Variational Bayesian Optimal Experimental Design

Introduction

- Paper talk about concept of, “how to designing experiments ?”
- Goal is to maximize the information gathering from experiments
- Bayesian optimal experimental design (BOED) is a principled framework for making efficient use of limited experimental resources
- Applicability of BOED is hampered by the difficulty of obtaining accurate (EIG) estimates of the expected information gain of an experiment.
- Writers have introduced several classes of fast EIG estimators by building on ideas from amortized variational inference.

What is the need?

- To have a predictive model
- To know the expected information gain (EIG) from an experiment.
- To compute the EIG before the experiment is conducted.
- EIG is computed in context of a given design *d*.
- Evaluate and choose the design for its outcome and EIG in parameters of interest θ

How is it done ?

- Predictive model $p(y|\theta, d)$ is constructed
- Model is created for possible outcomes y
- Given a design d
- Given a value of parameters of interest θ

Formulation

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

- EIG(d) : Expected Information Gain when an experiment is performed, or data point *d* is collected

- Ep(y|d) : Expected value over the probability distribution p(y|d)

- H(p(θ)): Entropy of the prior distribution over the parameters θ.

- H(p(θ|y, d)): Entropy of the posterior distribution of the parameters θ given the outcome y and the design *d*.

Features

- BOED framework is particularly powerful when it allows the results of previous experiments to be used in guiding the design for future experiments.

  *we ask a participant a series of questions in a psychology trial, we can use the information gathered from previous responses to ask more pertinent questions in the future, that will, in turn, return more information

- The ability to design experiments that are self-adaptive can substantially increase their efficiency: fewer iterations are required to uncover the same level of information

- In practice BOED is often hampered by the difficulty of obtaining fast and high-quality estimates of EIG, due to intractability of the posterior p(θ|y,d), it constitutes a nested expectation problem, so conventional Monte Carlo (MC) estimation methods cannot be applied.

  *Nested MC(NMC) can only achieve, at best, a rate of O(T−1/3) in the total computational cost T [33], compared with O(T−1/2) for conventional MC.

Approach taken by Paper

- To address this challenge :
  - Writers have proposed a variational BOED approach that sidesteps the double intractability of the EIG in a principled manner and yields estimators with convergence rates in line with those for conventional estimation problems.
  - Introduced four estimators for EIG, with different advantages.
    I. Variational Posterior
    II. Variational Marginal
    III. Variational NMC
    IV. Implicit likelihood

Approach taken by Paper

- Theoretically, this paper is showing that they all have a convergence rate of O(T−1/2) when the variational family contains the target distribution.

- These estimators can provide significant empirical gains in EIG estimation over previous methods and that these gains lead, in turn, to improved end-to-end performance

Background

- It is a Model-based approach for choosing an experiment design *d* in a manner that optimizes the information gained about some parameters of interest θ from the outcome *y* of the experiment :

- Example of implementation :

    - Choose the question *d* in a psychology trial to maximize the information gained about an underlying psychological property of the participant θ from their answer *y* to the question. The framework with a prior p(θ) and a predictive model p(y|θ,d)

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

    - In order to define a metric to assess the utility of the design *d* take the expectation of IG(y,d) under the marginal distribution over outcomes

    - p(y|d) = Ep(θ)[p(y|θ,d)] as per (1).
    - * The difference of integrals : EIG is difference between two terms
        - The expectation of log p(θ) under the joint distribution, p(θ ,y | d)
        - The expectation of log p(θ | y, d) under the same joint distribution.

Background

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

- Using the properties of logarithms and probabilities, log p(θ|y,d) can be decomposed into log p(y |θ,d) – log p(y|d) as p(θ|y,d) is proportional to p(y|θ,d) p(θ) by Baye's theorm and p(y|d) is the normalizing constant (marginal likelihood).

- With the result of following equation EIG can also be interpreted as the mutual information between θ and *y* given *d*

- Computing the EIG is challenging since neither p(θ|y,d) or p(y|d) can in general be found in closed form

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

Nested Monte Carlo

- One common way of getting around this is to employ a nested MC (NMC) estimator

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

Variational Estimators

- Though consistent , the convergence rate of the NMC (Nested Monte Carlo) is slow for multiple practical problems.

- This paper show how the idea from amortized variational inference can be used to resolve the double intractability of EIG and getting faster convergence rate.

- Variational approaches introduced in this paper looks to directly learn a functional approximation.

  Example :
  compute an approximation of y → p(y|d) and evaluate this approximation at multiple points to estimate the integral (sharing information across different values of y).

  For M evaluations made in the learning approximation is O(n) “Big O”
  Where n is Input data size and Number of Operations have complexity linear with size of the input

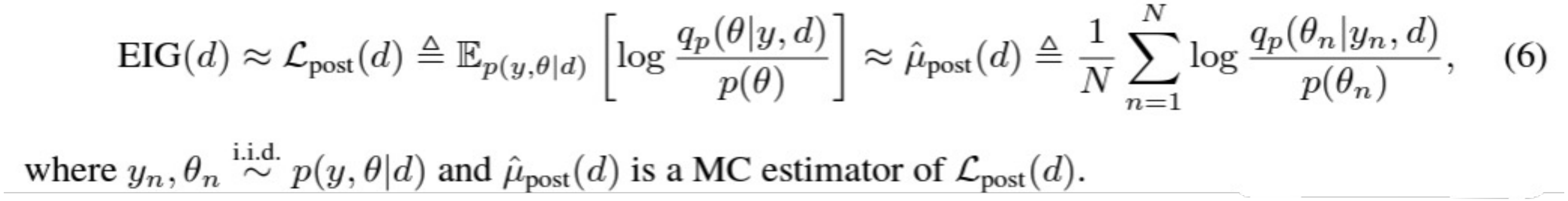
Variational Estimators

- **Variational posterior** : $\hat{\mu}_{\textbf{post}}$

  - It is based on learning an amortized approximation $q_p(\theta|y,d)$ to the posterior p(θ|y,d) and then using this to estimate the EIG :

    * The term "amortization" comes from finance and refers to spreading out a cost over time. In this context, it means spreading the computational cost of the optimization across all data points

    * The term “approximation” refers to the process of finding a simpler or more computationally tractable distribution that is close to true.



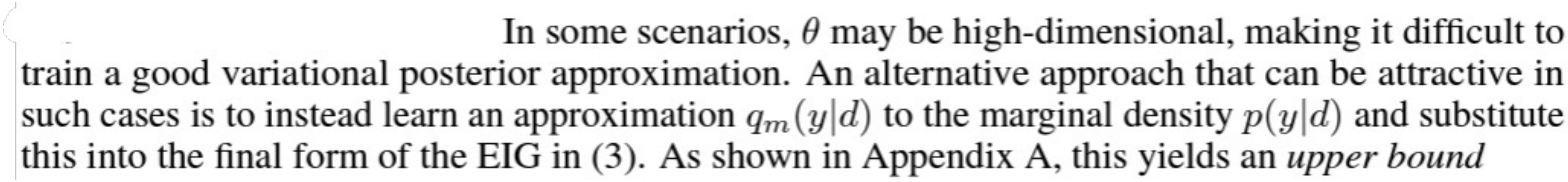
$$\mathrm{EIG}(d) \approx \mathcal{L}_{\text{post}}(d) \triangleq \mathbb{E}_{p(y,\theta|d)}\left[\log \frac{q_p(\theta|y,d)}{p(\theta)}\right] \approx \hat{\mu}_{\text{post}}(d) \triangleq \frac{1}{N}\sum_{n=1}^{N} \log \frac{q_p(\theta_n|y_n,d)}{p(\theta_n)}, \quad (6)$$

where $y_n, \theta_n \overset{\text{i.i.d.}}{\sim} p(y,\theta|d)$ and $\hat{\mu}_{\text{post}}(d)$ is a MC estimator of $\mathcal{L}_{\text{post}}(d)$.

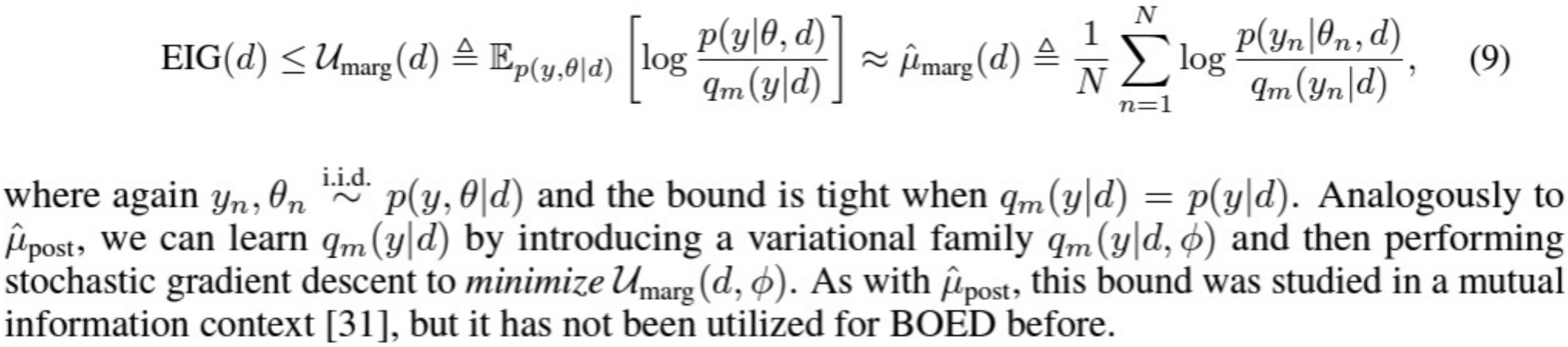
- We draw samples of p(y,θ|d) by sampling θ ~ p(θ) and then y|θ ~ p(y|θ,d). We can think of this approach as amortizing the cost of the inner expectation, instead of running inference separately for each y.

Variational Estimators

- **Variational marginal : $\hat{\mu}_{\mathbf{marg}}$**



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



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

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

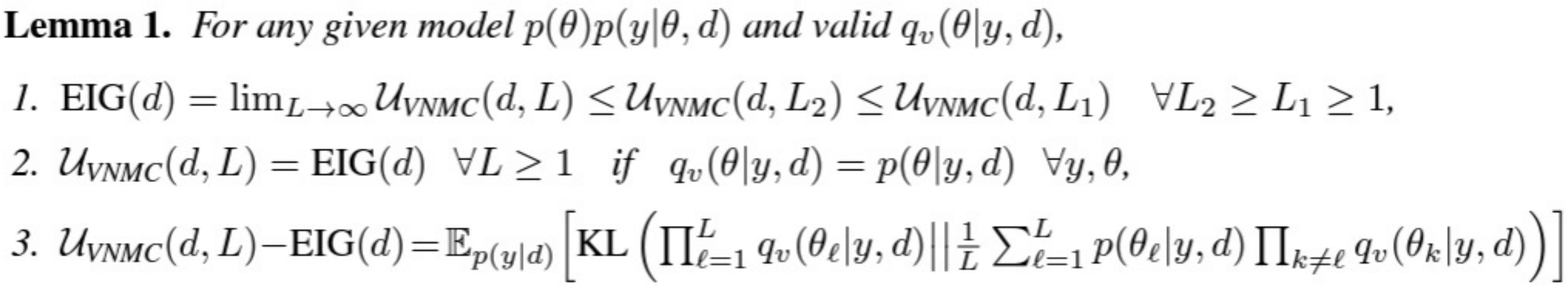
Variational Estimators

- **Variational NMS :** $\hat{\mu}_{\mathbf{VNMC}}$
  - Variational posterior and Variational Marginal can provide substantially faster convergence rate than NMC. However, to address the problem of biased estimation (if variational family does not contain the target distribution), EIG estimator NMC is proposed.
  - This estimator allows user to trade-off resources between fast learning of biased estimator permitted by variational approaches and ability of NMC to eliminate this bias.
  - Think of NMC estimator as approximating p(y|d) using M samples from prior.
  - VNMC is based around learning a proposal qv(θ|y,d) and then using samples from this proposal to make an importance sampling estimate of p(y|d), potentially requiring far fewer samples than NMC.

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

Variational Estimators

- **Variational NMS :** $\hat{\mu}_{\textbf{VNMC}}$
  - Important features of UVNMC(d,L) are summarized in the following lemma;



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

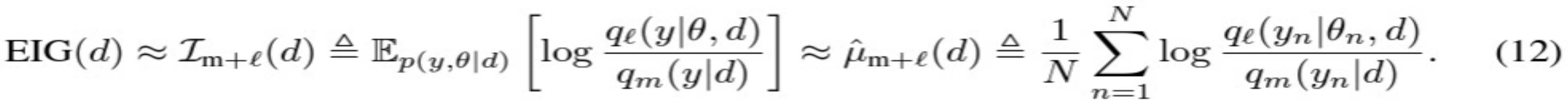
1. $\mathrm{EIG}(d) = \lim_{L\to\infty} \mathcal{U}_{VNMC}(d, L) \leq \mathcal{U}_{VNMC}(d, L_2) \leq \mathcal{U}_{VNMC}(d, L_1) \quad \forall L_2 \geq L_1 \geq 1,$

2. $\mathcal{U}_{VNMC}(d, L) = \mathrm{EIG}(d) \;\; \forall L \geq 1 \quad \textit{if} \quad q_v(\theta|y, d) = p(\theta|y, d) \;\; \forall y, \theta,$

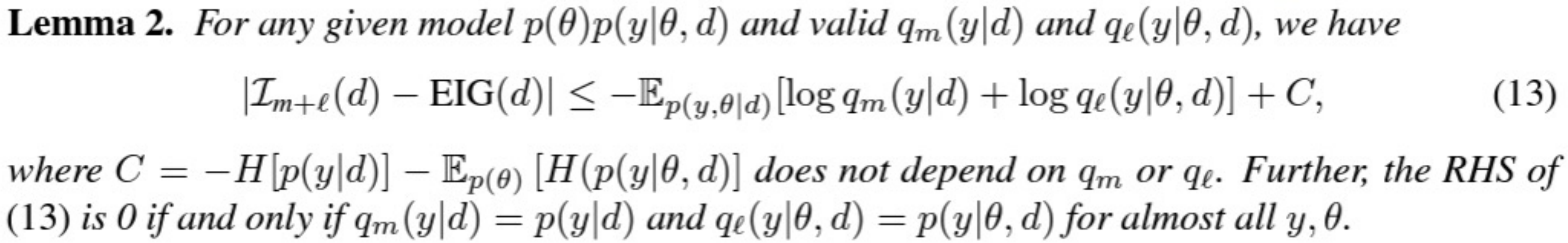
3. $\mathcal{U}_{VNMC}(d, L) - \mathrm{EIG}(d) = \mathbb{E}_{p(y|d)}\left[\mathrm{KL}\left(\prod_{\ell=1}^{L} q_v(\theta_\ell|y, d) \middle|\middle| \frac{1}{L}\sum_{\ell=1}^{L} p(\theta_\ell|y, d) \prod_{k\neq\ell} q_v(\theta_k|y, d)\right)\right]$

Variational Estimators

- **Implicit likelihood** : $\hat{\mu}_{\mathbf{m}+\ell}$
  - Many models of interest have implicit likelihoods from which we can draw samples, but not evaluate directly. For example, models with nuisance latent variables ψ (such as a random effect models) are implicit likelihood models because p(y|θ,d) = Ep(ψ|θ)[p(y|θ,ψ,d)] is intractable, but can still be straightforwardly sampled from.
  - Although variational marginal is not directly applicable in this setting, it can be modified to accommodate implicit likelihoods. Specifically, we can utilize two approximate densities: qm(y|d) for the marginal and q(y|θ,d) for the likelihood. We then form the approximation



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



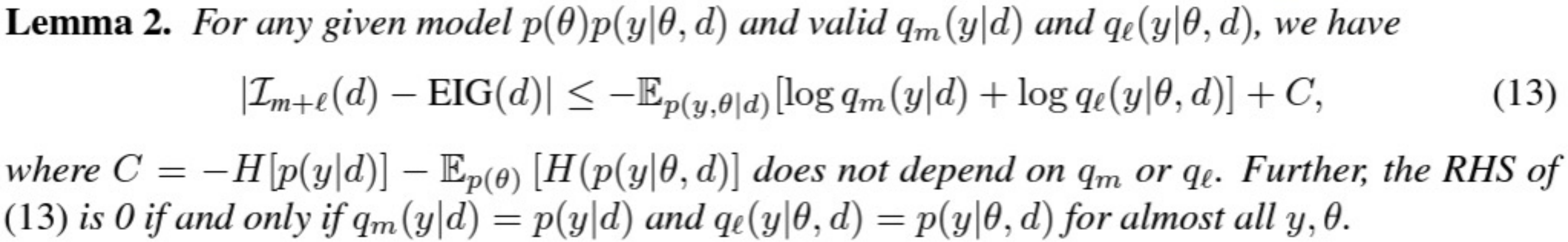
Lemma 2. *For any given model $p(\theta)p(y|\theta,d)$ and valid $q_m(y|d)$ and $q_\ell(y|\theta,d)$, we have*

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

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

Variational Estimators

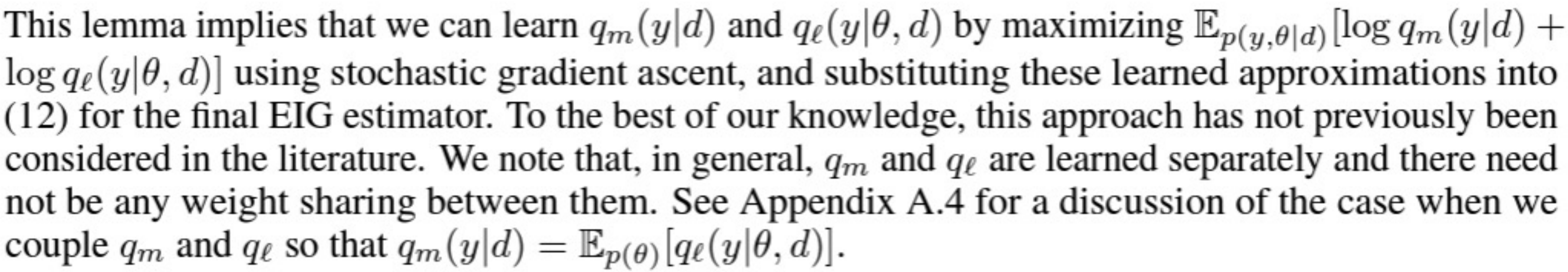
- **Implicit likelihood** : $\hat{\mu}_{\mathbf{m}+\ell}$
  - The following lemma shows that we can bound the EIG estimation error of $\mathcal{I}_{\mathrm{m}+\ell}$.



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

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

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



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

Related work

Alternative approaches to EIG estimation for BOED that will form the baseline for empirical comparisons.

- Nested Monte Carlo (NMC)

- Laplace approximation to the posterior :
  - This approach is fast but is limited to continuous variable and can exhibit bias.

- Likelihood-free inference by Ratio Estimation (LFIRE)

- Donsker-Varadhan (DV) :
  - representation of KL divergence as used by Belghazi for mutual information estimation. Included this as a baseline for illustrative purposes.

Experiments – EIG estimation accuracy

Four experiment design scenarios inspired by applications of Bayesian data analysis in science and industry :

1. **A/B testing** is used across marketing and design to study population traits.
   - The design is the choice of A and B group sizes, and the Bayesian model is a Gaussian linear model.

2. Revealed **preference** is used in economics to understand the consumer behavior.
   - We Consider an experiment design setting in which we aim to learn the underlying utility function of an economic agent by presenting them with proposal (such as offering them a price for commodity) and observing their revealed preference.

3. Fixed effects and random effects (nuisance variables) are combined in **mixed effect** models.
   - We Consider an example inspired by item-response theory [13] in psychology. We seek information only about the fixed effect, making this an implicit likelihood problem.

4. Labelled data from one region of design space must be used to predict labels in target region by **extrapolation.**

Summary : Two model with explicit likelihoods (A/B testing, preference) and two that are implicit (mixed effect, extrapolation)

Experiments – EIG estimation accuracy

Estimated the EIG across a grid of designs with fixed computational budget for each estimator and calculated the true EIG analytically or with brute force computation as appropriate.

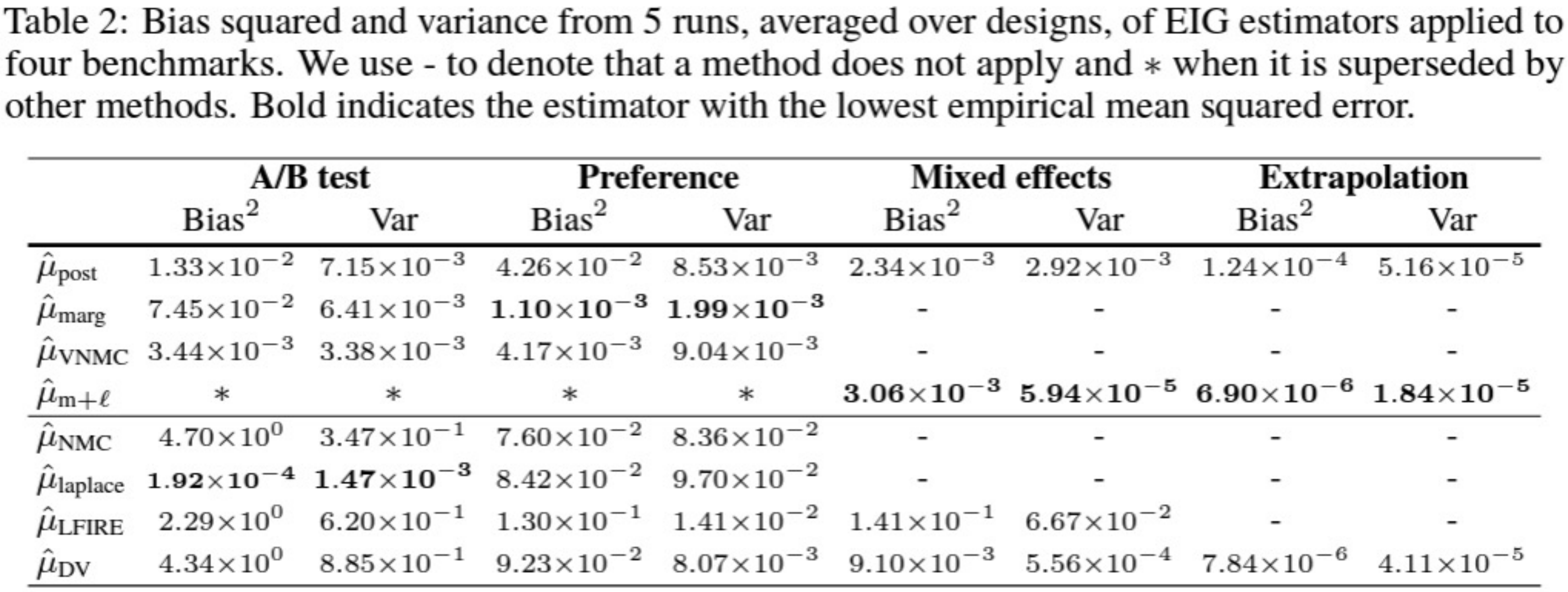
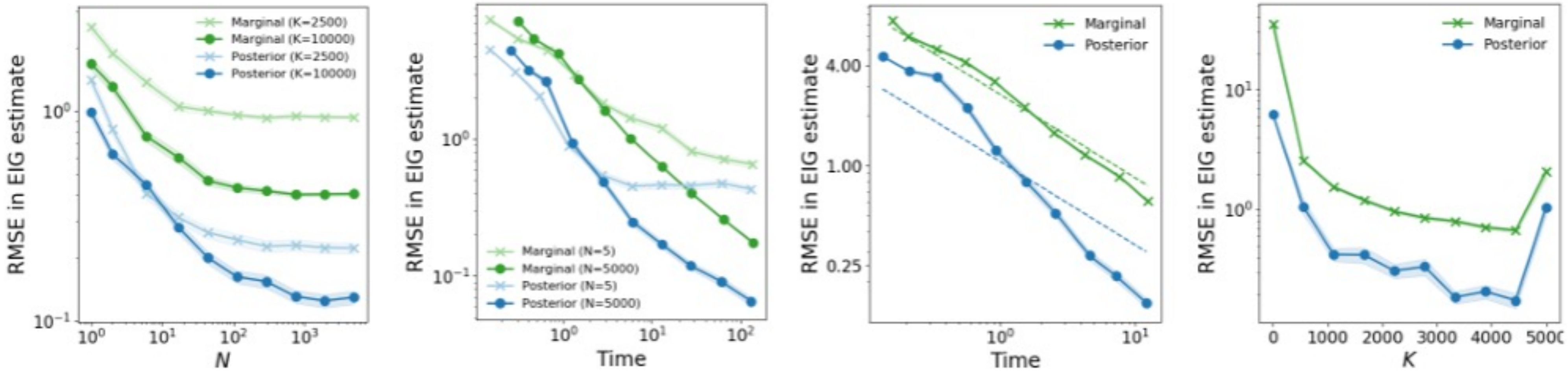


Table 2: Bias squared and variance from 5 runs, averaged over designs, of EIG estimators applied to four benchmarks. We use - to denote that a method does not apply and $*$ when it is superseded by other methods. Bold indicates the estimator with the lowest empirical mean squared error.

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

Experiments – EIG estimation accuracy

Laplace method, performed best for Gaussian linear model where its approximation becomes exact.
All methods outperformed NMC.

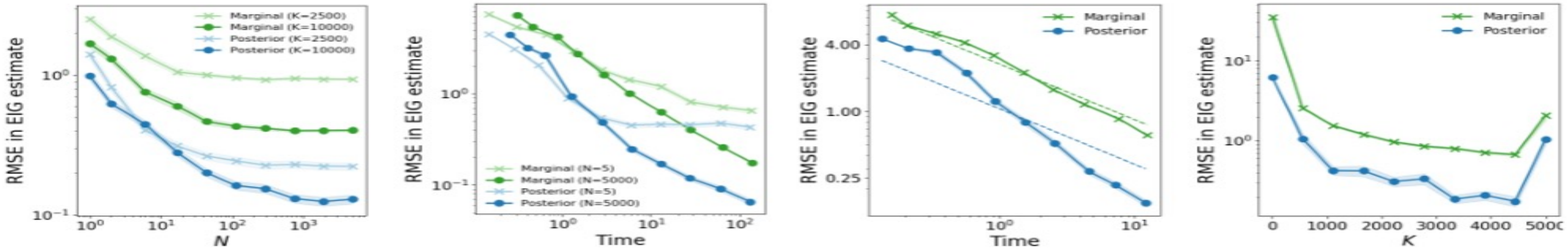


(a) Convergence in N (b) Convergence in K (c) Convergence $N = K$ (d) Fixed budget $N + K$

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

Experiments – Convergence rates

1. Consider the convergence in N after a fixed number of K updates to the variational parameters.
2. RMSE initially decreases as we increase N, before plateauing due to the bias in the estimator.
3. ˆ µpost substantially outperforms ˆ µmarg.
4. The errors decrease with time and that when a small value of N = 5 is taken, we again see a plateauing effect, with the variance of the final MC estimator now becoming the limiting factor
5. In Figure 1c we take N = K and increase both, obtaining the predicted convergence rate O(T−1/2) (shown by the dashed lines). We conjecture that the better performance of ˆ µpost is likely due to θ being lower dimensional (dim = 2) than y (dim = 10).
6. In Figure 1d, we instead fix T =N+Ktoinvestigate the optimal trade-off between optimization and MC error: it appears the range of K/T between 0.5 and 0.9 gives the lowest RMSE



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

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

Experiments – Convergence rates

- ˆ µVNMC can improve over NMC by using an improved variational proposal for estimating p(y|d).

- In Figure 2, plot the EIG estimates obtained by first running K steps of stochastic gradient with L = 1 to learn qv(θ|y,d), before increasing M and N.

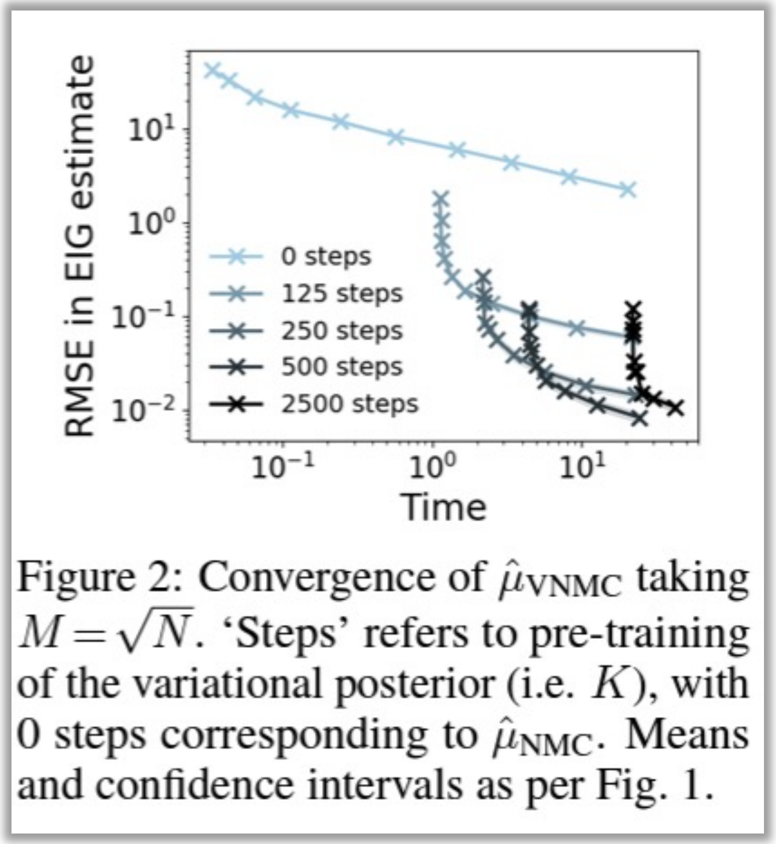


Figure 2: Convergence of $\hat{\mu}_{\text{VNMC}}$ taking $M = \sqrt{N}$. ‘Steps’ refers to pre-training of the variational posterior (i.e. K), with 0 steps corresponding to $\hat{\mu}_{\text{NMC}}$. Means and confidence intervals as per Fig. 1.

Experiments – Convergence rates

- Spending some of our time budget training qv(θ|y,d) leads to noticeable improvements in the estimation, but also that it is important to increase N and M. Rather than plateauing like ˆ µpost and ˆ µmarg, ˆ µVNMC continues to improve after the initial training period as, albeit at a slower O(T−1/3) rate. RMSE

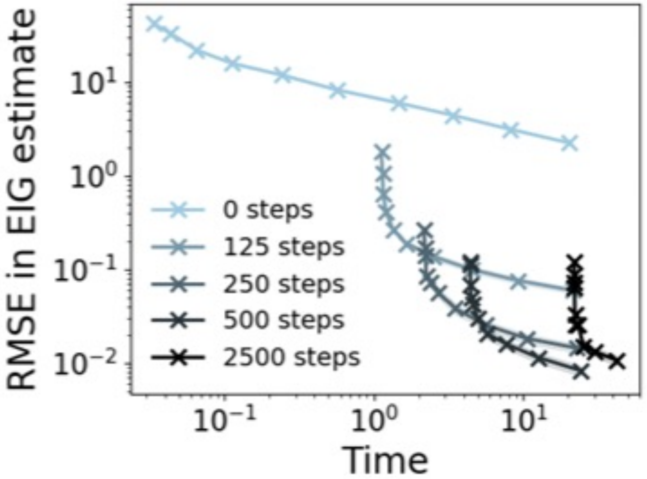
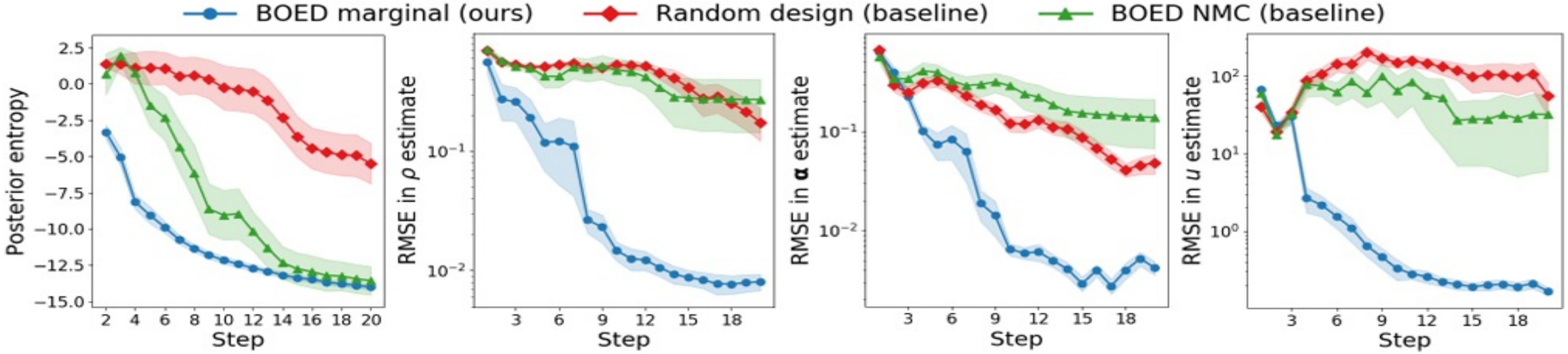


Figure 2: Convergence of $\hat{\mu}_{\text{VNMC}}$ taking $M=\sqrt{N}$. 'Steps' refers to pre-training of the variational posterior (i.e. K), with 0 steps corresponding to $\hat{\mu}_{\text{NMC}}$. Means and confidence intervals as per Fig. 1.

Experiments – End-to-end sequential experiments



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

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

Experiments – End-to-end sequential experiments

- Despite the relative simplicity of the design problem (with 36 possible designs) using BOED with ˆ µm+ leads to a more certain (i.e. lower entropy) posterior than random design.

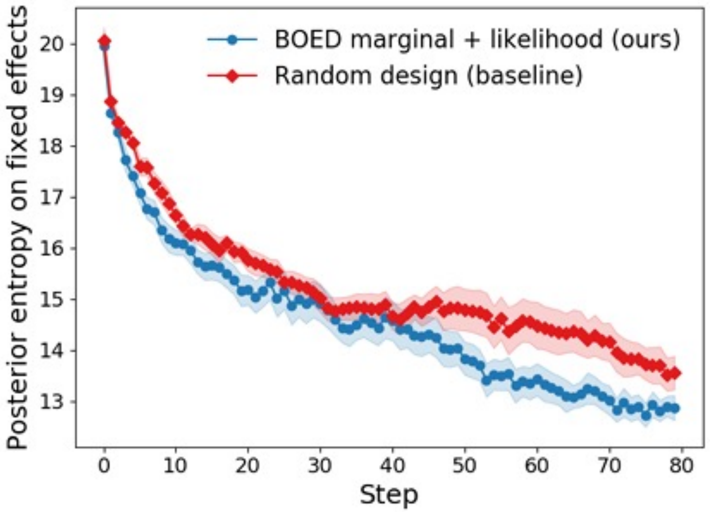


Figure 3: Evolution of the posterior entropy of the fixed effects in the Mechanical Turk experiment in Sec. 6.3. We depict the mean and ±1 std. err. from 10 experimental trials.

Selecting an estimator

- How to choose between estimators in practice.
  - First, Variational marginal and Implicit likelihood rely on approximating a distribution over y; ^ post and ^ VNMC approximate distributions over θ;
  - Second, ^marg and ^VNMC require an explicit likelihood whereas ^post and ^m+` do not.
  - If an explicit likelihood is available, it typically makes sense to use it—one would never use ^m+` over ^marg for example.

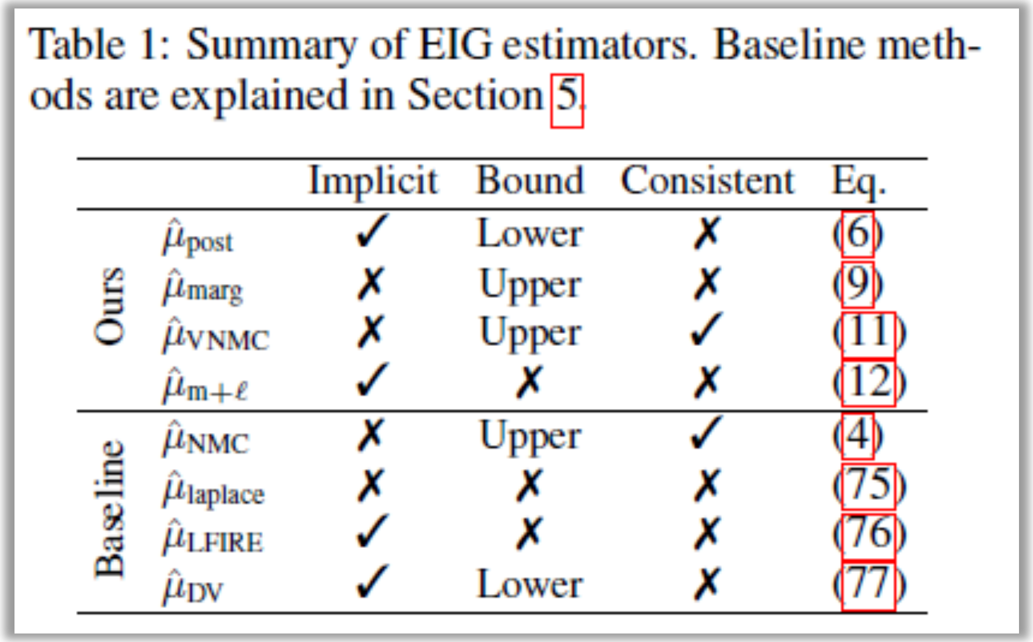


Table 1: Summary of EIG estimators. Baseline methods are explained in Section 5.

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

Selecting an estimator

- How to choose between estimators in practice.

  - Finally, if the variational families do not contain the target densities, ^VNMC is the only method guaranteed to converge true EIG(d) in the limit as the computational budget increases.

  - So, we might prefer ^VNMC when computation time and cost are not constrained.

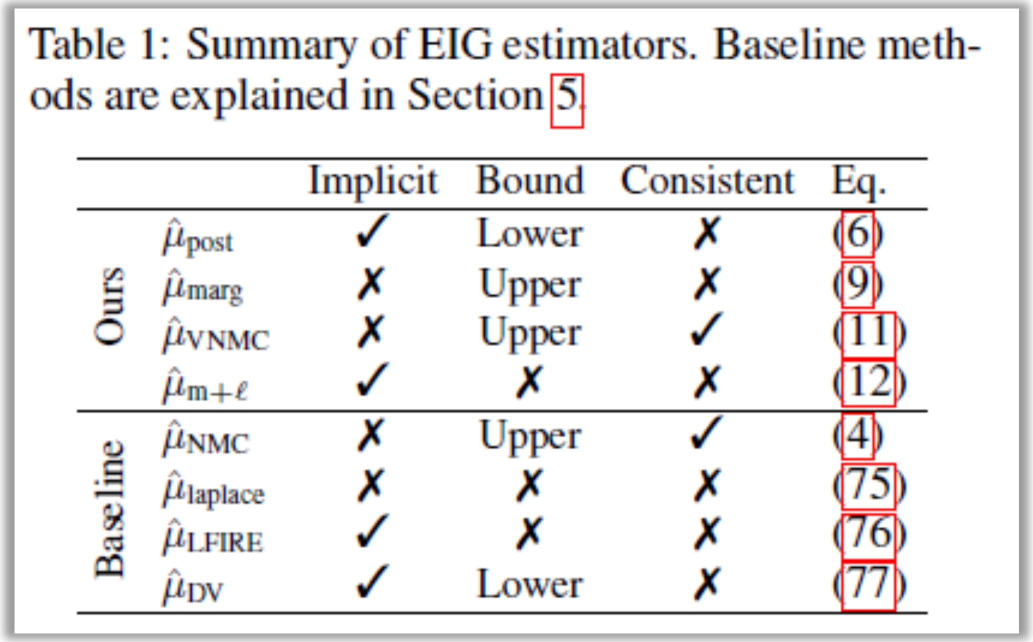


Table 1: Summary of EIG estimators. Baseline methods are explained in Section 5.

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

Conclusion

- We have developed efficient EIG estimators that are applicable to a wide range of experimental design problems.

- By tackling the double intractability of the EIG in a principled manner, they provide substantially improved convergence rates relative to previous approaches, and our experiments show that these theoretical advantages translate into significant practical gains

Conclusion

- Our estimators are well suited to modern deep probabilistic programming languages, and we have provided an implementation in Pyro.

- We note that the interplay between variational and MC methods in EIG estimation is not directly analogous to those in standard inference settings because the NMC EIG estimator is itself inherently biased

- Our ˆ µVNMC estimator allows one to play off the advantages of these approaches, namely the fast learning of variational approaches and asymptotic consistency of NMC.

END

Experiments – End-to-end sequential experiments

- We demonstrate the Utility of methods for designing sequential experiments.
- Variational estimators are sufficiently robust and fast to be used for adaptive experiments with a class of models that are of practical importance in many scientific disciplines.
- We run an adaptive psychology experiment with human participants recruited from Amazon Mechanical Turk to study how humans respond to features of stylized faces.
- To account for fixed effects—those common across the population—as well as individual variations that we treat as nuisance variables, we use the mixed effects regression model introduced in Sec. 6.1
- To estimate the EIG for different designs, we use $\hat{\mu}_{\mathrm{m}+\ell},$ since it yields the best performance on our mixed effects model benchmark (see Table 2).
- Our EIG estimator is integrated into a system that presents participants with a stimulus, receives their response, learns an updated model, and designs the next stimulus, all online

Experiments – End-to-end sequential experiments

- We consider a more challenging scenario in which a random design strategy gleans very little.
- We compare random design against two BOED strategies: ˆ µmarg and ˆ µNMC. Building on the revealed preference example in Sec. 6.1, we consider an experiment to infer an agent’s utility function which we model using the Constant Elasticity of Substitution (CES) model [2] with latent variables ρ,α,u.
- We seek designs for which the agent’s response will be informative about θ = (ρ,α,u)
- We estimate the EIG using ˆ µmarg because the dimension of y is smaller than that of θ, and select designs d ∈ [0,100]6 using Bayesian optimization
- To investigate parameter recovery, we simulate agent responses from the model with fixed values of ρ,α,u.