

Bayesian Decision-Theoretic Design of Experiments Under an Alternative Model*

Antony Overstall[†] and James McGree[‡]

Abstract. Traditionally Bayesian decision-theoretic design of experiments proceeds by choosing a design to minimise expectation of a given loss function over the space of all designs. The loss function encapsulates the aim of the experiment, and the expectation is taken with respect to the joint distribution of all unknown quantities implied by the statistical model that will be fitted to observed responses. In this paper, an extended framework is proposed whereby the expectation of the loss is taken with respect to a joint distribution implied by an alternative statistical model. Motivation for this includes promoting robustness, ensuring computational feasibility and for allowing realistic prior specification when deriving a design. To aid in exploring the new framework, an asymptotic approximation to the expected loss under an alternative model is derived, and the properties of different loss functions are established. The framework is then demonstrated on a linear regression versus full-treatment model scenario, on estimating parameters of a non-linear model under model discrepancy and a cubic spline model under an unknown number of basis functions.

Keywords: cubic spline basis, expected loss function, full-treatment model, model discrepancy, non-linear model, normal linear model.

1 Introduction

The Bayesian decision-theoretic approach (Chaloner and Verdinelli, 1995) is a natural framework to plan experiments in many fields of science and engineering. It starts with specification of a loss function representing the aim of the experiment. A Bayesian design then minimises the expectation of the loss over the space of all possible designs where expectation is with respect to a probability distribution over all unknown quantities implied by the statistical model that will be fitted upon observation of the experimental responses.

The approach is scientifically appealing since the precise aim of the experiment is encoded by the loss function and, through specification of a probability distribution over all unknown quantities, pre-experimental (lack of) knowledge is incorporated. In practice, finding a Bayesian design by minimising the expected loss is a challenging computational problem (Ryan et al., 2016b). The expected loss is typically analytically intractable and the space of all designs potentially high-dimensional. However, recently

*JM was supported by an Australian Research Council Discovery Project (DP200101263).

[†]School of Mathematical Sciences, University of Southampton, Southampton, United Kingdom, A.M.Overstall@soton.ac.uk

[‡]School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia, james.mcgree@qut.edu.au

there has been significant progress in the development of novel computational methodology for finding Bayesian designs originating in diverse fields of study (e.g. Long et al., 2013; Ryan et al., 2016b; Foster et al., 2019; Beck et al., 2020, and references therein).

In this paper, we propose an extended framework for Bayesian decision-theoretic design of experiments. This is achieved by defining the expected loss by taking expectation with respect to a probability distribution implied by an alternative statistical model (termed a designer model). Etzioni and Kadane (1993) considered the case where the fitted and designer prior distributions were different but the joint distributions for the responses were identical. We extend this by considering the case where both the prior *and* distribution for responses can be different. There are several reasons why a design may be sought under a different model to the fitted model. For example, the fitted model may emulate a partially observed process, but by considering the whole data-generating process, robustness is introduced into the design procedure. Assuming a simpler fitted model can also be used to induce robustness to model misspecification (Joseph et al., 2009) or may be necessary to ensure computational feasibility of finding Bayesian designs (Ryan et al., 2016a). Conversely, reliable prior specification can be assisted by the designer model being simpler than the fitted model. Accordingly, we provide the framework and supporting methodology to enable Bayesian design within such settings, and apply these methods to typical design problems found in the literature.

The outline of the paper is as follows. In Section 2, the extended framework for Bayesian design of experiments under an alternative model is described and justified, and the choice of loss function considered. In Section 3, an asymptotic approximation to the expected loss under the extended framework is developed to explore the consequences of designing under an alternative model. Under the traditional approach, designs found under a loss, and its posterior expectation, are equivalent. This is no longer the case under the proposed extended framework. We consider the choice between these pairs of loss functions and show that there is essentially a bias-variance trade off to be made. To demonstrate the extended framework, three examples are considered in Section 4 and Section SM6 in the Supplementary Material (Overstall and McGree, 2021). The first considers a linear regression versus full-treatment model scenario where application of the extended framework results in objective functions sharing properties of classical design criteria that promote increased replication (Gilmour and Trinca, 2012). The second considers estimating parameters of non-linear models under model discrepancy, and finds that the resulting designs resemble a compromise of designs found under the traditional Bayesian decision-theoretic framework and space-filling designs. Finally, the third (in the Supplementary Material), considers estimation of the parameters of a model formed using cubic splines, with an unknown number of basis functions.

The Supplementary Material includes proofs and derivations of the results, and the cubic spline model example.

2 Bayesian design under an alternative model

2.1 Internal and external expected loss

Suppose the ultimate aim of the experiment is to learn the relationship between k controllable variables and a response. The experiment consists of n runs where, for $i = 1, \dots, n$, the i th run involves specifying a treatment-level combination, i.e. values for the k controllable variables $\Delta_i = (\Delta_{i1}, \dots, \Delta_{ik})^\top$, and subsequent observation of a response y_i . Let Δ denote the design; the $n \times k$ matrix with i th row given by Δ_i^\top and let $\mathbf{y} = (y_1, \dots, y_n)^\top$ denote the $n \times 1$ vector of observed responses.

On the completion of the experiment, a statistical model is assumed describing the relationship between the controllable variables and response. That is, \mathbf{y} is assumed to be a realisation from a multivariate probability distribution with density/mass function $\pi(\mathbf{y}|\boldsymbol{\beta}, F, \Delta)$ completely known up to a $p \times 1$ vector of unknown parameters $\boldsymbol{\beta}$ with prior density $\pi(\boldsymbol{\beta}|F, \Delta)$. We refer to $\pi(\mathbf{y}|\boldsymbol{\beta}, F, \Delta)$ and $\pi(\boldsymbol{\beta}|F, \Delta)$ as the *fitted likelihood* and *prior*, respectively, and collectively as the *fitted model*, where the conditioning on F is to make it clear that this is under the fitted model. Using Bayes' theorem, the *fitted posterior* distribution is $\pi(\boldsymbol{\beta}|\mathbf{y}, F, \Delta) \propto \pi(\mathbf{y}|\boldsymbol{\beta}, F, \Delta)\pi(\boldsymbol{\beta}|F, \Delta)$. Note that $\boldsymbol{\beta}$ refers to parameters of interest. There may be additional nuisance parameters present but these have been marginalised, with respect to a prior distribution, to obtain $\pi(\mathbf{y}|\boldsymbol{\beta}, F, \Delta)$.

Since, apart from the values of $\boldsymbol{\beta}$, the fitted model completely specifies the relationship between the controllable variables and response, the experimental aim reduces to determining the value of $\boldsymbol{\beta}$. This aim can manifest itself in various different ways. Bayesian decision-theoretic design of experiments starts by encoding the exact aim in a loss function, $\lambda(\boldsymbol{\beta}, \mathbf{y}, F, \Delta)$, giving the loss of estimating $\boldsymbol{\beta}$ via the fitted posterior distribution conditional on \mathbf{y} . The choice of loss function is considered in more detail in Section 2.2.

Prior to experimentation, the aim is to specify Δ to best estimate $\boldsymbol{\beta}$ as measured by the loss function. Traditionally, a Bayesian design minimises the expected loss over the space of all designs, where expectation is with respect to the joint distribution of $\boldsymbol{\beta}$ and \mathbf{y} as implied by the fitted model (e.g. Lindley, 1972). Mathematically, the expected loss is

$$L_{FF}(\Delta) = \mathbb{E}_{\boldsymbol{\beta}, \mathbf{y}|F, \Delta}[\lambda(\boldsymbol{\beta}, \mathbf{y}, F, \Delta)] \quad (1)$$

and referred to, in this paper, as the *internal expected loss* under the fitted model. Unless otherwise specified; if referring to internal expected loss, it is under the fitted model.

Instead, consider minimising the expectation of the loss with respect to an alternative model: the *designer model*. Here we suppose that \mathbf{y} is a realisation of a probability distribution with density/mass $\pi(\mathbf{y}|\boldsymbol{\theta}, D, \Delta)$, where $\boldsymbol{\theta}$ is a $p_\theta \times 1$ vector of parameters with prior density $\pi(\boldsymbol{\theta}|D, \Delta)$. We refer to $\pi(\mathbf{y}|\boldsymbol{\theta}, D, \Delta)$ and $\pi(\boldsymbol{\theta}|D, \Delta)$ as the *designer likelihood* and *prior*, respectively, and the conditioning on D is to make it clear that this is under the designer model. Similar to the fitted model, any additional nuisance parameters have been marginalised to obtain $\pi(\mathbf{y}|\boldsymbol{\theta}, D, \Delta)$.

Let $\boldsymbol{\beta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\theta}^\top)^\top$, where $\boldsymbol{\gamma}$ is a $p_\gamma \times 1$ vector of parameters only present in the fitted model and $\boldsymbol{\theta}$ are parameters common to both models. If the fitted and designer

models share no common parameters, i.e. $\beta = \gamma$, then the two models are referred to as *disjoint*. Conversely, if all parameters of interest are common to both models, i.e. $\beta = \theta$, then the models are referred to as *compatible*. As such, prior to marginalisation, the fitted model can be nested within the designer model, vice-versa, or no nesting may be present.

A design found under the proposed extended framework is given by minimising the *external expected loss* defined as

$$L_{DF}(\Delta) = E_{\theta, \mathbf{y}|D, \Delta} \{E_{\gamma|\theta, \mathbf{y}, F, \Delta} [\lambda(\beta, \mathbf{y}, F, \Delta)]\}. \quad (2)$$

The external expected loss is formed by assuming that once responses are observed, we proceed as if the fitted model represents the true data-generating process, i.e. under the M-closed paradigm (Bernardo and Smith, 1994, page 384). Initially, expectation is taken with respect to the fitted posterior distribution of γ conditional on θ . This is because these parameters are only present in the fitted model (if the models are non-compatible). Then expectation is taken with respect to the remaining unknowns under the designer model.

Note that if the fitted and the designer models are the same, then the external and internal expected loss functions are equal, i.e. $L_{DF}(\Delta) = L_{FF}(\Delta)$.

There are several reasons to consider finding a design under an alternative model. These can be separated into two cases: when the designer model is simpler than the fitted model and the vice-versa. In Section 4 and Section SM6 in the Supplementary Material we provide examples of both cases.

First consider the case where the fitted model actually represents the, *a-priori*, current best representation of the data-generating process, and the designer model is actually a proxy simplification. This can occur when the fitted model is complex with a large number of parameters and it is necessary to use a fitted prior distribution representing weak prior information. It is known (e.g. Ryan et al., 2016b) that in these cases, the vaguery of the fitted prior distribution combined with the large number of parameters leads to the internal expected loss being relatively flat and, therefore, non-trivial to (numerically) minimise. However it may be possible to elicit a prior distribution for a simpler designer model and use this to take expectation of the loss function formed under the fitted model.

Now consider the case where the designer model actually represents the best representation of the data-generating process and the fitted model is a simplification. In this case, it initially would make sense to find a design by minimising the internal expected loss under the designer model, i.e.

$$L_{DD}(\Delta) = E_{\theta, \mathbf{y}|D, \Delta} [\lambda(\theta, \mathbf{y}, D, \Delta)]. \quad (3)$$

For certain models it may not be possible to represent the experimental aim by a suitable loss function, $\lambda(\theta, \mathbf{y}, D, \Delta)$, depending on the designer posterior distribution. Instead, an alternative fitted model is used for which it is possible to form a suitable loss function. However it is prudent to take expectation under the original designer model

since this is the best representation of the data-generating process. For example, suppose the designer model is a full-treatment model (e.g. Gilmour and Trinca, 2012) with each unique treatment-level combination permitting a unique mean response. This represents a plausible data-generating process but for convenience and interpretability a regression model (positing a relationship between the numerical values of the controllable variables and mean response) is usually assumed as the fitted model.

Alternatively, suppose the designer model is based on scientific reasoning through a mechanistic or phenomenological model. Again it would make sense to find the design via the internal expected loss under the designer model (3). Assuming the fitted and designer models are compatible, consider Monte Carlo approximations to the external and internal expected loss functions given by $\hat{L}_{DF}(\Delta) = B^{-1} \sum_{b=1}^B \lambda(\boldsymbol{\theta}_b, \mathbf{y}_b, F, \Delta)$ and $\hat{L}_{DD}(\Delta) = B^{-1} \sum_{b=1}^B \lambda(\boldsymbol{\theta}_b, \mathbf{y}_b, D, \Delta)$, respectively, where $\{\boldsymbol{\theta}_b, \mathbf{y}_b\}_{b=1}^B$ is a sample generated from the joint distribution of $\boldsymbol{\theta}$ and \mathbf{y} under the designer model. There are models, e.g. so-called intractable likelihood models, for which it is straightforward to generate such a sample but extremely difficult to evaluate the designer posterior distribution and hence the loss function $\lambda(\boldsymbol{\theta}_b, \mathbf{y}_b, D, \Delta)$, for $b = 1, \dots, B$. However by forming an approximating fitted model it becomes computationally feasible to find a Bayesian design via the external expected loss (e.g., see the indirect inference approach of Ryan et al., 2016a). In Section 2.2, we show, under a certain class of loss function, that the external expected loss is an upper bound on the internal expected loss under the designer model, providing formal justification for using such approaches.

2.2 Loss functions

In Section 2.1, the loss function is given informally by $\lambda(\boldsymbol{\beta}, \mathbf{y}, F, \Delta)$, stating that it gives the loss of estimating $\boldsymbol{\beta}$ via the fitted posterior (conditional on \mathbf{y}). We now consider the specification of the loss and the implications of such a choice when designing an experiment under an alternative model. For this, we consider two exemplar loss functions as defined in Table 1 (left-hand column); self-information and squared error. Note that $\|\mathbf{v}\|_{2,A}^2 = \mathbf{v}^T A \mathbf{v}$ and I_p is the $p \times p$ identity matrix. These loss functions have been chosen since under a normal linear model with non-informative prior distributions, the internal expected loss reduces to the objective function under D- and A-optimality, respectively; two widely-studied classical optimal design criteria.

Optimal loss functions

First consider estimating $\boldsymbol{\beta}$ based on observed responses \mathbf{y} . Suppose the loss for this aim is $\lambda_0(\boldsymbol{\beta}, \mathbf{b})$ where $\mathbf{b} \in \mathcal{B}$ is an action for estimating $\boldsymbol{\beta}$ and \mathcal{B} is the action space (e.g. \mathbf{b} is a point estimate of $\boldsymbol{\beta}$ in $\mathcal{B} = \mathbb{R}^p$). The optimal action minimises the fitted posterior expectation of $\lambda_0(\boldsymbol{\beta}, \mathbf{b})$, i.e. $\hat{\mathbf{b}}_F = \arg \min_{\mathbf{b} \in \mathcal{B}} E_{\beta|\mathbf{y}, F, \Delta} [\lambda_0(\boldsymbol{\beta}, \mathbf{b})]$.

The loss function, as written in Section 2.1, is given by $\lambda(\boldsymbol{\beta}, \mathbf{y}, F, \Delta) = \lambda_0(\boldsymbol{\beta}, \hat{\mathbf{b}}_F)$ and is optimal under the fitted posterior distribution. In this paper, only optimal loss functions are considered. Indeed, the self-information and squared error loss functions are optimal under logarithmic scoring and $\lambda_0(\boldsymbol{\beta}, \mathbf{b}) = \|\boldsymbol{\beta} - \mathbf{b}\|_{2, I_p}^2$, respectively. However,

Generator		Composite	
Name	Expression for $\lambda(\boldsymbol{\beta}, \mathbf{y}, F, \Delta)$	Name	Expression for $\lambda(\mathbf{y}, F, \Delta)$
Self-information (SI)	$\lambda_{SI}(\boldsymbol{\beta}, \mathbf{y}, F, \Delta) = -\log \pi(\boldsymbol{\beta} \mathbf{y}, F, \Delta)$	Entropy (E)	$\lambda_E(\mathbf{y}, F, \Delta) = \mathbb{E}_{\boldsymbol{\beta} \mathbf{y}, F, \Delta} [-\log \pi(\boldsymbol{\beta} \mathbf{y}, F, \Delta)]$
Squared error (SE)	$\lambda_{SE}(\boldsymbol{\beta}, \mathbf{y}, F, \Delta) = \ \boldsymbol{\beta} - \mathbb{E}_{\boldsymbol{\beta} \mathbf{y}, F, \Delta}(\boldsymbol{\beta})\ _{2, I_p}^2$	Trace variance (TV)	$\lambda_{TV}(\mathbf{y}, F, \Delta) = \text{tr} [\text{var}_{\boldsymbol{\beta} \mathbf{y}, F, \Delta}(\boldsymbol{\beta})]$

Table 1: Exemplar generator and composite loss functions.

the formulation in Section 2.1 allows consideration of sub-optimal loss functions, if deemed appropriate, e.g. $\lambda(\boldsymbol{\beta}, \mathbf{y}, F, \Delta) = \|\boldsymbol{\beta} - \mathbb{M}_{\boldsymbol{\beta}|\mathbf{y}, F, \Delta}(\boldsymbol{\beta})\|_{2, I_p}^2$, where $\mathbb{M}_{\boldsymbol{\beta}|\mathbf{y}, F, \Delta}(\boldsymbol{\beta})$ is the fitted posterior mode of $\boldsymbol{\beta}$.

Under optimal loss functions, the following result giving an inequality for the external expected loss in relation to the internal expected loss under the designer model. The proof is given in Section SM1 of the Supplementary Material.

Lemma 2.1. *Suppose fitted and designer models are compatible, i.e. $\boldsymbol{\beta} = \boldsymbol{\theta}$, with $\lambda(\boldsymbol{\beta}, \mathbf{y}, F, \Delta) = \lambda_0(\boldsymbol{\beta}, \hat{\mathbf{b}}_F)$ and $\lambda(\boldsymbol{\beta}, \mathbf{y}, D, \Delta) = \lambda_0(\boldsymbol{\beta}, \hat{\mathbf{b}}_D)$, then $L_{DF}(\Delta) \geq L_{DD}(\Delta)$, where $L_{DD}(\Delta)$ is the internal expected loss under the designer model given by (3).*

In Lemma 2.1, the difference between the external expected loss and the internal expected loss under the designer model is due to the loss of information caused by the fitted model being different to the model under which expectation is taken. Note that Lemma 2.1 only requires the compatible fitted and designer models to be different, not that the designer model is more complex.

By finding the design that minimises the external expected loss, we are minimising an upper bound on the internal expected loss under the designer model. A special case of Lemma 2.1 for the self-information loss is called the Barber-Agakov bound (Barber and Agakov, 2003) and has been used by Foster et al. (2019) to approximately find Bayesian designs. Lemma 2.1 now justifies the use of this approach to other loss functions as has previously been pursued for intractable likelihood models (for example, Ryan et al. 2016a, as discussed in Section 2.1).

Lemma 2.1 is uninformative on the relationship between the external expected loss and the internal expected loss under the fitted model (1).

Generator and composite loss functions

There are intuitive choices of loss function given as scalar summaries of the fitted posterior distribution, independent of $\boldsymbol{\beta}$, e.g. $\lambda(\mathbf{y}, F, \Delta) = |\text{var}(\boldsymbol{\beta}|\mathbf{y}, F, \Delta)|$, where $|\cdot|$ denotes

the determinant of a square matrix. In some cases, these are given as the fitted posterior expectation of a loss function $\lambda_G(\boldsymbol{\beta}, \mathbf{y}, F, \Delta)$, i.e. $\lambda_C(\mathbf{y}, F, \Delta) = E_{\boldsymbol{\beta}|\mathbf{y}, F, \Delta} [\lambda_G(\boldsymbol{\beta}, \mathbf{y}, F, \Delta)]$. For such pairs, we term $\lambda_C(\mathbf{y}, F, \Delta)$ as composite and $\lambda_G(\boldsymbol{\beta}, \mathbf{y}, F, \Delta)$ as generator. The right-hand column of Table 1 shows the entropy and trace variance loss functions which are composite to the generator loss functions of self-information and squared error, respectively, in the left-hand column.

The internal expected composite loss is equal to the corresponding internal expected generator loss. Apart from when fitted and designer models are disjoint, i.e. $\boldsymbol{\beta} = \boldsymbol{\gamma}$, this equivalence is not necessarily true when considering the external expected loss. This leads to a question on how to choose between a pair of loss functions equivalent under internal expected loss, e.g. how should we choose between self-information and entropy loss functions? In the next section, to answer this question, we develop an asymptotic approximation to the external expected loss. This is used to explore differences in behaviour between the exemplar generator and composite loss functions shown in Table 1. We find that a composite loss solely focuses on fitted posterior precision. On the other hand, a generator loss also incorporates point estimation of common parameters $\boldsymbol{\theta}$. This is essentially a bias-variance trade-off. Considering this trade-off is unnecessary under internal expected loss because the fitted model is, in a sense, “unbiased” as it is identical to the designer model.

3 Understanding the external expected loss

In this section, an asymptotic approximation to the external expected loss is derived. This approximation is used to provide insight into properties of exemplar loss functions, in particular, to investigate the difference between external expected generator and composite loss functions.

3.1 Asymptotic approximation of external expected loss

In general, the external expected loss (2) is not available in closed form and use of an approximation is therefore necessary. Recent developments in computational methodology for approximating (and minimising) the expected loss (example references in Section 1) can actually be used for this task with little modification. Indeed, in the examples in Section 4.2 and Section SM6 in the Supplementary Material we describe Monte Carlo approximations to the external expected loss function.

However, in this section, to gain understanding of the behaviour of the external expected loss, we develop an asymptotic approximation. The resulting approximations to the external expected loss under the exemplar loss functions in Table 1 are derived and compared to the corresponding expressions under the internal expected loss.

First, some preliminary definitions are required. Let $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\gamma}}^T, \tilde{\boldsymbol{\theta}}^T)^T$ be the $p \times 1$ vector of parameter values for $\boldsymbol{\beta}$ that minimise the Kullback-Leibler divergence between the designer and fitted likelihoods for a given value of $\boldsymbol{\theta}$ under the designer model, i.e.

$\tilde{\boldsymbol{\beta}}$ maximises

$$d(\mathbf{t}) = \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}, D, \Delta} [\log \pi(\mathbf{y}|\mathbf{t}, F, \Delta)]. \quad (4)$$

Define \mathcal{I} and $\tilde{\mathcal{I}}$ to be

$$\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}, D, \Delta} \left[-\frac{\partial^2 \log \pi(\mathbf{y}|\mathbf{t}, F, \Delta)}{\partial \mathbf{t} \partial \mathbf{t}^\top} \right]$$

evaluated at $\mathbf{t} = \boldsymbol{\theta}$ and $\mathbf{t} = \tilde{\boldsymbol{\theta}}$, respectively. Similarly, define $\tilde{\mathcal{J}}$ to be

$$\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}, D, \Delta} \left[\frac{\partial \log \pi(\mathbf{y}|\mathbf{t}, F, \Delta)}{\partial \mathbf{t}} \frac{\partial \log \pi(\mathbf{y}|\mathbf{t}, F, \Delta)}{\partial \mathbf{t}^\top} \right],$$

evaluated at $\mathbf{t} = \tilde{\boldsymbol{\theta}}$, respectively. Let $\tilde{\mathcal{K}} = \tilde{\mathcal{I}}^{-1} \tilde{\mathcal{J}} \tilde{\mathcal{I}}^{-1}$ with $\tilde{\mathcal{I}}$ and $\tilde{\mathcal{K}}$ partitioned as

$$\tilde{\mathcal{I}} = \begin{bmatrix} \tilde{\mathcal{I}}_{\gamma\gamma} & \tilde{\mathcal{I}}_{\gamma\theta} \\ \tilde{\mathcal{I}}_{\theta\gamma} & \tilde{\mathcal{I}}_{\theta\theta} \end{bmatrix}, \quad \tilde{\mathcal{K}} = \begin{bmatrix} \tilde{\mathcal{K}}_{\gamma\gamma} & \tilde{\mathcal{K}}_{\gamma\theta} \\ \tilde{\mathcal{K}}_{\theta\gamma} & \tilde{\mathcal{K}}_{\theta\theta} \end{bmatrix},$$

where $\tilde{\mathcal{I}}_{\gamma\gamma}$ is the $p_\gamma \times p_\gamma$ sub-matrix corresponding to γ with similar definitions for $\tilde{\mathcal{I}}_{\theta\theta}$, $\tilde{\mathcal{I}}_{\theta\gamma} = \tilde{\mathcal{I}}_{\gamma\theta}^\top$ and the sub-matrices of $\tilde{\mathcal{K}}$.

Under regularity conditions (see Section SM2 in the Supplementary Material), an asymptotic approximation (Kleijn and van der Vaart, 2012) to the fitted posterior distribution of $\boldsymbol{\beta}$ is given by

$$\mathbb{N}(\hat{\boldsymbol{\beta}}, \tilde{\mathcal{I}}^{-1}), \quad (5)$$

where $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\gamma}}^\top, \hat{\boldsymbol{\theta}}^\top)^\top$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ under the fitted model.

The loss function is approximated by replacing fitted posterior quantities by the corresponding approximated fitted posterior quantities under (5). Such an approximation is denoted $\lambda^*(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}, \Delta)$. For example, the self-information loss is approximated by

$$\lambda_{SI}^*(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}, \Delta) = N_{SI} - \frac{1}{2} \log |\tilde{\mathcal{I}}| + \frac{1}{2} \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_{2, \tilde{\mathcal{I}}}^2,$$

where $N_{SI} = p \log(2\pi)/2$ is a constant not depending on the design.

To approximate the external expected loss (2), the parameters present only in the fitted model, $\boldsymbol{\gamma}$, are marginalised with respect to an approximate fitted posterior conditional on $\boldsymbol{\theta}$, induced from (5). Then expectation with respect to the distribution of \mathbf{y} conditional on $\boldsymbol{\theta}$ under the designer model can be approximated using asymptotic results for inference under an alternative (or misspecified) model (White, 1982).

The result is now formally stated with proof given in Section SM2 of the Supplementary Material.

Theorem 3.1. *Under conditions given in Section SM2 of the Supplementary Material, an asymptotic approximation to the external expected loss is given by*

$$L_{DF}^*(\Delta) = \mathbb{E}_{\boldsymbol{\theta}|D, \Delta} \left(\mathbb{E}_{\hat{\boldsymbol{\beta}}|\boldsymbol{\theta}, D, \Delta} \left\{ \mathbb{E}_{\boldsymbol{\gamma}|\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}, F, \Delta} \left[\lambda^*(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}, \Delta) \right] \right\} \right), \quad (6)$$

where the two inner expectations are taken with respect to

$$\gamma|\hat{\beta}, \theta, F, \Delta \sim \text{N}\left(\tilde{P}_1\hat{\beta} - \tilde{P}_2\theta, \tilde{\mathcal{I}}_{\gamma\gamma}^{-1}\right), \quad \text{and} \quad \hat{\beta}|\theta, D, \Delta \sim \text{N}\left(\tilde{\beta}, \tilde{\mathcal{K}}\right),$$

with $\tilde{P}_1 = (I_{p_\gamma}, \tilde{P}_2)$ and $\tilde{P}_2 = \tilde{\mathcal{I}}_{\gamma\gamma}^{-1}\tilde{\mathcal{I}}_{\gamma\theta}$.

In Theorem 3.1, $\tilde{\mathcal{K}}$ has the sandwich form of the ‘‘robust variance formula’’ (e.g. Pawitan, 2003, page 374). In inferential settings, the values of $\tilde{\mathcal{I}}$ and $\tilde{\mathcal{J}}$ are estimated from observed responses, whereas, in design settings, the values follow from the assumption of a designer model.

Due to the tractability of the normal distribution, the two inner expectations in (6) are available in closed form for many common loss functions, e.g. the exemplar loss functions in Table 1. Then the approximate external expected loss is written $L_{DF}^*(\Delta) = E_{\theta|D,\Delta}[\ell_G^*(\theta, \Delta)]$; the designer prior expectation of a function $\ell_G^*(\theta, \Delta)$, typically a function of $\tilde{\mathcal{I}}$ and $\tilde{\mathcal{K}}$.

For loss functions independent of β , including composite losses, the approximate loss is denoted $\lambda^*(\hat{\beta}, \Delta)$ and the approximate external loss given by $L_{DF}^*(\Delta) = E_{\theta|D,\Delta}[\ell_C^*(\theta, \Delta)]$, where $\ell_C^*(\theta, \Delta) = E_{\hat{\beta}|\theta,D,\Delta}[\lambda^*(\hat{\beta}, \Delta)]$ is often available in closed form.

If the fitted and designer models are identical, then $\mathcal{I}_D = \mathcal{J}_D$, and $\tilde{\beta} = \beta$. In this case, we obtain an asymptotic approximation to the internal expected loss given by $L_{FF}^*(\Delta) = E_{\beta|F,\Delta}[\ell_U^*(\beta, \Delta)]$, where $\ell_U^*(\beta, \Delta)$ is the expected value of $\lambda^*(\beta, \hat{\beta}, \Delta)$ with respect to (5).

3.2 Exemplar loss functions

Table 2 shows the functions $\ell_G^*(\theta, \Delta)$ and $\ell_C^*(\theta, \Delta)$ for the exemplar loss functions considered in this paper (see Table 1). Derivation of these expressions is given in Section SM3 of the Supplementary Material. Also shown is the function $\ell_U^*(\beta, \Delta)$ for the asymptotic approximation to the internal expected loss. Note that designs that minimise the fitted prior expectation of $\ell_U^*(\beta, \Delta)$ under self-information and squared error are commonly referred to as pseudo-Bayesian D- and A-optimal, respectively. In Table 2, $\tilde{\mathcal{T}}_{\theta\theta} = (\tilde{\mathcal{I}}_{\theta\theta} - \tilde{\mathcal{I}}_{\theta\gamma}\tilde{\mathcal{I}}_{\gamma\gamma}^{-1}\tilde{\mathcal{I}}_{\gamma\theta})^{-1}$ is the approximate marginal fitted posterior variance matrix of θ , under (5). Additionally, $\tilde{\mathcal{S}}_{\theta\theta} = I_{p_\theta} + \tilde{\mathcal{I}}_{\theta\gamma}\tilde{\mathcal{I}}_{\gamma\gamma}^{-1}\tilde{\mathcal{I}}_{\gamma\gamma}^{-1}\tilde{\mathcal{I}}_{\gamma\theta}$.

Note that ℓ_G^* , ℓ_C^* and ℓ_U^* all feature $-\log|\mathcal{I}|$ (under self-information/entropy) or $\text{tr}(\mathcal{I}^{-1})$ (under squared error/trace variance). Minimising these terms has the effect of (approximately) maximising posterior precision. For the pseudo-Bayesian objective functions these are averaged with respect to the fitted prior. Whereas for external expected loss, they are averaged with respect to the designer prior of θ through $\hat{\beta}$; a function of θ and the best possible estimate of β under the fitted model. The expressions for ℓ_C^* (i.e. under composite loss) give the entropy and trace of variance under the asymptotic distribution of β under a misspecified model (5).

	Self-information	Squared error
$\ell_G^*(\boldsymbol{\theta}, \Delta)$	$N_{SI} - \frac{1}{2} \log \tilde{\mathcal{I}} + \frac{1}{2} \ \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\ _{2, \tilde{\mathcal{T}}_{\theta\theta}^{-1}}^2 + \frac{1}{2} \text{tr} \left(\tilde{\mathcal{T}}_{\theta\theta}^{-1} \tilde{\mathcal{K}}_{\theta\theta} \right) + p_\gamma/2$	$\text{tr} \left(\tilde{\mathcal{I}}^{-1} \right) + \ \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\ _{2, \tilde{\mathcal{S}}_{\theta\theta}}^2 + \text{tr} \left[\tilde{\mathcal{S}}_{\theta\theta} \left(\tilde{\mathcal{K}}_{\theta\theta} - \tilde{\mathcal{T}}_{\theta\theta} \right) \right]$
	Entropy	Trace variance
$\ell_C^*(\boldsymbol{\theta}, \Delta)$	$N_{SI} - \frac{1}{2} \log \tilde{\mathcal{I}} + \frac{p}{2}$	$\text{tr} \left(\tilde{\mathcal{I}}^{-1} \right)$
	D-optimality	A-optimality
$\ell_U^*(\boldsymbol{\beta}, \Delta)$	$N_{SI} - \frac{1}{2} \log \mathcal{I} + \frac{p}{2}$	$\text{tr} \left(\mathcal{I}^{-1} \right)$

Table 2: Exemplar approximate external expected loss functions.

The expressions for $\ell_G^*(\boldsymbol{\theta}, \Delta)$ under the generator squared error and self-information loss functions feature two extra terms related to the quality of $\hat{\boldsymbol{\theta}}$ as a point estimate of the common parameters $\boldsymbol{\theta}$. The first of these, $\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{2, A}^2$ where A is either $\tilde{\mathcal{T}}_{\theta\theta}^{-1}$ or $\tilde{\mathcal{S}}_{\theta\theta}$, relates to the bias. This term penalises designs that lead to $\hat{\boldsymbol{\theta}}$ being (on average under the designer model) far from $\boldsymbol{\theta}$. The second term measures the difference between the variance of $\hat{\boldsymbol{\theta}}$ under the fitted model, i.e. $\tilde{\mathcal{T}}_{\theta\theta}$, and under the designer model, i.e. $\tilde{\mathcal{K}}_{\theta\theta}$.

4 Examples

Two examples are considered in this section where designs are found under our proposed design framework. In both cases, the fitted model is simpler than the designer model. As will be seen, the designs found provide robustness through replicated design points within a linear regression setting, and also through providing space filling properties for estimating parameters under a non-linear model. In Section SM6 in the Supplementary Material, a further example is provided where the fitted model is more complex (as measured by number of parameters) than the designer model.

4.1 Normal linear regression vs full-treatment model

In this example we explore the implications of assuming a full-treatment designer model within a fitted linear regression setting. Under the fitted model, it is assumed

$$\mathbf{y} | \boldsymbol{\gamma}, \sigma^2, F, \Delta \sim \text{N} \left(X\boldsymbol{\gamma}, \sigma^2 I_n \right),$$

where X is an $n \times p$ model matrix (a function of Δ), and $\sigma^2 > 0$ is a nuisance parameter. We assume a conjugate normal-inverse-gamma fitted prior distribution for $\boldsymbol{\gamma}$ and σ^2 such that $\boldsymbol{\gamma} | \sigma^2, F \sim \text{N} \left(\boldsymbol{\mu}_F, \sigma^2 V_F \right)$ and $\sigma^2 | F \sim \text{IG} \left(a_F/2, b_F/2 \right)$. For the designer model, at

this stage, it is only assumed that \mathbf{y} has finite mean $E(\mathbf{y}|D, \Delta) = \mathbf{m}_D$ and variance $\text{var}(\mathbf{y}|D, \Delta) = \Sigma_D$. Note that the fitted and designer models are disjoint, sharing no common parameters.

Under the fitted and designer models, it is possible to derive closed form expressions, to some extent, for the external expected loss. The fitted posterior distribution of $\boldsymbol{\gamma}$ is $t(\hat{\boldsymbol{\mu}}_F, \hat{b}_F \hat{V}_F / \hat{a}_F, \hat{a}_F)$, i.e. a multivariate t-distribution with mean $\hat{\boldsymbol{\mu}}_F = \hat{V}_F (X^\top \mathbf{y} + V_F^{-1} \boldsymbol{\mu}_F)$, scale matrix $\hat{b}_F \hat{V}_F / \hat{a}_F$, and $\hat{a}_F = a_F + n$ degrees of freedom, where

$$\hat{V}_F = (X^\top X + V_F^{-1})^{-1}; \quad \hat{b}_F = b_F + \|\mathbf{y} - X\boldsymbol{\mu}_F\|_{2, \Sigma_F^{-1}}^2; \quad \Sigma_F = I_n + X V_F X^\top.$$

The external expected self-information and squared error loss functions are given by

$$L_{DF,SI}(\Delta) = H_{SI,1} + \frac{1}{2} \log |\hat{V}_F| + \frac{p}{2} E_{\mathbf{y}|D,\Delta}(\log \hat{b}_F), \quad (7)$$

$$L_{DF,SE}(\Delta) = \frac{E_{\mathbf{y}|D,\Delta}(\hat{b}_F)}{\hat{a}_F - 2} \text{tr}(\hat{V}_F), \quad (8)$$

respectively, where $E_{\mathbf{y}|D,\Delta}(\hat{b}_F) = b_F + \|\mathbf{m}_D - X\boldsymbol{\mu}_F\|_{2, \Sigma_F^{-1}}^2 + \text{tr}(\Sigma_D \Sigma_F^{-1})$ and $H_{SI,1}$ is a constant not depending on the design. Derivations of (7) and (8) are provided in Section SM4 in the Supplementary Material. The expectation of $\log \hat{b}_F$ with respect to the designer model, appearing in the external expected self-information loss (7), is not available in closed form. In what follows, we use a delta approximation (for example Davison, 2003, pages 33–35) where $E_{\mathbf{y}|D,\Delta}(\log \hat{b}_F) \approx \log E_{\mathbf{y}|D,\Delta}(\hat{b}_F)$. However, note that by Jensen's inequality, i.e. $E_{\mathbf{y}|D,\Delta}(\log \hat{b}_F) \leq \log E_{\mathbf{y}|D,\Delta}(\hat{b}_F)$, the approximate external expected self-information loss (which we minimise to find designs) is actually an upper bound on the true external expected loss.

Compare (7) and (8) to the corresponding expressions for the internal expected self-information and squared error given by

$$L_{FF,SI}(\Delta) = H_{SI,2} + \frac{1}{2} \log |\hat{V}_F|, \quad \text{and} \quad L_{FF,SE}(\Delta) = \frac{b_F}{a_F - 2} \text{tr}(\hat{V}_F),$$

respectively, where $H_{SI,2}$ is a constant not depending on the design. The difference is that the external expected loss functions feature expectation of (a function of) \hat{b}_F under the designer model. O'Hagan and Forster (2004, page 319) describe how \hat{b}_F summarises the inadequacy of the fitted model so it is natural that this drives the difference, with the external expected losses favouring designs for which the expected value of \hat{b}_F is small.

Now assume a specific form for the designer model as a full-treatment model (e.g Gilmour and Trinca, 2012), i.e. where each unique treatment permits a unique mean response. Mathematically, the full-treatment designer model is

$$\mathbf{y}|\boldsymbol{\tau}, \sigma^2, D, \Delta \sim N(Z\boldsymbol{\tau}, \sigma^2 I_n); \quad \boldsymbol{\tau}|\sigma^2, D, \Delta \sim N(\boldsymbol{\mu}_D, \sigma^2 V_D); \quad \sigma^2|D \sim \text{IG}\left(\frac{a_D}{2}, \frac{b_D}{2}\right),$$

where $\boldsymbol{\tau}$ is a $q \times 1$ vector of mean treatment effects. The $n \times q$ designer model matrix Z is a function of Δ where q is the number of unique treatments. For $i = 1, \dots, n$, the i th row of Z is the vector \mathbf{e}_j , i.e. a vector of zeros except for the j th element which is one, indicating that run i receives unique treatment j . Therefore, $\mathbf{m}_D = Z\boldsymbol{\mu}_D$, $\Sigma_D = b_D(I_n + ZV_DZ^T)/(a_D - 2)$ and $E_{\mathbf{y}|D,\Delta}(\hat{b}_F) = b_F + \|Z\boldsymbol{\mu}_D - X\boldsymbol{\mu}_F\|_{2,\Sigma_F^{-1}}^2 + \text{tr}(\Sigma_D\Sigma_F^{-1})$.

Interpretation is simplified by assuming non-informative fitted and designer prior distributions where, for the fitted model $b_F = 0$, $\boldsymbol{\mu}_F = \mathbf{0}_p$ (the $p \times 1$ vector of zeros) and $V_F^{-1} = 0_{p \times p}$ (the $p \times p$ matrix of zeros), and, for the designer model, $\boldsymbol{\mu}_D = \mathbf{0}_q$. However, V_D needs to be finite for Σ_D to be finite and for the external expected loss to exist. It is assumed the elements of $\boldsymbol{\tau}$ are independent with $\tau_j \sim N(0, \kappa\sigma^2/n_j)$, where $\kappa > 0$ and n_j is the number of runs receiving the j th unique treatment, for $j = 1, \dots, q$. Under this prior specification, the designer posterior mean of τ_j is the sample mean of the corresponding responses shrunk toward zero by a common factor $\kappa/(1 + \kappa)$, with κ controlling the amount of shrinkage. Consequently, $V_D = \kappa(Z^T Z)^{-1}$.

Under these prior specifications, minimising the external expected self-information and squared error loss functions is equivalent to minimising

$$\bar{L}_{DF,SI}(\Delta) = p \log[(1 + \kappa)(n - p) - \kappa d] - \log |X^T X|, \quad (9)$$

$$\bar{L}_{DF,SE}(\Delta) = [(1 + \kappa)(n - p) - \kappa d] \text{tr}[(X^T X)^{-1}], \quad (10)$$

respectively, where $d = n - q$ is known as the pure error degrees of freedom, i.e. the number of repeated treatments. Derivations of expressions (9) and (10) are provided in Section SM5 of in the Supplementary Material. We refer to designs that minimise (9) and (10) as DE- and AE-optimal, respectively. The corresponding expressions to (9) and (10) under the internal expected loss are

$$\bar{L}_{FF,SI}(\Delta) = -\log |X^T X|, \quad \text{and} \quad \bar{L}_{FF,SE}(\Delta) = \text{tr}[(X^T X)^{-1}],$$

respectively, and designs that minimise $\bar{L}_{FF,SI}(\Delta)$ and $\bar{L}_{FF,SE}(\Delta)$ are called D- and A-optimal, respectively. Hence it can be seen that under the external expected loss there is a trade-off between precision of estimation of $\boldsymbol{\gamma}$, as measured by functions of $X^T X$, and pure error degrees of freedom, i.e. DE- and AE-optimal designs will tend to feature more replication than D- and A-optimal designs, respectively. A numerical example verifying this is now provided.

Consider Example 1 from Gilmour and Trinca (2012) involving an experiment with $n = 16$ runs and $k = 3$ design variables. The fitted model is a second-order model including an intercept, three first-order terms, three quadratic terms and three pairwise interactions, i.e. $p = 10$. We specify $\kappa = n$ leading to a unit-information prior for $\boldsymbol{\tau}$ (Smith and Spiegelhalter, 1980).

DE- and AE-efficiency of a design Δ are defined by

$$\text{Eff}_{DE}(\Delta) = \exp \left\{ \frac{1}{p} [\bar{L}_{DF,SI}(\Delta_{DF,SI}^*) - \bar{L}_{DF,SI}(\Delta)] \right\},$$

$$\text{Eff}_{AE}(\Delta) = \frac{\bar{L}_{DF,SE}(\Delta_{DF,SE}^*)}{\bar{L}_{DF,SE}(\Delta)},$$

respectively, where $\Delta_{DF,SI}^*$ and $\Delta_{DF,SE}^*$ are the DE- and AE-optimal designs, respectively, with similar expressions for D- and A-efficiency. The D-efficiency of the DE-optimal design is 85%, whereas the DE-efficiency of the D-optimal design is only 7%. The corresponding figures for squared error are 84% and 10%. Clearly, the D- and A-optimal designs are not robust to the external expected loss, whereas the DE- and AE-optimal designs are robust to the internal expected loss. The D-optimal design has $q = 16$ unique treatments compared to $q = 10$ for the DE-optimal design. The equivalent values for the A- and AE-optimal designs are $q = 14$ and $q = 10$, respectively. The minimum number of unique treatments to be able to estimate all elements of $\boldsymbol{\gamma}$ under the fitted model is $q = p = 10$. Thus the DE- and AE-optimal designs feature the maximum amount of replication possible whilst still being able to estimate $\boldsymbol{\gamma}$.

Expressions for the external expected loss under self-information (9) and squared error (10) should be compared to the following objective functions under the DP- and AP-criteria (Gilmour and Trinca, 2012)

$$\bar{L}_{DP}(\Delta) = p \log F_{p,d,1-\alpha} - \log |X^T X|, \quad (11)$$

$$\bar{L}_{AP}(\Delta) = F_{1,d,1-\alpha} \text{tr} \left[(X^T X)^{-1} \right], \quad (12)$$

respectively. Note that $\bar{L}_{DP}(\Delta)$ in (11) is actually the natural logarithm of the objective function in Gilmour and Trinca (2012) but the optimisation problem is equivalent due to log being a monotonically increasing function. In (11) and (12), $F_{p,d,1-\alpha}$ the $1 - \alpha$ -quantile of the $F_{p,d}$ distribution with $F_{p,d,1-\alpha}$ being a decreasing function of d . The motivation behind the DP- and AP-criteria is to minimise the volume of the confidence region or sum of confidence intervals, respectively, for $\boldsymbol{\gamma}$, where the response variance σ^2 is estimated under the full-treatment model; a model-independent estimate. In both cases, the objective functions are modified by a decreasing function of the pure error degrees of freedom, d . As shown above we are able to provide a different, Bayesian decision-theoretic, motivation for the same outcome.

4.2 Estimating parameters of a Michaelis-Menten model under model discrepancy

In this section we consider designing an experiment to estimate the parameters of a non-linear model. Finding designs for such models is widely featured in the design of experiments literature (see, e.g. Pronzato and Pázman 2013; Federov and Leonov 2014). Under such models, it is hypothesised that the expectation of y_i is $\eta(\boldsymbol{\theta}, \Delta_i)$; a mathematical model based on mechanistic or phenomenological arguments. As such the parameters $\boldsymbol{\theta}$ can have interpretable physical meaning.

As a specific example, consider the Michaelis-Menten model (e.g. Dette and Biedermann, 2003), commonly-used in the study of pharmacokinetics and chemical kinetics. A standard experiment has a response given by reaction velocity y_i measured for

$k = 1$ controllable variable: the scaled concentration of substrate $\Delta_i = x_i \in [0, 1]$, for $i = 1, \dots, n$. The model has

$$\eta(\boldsymbol{\theta}, x) = \theta_1 x L / (xL + \theta_2), \quad (13)$$

where $\theta_1 > 0$ is the maximum velocity of the chemical reaction, $\theta_2 > 0$ is the substrate concentration at which the reaction rate is half of θ_1 , and L is the maximum possible concentration of substrate, which we assume is $L = 400$.

Supposing a normal distribution, we assume the following fitted model

$$\mathbf{y} | \boldsymbol{\theta}, \sigma^2, F, \Delta \sim N(\boldsymbol{\eta}(\boldsymbol{\theta}, \Delta), \sigma^2 I_n), \quad (14)$$

where $\boldsymbol{\eta}(\boldsymbol{\theta}, \Delta) = (\eta(\boldsymbol{\theta}, x_1), \dots, \eta(\boldsymbol{\theta}, x_n))^T$ and $\sigma^2 > 0$ is the unknown response variance (a nuisance parameter, not of direct interest). The design problem is to choose the values of $\Delta = (x_1, \dots, x_n)^T$, i.e. the n concentrations.

The fitted model given by (14) is assumed for pragmatic convenience. However the reasoning behind the mathematical model is usually incomplete, meaning that there can be a systematic discrepancy between the true expected value of y_i and $\eta(\boldsymbol{\theta}, x_i)$. The Kennedy and O'Hagan (2001) framework aims to model this discrepancy using a Gaussian process and we assume this for the designer model. Specifically, the designer model is assumed to be

$$\mathbf{y} | \boldsymbol{\theta}, \sigma^2, \boldsymbol{\delta}, D, \Delta \sim N(\boldsymbol{\eta}(\boldsymbol{\theta}, \Delta) + \boldsymbol{\delta}, \sigma^2 I_n), \quad (15)$$

where

$$\boldsymbol{\delta} | \alpha, \rho, \sigma^2, D, \Delta \sim N(\mathbf{0}_n, \sigma^2 \rho C). \quad (16)$$

In (16), $\rho > 0$, $\tau^2 > 0$, and C is an $n \times n$ correlation matrix with ij th element $C_{ij} = c(|x_i - x_j|; \alpha)$ where $c(|x_i - x_j|; \alpha)$ is a correlation function depending on $\alpha > 0$ and the distance between x_i and x_j , with the property that $c(0; \alpha) = 1$. In what follows, we choose the Matérn ($\frac{5}{2}$) correlation function where

$$c(d; \alpha) = \left(1 + \frac{d}{\alpha} + \frac{d^2}{3\alpha^2}\right) \exp\left(-\frac{d}{\alpha}\right)$$

a commonly-used correlation function (e.g. Rasmussen and Williams, 2006, page 84). It follows from (15) and (16) that

$$\mathbf{y} | \boldsymbol{\theta}, \sigma^2, \alpha, \rho, D, \Delta \sim N[\boldsymbol{\eta}(\boldsymbol{\theta}, \Delta), \sigma^2 (I_n + \rho C)]. \quad (17)$$

We consider three designs found by minimising

- (i) external expected squared error loss;
- (ii) external expected trace variance loss;
- (iii) internal expected squared error loss.

In each case, the expected loss is not available in closed form. Indeed, since the fitted posterior is not of known form, the loss function itself is analytically intractable. To approach this problem, we use a nested Monte Carlo approximation (e.g. Rainforth et al., 2018). The following algorithm is used to approximate the external expected squared error loss.

1. Generate a sample $\{\boldsymbol{\theta}_b, \sigma_b^2, \alpha_b, \rho_b\}_{b=1}^B$ of size B from the designer prior distribution of $\boldsymbol{\theta}$, σ^2 , α and ρ .
2. For $b = 1, \dots, B$, generate \mathbf{y}_b from the designer likelihood

$$\mathbf{y}|\boldsymbol{\theta}_b, \sigma_b^2, \alpha_b, \rho_b, D, \Delta \sim N[\boldsymbol{\eta}(\boldsymbol{\theta}_b, \Delta), \sigma_b^2 (I_n + \rho_b C_b)].$$

3. Approximate the external expected squared error loss by the following Monte Carlo approximation

$$\hat{L}_{DF,SE}(\Delta) = \frac{1}{B} \sum_{b=1}^B \|\boldsymbol{\theta}_b - \hat{\mathbb{E}}(\boldsymbol{\theta}|\mathbf{y}_b, F, \Delta)\|_{2, I_2}^2. \quad (18)$$

In (18), $\hat{\mathbb{E}}(\boldsymbol{\theta}|\mathbf{y}_b, F, \Delta)$ is a Monte Carlo approximation to the fitted posterior mean of $\boldsymbol{\theta}$ conditional on \mathbf{y}_b . This approximation is formed as follows. Generate a sample $\{\tilde{\boldsymbol{\theta}}_j, \tilde{\sigma}_j^2\}_{j=1}^{\tilde{B}}$ of size \tilde{B} from the fitted prior distribution of $\boldsymbol{\theta}$ and σ^2 . Then

$$\hat{\mathbb{E}}(\boldsymbol{\theta}|\mathbf{y}_b, F, \Delta) = \frac{\sum_{j=1}^{\tilde{B}} \tilde{\boldsymbol{\theta}}_j \pi(\mathbf{y}_b|\tilde{\boldsymbol{\theta}}_j, \tilde{\sigma}_j^2, F, \Delta)}{\sum_{j=1}^{\tilde{B}} \pi(\mathbf{y}_b|\tilde{\boldsymbol{\theta}}_j, \tilde{\sigma}_j^2, F, \Delta)},$$

for $b = 1, \dots, B$.

Note that the sample generated in Step 3 is reused to calculate $\hat{\mathbb{E}}(\boldsymbol{\theta}|\mathbf{y}_b, F, \Delta)$ for all $b = 1, \dots, B$.

In a similar fashion, the external expected trace variance loss can be approximated by

$$\hat{L}_{DF,TV}(\Delta) = \frac{1}{B} \sum_{b=1}^B \text{tr}\{\text{vâr}(\boldsymbol{\theta}|\mathbf{y}_b, F, \Delta)\}$$

with

$$\text{vâr}(\boldsymbol{\theta}|\mathbf{y}_b, F, \Delta) = \hat{\mathbb{E}}(\boldsymbol{\theta}\boldsymbol{\theta}^T|\mathbf{y}_b, F, \Delta) - \hat{\mathbb{E}}(\boldsymbol{\theta}|\mathbf{y}_b, F, \Delta)\hat{\mathbb{E}}(\boldsymbol{\theta}|\mathbf{y}_b, F, \Delta)^T,$$

and

$$\hat{\mathbb{E}}(\boldsymbol{\theta}\boldsymbol{\theta}^T|\mathbf{y}_b, F, \Delta) = \frac{\sum_{j=1}^{\tilde{B}} \tilde{\boldsymbol{\theta}}_j \tilde{\boldsymbol{\theta}}_j^T \pi(\mathbf{y}_b|\tilde{\boldsymbol{\theta}}_j, \tilde{\sigma}_j^2, F, \Delta)}{\sum_{j=1}^{\tilde{B}} \pi(\mathbf{y}_b|\tilde{\boldsymbol{\theta}}_j, \tilde{\sigma}_j^2, F, \Delta)}.$$

Design	Efficiency			Number of unique concentrations
	External SE (i)	External TV (ii)	Internal SE (iii)	
External SE (i)	100	67.0	79.7	13
External TV (ii)	95.4	100.0	99.9	12
Internal SE (iii)	86.1	85.8	100	12

Table 3: Efficiencies and number of unique concentrations for the three designs found for the Michaelis-Menten model.

The internal expected squared error loss is approximated by, first, generating a sample $\{\boldsymbol{\theta}_b, \sigma_b^2\}_{b=1}^B$ of size B from the fitted prior distribution of $\boldsymbol{\theta}$ and σ^2 . For $b = 1, \dots, B$, generate \mathbf{y}_b from the fitted likelihood

$$\mathbf{y}|\boldsymbol{\theta}_b, \sigma_b^2, F, \Delta \sim \mathcal{N}(\boldsymbol{\eta}(\boldsymbol{\theta}_b, \Delta), \sigma_b^2 I_n).$$

Then

$$\hat{L}_{FF,SE}(\Delta) = \frac{1}{B} \sum_{b=1}^B \|\boldsymbol{\theta}_b - \hat{\mathbf{E}}(\boldsymbol{\theta}|\mathbf{y}_b, F, \Delta)\|_{2,I_2}^2.$$

We find designs with $n = 20$ concentrations. The fitted prior distribution is such that $\boldsymbol{\theta}$ and σ^2 are independent with $\theta_1, \theta_2 \stackrel{\text{iid}}{\sim} \text{U}[20, 200]$ and $\sigma^2 \sim \text{Exp}(1)$; an exponential distribution with mean one. The designer prior distribution is such that $\boldsymbol{\theta}$, σ^2 , α and ρ are independent with $\theta_1, \theta_2 \stackrel{\text{iid}}{\sim} \text{U}[20, 200]$, $\sigma^2, \rho \stackrel{\text{iid}}{\sim} \text{Exp}(1)$ and $\alpha \sim \text{Exp}(5)$.

To minimise the nested Monte Carlo approximations to the expected loss functions, we use the approximate coordinate exchange algorithm (Overstall and Woods, 2017). We use $B = \hat{B} = 20,000$ but it is noted that if $\|\boldsymbol{\theta} - \hat{\mathbf{E}}(\boldsymbol{\theta}|\mathbf{y}, F, \Delta)\|_{2,I_2}^2$ is continuously differentiable then the convergence rate can be slightly improved (Rainforth et al., 2018) by setting $B \propto \hat{B}^2$ (whilst holding $B\hat{B}$ fixed).

Table 3 shows the efficiencies of the three designs under the three expected loss functions and the number of unique concentrations. The top panel of Figure 1 shows $\mu(\boldsymbol{\theta}, x)$ plotted against x for 100 values generated from the designer prior of $\boldsymbol{\theta}$ and the bottom panel shows the location of the concentrations for each of the three designs, i.e. found by minimising (i) external expected squared error loss; (ii) external expected trace variance loss; and (iii) internal expected squared error loss. The size of the plotting symbol is an increasing function of the replication of each concentration. From Figure 1, the three designs have common clusters of concentrations around 0.2 and 1. The designs found by minimising the external expected trace variance and internal expected squared error loss functions are broadly similar. This is confirmed in Table 3 by the relatively high efficiencies for these designs. Compared to the other designs, the concentrations under the design that minimises the external expected squared error loss are more ‘‘spread out’’, i.e. a compromise between the designs found under external trace variance and internal squared error and a space-filling design.

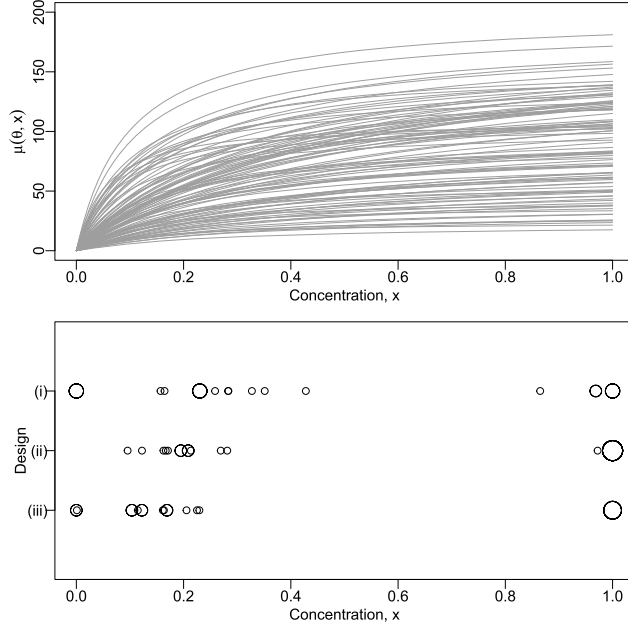


Figure 1: Designs for the Michaelis-Menten model. The top panel shows $\mu(\boldsymbol{\theta}, x)$ plotted against x for 100 values generated from the designer prior of $\boldsymbol{\theta}$. The bottom panel shows the location of the concentrations for each of the three designs, i.e. found by minimising (i) external expected squared error loss; (ii) external expected trace variance loss; and (iii) internal expected squared error loss. The size of the plotting symbol is an increasing function of the replication of each concentration.

The similarity of the designs found under the external expected trace variance (ii) and the internal expected squared error (iii) loss functions agrees with the asymptotic results from Section 4.2. In this example, the function $d(\mathbf{t})$ given by (4) is

$$\begin{aligned} d(\mathbf{t}_\theta, t_{\sigma^2}) &= \mathbb{E}_{\sigma^2, \rho, \alpha | D, \Delta} \left\{ \mathbb{E}_{\mathbf{y} | \boldsymbol{\theta}, \sigma^2, \rho, \alpha, D, \Delta} [\log \pi(\mathbf{y} | \mathbf{t}_\theta, t_{\sigma^2}, F, \Delta)] \right\} \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log t_{\sigma^2} - \frac{1}{2t_{\sigma^2}} \|\boldsymbol{\eta}(\boldsymbol{\theta}, \Delta) - \boldsymbol{\eta}(\mathbf{t}_\theta, \Delta)\|_{2, I_n}^2 - \frac{n\sigma_0^2(1 + \rho_0)}{2t_{\sigma^2}}, \end{aligned}$$

where \mathbf{t}_θ and t_{σ^2} are the elements of \mathbf{t} corresponding to $\boldsymbol{\theta}$ and σ^2 , respectively, and $\sigma_0^2 = \mathbb{E}_{\sigma^2 | D, \Delta}(\sigma^2)$ and $\rho_0 = \mathbb{E}_{\rho | D, \Delta}(\rho)$ are the designer prior means of σ^2 and ρ , respectively. Clearly, $d(\mathbf{t}_\theta, t_{\sigma^2})$ is maximised by $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and $\tilde{\sigma}^2 = \sigma_0^2(1 + \rho_0)$. It then follows from (5) that the fitted marginal posterior distribution of $\boldsymbol{\theta}$ is approximated by

$$\mathbb{N}\left(\boldsymbol{\theta}, \tilde{\mathcal{I}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\right),$$

where

$$\tilde{\mathcal{I}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} = \sigma_0^2(1 + \rho_0) \left[\sum_{i=1}^n \frac{\partial \eta(\mathbf{t}, \Delta_i)}{\partial \mathbf{t}} \Big|_{\mathbf{t}=\boldsymbol{\theta}} \frac{\partial \eta(\mathbf{t}, \Delta_i)}{\partial \mathbf{t}^\top} \Big|_{\mathbf{t}=\boldsymbol{\theta}} \right]^{-1}.$$

Therefore, from Table 2, the external expected trace variance loss is approximated by

$$L_{DF,TV}^*(\Delta) = E_{\theta|D,\Delta} \left[\text{tr} \left(\tilde{\mathcal{I}}_{\theta\theta}^{-1} \right) \right]. \quad (19)$$

By contrast, from Table 2, the internal expected squared error loss is approximated by

$$L_{FF,SE}^*(\Delta) = E_{\theta|F,\Delta} \left[\text{tr} \left(\mathcal{I}_{\theta\theta}^{-1} \right) \right], \quad (20)$$

where

$$\mathcal{I}_{\theta\theta}^{-1} = E_{\sigma^2|F,\Delta}(\sigma^2) \left[\sum_{i=1}^n \frac{\partial \eta(\mathbf{t}, \Delta_i)}{\partial \mathbf{t}} \Big|_{\mathbf{t}=\theta} \frac{\partial \eta(\mathbf{t}, \Delta_i)}{\partial \mathbf{t}^T} \Big|_{\mathbf{t}=\theta} \right]^{-1}.$$

Since $\sigma_0^2(1 + \rho_0)$ and $E_{\sigma^2|F,\Delta}(\sigma^2)$ do not depend on the design (for this example), the approximations (19) and (20) are minimised by the same design.

5 Discussion

An extended framework is proposed for Bayesian design whereby the expectation of the loss is taken with respect to the probability distribution implied by an alternative model (the designer model) to the one being fitted to the observed responses. Properties of the framework are explored. The key findings are that the external expected loss is an upper bound on the corresponding internal expected loss under the designer model, providing justification for using a simpler fitted model for the purposes of computing a design. We also observed that designs found under composite and generator loss functions are no longer equivalent. By developing an asymptotic approximation to the external expected loss we were able to show that the composite loss functions focus solely on maximising posterior precision, whereas generator loss functions also incorporate point estimation.

We encourage researchers consider the extended framework when designing experiments under the Bayesian decision-theoretic approach, if not to specify the design to be used in practice, but at least as a benchmark to assess the performance of that design. Designs can be found under the proposed framework with only minor modification of existing computational methods (see Section 4.2 and Section SM6 in the Supplementary Material).

In two of the examples, the designer model is more complex than the fitted model, and we are, in a sense, attempting to insert robustness into the design procedure. Robust Bayesian inference (i.e. inference under the M-open paradigm of Bernardo and Smith 1994, page 385) has become an important research field in recent years (e.g. Bissiri et al., 2016; Grünwald and van Ommen, 2017; Jewson et al., 2018). Related to this is a body of current work on the estimation of the parameters of a non-linear model (see Section 4.2) taking account of the incompleteness of the underlying mathematical model (e.g. Plumlee, 2017; Xie and Xu, 2021). Considering Bayesian decision-theoretic design of experiments under these approaches will be the focal point of future work.

Lastly, we have focused on an experimental aim of parameter estimation. In principle, the extended framework can be generalised to model selection and prediction aims. Exploring this will be a further avenue for future work.

Supplementary Material

Supplementary Material for “Bayesian decision-theoretic design of experiments under an alternative model” (DOI: [10.1214/21-BA1286SUPP](https://doi.org/10.1214/21-BA1286SUPP); .pdf).

References

- Barber, D. and Agakov, F. (2003). “The IM algorithm: a variational approach to information maximization.” *Advances in Neural Information Processing Systems*, 16: 202–208. [6](#)
- Beck, J., Dia, B., Espath, L., and Tempone, R. (2020). “Multilevel double loop Monte Carlo and stochastic collocation methods with importance sampling for Bayesian optimal experimental design.” *Numerical Methods in Engineering*, 121: 3482–3503. [MR4158428](#). doi: <https://doi.org/10.1002/nme.6367>. [2](#)
- Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. Wiley. [MR1274699](#). doi: <https://doi.org/10.1002/9780470316870>. [4, 18](#)
- Bissiri, P., Holmes, C., and Walker, S. (2016). “A general framework for updating belief distributions.” *Journal of the Royal Statistical Society Series B*, 78: 1103–1130. [MR3557191](#). doi: <https://doi.org/10.1111/rssb.12158>. [18](#)
- Chaloner, K. and Verdinelli, K. (1995). “Bayesian experimental design: a review.” *Statistical Science*, 10: 273–304. [MR1390519](#). [1](#)
- Davison, A. (2003). *Statistical Models*. Cambridge University Press. [MR1998913](#). doi: <https://doi.org/10.1017/CB09780511815850>. [11](#)
- Detle, H. and Biedermann, S. (2003). “Robust and efficient designs for the Michaelis-Menten model.” *Journal of the American Statistical Association*, 98: 679–686. [MR2011681](#). doi: <https://doi.org/10.1198/016214503000000585>. [13](#)
- Etzioni, R. and Kadane, J. (1993). “Optimal experimental design for another’s analysis.” *Journal of the American Statistical Association*, 88: 1404–1411. [MR1245377](#). [2](#)
- Federov, V. and Leonov, S. (2014). *Optimal Design for Nonlinear Response Models*. Chapman & Hall. [MR3113978](#). [13](#)
- Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y. W., Rainforth, T., and Goodman, N. (2019). “Variational Bayesian optimal experimental design.” *Advances in Neural Information Processing Systems (NeurIPS, spotlight)*. [2, 6](#)
- Gilmour, S. and Trinca, L. (2012). “Optimum design of experiments for statistical inference (with discussion).” *Journal of the Royal Statistical Society Series C*, 61: 345–401. [MR2914518](#). doi: <https://doi.org/10.1111/j.1467-9876.2011.01000.x>. [2, 5, 11, 12, 13](#)
- Grünwald, P. and van Ommen, T. (2017). “Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it.” *Bayesian Analysis*, 12: 1069–1103. [MR3724979](#). doi: <https://doi.org/10.1214/17-BA1085>. [18](#)

- Jewson, J., Smith, J., and Holmes, C. (2018). “Principles of Bayesian inference using general divergence criteria.” *Entropy*, 20: 442–466. MR3879894. doi: <https://doi.org/10.3390/e20060442>. 18
- Joseph, V. R., Ai, M., and Wu, C. (2009). “Bayesian-inspired minimum aberration two- and four-level designs.” *Biometrika*, 96: 95–106. MR2482137. doi: <https://doi.org/10.1093/biomet/asn062>. 2
- Kennedy, M. and O’Hagan, A. (2001). “Bayesian calibration of computer models (with discussion).” *Journal of the Royal Statistical Society Series B*, 63: 425–464. MR1858398. doi: <https://doi.org/10.1111/1467-9868.00294>. 14
- Kleijn, B. and van der Vaart, A. (2012). “The Bernstein-Von-Mises theorem under misspecification.” *Electronic Journal of Statistics*, 6: 354–381. MR2988412. doi: <https://doi.org/10.1214/12-EJS675>. 8
- Lindley, D. V. (1972). *Bayesian Statistics, A Review*, volume 2. SIAM. MR0329081. 3
- Long, Q., Scavino, M., Tempone, R., and Wang, S. (2013). “Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations.” *Computer Methods in Applied Mechanics and Engineering*, 259: 24–39. MR3063259. doi: <https://doi.org/10.1016/j.cma.2013.02.017>. 2
- O’Hagan, A. and Forster, J. (2004). *Kendall’s Advanced Theory of Statistics Volume 2B Bayesian Inference*. Wiley, 2nd edition. MR3237119. 11
- Overstall, A. M. and Woods, D. C. (2017). “Bayesian design of experiments using approximate coordinate exchange.” *Technometrics*, 59: 458–470. MR3740963. doi: <https://doi.org/10.1080/00401706.2016.1251495>. 16
- Overstall, A. and McGree, J. (2021). “Supplementary material for: Bayesian Decision-Theoretic Design of Experiments Under an Alternative Model.” *Bayesian Analysis*. doi: <https://doi.org/10.1214/21-BA1286SUPP>. 2
- Pawitan, Y. (2003). *In All Likelihood: Statistical Modelling and Inference using Likelihood*. Oxford University Press. 9
- Plumlee, M. (2017). “Bayesian calibration of inexact computer models.” *Journal of the American Statistical Association*, 112: 1274–1285. MR3735376. doi: <https://doi.org/10.1080/01621459.2016.1211016>. 18
- Pronzato, L. and Pázman, A. (2013). *Design of Experiments in Nonlinear Models*. Springer. MR3058804. doi: <https://doi.org/10.1007/978-1-4614-6363-4>. 13
- Rainforth, T., Cornish, R., Yang, H., Warrington, A., and Wood, F. (2018). “On Nesting Monte Carlo Estimators.” *Proceedings of the 35th International Conference on Machine Learning*. 15, 16
- Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press. MR2514435. 14
- Ryan, C., Drovandi, C., and Pettitt, A. (2016a). “Optimal Bayesian experimental design for models with intractable likelihoods using indirect inference applied to biological

- process models.” *Bayesian Analysis*, 11: 857–883. MR3543911. doi: <https://doi.org/10.1214/15-BA977>. 2, 5, 6
- Ryan, E., Drovandi, C., McGree, J., and Pettitt, A. (2016b). “A review of modern computational algorithms for Bayesian optimal design.” *International Statistical Review*, 84: 128–154. MR3491282. doi: <https://doi.org/10.1111/insr.12107>. 1, 2, 4
- Smith, A. and Spiegelhalter, D. (1980). “Bayes factors and choice criteria for linear models.” *Journal of the Royal Statistical Society Series B*, 42: 213–220. MR0583359. 12
- White, H. (1982). “Maximum likelihood estimation of misspecified models.” *Econometrica*, 50: 1–25. MR0640163. doi: <https://doi.org/10.2307/1912526>. 8
- Xie, F. and Xu, Y. (2021). “Bayesian projected calibration of computer models.” *Journal of the American Statistical Association*, To appear. 18

Acknowledgments

The authors would like to thank the associate editor and two anonymous reviewers for providing comments which led to substantial improvement in the paper. They would also like to thank members of the Design Study Group at the University of Southampton for initial discussions and feedback.