}$$
$$\Sigma_{\psi\psi} = \left((\tfrac{1}{4})^2\right) \tag{45}$$
$$\sigma^2 = 2^2 \tag{46}$$
$$X_{d,\theta} = (1) \tag{47}$$
$$X_{d,\psi} = (x) \text{ with } x \in [-30, 30] \tag{48}$$
$$\tag{49}$$

We also altered the model, adding a sigmoid transformation and censoring

$$y|\theta, \psi, d = \text{censor}(\text{sigmoid}(y')) \tag{50}$$

where y' is the output of a linear regression model. The censoring affects the output near the endpoints and maps $(0, \epsilon]$ to ϵ and $[1-\epsilon, 1)$ to $1-\epsilon$. This censoring both aids numerical stability and makes the problem more interesting by censoring information gained from values very near to 0 or 1.

We chose families of variational distributions as similar to LinReg as possible. For the amortised posterior, we set $\phi = (\Lambda, \delta, \Sigma_\text{p}, \mu_0, \Sigma_0, \mu_1, \Sigma_1)$. For $y \in (\epsilon, 1-\epsilon)$ we took

$$q_p(\theta|y, d; \phi) \sim N(\mu_\text{p}, \Sigma_\text{p}) \tag{51}$$
$$\text{where } \mu_\text{p} = (X_d^T X_d + \Lambda)^{-1} X_d^T (y + \delta) \tag{52}$$

as before. However, for $y = \epsilon$ we took

$$q_p(\theta|y, d; \phi) \sim N(\mu_0, \Sigma_0) \tag{53}$$

and for $y = 1 - \epsilon$ we took

$$q_p(\theta|y, d; \phi) \sim N(\mu_1, \Sigma_1). \tag{54}$$

This allowed us to correctly deal with the censoring.

For the marginal, we simple took $\phi = (\mu_\text{m}, \Sigma_\text{m})$ and

$$q_m(y'|d; \phi) \sim N(\mu_\text{m}, \Sigma_\text{m}) \tag{55}$$

followed by

$$y|d; \phi = \text{censor}(\text{sigmoid}(y')) \tag{56}$$

The marginal variational distribution was the same, except with the mean shifted by $X_{d,\theta}\theta$.

The exact parameter settings, to get about 80 seconds of computation for each method, were

NMC		Posterior				Marginal			
N	M	N_t	T	lr	N	N_t	T	lr	N
60^2	60	10	900	0.05	3200	10	600	0.05	500

Since no ground truth was available, we used NMC with a very large number of samples ($N = 160^2, M = 160$).

D.7 Figures

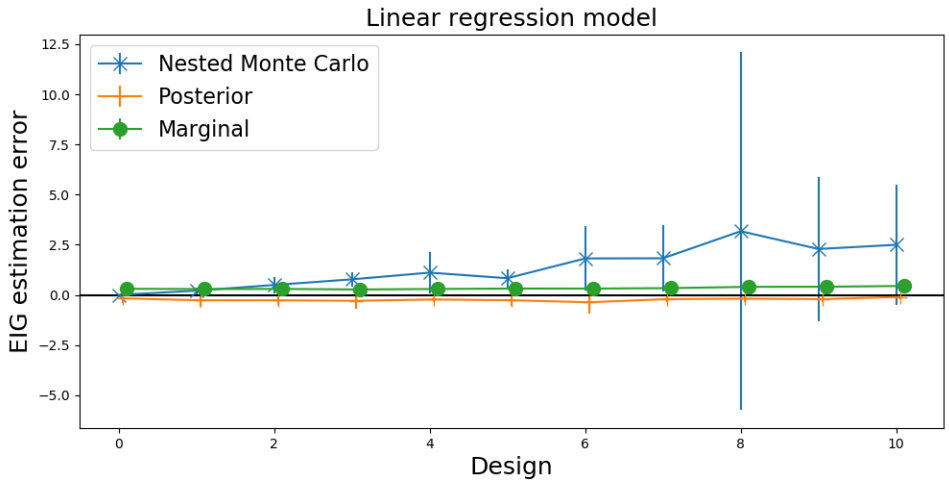

Figure 1: LinReg: EIG estimates for a linear regression model over 11 designs. We plot the mean and twice the standard deviation from 10 runs. Computational time was set to 10 seconds for comparison.

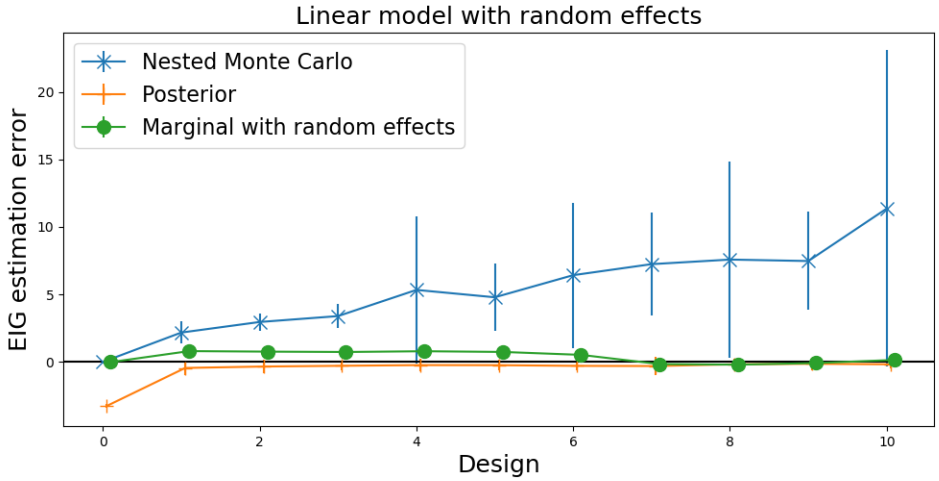

Figure 2: LinReg + RE: EIG estimates for a linear regression model with random effects. Settings as in Figure 1.

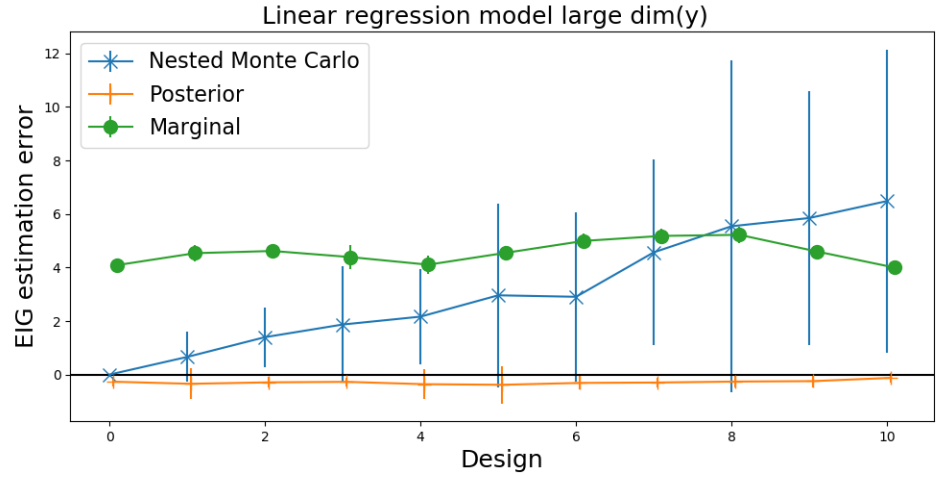

Figure 3: LinReg-HD: with settings as in Figure 1.

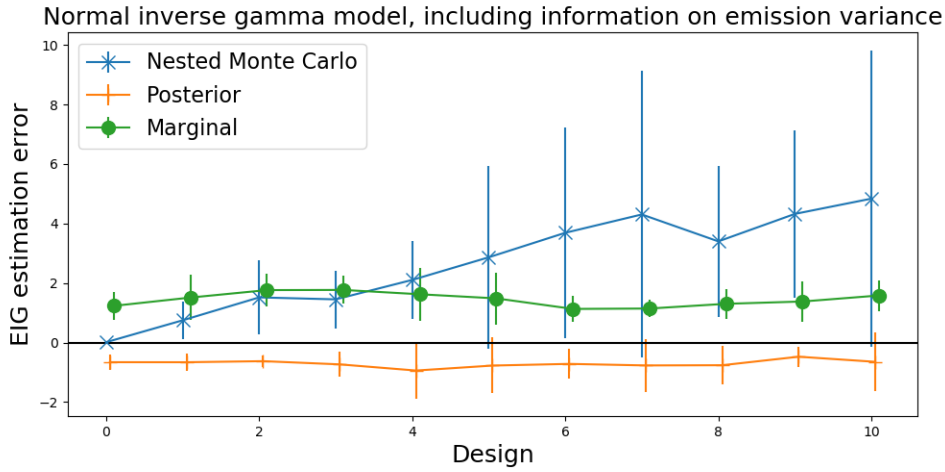

Figure 4: NIG^{-1}Reg: EIG estimates for a Normal inverse-Gamma model. Settings as in Figure 1.

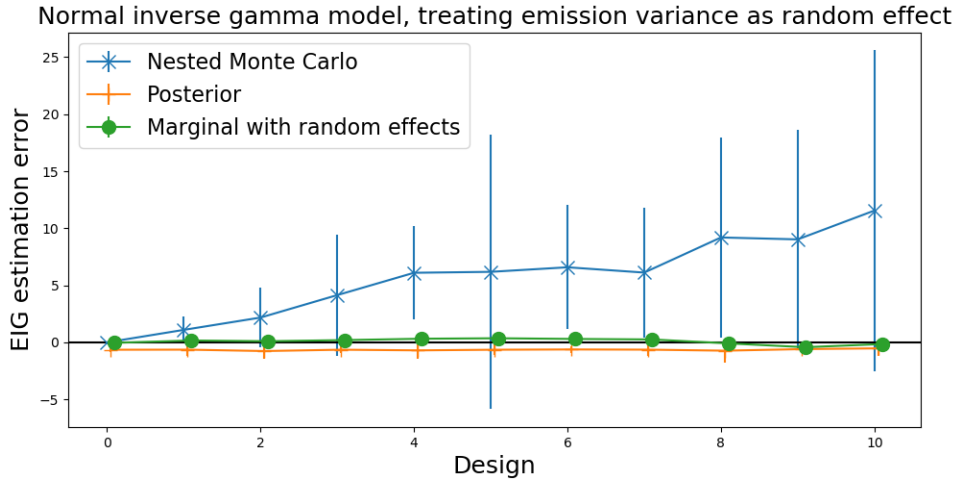

Figure 5: NIG^{-1}Reg + RE: EIG estimates for a Normal inverse-Gamma model treating σ^2 as a random effect. Settings as in Figure 1.

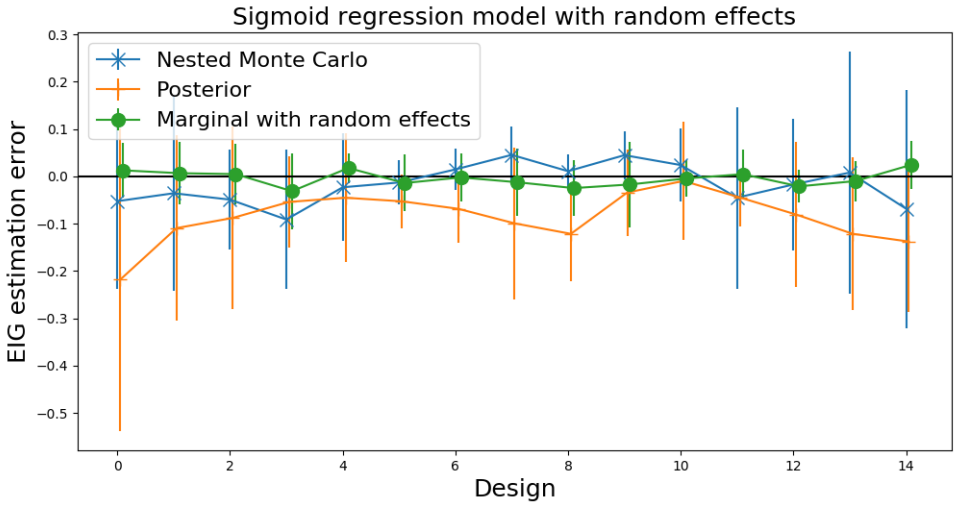

Figure 6: SigReg: EIG estimates for a sigmoid regression model over 15 designs. We plot the mean and twice the standard deviation from 10 runs. Computational time was set to 80 seconds for comparison.