Workshop at the 7th Symposium on Advances in Approximate Bayesian Inference (non-archival), 2025 1–14

Improving Robustness to Model Misspecification in Bayesian Experimental Design

Alexander J. Forster FORSTER@STATS.OX.AC.UK
Desi R. Ivanova DESI.IVANOVA@STATS.OX.AC.UK
Tom Rainforth RAINFORTH@STATS.OX.AC.UK
Department of Statistics, University of Oxford

Abstract

We propose a method to improve robustness to model misspecification in Bayesian experimental design (BED). Our approach introduces a flexible auxiliary model and jointly optimizes the expected information gain (EIG) in the original model parameters, the predictions of the auxiliary model, and a Bernoulli random variable indicating whether the original model is correct or misspecified. We show this balances learning about the original model, gathering data useful for general prediction, and assessing model fit. By leveraging the domain-specific knowledge embedded in the original model, we guide the design process while maintaining flexibility in the face of model misspecification. This is particularly important in adaptive design settings, where the original model informs early design decisions, but the auxiliary model enables adaptation when new data reveals model inaccuracies.

1. Introduction

Adaptive experimentation is central to a range of scientific disciplines as it enables efficient data acquisition through iterative refinement of experimental designs based on previously gathered information (MacKay, 1992; Myung et al., 2013). Bayesian experimental design (BED) offers a principled model-based framework for solving such adaptive design problems (Chaloner and Verdinelli, 1995; Ryan et al., 2016; Rainforth et al., 2024).

While BED is theoretically elegant, its practical effectiveness hinges on the correctness of the assumed outcome model, $p(y \mid \theta, \xi)$, where y represents the experiment outcome, θ the parameters of the model, and ξ the experiment design. In real-world scenarios, model misspecification is almost inevitable: the true data-generating process (DGP) rarely aligns perfectly with the assumed model. When the model is misspecified, BED can fail catastrophically, leading to poor design choices, biased inferences, and datasets that fail to reveal the model's flaws. For example, assuming a linear relationship when the true DGP is quadratic can result in designs that cluster in uninformative regions of the design space (Figure 1). Worse still, such datasets may not even reveal evidence of the model's incorrectness, perpetuating flawed assumptions.

Despite its importance, the problem of model misspecification in BED remains underexplored. Existing work tackling the issue includes Overstall and McGree (2022); Go and Isaac (2022); Feng et al. (2015), but there remains a critical gap in methods tailored to adaptive experimental design, where the interplay between model uncertainty and sequential decision-making is particularly challenging.

In this paper, we tackle model misspecification in BED by introducing a framework that explicitly accounts for the possibility of misspecification. Our method is built on the following principles:

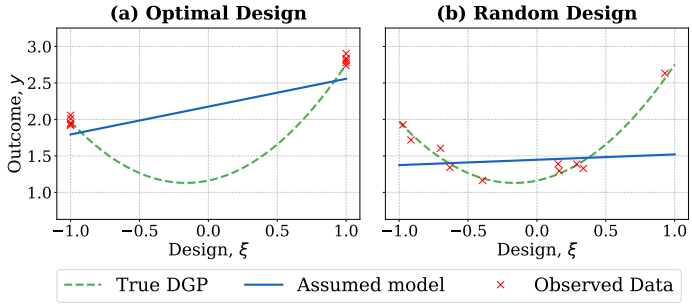


Figure 1: **BED with a misspecified model:** an illustrative example of a catastrophic failure. The assumed model is linear: $y \mid \theta, \xi \sim \mathcal{N}(\theta_0 + \theta_1 \xi, 0.1)$, whilst the true data generating process (DGP) is quadratic: $y \mid \theta, \xi \sim \mathcal{N}(\theta_0 + \theta_1 \xi + \theta_2 \xi^2, 0.1)$. The parameters are i.i.d. normal, $\theta_i \sim \mathcal{N}(0, 1)\ \forall i$, and the design $\xi \in [-1, 1]$. The green dashed curve shows the true DGP, whilst the solid blue line indicates the MAP fit under the assumed linear model based on T=10 design-outcome pairs (red crosses). **Panel (a)** shows the fit resulting from the optimal design strategy for the assumed linear model: lack of design diversity leads to a poor quality dataset and biased, inaccurate fit. **Panel (b)** shows the fit from a random design strategy, which, although still inaccurate, aligns more closely with the true DGP.

(i) Early guidance from the model: The framework leverages the model to guide data collection initially, capitalising on its structure.

(ii) Flexibility for robust data collection: Designs are encouraged to explore regions of the design space that are informative for both parameter learning and prediction, even under misspecification.

(iii) Model self-assessment: The method provides mechanisms to assess and adapt to potential model misspecification as data accumulates.

2. Background

2.1. Bayesian Experimental Design

Bayesian experimental design (BED) operates under the assumption that the data-generating process (DGP) can be described by a probabilistic model $p(y, \theta \mid \xi)$, where θ represents unknown parameters of interest, and ξ denotes design choices. BED selects designs by optimizing the Expected Information Gain (EIG):

$$\mathrm{EIG}_\theta(\xi) = \mathbb{E}_{p(y \mid \xi)} \left[\mathrm{H}[p(\theta)] - \mathrm{H}\left[p(\theta \mid \xi, y)\right] \right], \tag{1}$$

where $p(y \mid \xi) = \mathbb{E}_{p(\theta)}[p(y \mid \theta, \xi)]$ is the prior-predictive marginal distribution, and $H[\cdot]$ denotes Shannon entropy. In adaptive settings, this process is extended to sequentially update the posterior distribution $p(\theta \mid h_{t-1})$, where $h_{t-1} = \{(\xi_i, y_i)\}_{i=1}^{t-1}$ is the data collected up to time $t-1$. At each iteration, the next design ξ_t is chosen to maximize the conditional EIG

$$\mathrm{EIG}_{\theta \mid h_{t-1}}(\xi_t) = \mathbb{E}_{p(y_t \mid \xi_t, h_{t-1})} \left[\mathrm{H}[p(\theta \mid h_{t-1})] - \mathrm{H}\left[p(\theta \mid h_t)\right] \right]. \tag{2}$$

More recently, non-myopic, policy-based, methods have also been developed that instead learn a policy network π that maps from the history to the next design decision, such that $\xi_t = \pi(h_{t-1})$ (Foster et al., 2021; Ivanova et al., 2021; Blau et al., 2022; Huan et al., 2024).

2.2. Model Misspecification

In the standard Bayesian framework, uncertainty in the parameters θ is well-characterised, but uncertainty regarding the *model itself* is typically not addressed. This becomes problematic when the assumed model fails to capture the true underlying process—a scenario known as model misspecification. More formally, we say a model is *misspecified* if there exists no parameter setting θ^* such that the predictive distribution $p(y \mid \theta = \theta^*, \xi)$ matches the true data-generating distribution of $y \mid \xi$, subsequently denoted $p_{\text{true}}(y \mid \xi)$.

Whilst misspecification is typically inevitable (Box, 1982), Bayesian modelling can often still produce meaningful results under misspecification. In particular, if designs are chosen i.i.d. from some distribution $p(\xi^*)$, Kleijn and van der Vaart (2012) showed that the posterior will concentrate around $\tilde{\theta}$ in the limit of large data, where

$$\tilde{\theta} = \arg\min_{\theta \in \Theta} \mathbb{E}_{p(\xi^*)} \mathrm{KL}[p_{\text{true}}(y^* \mid \xi^*) \parallel p_{\text{model}}(y^* \mid \theta, \xi^*)], \tag{3}$$

and p_{true} is the true underlying data generating process. If the model is well–specified then $\tilde{\theta} = \theta^*$ regardless of $p(\xi^*)$, including when designs are chosen adaptively using BED (Paninski, 2005). However, when it is misspecified, $\tilde{\theta}$ explicitly depends on $p(\xi^*)$ and our conclusions become entwined with our design policy, whether this is adaptive or not (c.f. Figure 1).

Consequently, model misspecification can present an even deeper challenge to experimental design than it does for modelling already collected data. For BED we are now using the model not only to fit existing data, but also to guide the collection of new data (Rainforth et al., 2024). In particular, misspecification can cause catastrophic failures where deficiencies in the model lead to poor design decisions, which in turn means we collect data that hides the deficiencies in the model (see Appendix A for further discussion). Figure 1 presents a clear example of this with the designs chosen by BED all at ± 1. This in turn means that we cannot actually reject the hypothesis of a linear fit from the data collected, even though this is easily falsified in the case where data was sampled uniformly.

3. Robust Bayesian Experimental Design

BED assumes the data is generated by a parametric model with parameters θ, which from here onwards is referred to as p_{model}. Explicitly, the assumed data generating process (DGP) is: $\theta \sim p_\theta(\cdot), \forall t = 1, \ldots, T, \xi_t = \pi_t(h_{t-1}), y_t \sim p_{\text{model}}(\cdot \mid \theta, \xi_t)$, where each π_t is some (potentially random) policy, which we abstractly define as the mechanism for choosing experimental designs. For example, if designs are being chosen by sequentially maximising Equation 2, then $\pi_t(h_{t-1})$ is the maximising value.

As discussed above, the assumption the data are generated in this way can cause BED to fail catastrophically if p_{model} is misspecified. That is, if the true underlying DGP (denoted p_{true}) is $\xi_t = \pi_t(h_{t-1}), y_t \sim p_{\text{true}}(\cdot \mid \xi_t)$, then the model is misspecified when there exists no θ^* such that $p_{\text{true}}(y \mid \xi) = p_{\text{model}}(y \mid \theta^*, \xi)$ for all $y \in \mathcal{Y}$ and $\xi \in \mathcal{X}$.

3.1. Modelling the misspecification

Our method addresses this issue by explicitly modelling the mechanism of misspecification. We represent model misspecification as a mixture between the original parametric distribution (p_{model}) and an auxiliary distribution, p_{aux}. Specifically, p_{model} is treated as "correct" with probability $1 - \epsilon$ and as misspecified with probability ϵ.

The auxiliary model p_{aux} is chosen to capture plausible forms of model misspecification, and thus should be more flexible than p_{model}. Critically though, we generally do not want to just replace p_{model} with p_{aux}, because using an overly flexible model in BED undermines our ability to utilise our prior information to make targeted design decisions which we believe will yield particularly informative data. Particularly at early experiment iterations, we thus still wish to utilize p_{model} to efficiently investigate our initial beliefs and quickly contract our model's posterior. But we also want to make sure that as more data is acquired we can unearth deficiencies in p_{model} and, in turn, avoid the pathologies that can occur when performing BED with misspecified models.

Let ψ denote the (potentially infinite-dimensional) parameters of p_{aux}, with prior distribution $p_\psi(\cdot)$. Our method now posits using the following extended model

$$\begin{gathered} \theta \sim p_\theta(\cdot), \quad \psi \sim p_\psi(\cdot), \quad Z \sim \text{Bernoulli}(1-\epsilon), \\ \forall t = 1, \dots, T \quad \xi_t = \pi_t(h_{t-1}) \quad y_t \sim \begin{cases} p_{\text{model}}(\cdot \mid \theta, \xi_t) & \text{if } Z = 1 \\ p_{\text{aux}}(\cdot \mid \psi, \xi_t) & \text{if } Z = 0\,. \end{cases} \end{gathered} \tag{4}$$

Note that this formulation assumes that with probability $1-\epsilon$, *all* the data is distributed as p_{model}, and with probability ϵ, *all* the data is distributed as p_{aux}.

This extended DGP defines a new model, $p_{\text{ext}}(y \mid \Phi, \xi)$, with parameters $\Phi \coloneqq \{\theta, \psi, Z\}$. Experimental designs can then be chosen by maximising the EIG under the extended model. By using an auxiliary model that is more flexible than the original model, this extension serves to regularise the EIG. For example, we demonstrate later that using a Gaussian process auxiliary model results in a regularisation that penalises experimental designs that are too similar to previous designs and rewards greater exploration of the design space, mitigating the adverse effects of misspecification.

However, targeting information in Φ directly could be problematic, especially if the extended model is non-parametric, as the potential information gain in ψ can dominate that in θ or Z. Moreover, we do not really directly care about information gain in ψ itself, we simply want to collect data that ensures that the auxiliary model makes effective predictions. In other words, we want to ensure robustness by making sure that an effective predictive model can be trained from the gathered data.

To account for this, we leverage ideas about the so called expected predictive information gain introduced by Bickford Smith et al. (2023). Namely, we introduce some test–time "input distribution" $p(\xi^*)$ that represents downstream distribution we want to ensure we can make predictions for. We then target the expected information gain in $\Omega \coloneqq \{\theta, (\xi^*, y^*), Z\}$ where y^* is the predicted outcome associated with using the design ξ^*. Here there is no marginal information to be learned about the distribution of ξ^*, but what we are trying to learn about (alongside θ and Z) is how to predict $y^*|\xi^*$ across a distribution of possible ξ^*. If we are able to learn this effectively, it indicates that we have gathered data that allows us to predict outcomes with respect to other possible design setups that one could have used, not just the precise designs that were chosen. In turn, this provides robustness to the overall process, by reducing sensitivity to the precise design decisions that were made.

We refer to EIG_Ω as the ***robust* EIG**. The following decomposition now shows how we can target the robust EIG, and provides insights into how it behaves. A proof for the result is given in Appendix D.

Theorem 1 (EIG decomposition) *The expected information gain in Ω, conditional on history $h_{t-1} = \{(y_i, \xi_i)\}_{i=1}^{t-1}$, is given by*

$$\begin{aligned} \mathrm{EIG}_\Omega(\xi_t \mid h_{t-1}) = {} & \mathbb{P}(Z=1 \mid h_{t-1})\, \mathrm{EIG}^{\mathrm{model}}(\xi_t \mid h_{t-1}) \\ & + \mathbb{P}(Z=0 \mid h_{t-1})\, \mathrm{EPIG}^{\mathrm{aux}}(\xi_t \mid h_{t-1}) + \mathrm{EIG}_Z(\xi_t \mid h_{t-1}), \end{aligned} \tag{5}$$

where

$$\mathrm{EIG}^{\mathrm{model}}(\xi_t \mid h_{t-1}) \coloneqq \mathbb{E}_{p(\theta \mid h_{t-1}) p_{\mathrm{model}}(y_t \mid \theta, \xi_t)} \left[\log \frac{p_{\mathrm{model}}(y_t \mid \theta, \xi_t)}{p_{\mathrm{model}}(y_t \mid h_{t-1}, \xi_t)} \right], \tag{6}$$

$$\mathrm{EPIG}^{\mathrm{aux}}(\xi_t \mid h_{t-1}) \coloneqq \mathbb{E}_{p(\xi^*) p_{\mathrm{aux}}(y_t, y^* \mid h_{t-1}, \xi_t, \xi^*)} \left[\log \frac{p_{\mathrm{aux}}(y^* \mid h_{t-1}, \xi_t, y_t, \xi^*)}{p_{\mathrm{aux}}(y^* \mid h_{t-1}, \xi^*)} \right], \tag{7}$$

$$\mathrm{EIG}_Z(\xi_t \mid h_{t-1}) \coloneqq \mathrm{I}(y_t; Z \mid h_t, \xi_t). \tag{8}$$

Theorem 1 shows that the robust EIG adapts to the data as it is collected: if the data supports the original model being well specified (i.e. if $\mathbb{P}(Z=1 \mid h_{t-1})$ is close to 1), then it targets $\mathrm{EIG}^{\mathrm{model}}$ – the original reward function. On the other hand, if the data gives evidence of misspecification (i.e. if $\mathbb{P}(Z=0 \mid h_{t-1})$ is close to 1), then it abandons gaining information in θ and instead targets $\mathrm{EPIG}^{\mathrm{aux}}$. It also rewards gaining information in Z, i.e. designs that help determine whether $Z=1$ (that the original model p_{model} is well specified) or $Z=0$ (that the original model p_{model} is misspecified). Further discussion of the approach is given in Appendix A, additional theoretical arguments are provided in Appendix C.

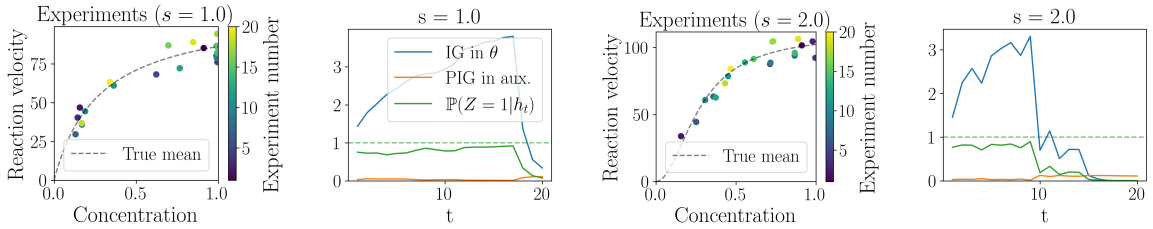
4. Experiments

We now test the efficacy of our approach using the Michaelis-Menten kinetics model from the field of chemical kinetics. It has been studied in the context of BED in e.g. Overstall and McGree (2022) and Dette and Biedermann (2003). In a chemical experiment, it models the reaction velocity y in terms of substrate concentration $\xi \in [0, 400]$ and parameters of interest $\theta = (V_{\max}, V_{0.5})$. The model posits the mean response $\eta(\theta, \xi) \coloneqq V_{\max}\xi / (\xi + V_{0.5})$ and assumes Gaussian noise $y \mid \theta, \xi \sim N\left(\eta(\theta, \xi), 5^2\right)$. The prior distribution is taken to be $V_{\max}, V_{0.5} \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(20, 200)$.

While the above model is what we will use as p_{model}, we also need a distinct true data-generating process to test against. For this, we use a generalisation of the Michaelis-Menten model inspired by the Hill equation (Cornish-Bowden, 2012); it is as above, but with mean response $\eta_s(\theta, \xi) \coloneqq V_{\max}\xi^s / (\xi^s + V_{0.5}^s)$. By choosing different values of s, we can then control the degree of misspecification in our model. Our method further requires choosing an auxiliary model; for this we choose $f \sim \mathrm{GP}(\mu, k)$, $y \mid f, \xi \sim N\left(f(\xi), 5^2\right)$, where the mean function μ and kernel function k are given in Appendix E.

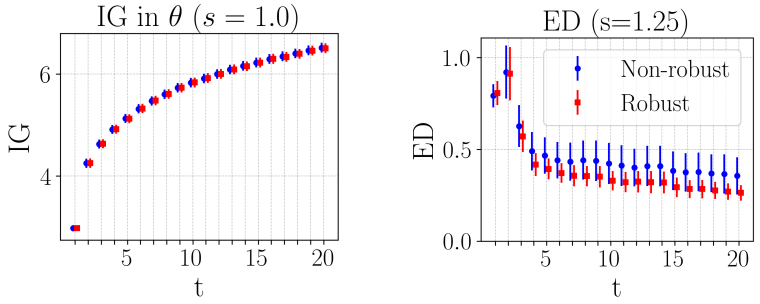
To test our method, we compare two data-gathering regimes: (i) a baseline consisting of sequentially maximising the conditional EIG in the *model* parameters, θ, (Equation 2), and (ii) our method, consisting of sequentially maximising the robust EIG, EIG_Ω.

We first demonstrate that our method can detect model misspecification, and having done so, selects designs that help prediction throughout the design space. In Figure 2 we show the designs that are chosen in a case where the model is correctly specified, $s=1$, and a case where it is misspecified, $s=2$. We see that our method is able to produce sensible designs in both cases, with the different components from the decomposition in Theorem 1 being differently weighted as the experiments progress.



(a) (b) (c) (d)

Figure 2: (a) A single rollout of 20 designs chosen sequentially using our method. Observe there are two primary design clusters near $\xi = 0.1$ and $\xi = 1.0$. (b) For $t = 1, \ldots, 20$, the blue line shows the *information gain* in θ, the orange line shows the *predictive information gain* in the auxiliary model, and the green line shows the posterior quantity $\mathbb{P}(Z = 1 \,|\, h_t)$. (The dashed green line at 1 is for reference.) (c) The same as (a) but the underlying DGP has $s = 2$. Observe the clusters no longer appear. (d) As per (b) with $s = 2$. Observe that at the 10th design, $\mathbb{P}(Z = 1 \,|\, h_t)$ falls dramatically, and continues towards 0 – the model misspecification has been detected, and the design-choosing behaviour changes with it.



(a) (b)

Figure 3: (a) The underlying DGP has s=1 (i.e. well-specified). The blue points show the IG when designs are chosen to maximise the conditional EIG, averaged over 100 runs (error bars represent 1 s.e.). The red points show the same for designs from the robust EIG. As expected, in this well-specified context our method incurs a small penalty for protecting against misspecification and performs marginally worse. (b) The underlying DGP has $s = 1.25$ (i.e. the model is misspecified). The blue points show the mean ED over 100 runs for designs chosen according to the baseline method, with error bars showing 1 s.e. The red points shows the same for our method. Our method produces designs resulting in a lower ED.

We then compare our method to standard BED, showing that it produces data from which we can better capture the true underlying data distribution. Specifically, we compare the two approaches in terms of their achieved information gain and the expected KL divergence (ED) from the true data distribution to that of our model

$$\mathrm{ED}(h_t) = \mathbb{E}_{p(x^*)} \left[\mathrm{KL}[p_{\text{true}}(y^* \,|\, x^*) \parallel p_{\text{model}}(y^* \,|\, x^*, h_t)] \right]. \tag{9}$$

As we see in Figure 3, using the robust EIG leads to less information gain in the true model when things are well-specified, but we get a better ED from the true data-generating distribution when misspecified, thereby providing the desired improvement in robustness.

Acknowledgements

Alex Forster acknowledges support from the EPSRC CDT in Modern Statistics and Statistical Machine Learning (EP/S023151/1). Tom Rainforth is supported by the UK EPSRC grant EP/Y037200/1.

References

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning. In *International Conference on Artificial Intelligence and Statistics*, pages 7331–7348. PMLR, 2023.

Tom Blau, Edwin V Bonilla, Iadine Chades, and Amir Dezfouli. Optimizing sequential experimental design with deep reinforcement learning. In *International conference on machine learning*, pages 2107–2128. PMLR, 2022.

George EP Box. Choice of response surface design and alphabetic optimality. Technical report, University of Wisconsin—Madison, 1982.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: a review. *Statistical Science*, 1995.

Athel. Cornish-Bowden. *Fundamentals of enzyme kinetics*, chapter 12: Regulation of Enzyme Activity. Wiley-Blackwell, 4th, completely rev. and greatly enl. ed. edition, 2012.

Holger Dette and Stefanie Biedermann. Robust and efficient designs for the michaelis–menten model. *Journal of the American Statistical Association*, 98(463):679–686, 2003.

Chi Feng et al. Optimal bayesian experimental design in the presence of model error. Master's thesis, Massachusetts Institute of Technology, 2015.

Adam Foster, Desi R Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential bayesian experimental design. *Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR 139*, 2021.

Jinwoo Go and Tobin Isaac. Robust expected information gain for optimal bayesian experimental design using ambiguity sets. In *Uncertainty in Artificial Intelligence*, pages 728–737. PMLR, 2022.

Xun Huan, Jayanth Jagalur, and Youssef Marzouk. Optimal experimental design: Formulations and computations. *Acta Numerica*, 33:715–840, 2024.

Desi R Ivanova, Adam Foster, Steven Kleinegesse, Michael Gutmann, and Tom Rainforth. Implicit Deep Adaptive Design: Policy–Based Experimental Design without Likelihoods. In *Advances in Neural Information Processing Systems*, volume 34, pages 25785–25798. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/d811406316b669ad3d370d78b51b1d2e-Paper.pdf.

Bas JK Kleijn and Aad W van der Vaart. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 2012.

David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

Jay I Myung, Daniel R Cavagnaro, and Mark A Pitt. A tutorial on adaptive design optimization. *Journal of mathematical psychology*, 57(3-4):53–67, 2013.

Antony Overstall and James McGree. Bayesian decision-theoretic design of experiments under an alternative model. *Bayesian Analysis*, 17(4):1021–1041, 2022.

Liam Paninski. Asymptotic theory of information-theoretic experimental design. *Neural Computation*, 17(7):1480–1507, 2005.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Elizabeth G Ryan, Christopher C Drovandi, James M McGree, and Anthony N Pettitt. A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, 2016.

Appendix A. Discussion

The effectiveness of BED relies critically on the assumed probabilistic model $p(y \mid \theta, \xi)$ accurately representing the true data-generating process (DGP). Some particularly notable risks for BED in this setting include:

1. Ineffective Designs: Optimizing EIG under a misspecified model may cause BED to select designs that explore irrelevant regions of the design space or fail to capture critical aspects of the true DGP.

2. Unrepresentive Data: If misspecification causes BED to fail to properly explore the design space or overemphasize certain regions, as in the case in Figure 1, it may lead to highly unrepresentative data being gathered.

3. Reinforced Errors: In adaptive settings, model misspecification can lead to a feedback loop, where designs systematically reinforce incorrect assumptions, leaving practitioners unaware of the model's deficiencies. In particular, we can collect data which is informative if the model is correct, but from which it is difficult to perform effective model checking.

Model misspecification is an unavoidable reality in practice, as the true data-generating process is rarely fully captured by the models we employ. For example, we may require our model to make simplifying assumptions for computational feasibility, interpretability, difficulty in encapsulating our true beliefs, or simply to try and ensure it has sufficient predictive power. For example, in clinical trials we often care primarily about simple statistics about drug efficacy, even though we do not think the true underlying mechanism is itself so simple.

These considerations give rise to a critical question in BED: What is the role of a model when it is likely to be misspecified? Specifically, when collecting data using a BED approach, why not instead employ a more flexible model—one with high, or even infinite, dimensionality—so that the model can better accommodate the data and minimize misspecification?

Our findings, as illustrated in section 4, reveal that adopting highly flexible models, such as those incorporating Gaussian process priors (i.e., models with infinite-dimensional parameters), tends to produce designs that uniformly fill the design space. While this approach avoids severe misspecification, it leads to designs that are generic rather than tailored to the specific experiment. Consequently, experimental designs may focus on regions of the design space that are known, a priori, to be uninformative.

Conversely, parametric models, even when misspecified, provide meaningful guidance early in the adaptive BED process. These models are capable of generating experimental designs that yield informative data about the unknown parameters at the outset. However, as demonstrated in Figure 1, this approach can overly constrain the design space, limiting exploration and potentially missing critical regions of interest.

Given the inevitability of model misspecification, we aim to strike a balance between these two extremes. Specifically, we seek a design mechanism that satisfies the following criteria: (i) In the early stages of the adaptive BED process, it prioritizes experimental designs that are highly informative about the parameters of interest. (ii) Over the medium

term, it promotes systematic exploration of the design space to mitigate the risks of local overfitting and gather data that can be used not only for model fitting, but for model checking as well. (iii) In the long run, the mechanism adapts to the quality of the model: if the model fits well, it reverts to the classic adaptive BED approach, operating under the assumption of no model misspecification; if the model fits poorly, it defaults to producing uniform designs.

Appendix B. Gaussian Process Auxiliary Model

For the case where observations y are in $\mathbb{R}$ it is illustrative to consider a Gaussian process auxiliary model, where $p_{\text{aux}}(y \mid \psi, \xi)$ and $p(\psi)$ are such that

$$\psi \sim \text{GP}(\mu, k), \tag{10}$$

$$y \mid \psi, \xi \sim N(\psi(\xi), \sigma^2), \tag{11}$$

for some mean function μ and kernel function k. Denote $\boldsymbol{\xi}_t \coloneqq (\xi_i)_{i=1}^t$, and for $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}' \in \mathbb{R}^m$, let $\mathbf{k}(\mathbf{x}, \mathbf{x}') \in \mathbb{R}^n \times \mathbb{R}^m$ have $\mathbf{k}(\mathbf{x}, \mathbf{x}')_{ij} = k(\mathbf{x}_i, \mathbf{x}'_j)$. Further, let the variance of $y^* \mid h_{t-1}, \xi_t, y_t, \xi^*$ be denoted $\sigma^2_{\boldsymbol{\xi}_{t-1}}(\xi_t, \xi^*)$ $(= k(\xi^*, \xi^*) + \sigma^2 - \mathbf{k}(\xi^*, \boldsymbol{\xi}_t)(\mathbf{k}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_t) + \sigma^2 I_t)^{-1}\mathbf{k}(\boldsymbol{\xi}_t, \xi^*))$. In this case, maximising $\text{EPIG}^{\text{aux}}(\xi_t \mid h_{t-1})$ is equivalent to minimising

$$\mathbb{E}_{p(\xi^*)}\left[\log \sigma^2_{\boldsymbol{\xi}_{t-1}}(\xi_t, \xi^*)\right] \tag{12}$$

over $\xi_t \in \mathcal{X}$. Therefore the optimal design seeks to minimise the posterior variance of $y^* \mid h_{t-1}, \xi_t, y_t, \xi^*$ over $\xi^* \in \mathcal{X}$, weighted by $p(\xi^*)$. It is intuitive that this is often achieved by ξ_t that is far away from the previous designs. Such space-filling behaviour is desired when there is potential model misspecification (Huan et al., 2024).

Acknowledging this, we introduce a flexible auxiliary model $p_{\text{aux}}(y \mid \psi, \xi)$, where ψ is a (potentially infinite-dimensional) parameter with a prior distribution $p(\psi)$, designed to capture plausible misspecification in the original model p_{model}, but perhaps with no natural interpretation.

Appendix C. Expected Divergence

A natural goal is to collect data (ξ, y) such that

$$\text{KL}[p_{\text{true}}(y^* \mid \xi^*) \parallel p_{\text{aux}}(y^* \mid \xi, y, \xi^*)] \tag{13}$$

is small over $\xi^* \in \mathcal{X}$ (i.e. the posterior predictive distribution is "close" to the truth). Let $p(\xi^*)$ be a distribution over $\mathcal{X}$ that weights the importance of areas of the design space: $p(\xi^*)$ is large for ξ^* where it is more important for p_{aux} to fit well to the truth. Taking expectation over $p(\xi^*)$ and $p_{\text{true}}(y \mid \xi)$ leads to the utility function

$$U(\xi) \coloneqq -\,\mathbb{E}_{p(\xi^*)p_{\text{true}}(y \mid \xi)}\left[\text{KL}[p_{\text{true}}(y^* \mid \xi^*) \parallel p_{\text{aux}}(y^* \mid \xi, y, \xi^*)]\right] \tag{14}$$

$$= \mathbb{E}_{p(\xi^*)p_{\text{true}}(y, y^* \mid \xi, \xi^*)}\left[\log \frac{p_{\text{aux}}(y^* \mid \xi, y, \xi^*)}{p_{\text{aux}}(y^* \mid \xi^*)}\right] + \text{const.} \tag{15}$$

This cannot be estimated because it involves an expectation over the unknown p_{true}. However, p_{aux} is defined such that $p_{\text{true}}(y \mid \xi) \approx p_{\text{aux}}(y \mid \xi, \psi)$ for *some* ψ. Therefore, we can accurate approximate p_{true} by p_{aux} for some unknown ψ. Plugging in this approximation and dropping the constant in Equation 15, thus yields the approximation

$$U_\psi(\xi) \coloneqq \mathbb{E}_{p(\xi^*)p_{\text{aux}}(y,y^* \mid \xi,\xi^*,\psi)}\left[\log \frac{p_{\text{aux}}(y^* \mid \xi, y, \xi^*)}{p_{\text{aux}}(y^* \mid \xi^*)}\right] \tag{16}$$

which should accurately approximate $U(\xi)$ for some currently unknown choice of ψ. While the ψ for which p_{aux} best approximates p_{true} is unknown, $p(\psi)$ describes our current beliefs for which ψ will be best (replacing this with $p(\psi|h_t)$ as data is observed). Taking an expectation over $p(\psi)$ thus gives

$$\mathbb{E}_{p(\psi)}[U_\psi(\xi)] = \mathbb{E}_{p(\xi^*)p_{\text{aux}}(y,y^* \mid \xi,\xi^*)}\left[\log \frac{p_{\text{aux}}(y^* \mid \xi, y, \xi^*)}{p_{\text{aux}}(y^* \mid \xi^*)}\right] \tag{17}$$

$$=: \text{EPIG}^{\text{aux}}(\xi)\,, \tag{18}$$

the *expected predictive information gain* (EPIG), recently introduced by Bickford Smith et al. (2023). Thus we can interpret EPIG as a Bayes estimator for our true utility function based on the auxiliary model.

Although this reward function has merit itself, it has no link to the original model; using this loss function would be rejecting p_{model} in choosing designs. However, in the case where p_{model} can closely represent the truth, it is still reasonable to target EIG in the parameters θ of p_{model}:

$$\text{EIG}^{\text{model}}(\xi) = \mathbb{E}_{p(\theta)p_{\text{model}}(y \mid \xi,\theta)}\left[\log \frac{p_{\text{model}}(y \mid \xi, \theta)}{p_{\text{model}}(y \mid \xi)}\right]. \tag{19}$$

In particular, in the adaptive experimental design setting, if there was strong evidence for p_{model} after collecting data $h_t = \{(\xi_i, y_i)\}_{i=1}^t$ then it would be reasonable to target $\text{EIG}^{\text{model}}$, but if there was strong evidence against p_{model}, it would be reasonable to target EPIG^{aux} instead. This reasoning suggests using a weighted sum of these two reward functions. The desire to also gather data that allows effective model checking to be performed, then provides high level motivation for the final $\text{EIG}_Z(\xi)$ term that appears in our objective.

Appendix D. Proof of Theorem 1

For simplicity of notation, we will consider a static design setting where $\xi_t = \xi$ and $h_{t-1} = \emptyset$. The results then trivially extend to the sequential case by conditioning all distributions on a non-empty h_{t-1}.

The robust EIG, EIG_Ω, is equivalent to the mutual information between y and $\Omega = \{\theta, (\xi^*, y^*), Z\}$, $I(y; \Omega)$. The joint distribution over (y, Ω) for a given ξ is

$$p(y, \Omega) = p(\xi^*)p(\theta)p(Z)p(y, y^*|\xi, \xi^*, \theta, Z) \tag{20}$$

where

$$p(y, y^*|\xi, \xi^*, \theta, Z = 0) = p_{\text{aux}}(y, y^*|\xi, \xi^*), \tag{21}$$

$$p(y, y^*|\xi, \xi^*, \theta, Z = 1) = p_{\text{model}}(y, y^*|\xi, \xi^*, \theta). \tag{22}$$

Now using standard results for mutual information we have

$$I(y;\{\theta,(\xi^*,y^*),Z\}) = I(y;Z) + \mathbb{E}_{p(Z)}\left[I(y;\theta|Z)\right] + \mathbb{E}_{p(\theta,Z)}\left[I(y;(\xi^*,y^*)|\theta,Z)\right]. \tag{23}$$

Now $I(y;\theta|Z=0)=0$, as the observations tell us nothing about our model parameters when we reject the model. Similarly, $I(y;(\xi^*,y^*)|\theta,Z=1)=0$, as $Z=1$ indicates rejection of the auxiliary model, while (ξ^*,y^*) provides no new information about y under the original model if we already know θ.[1] Thus expanding the expectations we have

$$\begin{aligned}
&I(y;\{\theta,(\xi^*,y^*),Z\}) \\
&\quad= I(y;Z) + p(Z=1)I(y;\theta|Z=1) + p(Z=0)\mathbb{E}_{p(\theta|Z=0)}\left[I(y;(\xi^*,y^*)|\theta,Z=0)\right], &(24)\\
&\quad= I(y;Z) + p(Z=1)I(y;\theta|Z=1) + p(Z=0)I(y;(\xi^*,y^*)|Z=0), &(25)\\
&\quad= \mathrm{EIG}_Z(\xi) + \mathbb{P}(Z=1)\,\mathrm{EIG}^{\mathrm{model}}(\xi) + \mathbb{P}(Z=0)\,\mathrm{EPIG}^{\mathrm{aux}}(\xi). &(26)
\end{aligned}$$

as required.

Appendix E. Experimental Details

It is natural to choose the mean kernel functions for our problem such that they resemble the mean and covariance of the function $\eta(\theta,\cdot)$ over $p(\theta)$. We have

$$\mathbb{E}_{p(\theta)}\left[\eta(\theta,\xi)\right] = \frac{11}{18}\xi\log\frac{200+\xi}{20+\xi} \tag{27}$$

and

$$\mathbb{E}_{p(\theta)}\left[\eta(\theta,\xi)\eta(\theta,\xi')\right] = \begin{cases} \frac{14800\xi^2}{(20+\xi)(200+\xi)} & \text{if } \xi=\xi' \\ \frac{740}{9}\frac{\xi\xi'}{\xi'-\xi}\log\frac{(200+\xi)(20+\xi')}{(200+\xi')(20+\xi)} & \text{otherwise} \end{cases} \tag{28}$$

allowing to compute the mean and covariance of $\eta(\theta,\cdot)$ analytically. For our GP we thus use

$$\mu(\xi) = \mathbb{E}_{p(\theta)}\left[\eta(\theta,\xi)\right] \tag{29}$$

$$k(\xi,\xi') = \mathrm{Cov}_{p(\theta)}(\eta(\theta,\xi),\eta(\theta,\xi')) + K_{\mathrm{M5/2}}(\xi,\xi') \tag{30}$$

where $K_{\mathrm{M5/2}}$ is the Matern 5/2 kernel with length scale 80 and vertical scale 5. The resulting GP has the same mean as η but a covariance function with increased flexibility due to the Matern 5/2 kernel.

Appendix F. Full Set of Experimental Results Plots

1. Note here that our setup has assumed a model where the outcomes are conditionally independent given θ, which can be violated in some cases (e.g. if the model has nuisance parameters alongside θ). However, even when this is not the case, the result still holds as long as we interpret y^* as the prediction made by the *auxiliary* model instead of the extended model.

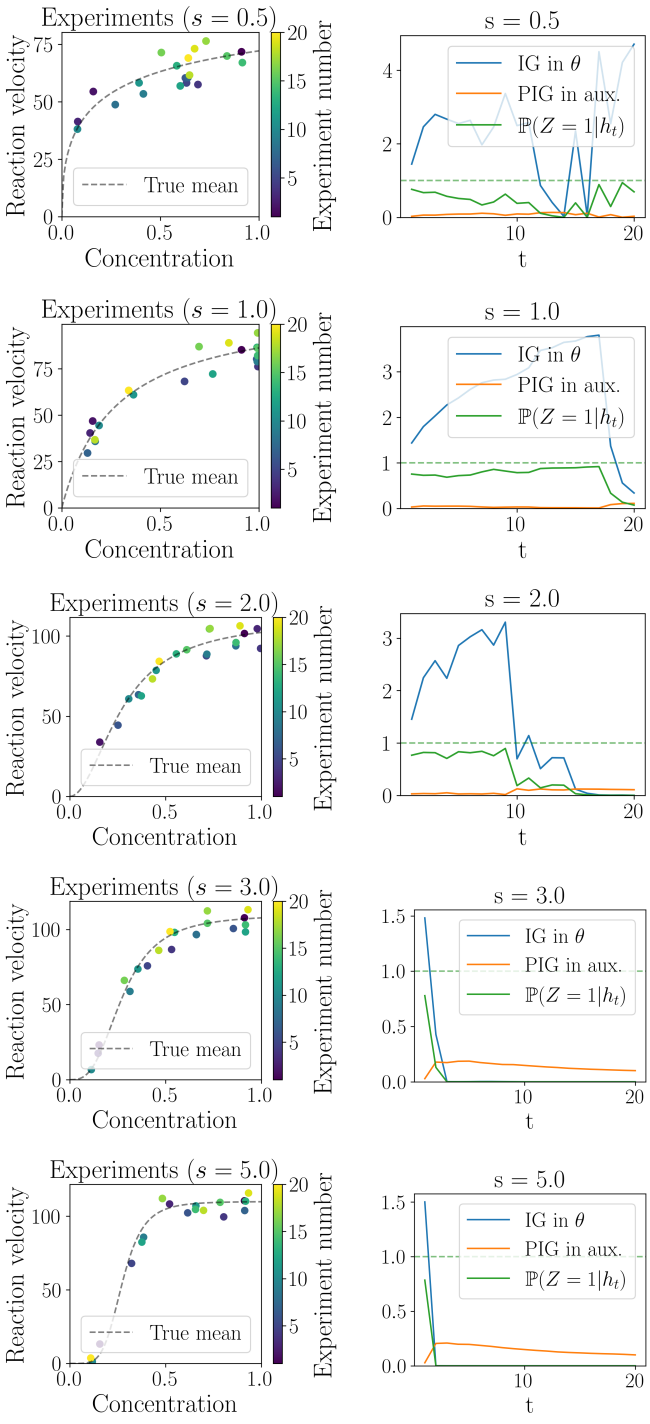


Figure 4: Full set of results plots from Figure 2.

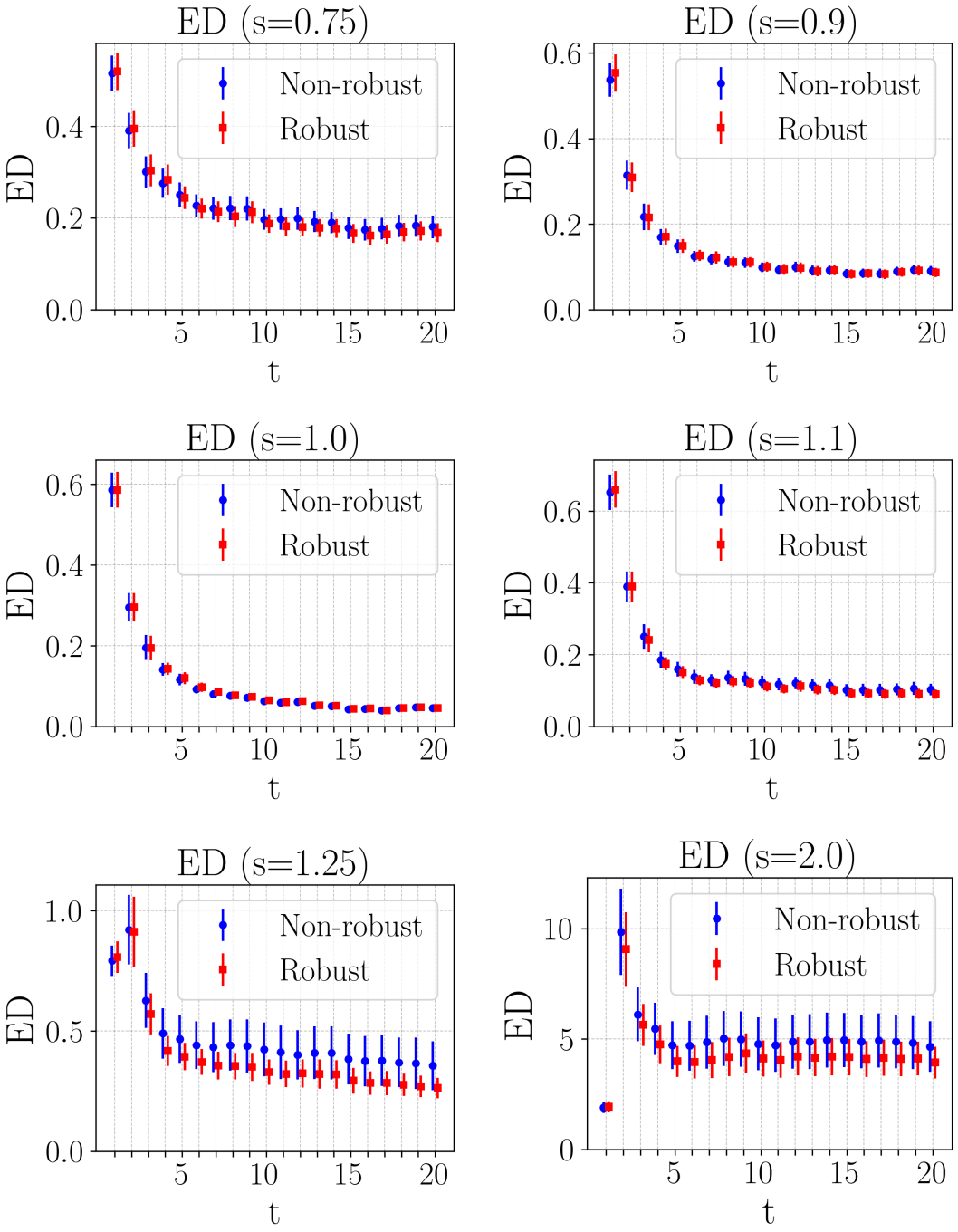


Figure 5: Full set of results plots from Figure 3.