Variational Bayesian Optimal Experimental Design

Adam Foster† Martin Jankowiak‡ Eli Bingham‡ Paul Horsfall‡ Yee Whye Teh† Tom Rainforth† Noah D. Goodman‡§

†Department of Statistics, University of Oxford ‡Uber AI §Stanford University

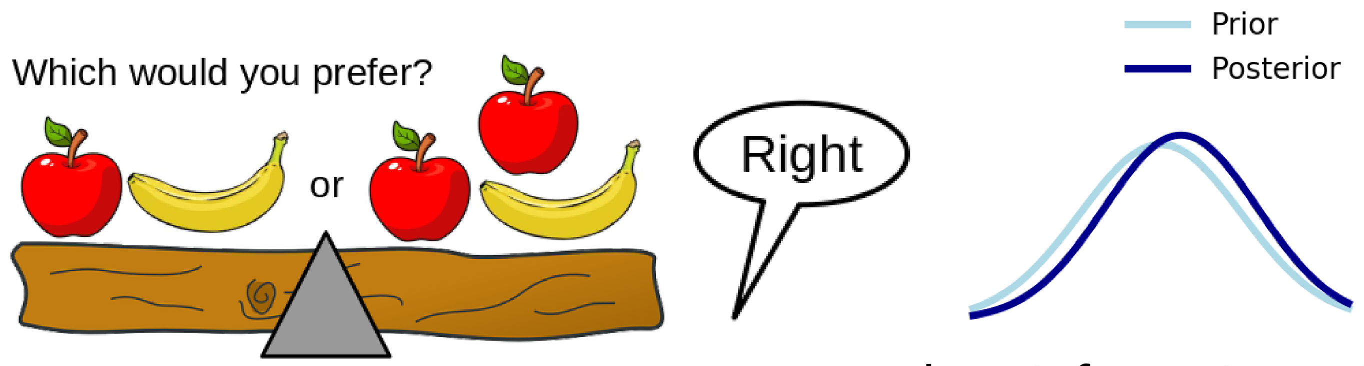
Overview

- In Bayesian experimental design, we want to find experiment designs that have maximal expected information gain (EIG)
- Computing EIG is challenging because it is the expectation of an intractable integrand
- We use an **amortized variational** approximations to the integrand to avoid recomputing similar integrals again and again
- We establish EIG bounds to train the variational approximations

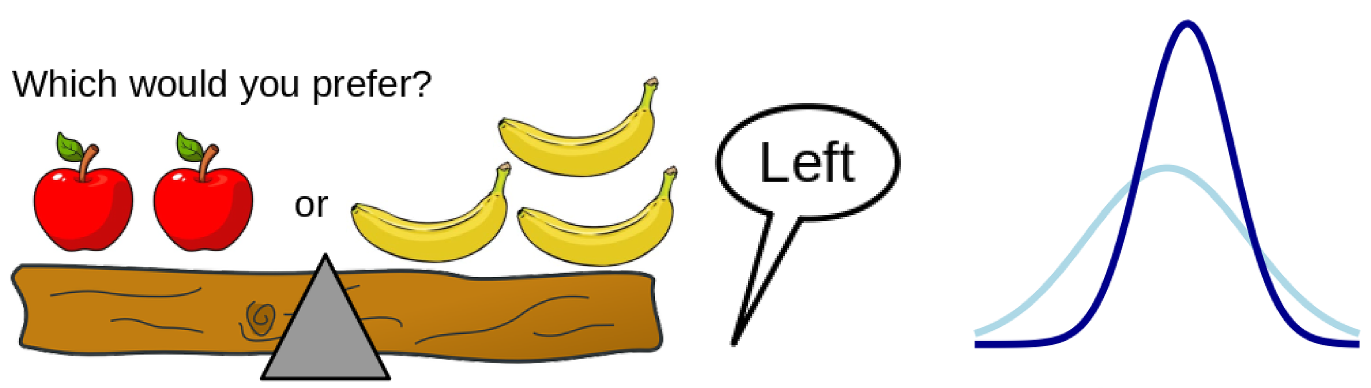
Bayesian optimal experimental design (BOED)

Motivation How do we design an experiment, e.g. pose a question to a participant, that will lead to high quality inference?

Design choose d	Experiment obtain y	Inference on θ

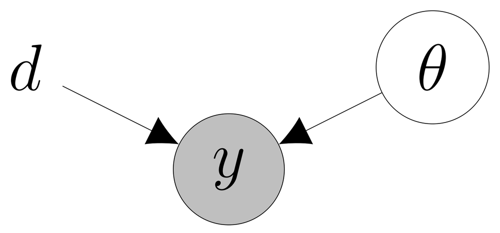


Low information gain



High information gain

Bayesian model for experimental design: **design** d, **outcome** y and **latent variable** of interest θ



with prior $p(\theta)$ and likelihood $p(y|\theta,d)$.
The **information gain** of the experiment is the reduction in entropy from the prior to the posterior

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

The **expected information gain (EIG)** [1] is

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

The optimal design d^* is the one which maximizes EIG over the space of feasible designs

Estimating EIG

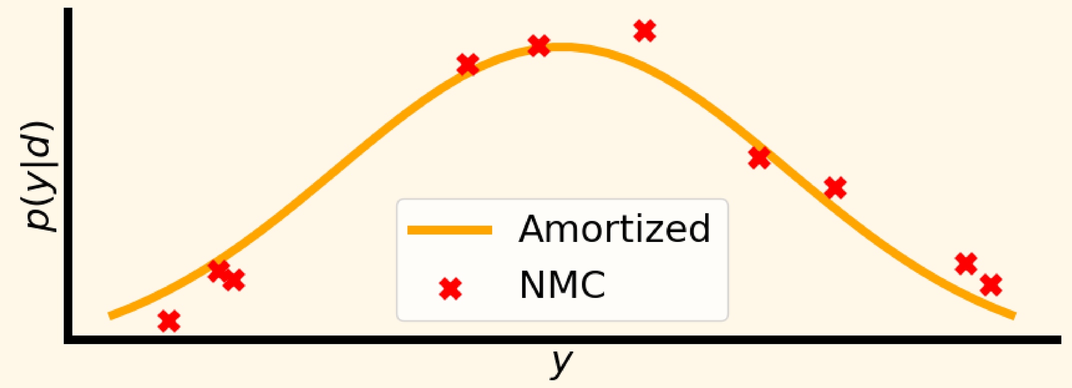
$$\text{EIG}(d) = \mathbb{E}_{\theta,y\sim p(\theta)p(y|\theta,d)}\left[\log\frac{\boxed{p(\theta|y,d)}}{p(\theta)}\right] = \mathbb{E}_{\theta,y\sim p(\theta)p(y|\theta,d)}\left[\log\frac{p(y|\theta,d)}{\boxed{p(y|d)}}\right].$$

We do not know $p(y|d)$ or $p(\theta|y,d)$ $\implies$ no standard Monte Carlo
Traditional solution is Nested Monte Carlo (NMC) [2]

$$\text{EIG}(d) \approx \frac{1}{N}\sum_{n=1}^{N}\log\frac{p(y_n|\theta_n,d)}{\frac{1}{M}\sum_{m=1}^{M}p(y_n|\theta_m,d)}$$

where $\theta_n, y_n \overset{\text{i.i.d.}}{\sim} p(\theta)p(y|\theta,d)$ and $\theta_m \overset{\text{i.i.d.}}{\sim} p(\theta)$.

Use amortization: learn the functional form of the integrand instead of repeatedly re-estimating $p(y|d)$



To learn this functional form, we use variational bounds on EIG

Marginal upper bound: $\text{EIG}(d) \le \mathbb{E}_{p(\theta)p(y|\theta,d)}\left[\log\frac{p(y|\theta,d)}{q_m(y|d,\phi)}\right]$, or

Posterior lower bound: $\text{EIG}(d) \ge \mathbb{E}_{p(\theta)p(y|\theta,d)}\left[\log\frac{q_p(\theta|y,d,\phi)}{p(\theta)}\right]$.

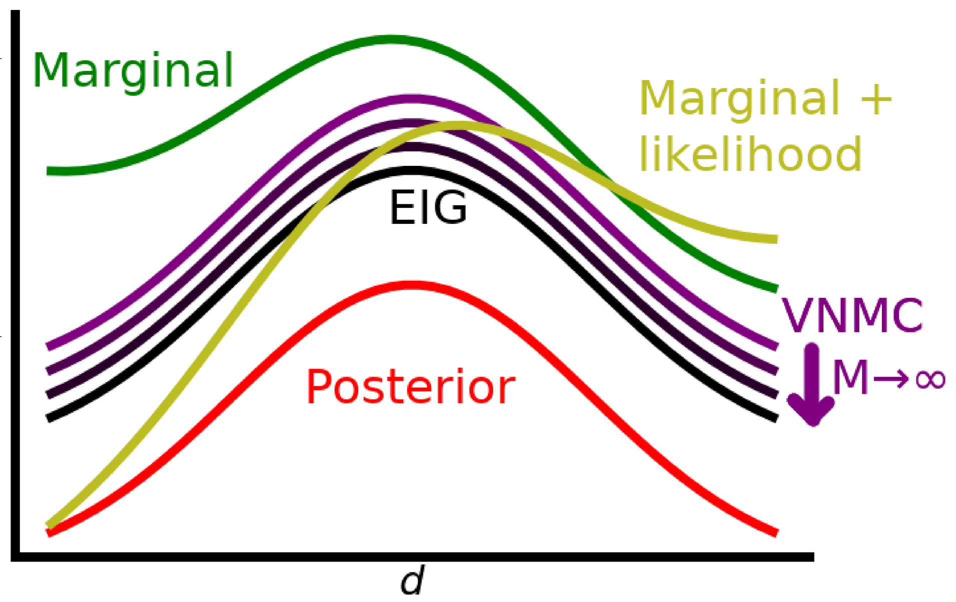
We train ϕ by stochastic gradient minimization or maximization.
Extend posterior to get a bound that becomes arbitrarily tight as $M\to\infty$

Variational NMC: $\text{EIG}(d) \le \mathbb{E}_{p(\theta_0)p(y|\theta_0,d)q_v(\theta_{1:M}|y,\phi)}\left[\log\frac{p(y|\theta_0,d)}{\frac{1}{M}\sum_{m=1}^{M}\frac{p(\theta_m)p(y|\theta_m,d)}{q_v(\theta_m|y,d,\phi)}}\right]$

Extend marginal for the **implicit likelihood** scenario

Marginal + likelihood: $\text{EIG}(d) \approx \mathbb{E}_{p(\theta)p(y|\theta,d)}\left[\log\frac{q_\ell(y|\theta,d,\psi)}{q_m(y|d,\phi)}\right]$

		Implicit	Bound	Consistent
Ours	Posterior	✓	Lower	✗
	Marginal	✗	Upper	✗
	VNMC	✗	Upper	✓
	M. + ℓ.	✓	✗	✗
Baseline	NMC	✗	Upper	✓
	Laplace	✗	✗	✗
	LFIRE	✓	✗	✗
	DV	✓	Lower	✗

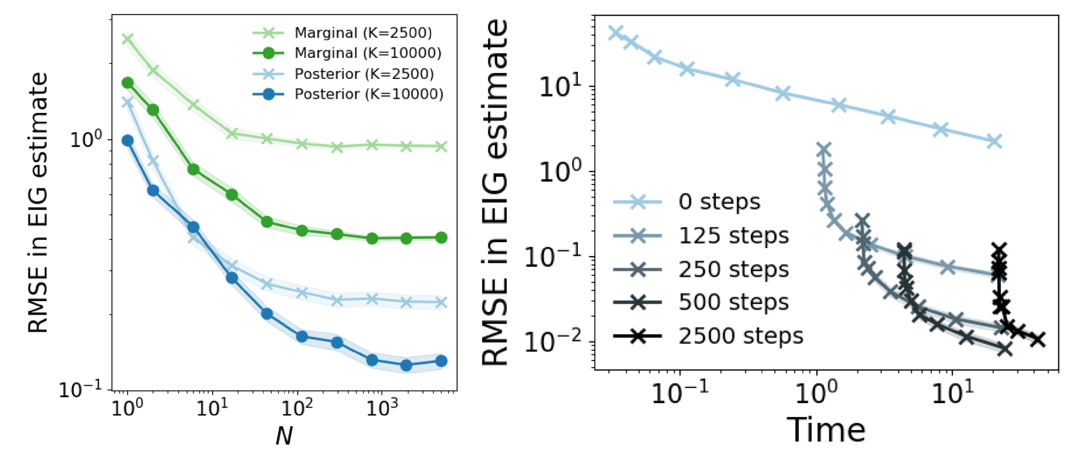


Convergence rates

For any one of our variational bounds $\mathcal{B}$, we train ϕ for K steps and then take an N-sample Monte Carlo estimate $\hat{\mu}$ of the bound. We decompose the EIG estimation error using the triangle inequality

$$\begin{aligned}\|\hat\mu(d,\phi_K)-\text{EIG}(d)\|_2 \le{}& \|\hat\mu(d,\phi_K)-\mathcal{B}(d,\phi_K)\|_2 && \text{Monte Carlo error} && \mathcal{O}(N^{-1/2})\\ &+\|\mathcal{B}(d,\phi_K)-\mathcal{B}(d,\phi^*)\|_2 && \text{Optimization error} && \mathcal{O}(K^{-1/2})\\ &+|\mathcal{B}(d,\phi^*)-\text{EIG}(d)| && \text{Asymptotic bias} && \mathcal{O}(M^{-1}) \quad \text{for VNMC [2]}\end{aligned}$$

For VNMC, the total cost of the estimator is $T = MN + K$. If we take $M=\sqrt{N}$ we obtain an overall convergence rate $\mathcal{O}(T^{-1/3})$, but unlike NMC we make use of fast variational training.



(Left) Posterior and marginal with K training steps, then N samples

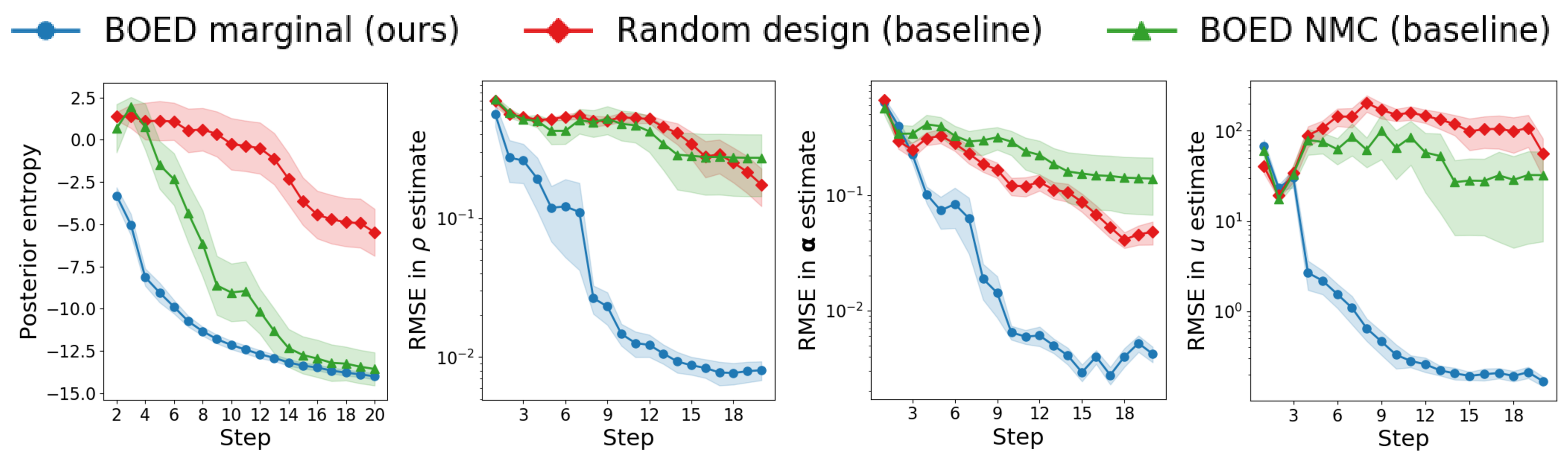
(Right) VNMC with K training steps, then N and $M=\sqrt{N}$ samples

Experiments

Bias-variance of EIG estimators

	A/B test		Econ. preference		Mixed effects	
	Bias^2	Var	Bias^2	Var	Bias^2	Var
Posterior	1.33×10^{-2}	7.15×10^{-3}	4.26×10^{-2}	8.53×10^{-3}	2.21×10^{-3}	2.70×10^{-3}
Marginal	7.45×10^{-2}	6.41×10^{-3}	$\mathbf{1.10\times10^{-3}}$	$\mathbf{1.99\times10^{-3}}$	-	-
VNMC	$\mathbf{3.44\times10^{-3}}$	$\mathbf{3.38\times10^{-3}}$	4.17×10^{-3}	9.04×10^{-3}	-	-
M. + ℓ.	-	-	-	-	$\mathbf{3.05\times10^{-3}}$	$\mathbf{7.72\times10^{-5}}$
NMC	4.70×10^{0}	3.47×10^{-1}	7.60×10^{-2}	8.36×10^{-2}	-	-
Laplace	$\mathbf{1.92\times10^{-4}}$	$\mathbf{1.47\times10^{-3}}$	8.42×10^{-2}	9.70×10^{-2}	-	-
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}
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}

End-to-end sequential experiment



(a) Entropy (b) RMSE of ρ (c) RMSE of α (d) RMSE of u

Acknowledgements

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

References

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

[2] 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.