Noname manuscript No.
(will be inserted by the editor)

Approximate Laplace importance sampling for the estimation of expected Shannon information gain in high-dimensional Bayesian design for nonlinear models

Yiolanda Englezou · Timothy W. Waite · David C. Woods

Received: date / Accepted: date

Abstract One of the major challenges in Bayesian optimal design is to approximate the expected utility function in an accurate and computationally efficient manner. We focus on Shannon information gain, one of the most widely used utilities when the experimental goal is parameter inference. We compare the performance of various methods for approximating expected Shannon information gain in common nonlinear models from the statistics literature, with a particular emphasis on Laplace Importance Sampling (LIS) and approximate Laplace Importance Sampling (ALIS), a new method that aims to reduce the computational cost of LIS. Specifically, in order to centre the importance distributions LIS requires computation of the posterior mode for each of a large number of simulated possibilities for the response vector. ALIS substantially reduces the amount of numerical optimization that is required, in some cases eliminating all optimization, by centering the importance distributions on the data-generating parameter values wherever possible. Both methods are thoroughly compared with existing approximations including Double Loop Monte Carlo, nested importance sampling, and Laplace approximation. It is found that LIS and ALIS both give an efficient trade-off between mean squared error and computational cost for utility estimation, and ALIS can be up to 70% cheaper than LIS. Usually ALIS gives an approximation that is cheaper but less accurate than LIS, while still being efficient, giving a useful addition to the suite of efficient methods. However, we observed one case where ALIS is both cheaper and more accurate. In addition, for the first time we show that LIS and ALIS yield superior designs to existing methods in problems with large numbers of model parameters when combined with the approximate co-ordinate exchange algorithm for design optimization.

Keywords Optimal design · Monte Carlo · Importance sampling

Yiolanda Englezou
KIOS Research and Innovation Center of Excellence
University of Cyprus
Nicosia
Cyprus

Timothy W. Waite
Department of Mathematics
University of Manchester
Manchester
United Kingdom
E-mail: timothy.waite@manchester.ac.uk

David C. Woods
Statistical Sciences Research Institute
University of Southampton
Southampton
United Kingdom

1 Introduction

When designing experiments for nonlinear models there is usually uncertainty about the model parameters, $\psi \in \Psi$, and often also in the structural form of the model itself. A Bayesian approach enables this uncertainty to be taken into account coherently when choosing the variable settings to be applied in the experiment.

In contrast, frequentist optimal designs such as locally optimal designs (Chernoff 1953) and minimax designs have a less satisfactory approach to a priori parameter uncertainty. Locally optimal designs are tailored for a specific set of assumed parameter values and may perform poorly if the assumed values differ from the truth. Minimax designs optimize worst-case performance, potentially at the expense of reduced efficiency in the most likely parameter scenarios.

Suppose that the design is denoted by $\boldsymbol{\xi} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})^{\mathrm{T}} \in \mathbb{R}^q$ is a vector that defines the settings of the q controllable variables to be applied to the ith experimental unit, with corresponding

response y_i ($i = 1, \ldots, n$). A design $\boldsymbol{\xi}^*$ is *Bayesian optimal* if it maximizes the expected utility,

$$U(\boldsymbol{\xi}) = \int_{\mathbb{R}^n} \int_{\Psi} u(\boldsymbol{\xi}, \boldsymbol{\psi}, \mathbf{y}) f_R(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\xi}) f_B(\boldsymbol{\psi}) d\boldsymbol{\psi} d\mathbf{y}, \quad (1)$$

with respect to $\boldsymbol{\xi} \in \Xi$, where Ξ denotes the set of possible designs. Above $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, with $f_B(\boldsymbol{\psi})$ denoting the prior probability density of the parameters and $f_R(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\xi})$ denoting the conditional probability density of the response vector under the assumed model.

The utility function u is chosen to reflect the goal of the experiment, such as point estimation of $\boldsymbol{\psi}$ or hypothesis testing. We will focus on the case where the goal is to report all knowledge about the parameters via the full posterior distribution, with density $f_A(\boldsymbol{\psi}|\mathbf{y}, \boldsymbol{\xi}) \propto f_R(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\xi}) f_B(\boldsymbol{\psi})$, ensuring that this is as concentrated as possible. Here a commonly recommended utility is

$$\begin{aligned} u(\boldsymbol{\xi}, \boldsymbol{\psi}, \mathbf{y}) &= \log \frac{f_A(\boldsymbol{\psi}|\mathbf{y}, \boldsymbol{\xi})}{f_B(\boldsymbol{\psi})} \\ &= \log f_R(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\xi}) - \log f_E(\mathbf{y}|\boldsymbol{\xi}), \end{aligned} \quad (2)$$

involving the model evidence, defined via $f_E(\mathbf{y}|\boldsymbol{\xi}) = \int_{\Psi} f_R(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\xi}) f_B(\boldsymbol{\psi}) d\boldsymbol{\psi}$. The above is the unique utility corresponding to a local proper scoring rule. A Bayesian optimal design for utility (2) maximizes the expected Kullback-Leibler divergence, or equivalently the expected Shannon information gain (SIG), between the prior and posterior distributions (Lindley et al. 1956, Bernardo 1979, Chaloner & Verdinelli 1995).

Note that the role of the subscripts above is to ensure that different functions have different names, e.g. $f_A(\cdot|\cdot, \cdot)$ is the posterior of $\boldsymbol{\psi}$ and $f_B(\cdot)$ is the prior for $\boldsymbol{\psi}$. This is more precise than the simpler notation more commonly used in Bayesian statistics in which both density functions would be denoted by f and distinguished purely by their arguments; it is also shorter than the more formal probabilistic notation in which the two functions would be denoted $f_{\Theta|Y,\Xi}(\cdot|\cdot, \cdot)$ and $f_{\Theta}(\cdot)$. The more precise notation will be important later, when we wish to substitute other quantities, e.g. one denoted $\hat{\boldsymbol{\mu}}$, into the posterior density of $\boldsymbol{\psi}$. The simpler notation is considered an 'abuse of notation' by mathematicians (e.g. Gelman et al. 2013, p.6), though it is often expedient.

Despite the apparent simplicity of the above theory, until recently it was all but impossible to compute a Bayesian optimal design in practice for realistically complex experiments. This is due to the presence of two main challenges. Firstly, the (potentially high-dimensional) integrals involved in (1) and (2) are analytically intractable except for linear models with normally-distributed response. Thus, in general the expected utility can only be evaluated approximately using numerical integration. Typically the outer integral in (1) is estimated via Monte Carlo. The inner integral in the model evidence in (2) can be estimated stochastically, giving Double Loop Monte Carlo (Ryan 2003) or nested Importance Sampling (Feng 2015). Alternatively, deterministic estimates such as Laplace approximations can be used (Long et al. 2013, Overstall et al. 2018). Earlier approaches such as Bayesian *D*-optimality relied more heavily on asympototic approximations (Chaloner & Verdinelli 1995).

The second challenge is numerical maximization of the approximately evaluated utility. This is difficult as a result of the high dimension of the design space. In addition, due to the use of Monte Carlo, the approximate evaluations of the objective function are computationally expensive, noisy, and non-smooth. This precludes the use of standard optimization algorithms such as quasi-Newton methods or co-ordinate exchange algorithms. Instead, more sophisticated optimization techniques have been developed, one of the most promising being approximate co-ordinate exchange (ACE; Overstall & Woods 2017). Alternative methods include stochastic approximation (Huan & Marzouk 2013) and sampling-based methods (Müller et al. 2004).

The idea of the ACE algorithm is to optimize one co-ordinate of the design at a time using a Gaussian process emulator to form a smooth estimate of the expected utility as a function of the current co-ordinate. To ensure robustness to the quality of the emulator, each proposed change to a co-ordinate is subject to an independent emulator-free acceptance-rejection step. After making several passes through the design matrix using this process, the design points are consolidated using a point exchange procedure. An implementation is available in the R package `acebayes` (Overstall et al. 2019).

This paper makes several contributions. First, we introduce a new method for the approximation of the expected SIG utility, called Approximate Laplace Importance Sampling (ALIS). Our method is computationally cheaper (in some cases up to 70%) than the Laplace Importance Sampling (LIS) method used by Beck et al. (2018) to find low-dimensional designs for partial differential equation models, and by Senarathne et al. (2020) for sequential design. Second, we conduct a thorough comparison of ALIS and LIS with a number of other algorithms in the context of nonlinear models familiar from the statistics literature. Third, we discuss approximations to the expected SIG utility in the common case where there are nuisance parameters (cf. Feng & Marzouk 2019). Finally, we demonstrate that the use of ALIS and LIS in conjunction with the ACE optimization algorithm gives better designs than previous approximations in some models with a large number of parameters.

2 Existing approximations for expected Shannon information gain

All of the methods considered in this paper use Monte Carlo integration to estimate the (outer) integral in (1), giving an approximation of the form

$$\tilde{U}(\boldsymbol{\xi}) = \frac{1}{M_1}\sum_{h=1}^{M_1}\left[\log f_R(\mathbf{y}_h \,|\, \boldsymbol{\psi}_h, \boldsymbol{\xi}) - \log \tilde{f}_E^h\right], \tag{3}$$

where $(\boldsymbol{\psi}_h, \mathbf{y}_h)$, $h = 1, \ldots, M_1$, are independent random samples from the joint prior density, i.e. $f_J(\boldsymbol{\psi}, \mathbf{y} \,|\, \boldsymbol{\xi}) = f_B(\boldsymbol{\psi}) f_R(\mathbf{y} \,|\, \boldsymbol{\psi}, \boldsymbol{\xi})$, and $\tilde{f}_E^h$ is an estimate of the evidence $f_E(\mathbf{y}_h|\boldsymbol{\xi})$ in (2).

The main difference between the various methods is in the choice of the estimate of the evidence in (3), which affects both accuracy and computational expense. The primary distinction is whether a second Monte Carlo estimate is used, giving a 'Nested Monte Carlo' method, or a deterministic estimate such as the Laplace approximation. We detail these different methods below.

2.1 Naïve Monte Carlo

The simplest way to approximate the evidence in (3) is via $\tilde{f}_E^h = \frac{1}{M_2}\sum_{k=1}^{M_2} f_R(\mathbf{y}_h \,|\, \tilde{\boldsymbol{\psi}}_{hk}, \boldsymbol{\xi})$, where the 'inner sample' $\tilde{\boldsymbol{\psi}}_{hk}$, $k = 1, \ldots, M_2$, is another independent random sample from the prior density, $f_B(\boldsymbol{\psi})$. The inner sample is chosen independently of the 'outer sample', $(\boldsymbol{\psi}_h, \mathbf{y}_h)$. This gives an overall approximation

$$\tilde{U}_{\mathrm{nMC}}(\boldsymbol{\xi}) = \frac{1}{M_1}\sum_{h=1}^{M_1}\left[\log f_R(\mathbf{y}_h \,|\, \boldsymbol{\psi}_h, \boldsymbol{\xi}) - \log\left(\frac{1}{M_2}\sum_{k=1}^{M_2} f_R(\mathbf{y}_h \,|\, \tilde{\boldsymbol{\psi}}_{hk}, \boldsymbol{\xi})\right)\right].$$

We refer to the above approximation as *naïve Monte Carlo* (nMC); it is known elsewhere in the literature as Double Loop Monte Carlo (DLMC). The estimator $\tilde{U}_{\mathrm{nMC}}(\boldsymbol{\xi})$ has variance of asymptotic order $O(1/M_1)$ and positive asymptotic bias $C(\boldsymbol{\xi})/M_2$, where

$$C(\boldsymbol{\xi}) = \frac{1}{2}\mathbb{E}\left[\mathrm{Var}\left(\frac{f_R(\mathbf{y} \,|\, \boldsymbol{\psi}, \boldsymbol{\xi})}{f_E(\mathbf{y} \,|\, \boldsymbol{\xi})}\,\middle|\,\mathbf{y}\right)/f_E(\mathbf{y} \,|\, \boldsymbol{\xi})^2\right]$$

(Ryan 2003). Thus, the variance can be reduced by increasing the outer sample size, and the bias can be reduced by increasing the inner sample size.

Despite its good asymptotic properties, for practical inner sample sizes the naïve Monte Carlo estimator commonly suffers from problems with numerical underflow. When this happens one obtains a numerically negligible estimate for the evidence and a numerical estimate of infinity for the expected utility. The latter is clearly unreasonable, making it questionable whether the method can be reliably used to compare designs when M_1 and M_2 are small. This *zero evidence problem* is particularly acute when the posterior is highly concentrated relative to the prior. In this case the likelihood $f_R(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\xi})$ is numerically negligible throughout the majority of the parameter space, except on a very small neighbourhood around the maximum likelihood estimate. It is thus highly likely that all of the $\tilde{\boldsymbol{\psi}}_{hk}$, which are sampled from the prior, will lie outside of this neighbourhood, giving a numerically negligible estimate of the evidence.

2.2 Reuse estimator

To alleviate the numerical stability problems of the Naïve Monte Carlo estimator, Huan & Marzouk (2013) proposed the *reuse* approximation,

$$\tilde{U}_{\mathrm{reuse}}(\boldsymbol{\xi}) = \frac{1}{M_1}\sum_{h=1}^{M_1}\left[\log f_R(\mathbf{y}_h \,|\, \boldsymbol{\psi}_h, \boldsymbol{\xi}) - \log\left(\frac{1}{M_1}\sum_{k=1}^{M_1} f_R(\mathbf{y}_h \,|\, \boldsymbol{\psi}_k, \boldsymbol{\xi})\right)\right],$$

which uses the same parameter sample in both the inner and outer summation. The asymptotic bias of the reuse estimator has the same order of magnitude as that of the naïve Monte Carlo method. However the reuse estimator is more numerically stable for small Monte Carlo sample sizes. In particular, it will usually give a finite estimate of the expected utility gain because each inner sum contains the term $f_R(\mathbf{y}_h \,|\, \boldsymbol{\psi}_h, \boldsymbol{\xi})$, which is non-negligible as $\boldsymbol{\psi}_h$ is the parameter vector used to generate $\mathbf{y}_h$ in the simulation.

2.3 Laplace approximations

The literature contains two methods for using Laplace approximations to avoid nested Monte Carlo integration. For the first method, considered by Overstall et al. (2018) and denoted LA1 here, equation (3) is used with the standard Laplace approximation to the evidence, giving

$$\tilde{U}_{\mathrm{LA1}}(\boldsymbol{\xi}) = \frac{1}{M_1}\sum_{h=1}^{M_1}\left[\log f_R(\mathbf{y}_h \,|\, \boldsymbol{\psi}_h, \boldsymbol{\xi}) - \log \tilde{f}_A(\hat{\boldsymbol{\psi}}_h|\mathbf{y}_h, \boldsymbol{\xi}) - \frac{p}{2}\log 2\pi + \frac{1}{2}\log|\mathbf{H}_h|\right],$$

where $\tilde{f}_A(\boldsymbol{\psi}|\mathbf{y}, \boldsymbol{\xi}) = f_R(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\xi}) f_B(\boldsymbol{\psi}) = f_J(\boldsymbol{\psi}, \mathbf{y}|\boldsymbol{\xi})$ denotes the unnormalized posterior. In addition $\hat{\boldsymbol{\psi}}_h = \arg\max_{\boldsymbol{\psi}} \tilde{f}_A(\boldsymbol{\psi}|\mathbf{y}_h, \boldsymbol{\xi})$ denotes the posterior mode for the hth response realization, while

$$\mathbf{H}_h = -\frac{\partial^2 \log \tilde{f}_A(\boldsymbol{\psi}|\mathbf{y}_h, \boldsymbol{\xi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^{\mathrm{T}}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}_h}$$

denotes the Hessian of the negative log-posterior at the mode. This asymptotic approximation should be accurate provided n is large enough for the posterior to be approximately normal, and can be used to find efficient designs for a wide range of sample sizes.

A second method, denoted LA2 here, was considered by Long et al. (2013). This requires additionally that the sample size is large enough for the posterior to be highly concentrated around the posterior mode. In this case, both the log-posterior and the log-*prior* can be approximated within the region of highest posterior density by a second-order Taylor expansion, giving

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y}}\mathbb{E}_{\boldsymbol{\psi}|\mathbf{y}}\left[\log f_A(\boldsymbol{\psi}|\mathbf{y}) - \log f_B(\boldsymbol{\psi})\right] \\
&\approx \mathbb{E}_{\mathbf{y}}\mathbb{E}_{\boldsymbol{\psi}|\mathbf{y}}\Big[\frac{1}{2}\log|\mathbf{H}_{\mathbf{y}}| - \frac{p}{2}\log 2\pi \\
&\quad - \frac{1}{2}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}}_{\mathbf{y}})^{\mathrm{T}}\mathbf{H}_{\mathbf{y}}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}}_{\mathbf{y}}) - \log f_B(\hat{\boldsymbol{\psi}}_{\mathbf{y}}) \\
&\quad - \nabla_{\boldsymbol{\psi}}\log f_B(\hat{\boldsymbol{\psi}}_{\mathbf{y}})(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}}_{\mathbf{y}}) \\
&\quad - \frac{1}{2}(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}}_{\mathbf{y}})^{\mathrm{T}}\nabla^2_{\boldsymbol{\psi}}\log f_B(\hat{\boldsymbol{\psi}}_{\mathbf{y}})(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}}_{\mathbf{y}})\Big] \\
&\quad \approx \mathbb{E}_{\mathbf{y}}\Big[\frac{1}{2}\log|\mathbf{H}_{\mathbf{y}}| - \frac{p}{2}(\log 2\pi + 1) \\
&\quad - \log f_B(\hat{\boldsymbol{\psi}}_{\mathbf{y}}) - \frac{1}{2}\operatorname{tr}(\nabla^2_{\boldsymbol{\psi}}\log f_B(\hat{\boldsymbol{\psi}}_{\mathbf{y}})\mathbf{H}_{\mathbf{y}}^{-1})\Big],
\end{aligned}$$

where $\hat{\boldsymbol{\psi}}_{\mathbf{y}}$ denotes the posterior mode given response vector $\mathbf{y}$ and $\mathbf{H}_{\mathbf{y}}$ the corresponding Hessian of the negative log-posterior. Above, the last line has been obtained using the elementary fact that if $\boldsymbol{\psi} \sim N(\boldsymbol{\mu}, \mathbf{V})$ then $\mathbb{E}[(\boldsymbol{\psi}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{Q}(\boldsymbol{\psi}-\boldsymbol{\mu})] = \operatorname{tr}\mathbf{Q}\mathbf{V}$. Monte Carlo estimation of the above gives

$$\begin{aligned}
\tilde{U}_{\mathrm{LA2}}(\boldsymbol{\xi}) = \frac{1}{M_1}\sum_{h=1}^{M_1}\Big[&\frac{1}{2}\log|\mathbf{H}_h| - \frac{p}{2}(\log 2\pi + 1) \\
&- \log f_B(\hat{\boldsymbol{\psi}}_h) - \frac{1}{2}\operatorname{tr}(\nabla^2_{\boldsymbol{\psi}}\log f_B(\hat{\boldsymbol{\psi}}_h)\mathbf{H}_h^{-1})\Big].
\end{aligned}$$

3 Approximate Laplace Importance Sampling

3.1 Importance sampling

Another Monte Carlo method for estimating the evidence is importance sampling, i.e.

$$\tilde{f}_E^h = \frac{1}{M_2}\sum_{k=1}^{M_2}\frac{f_R(\mathbf{y}_h\,|\,\tilde{\boldsymbol{\psi}}_{hk},\boldsymbol{\xi})f_B(\tilde{\boldsymbol{\psi}}_{hk})}{q_h(\tilde{\boldsymbol{\psi}}_{hk})}, \tag{4}$$

where $\tilde{\boldsymbol{\psi}}_{hk}$, $k = 1,\ldots,M_2$ is an independent sample from the importance density q_h. Note that nMC corresponds to the special case where the prior is chosen as the importance density. By standard theory (e.g. Lemieux 2009, p.114) the optimal importance density is $q_h^*(\boldsymbol{\psi}) \propto f_R(\mathbf{y}_h\,|\,\boldsymbol{\psi},\boldsymbol{\xi})f_B(\boldsymbol{\psi}_{hk})$, i.e. q_h^* is the posterior density of $\boldsymbol{\psi}$ given $\mathbf{y}_h$. This gives a zero variance unbiased (i.e. error-free) estimator; unfortunately the optimal importance density cannot be used in practice as it requires knowledge of the evidence, the quantity we are trying to estimate.

The above discussion suggests that a good choice of importance density would be a computationally cheap approximation to the posterior, such as $N(\boldsymbol{\psi};\hat{\boldsymbol{\mu}}_h,\hat{\boldsymbol{\Sigma}}_h)$ or $t_\nu(\boldsymbol{\psi};\hat{\boldsymbol{\mu}}_h,\hat{\boldsymbol{\Sigma}}_h)$, where $\hat{\boldsymbol{\mu}}_h$ and $\hat{\boldsymbol{\Sigma}}_h$ are approximations to the mean vector and variance matrix of $f_A(\boldsymbol{\psi}|\mathbf{y}_h,\boldsymbol{\xi})$. Here $N(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ and $t_\nu(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ denote respectively the probability density function of a multivariate normal and multivariate t distribution with mean $\boldsymbol{\mu}$, variance matrix $\boldsymbol{\Sigma}$, and degrees of freedom ν for the latter. Below we discuss different methods for choosing $\hat{\boldsymbol{\mu}}_h$ and $\hat{\boldsymbol{\Sigma}}_h$.

3.2 Laplace-type Importance Sampling Methods

Laplace-type importance sampling methods set the variance of the importance distribution as $\hat{\boldsymbol{\Sigma}}_h = \mathbf{H}_h(\hat{\boldsymbol{\mu}}_h)^{-1}$ where $\mathbf{H}_h(\hat{\boldsymbol{\mu}}_h) = -\frac{\partial^2\log\tilde{f}_A(\boldsymbol{\psi}\,|\,\mathbf{y}_h,\boldsymbol{\xi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathrm{T}}}\big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\mu}}_h}$.The two variants, Laplace Importance Sampling (LIS), and Approximate Laplace Importance Sampling (ALIS), are distinguished via the choice of mean $\hat{\boldsymbol{\mu}}_h$.

3.2.1 LIS

With LIS, the mean is approximated using $\hat{\boldsymbol{\mu}}_h = \hat{\boldsymbol{\psi}}_h = \arg\max_{\boldsymbol{\psi}\in\Psi}\tilde{f}_A(\boldsymbol{\psi}\,|\,\mathbf{y}_h,\boldsymbol{\xi})$. This necessitates a total of M_1 potential costly numerical optimizations to find the mode, $\hat{\boldsymbol{\psi}}_h$ ($h = 1,\ldots,M_1$), of the posterior distribution: one for each the M_1 simulated response vectors, $\mathbf{y}_1,\ldots,\mathbf{y}_{M_1}$, in the outer sample. However, provided the search for $\hat{\boldsymbol{\psi}}_h$ is initialized at the data-generating parameter values $\boldsymbol{\psi}_h$, it typically converges in a small number of iterations. We performed these optimizations using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. Algorithm 1 gives a detailed description of the LIS method and our proposed new variant, ALIS. See Ryan et al. (2015) and Beck et al. (2018) for low-dimensional examples of design selection with LIS.

3.2.2 ALIS

The key observation underpinning ALIS is that the posterior mode used to center the importance distribution in LIS is frequently close to the data-generating values, i.e. it is often the case that $\hat{\boldsymbol{\psi}}_h \approx \boldsymbol{\psi}_h$. Given this, an obvious question is whether it is possible to reduce the computational cost of the LIS method by removing some of the optimization steps, setting $\hat{\boldsymbol{\mu}}_h = \boldsymbol{\psi}_h$ for some h. For this choice to work we require at a minimum that $\mathbf{H}_h(\boldsymbol{\psi}_h)$ is positive definite, since without this $\hat{\boldsymbol{\Sigma}}_h$ would not be a valid covariance matrix and it would be impossible to sample from the importance distribution q_h. The ALIS

Algorithm 1: ALIS/LIS Algorithm

Generate a sample ψ_h, $h = 1,\dots,M_1$, from $f_B(\psi)$;
for $h = 1,\dots,M_1$ **do**
  Generate a response $\mathbf{y}_h$ from $f_R(\mathbf{y}|\psi_h,\boldsymbol{\xi})$;
  Compute mean and variance of importance distribution $q_h(\psi)$:
  if *method=='ALIS' and* $\mathbf{H}_h(\psi_h)$ *is positive-definite* **then**
    Set $\hat{\boldsymbol{\mu}}_h = \psi_h$ and $\hat{\boldsymbol{\Sigma}}_h = \mathbf{H}_h(\psi_h)^{-1}$
  else
    Calculate the posterior mode $\hat{\psi}_h$ of $f_J(\psi,\mathbf{y}_h|\boldsymbol{\xi}) = f_R(\mathbf{y}_h|\psi,\boldsymbol{\xi})f_B(\psi)$, e.g. via BFGS
    Set $\hat{\boldsymbol{\mu}}_h = \hat{\psi}_h$ and $\hat{\boldsymbol{\Sigma}}_h = \mathbf{H}_h(\hat{\boldsymbol{\mu}}_h)^{-1}$
  Generate a sample $\{\tilde{\psi}_{hk}\}_{k=1}^{M_2}$, from the importance density $q_h(\psi)$;
  for $k = 1,\dots,M_2$ **do**
    Calculate $\tilde{u}_{hk} = \frac{f_R(\mathbf{y}_h|\tilde{\psi}_{hk},\boldsymbol{\xi})f_B(\tilde{\psi}_{hk})}{q_h(\tilde{\psi}_{hk})}$;
  Estimate the evidence $f_E(\mathbf{y}_h|\boldsymbol{\xi})$ via $\tilde{f}_E^h = \frac{1}{M_2}\sum_{k=1}^{M_2}\tilde{u}_{hk}$;
  Calculate $\tilde{u}_h = \log f_R(\mathbf{y}_h|\psi_h,\boldsymbol{\xi}) - \log \tilde{f}_E^h$;
Estimate the expected Shannon information gain utility via $\tilde{U}(\boldsymbol{\xi}) = \frac{1}{M_1}\sum_{h=1}^{M_1}\tilde{u}_h$;

importance distribution is thus centred at

$$\hat{\boldsymbol{\mu}}_h = \begin{cases} \psi_h & \text{if } \mathbf{H}_h(\hat{\boldsymbol{\mu}}_h) \text{ is numerically} \\ & \text{positive-definite}, \\ \hat{\psi}_h & \text{otherwise.} \end{cases}$$

We show in Section 4.2 that this choice gives a method with comparable accuracy but lower computational cost.

3.3 Nested importance sampling

In nested importance sampling (nIS; Feng 2015), the posterior mean $\boldsymbol{\mu}_h$ and variance $\boldsymbol{\Sigma}_h$ are approximated via self-normalized importance sampling using the outer sample. This gives

$$\hat{\boldsymbol{\mu}}_h = \sum_{k=1}^{M_1} \bar{w}_{hk}\psi_k\,,$$
$$\hat{\boldsymbol{\Sigma}}_h = \sum_{k=1}^{M_1} \bar{w}_{hk}(\psi_h - \hat{\boldsymbol{\mu}}_h)(\psi_h - \hat{\boldsymbol{\mu}}_h)^{\mathrm{T}}\,,$$

where $\bar{w}_{hk} = f_R(\mathbf{y}_h\,|\,\psi_k,\boldsymbol{\xi})/\sum_{l=1}^{M_1} f_R(\mathbf{y}_h\,|\,\psi_l,\boldsymbol{\xi})$. This approach has the potential to suffer from low effective sample size. To counter this, Feng proposed to revert to the original naïve Monte Carlo estimate of the evidence if $\mathrm{ESS}_h = 1/(\sum_{k=1}^{M_1}\bar{w}_{hk}^2)$ drops below a prespecified minimum effective sample size.

4 Performance comparison

4.1 Models for performance assessment

In this section we compare the performance of the methods from Sections 2 and 3. Results are given for two tasks: (i) evaluation of the utility function and (ii) design selection, in Sections 4.2 and 4.3 respectively. Three different nonlinear models are considered, all of the form

$$y_i = \eta(\mathbf{x}_i,\theta) + \varepsilon_i\,, \tag{5}$$

with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma_\varepsilon^2)$ for $i = 1,\dots,n$. The models differ with respect to their mean functions η, parameters, and priors, with details given below.

Michaelis-Menten model

The Michaelis-Menten model has mean function

$$\eta(x,\theta) = \theta_1 x/(\theta_2 + x)\,,$$

with unknown parameters θ_1, θ_2, both positive. It is assumed a priori that $\log\theta_1 \sim N(4.38, 0.07^2)$, $\log\theta_2 \sim N(1.19, 0.84^2)$, and $\sigma_\varepsilon^2 \sim$ Inverse-Gamma$(3,2)$ independently.

The prior on θ_2 is relatively diffuse, implying a wide range of possible shapes of the response curve. The prior on σ_ε^2 was chosen to imply a low noise-to-signal ratio as this leads to a relatively concentrated posterior, the most demanding scenario for methods such as naïve Monte Carlo and nested importance sampling. Specifically, the 10% and 90% quantiles of $\sigma_\varepsilon/\eta(400,\theta)$ are 0.009 and 0.02 respectively, where the denominator is the maximum value of η over the design region $[0,400]$.

In order to preserve the positivity constraint on θ_1 and θ_2, we reparameterize the model in terms of $\vartheta_1 = \log\theta_1$ and $\vartheta_2 = \log\theta_2$. The resulting normal ALIS/LIS importance distribution on the ϑ scale effectively implies a log-normal importance distribution on the θ scale. Note that reparameterization does not change the value of the expected Shannon information gain, but it does change our numerical estimate thereof due to the modified importance distributions. Here the noise variance σ_ε^2 can be integrated out analytically owing to its conjugate prior. See the appendix for full technical details of the calculations for the marginal likelihood and its derivatives.

Parameter	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
Mean	1054.54	206.55	1.46	−0.26	0.02	0.40	0.04	57.40	−0.48	−1.50
Std. dev	24.63	5.29	0.04	0.01	0.002	0.03	0.001	2.37	0.075	0.10

Table 1 Prior means and standard deviations for the lubricant kinematic viscosity model

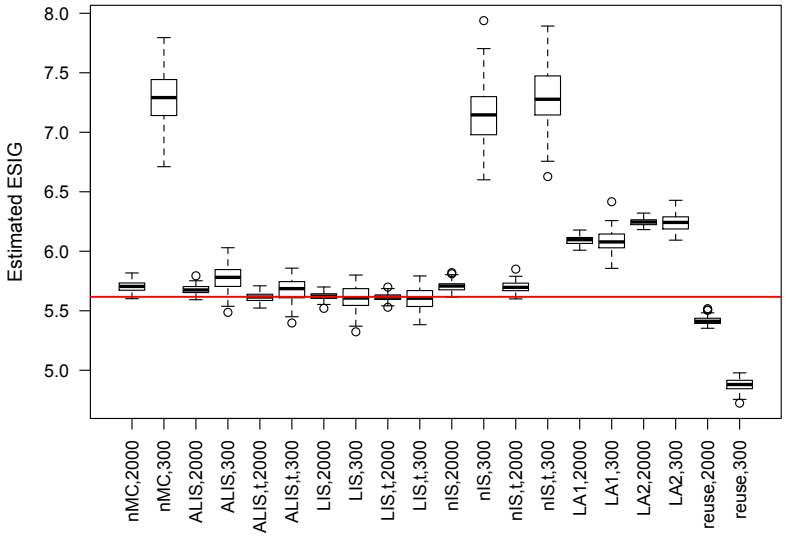

Fig. 1 Distribution of the estimator of expected utility from different methods, using the Michaelis-Menten model and a space-filling design with $n = 5$ runs. Empirical distributions of the estimator are based on 100 evaluations for each method. The 'true' expected Shannon information gain given by the reference approximation is indicated by the red line. Numbers after the method name indicate the outer Monte Carlo sample size M_1. The pairs $(M_1, M_2) = (300, 300)$ and $(2000, 10000)$ were used.

	BOD			Michaelis-Menten			Average
Method/M_1	$n = 6$	$n = 10$	$n = 20$	$n = 5$	$n = 10$	$n = 20$	RRMSE
LIS,2000	1.2	0.6	0.4	0.6	0.5	0.5	0.6
ALIS,2000	1.0	0.6	0.5	1.2	0.5	0.6	0.7
LIS,300	2.3	1.5	1.0	1.7	1.1	1.2	1.5
ALIS,300	2.1	1.7	0.9	3.3	1.5	1.5	1.8
nMC,2000	2.1	2.6	5.7	1.7	3.8	6.4	3.7
nIS,2000	1.4	2.4	7.5	1.8	4.7	8.5	4.4
LA1,2000	7.9	4.1	1.6	8.5	3.7	3.5	4.9
LA1,300	8.3	4.5	1.8	8.5	4.0	3.7	5.1
reuse,2000	3.1	4.8	9.4	3.6	7.3	9.2	6.2
LA2,2000	15.2	7.8	2.8	11.2	5.3	4.4	7.8
LA2,300	15.5	8.1	2.9	11.2	5.3	4.5	7.9
reuse,300	13.8	17.7	31.4	13.3	22.7	25.9	20.8
nIS,300	22.7	39.3	99.1	27.7	71.2	109.1	61.5
nMC,300	25.5	39.9	99.1	30.0	72.0	110.5	62.8

Table 2 Percentage RRMSE of the estimator of expected Shannon information gain, for different combinations of model, sample size, and approximation method. Methods are ordered according to the average RRMSE across all examples.

	BOD			Michaelis-Menten		
Example	$n = 6$	$n = 10$	$n = 20$	$n = 5$	$n = 10$	$n = 20$
% optimizations avoided	75.9	91.8	99.3	98.0	99.7	100.0

Table 3 ALIS method: long-run percentage of outer loop iterations for which the data-generating values of ψ are used to centre the importance distribution, i.e. the percentage of outer loop iterations for which numerical optimization to find the posterior mode is avoided. Percentages were estimated by simulating 10,000 outer loop iterations. For smaller values of M_1 the actual number of outer loop iterations for which numerical optimization is avoided will follow a Binomial distribution with success probabilities approximately equal to the above percentages.

			Detailed timings (s)					
	Mean performance		BOD			Michaelis-Menten		
Method	rRMSE (%)	Time (s)	$n=6$	$n=10$	$n=20$	$n=5$	$n=10$	$n=20$
LA1,300	5.1	0.007	0.002	0.003	0.007	0.005	0.005	0.019
LA2,300	7.9	0.009	0.005	0.005	0.012	0.006	0.005	0.019
LA1,2000	4.9	0.050	0.021	0.022	0.049	0.038	0.038	0.130
LA2,2000	7.8	0.059	0.035	0.036	0.080	0.040	0.038	0.127
reuse,300	20.8	0.062	0.040	0.042	0.136	0.023	0.026	0.103
nMC,300	62.8	0.062	0.040	0.042	0.135	0.028	0.025	0.103
ALIS,300	1.8	0.116	0.089	0.092	0.144	0.103	0.092	0.178
LIS,300	1.5	0.122	0.092	0.094	0.152	0.099	0.102	0.192
nIS,300	61.5	0.137	0.090	0.077	0.201	0.103	0.062	0.290
reuse,2000	6.2	2.687	1.671	1.825	5.949	1.109	1.136	4.434
nMC,2000	3.7	13.790	8.425	9.278	30.423	5.488	5.656	23.470
ALIS,2000	0.7	25.295	19.375	20.490	31.864	20.200	20.855	38.988
LIS,2000	0.6	25.632	19.805	20.878	32.563	20.463	21.143	38.939
nIS,2000	4.4	33.159	24.298	24.801	44.066	24.868	23.232	57.686

Table 4 Computational expense and accuracy of different methods. The left part of the table shows, for each method, the mean rRMSE and mean time to produce one evaluation of the utility function. The mean is an average across ten repeats of all examples for the Michaelis-Menten and BOD models. Methods are sorted from least expensive to most expensive. The right part of the table shows a detailed breakdown of the mean evaluation time of the utility function for each example, averaged across 10 repeats.

Model	n	M_1	M_2	Time (LIS, s)	Reduction (ALIS, %)	rRMSE (LIS)	Increase (ALIS)
BOD	6	2000	50	0.110	10.0	2.15	-0.06
BOD	10	2000	50	0.113	11.4	0.91	0.28
BOD	6	300	30	0.011	14.1	2.69	0.07
BOD	10	300	30	0.011	16.8	1.69	0.22
BOD	20	2000	50	0.189	19.0	0.41	0.18
MM	5	2000	50	0.128	19.3	0.65	0.19
MM	10	2000	50	0.128	22.7	0.49	0.07
BOD	20	300	30	0.019	24.3	1.05	0.18
BOD	6	300	10	0.005	26.2	3.27	0.41
MM	5	300	30	0.014	30.5	1.69	0.13
BOD	10	300	10	0.005	32.5	1.79	1.07
MM	10	300	30	0.014	33.5	1.18	0.17
MM	20	2000	50	0.287	33.9	0.46	0.20
BOD	20	300	10	0.010	45.3	1.07	1.00
MM	20	300	30	0.032	45.6	1.19	0.17
MM	5	300	10	0.008	53.8	1.67	0.99
MM	10	300	10	0.008	54.8	1.22	0.90
MM	20	300	10	0.022	69.5	1.17	1.24

Table 5 Comparison between ALIS and LIS for smaller inner loop sample sizes. Cost reduction is shown as a percentage, while rRMSE increase is shown as the absolute increase (which is a difference in percentage points). The table is ordered according to the magnitude of the cost reduction from ALIS.

Biochemical oxygen demand (BOD) model

Bates & Watts (1988, Chapter 2) modelled biochemical oxygen demand y (mg/L) with the mean function

$$\eta(x,\theta) = \theta_1\{1-\exp(-\theta_2 x)\},$$

where x is time (in days). We adopt the following independent priors:

$$\begin{aligned}\log\theta_1 &\sim N(3.38, 0.20^2),\\ \log\theta_2 &\sim N(1.098, 1.12^2),\\ \pi_b(\sigma_\varepsilon) &\propto \sigma_\varepsilon^{-1}.\end{aligned}$$

The prior means for θ_1, θ_2 were chosen to match the means given by DiCiccio et al. (1997), while the variances were chosen to illustrate the differences between the methods (smaller and larger variances resulted in more similar performance). Similar to the Michaelis-Menten model, we reparameterize in terms of $\vartheta_j = \log\theta_j$ when carrying out utility approximations, and σ_ε^2 is integrated out analytically. The design region is $[0,7]$.

Lubricant kinematic viscosity model

Bates & Watts (1988, Chapter 3) modelled the kinematic viscosity of a lubricant using the following mean function, depending on temperature, x_1 (°C) and pressure, x_2 (atm):

$$\begin{aligned}\eta(\mathbf{x},\theta) = \frac{\theta_1}{\theta_2+x_1} + \theta_3 x_2 + \theta_4 x_2^2 + \theta_5 x_2^3\\ + (\theta_6+\theta_7 x_2^2)x_2 \exp\left\{-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right\}.\end{aligned}$$

Defining $\theta_{10} = \log\sigma_\varepsilon$, we adopt independent normal priors on θ_j, $j = 1,\ldots,10$, with means and standard

deviations equal to the maximum likelihood estimates and their standard errors based on the data from Bates & Watts (1988) (see Table 1). Unlike the previous models, no reparameterization is used for the θ_j. Moreover the noise variance is treated as an interest parameter. The design region for (x_1, x_2) is $[0,100] \times [0,7]$.

4.2 Utility evaluation results

For utility evaluation, we compare the methods in terms of accuracy and computational expense. To assess accuracy, we need an approximation with negligible error to serve as a reference. For the Michaelis-Menten and BOD models we were able to obtain such a reference approximation by using naïve Monte Carlo with $M_1 = M_2 = 10^6$, though this approximation is too computationally expensive for routine use. However for the lubricant model, whose parameter space is of substantially larger dimension, nMC yields unstable estimates even with such a large Monte Carlo sample size. Thus we restrict our attention to the Michaelis-Menten and BOD models in this section, though our results will suggest that expected utility can be reliably estimated for the lubricant model using the LIS and ALIS approximations.

To investigate how performance depends on the number of experimental runs, results are obtained for a variety of experiment sizes. For the Michaelis-Menten model space-filling designs with $n = 5$, 10, and 20 are considered, while for the BOD model the design from Bates & Watts (1988) with $n = 6$ is considered alongside space-filling designs with $n = 10$ and 20. The type of space-filling design used throughout is a random Latin Hypercube design. For consistency the same specific design realisation was used throughout, so that differences between different sampled designs of the same size are not a factor in the results.

Figure 1 shows how the distribution of the estimator of expected Shannon information varies across different methods and different combinations of inner and outer Monte Carlo sample sizes. The results shown are for the Michaelis-Menten model and a space-filling design with $n = 5$ runs. The reference value of the expected utility is indicated by the red horizontal line. It is seen that ALIS and LIS have small bias and variance compared to all other methods with similar Monte Carlo sample size, even for small M_1 and M_2. The nMC, nIS, and reuse methods give highly biased and variable estimators for small Monte Carlo sample size, but increasing M_1 and M_2 reduces both the variance and bias, and with $(M_1, M_2) = (2000, 10000)$ both the nMC and nIS methods give comparable utility values to LIS and ALIS. In contrast, for the Laplace approximations, increasing the Monte Carlo sample size only reduces the variance, not the bias, as these methods are intrinsically biased due to the poor quality of the asymptotic approximation when $n = 5$. This figure is quite representative of the general picture, but further insight can be obtained by combining results across several examples.

Table 2 shows the accuracy of the methods across models, Monte Carlo sample sizes, and numbers of experimental runs. Accuracy is measured by the percentage relative root mean squared error (RRMSE), i.e. $100 \times \sqrt{\text{MSE}[\tilde{U}(\boldsymbol{\xi})]}/U(\boldsymbol{\xi})$. It is seen that the most accurate methods overall are LIS and ALIS; these have excellent performance even with low Monte Carlo sample size ($M_1 = M_2 = 300$). Moreover, the accuracy of LIS and ALIS remains stable or even improves as the number of experimental runs increases. nMC is the next most accurate method when Monte Carlo sample size is large but it performs poorly when Monte Carlo sample size is small, i.e. when $(M_1, M_2) = (2000, 10000)$ and $(300, 300)$ respectively. Moreover, the performance of nMC degrades as the number of experimental runs increases. This result is intuitive: as the number of experimental runs increases, the posterior will become more concentrated and the prior will become a worse importance distribution. nIS has similar performance and caveats to nMC. The accuracy of LA1 is good when n is large, poor when n is small, and fairly insensitive to M_1. Similar comments apply to LA2, but with slightly worse accuracy overall. The reuse estimator has relatively poor performance even with large Monte Carlo sample sizes and is not recommended.

The left part of Table 4 shows the relationship between the accuracy of the different methods and their computational cost. The timings show that the most efficient methods are LA1, ALIS, and LIS: all other methods have worse accuracy than another method with lower computational cost. In particular LIS and ALIS with $M_1 = M_2 = 300$ give a good trade-off between accuracy and computational expense. Increasing the Monte Carlo sample size to $(M_1, M_2) = (2000, 10000)$ gives only a small increase in accuracy for a very large increase in cost. The right part of Table 4 shows how the utility evaluation time varies across methods, models and experiment sizes. It is clear that larger n results in increased evaluation time, though the relative timings of the different methods are similar for all examples.

The difference between LIS and ALIS can be considered in more detail. Table 3 shows that a high percentage of the numerical optimizations required in the LIS method can be avoided through ALIS. The percentage is higher for large n, which is intuitive since we would expect in that case that the posterior mode would be closer to the data-generating values. The percentage of avoided optimizations is also higher for the Michaelis-Menten examples than for those using the BOD model. This is consistent with the fact that the priors for the Michaelis-Menten example were chosen to have a low noise-to-signal ratio, which is anticipated

to give a posterior that is relatively highly concentrated around the data-generating values.

Although ALIS greatly reduces the amount of optimization required compared to LIS, the computational cost saving in Table 4 is modest: approximately 5% for $(M_1, M_2) = (300, 300)$ and 1.4% for $(M_1, M_2) = (2000, 10000)$. This is due to the values of M_2, which are large enough that the optimization cost is relatively small compared to the cost of the inner loop sampling and averaging. However, for smaller inner loop sample sizes the cost savings due to ALIS are much larger; Table 5 shows cost savings of 10–70% from ALIS compared to LIS when M_2 ranges from 10 to 50. The computational cost saving for ALIS usually comes at the expense of a small decrease in accuracy, though in one case (BOD, $n = 6$, $M_1 = 2000$, $M_2 = 50$) ALIS is both cheaper and more accurate than LIS.

The smaller values of M_2 in Table 5 are more than an intellectual curiosity; there is empirical and theoretical evidence that smaller values may sometimes be a more efficient choice than setting $M_1 = M_2$. E.g. for naïve Monte Carlo asymptotic results suggest it is optimal to take $M_2 = O(\sqrt{M_1})$ (Beck et al. 2018). Empirically, taking as an example the BOD model with $n = 20$, we find that ALIS and LIS with $(M_1, M_2) = (300, 30)$ are both cheaper and more accurate than the Laplace approximation with $M_1 = 2000$.

Clearly such timings will depend on the implementation language and hardware involved, but they nonetheless give a useful idea of the relative cost of the different methods. We used C++ in R via the Rcpp and RcppArmadillo libraries (Eddelbuettel et al. 2011, Eddelbuettel & Sanderson 2014) to obtain high-performance code. Timings were carried out on a 2018 Mac Mini with a 3GHz 6-core Intel i5 processor and 8GB RAM; the calculations took place on a single core.

A common technique to gain better estimates in importance sampling is to inflate the tails of the importance distribution, e.g. by using a t-distribution. We obtained results for t importance distributions, but for brevity the results are omitted here as there was not a substantial difference in performance from the multivariate normal importance distributions. For full details see the first author's PhD thesis (Englezou 2018).

4.3 Design optimization results

In this section we compare the performance of the different expected utility approximation methods for the purpose of design optimization. To enable the comparison we found (near-)optimal designs for each of the different methods discussed in Sections 2 and 3. This was done for the Michaelis-Menten, BOD, and lubricant models discussed in Section 4.1 using the ACE algorithm to perform utility optimization. The experiment sizes considered were as follows: $n = 5$, 10, and 20 for the Michaelis-Menten; $n = 6$, 10, and 20 for BOD; and $n = 20$ and $n = 53$ for the lubricant model. The cases $n = 6$ for BOD and $n = 53$ for the lubricant model correspond to designs in the literature.

Different runs of the ACE algorithm may result in multiple different near-optimal designs being found for the same design problem. This can arise due to different starting designs being used and also the stochastic nature of the expected utility estimates. To obtain more stable results we therefore ran the ACE algorithm 10 times with a different random starting design for each problem, i.e. each combination of approximation method, model, and design size. The best design resulting from these random starts, judged via an independent estimate of the expected utility, was chosen as our estimate of the overall (near-)optimal design for that problem.

The designs obtained from each method were compared using an independent estimate of the expected utility calculated using ALIS with $M_1 = M_2 = 300$. This calculation was repeated 100 times for each design to form an empirical distribution for the estimator, thereby giving an indication of the variability of the expected utility estimate. Comparisons with a 'naïve' (i.e. non-optimal) design were also included for each example. This naïve design was taken from the literature where available, that is when $n = 6$ for BOD and $n = 53$ for the lubricant model. For other cases a space-filling design was used as the naïve design, namely a random Latin Hypercube design. Figures 2 and 3 show typical figures resulting from this process for the BOD and lubricant models with $n = 20$. Similar figures for other examples are given in the first author's PhD thesis (Englezou 2018).

Within ACE, separate Monte Carlo sample sizes must be specified for the emulator-building and accept-reject steps. These were chosen as follows. In the accept-reject step $B = M_1 = M_2 = 10000$ was used throughout. For the emulator-building step, in the Michaelis-Menten and BOD models we used $M_1 = 2000$ for the 'single-loop' methods LA1 and LA2; $M_1 = M_2 = 2000$ was used for all other 'double loop' methods. For the lubricant model, we used $M_1 = M_2 = 300$ for LIS and ALIS, $M_1 = 300$ for LA1 and LA2, and a larger sample size of $M_1 = 2000$, $M_2 = 10000$ for nMC and nIS. The latter was needed to avoid failure of the evaluation due to the zero evidence problem discussed in Section 2.1.

Combining results from across the different examples several observations can be made. First, as anticipated, the optimized designs are better than naïve comparator designs in all but one case. The single exception is that, as seen in Figure 2, the design from the reuse estimator is worse than a space-filling design for the BOD model when $n = 20$. Second, for the two-parameter models, in most cases designs from the different approximations have similar expected utility, aside from the

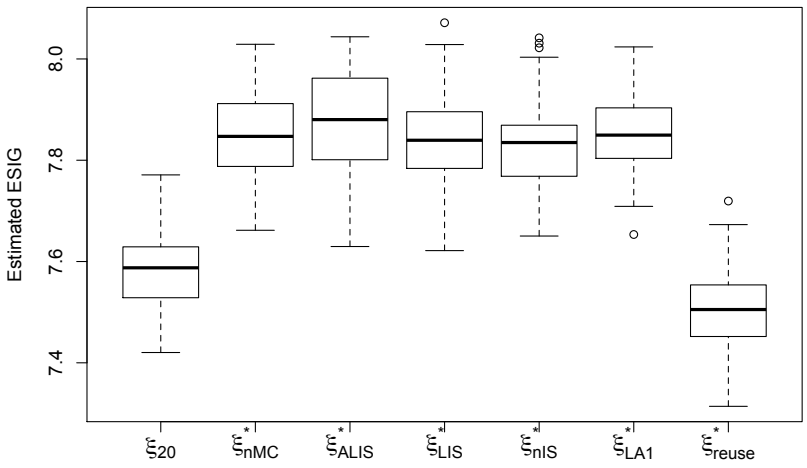

Fig. 2 Comparison of (near-)optimal designs found from the different utility approximation methods for the BOD model with $n = 20$. Each boxplot corresponds to the best design found from 10 random starts of the ACE algorithm using a particular method, and shows the distribution of 100 independent evaluations of the ALIS estimator of expected Shannon information gain, obtained with $M_1 = M_2 = 300$. ξ_{20} refers to a 20-run space filling design.

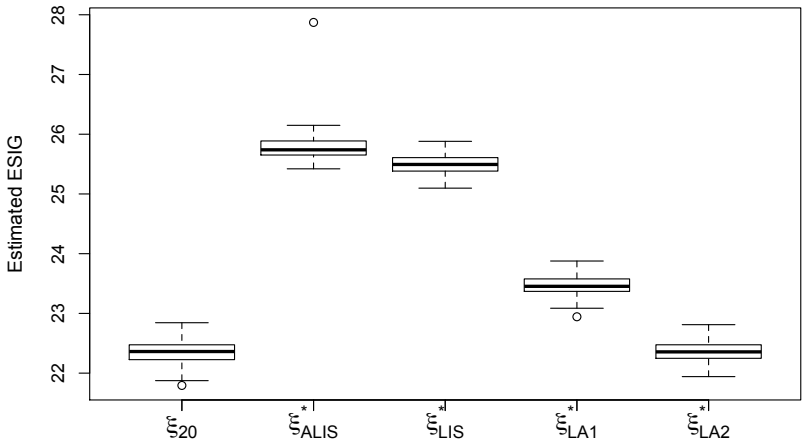

Fig. 3 Comparison of (near-)optimal designs found from the different utility approximation methods for the lubricant model with $n = 20$. Each boxplot corresponds to the best design found from 10 random starts of the ACE algorithm using a particular method, and shows the distribution of 100 independent evaluations of the ALIS estimator of expected Shannon information gain, obtained with $M_1 = M_2 = 300$. ξ_{20} refers to a 20-run space filling design.

reuse and LA2 designs which appear somewhat worse for $n = 10$ and 20. Aside from these special cases for two-parameter models the utility differences between the designs from different methods are usually smaller than the variability of the utility estimator for a fixed design.

Bigger differences in the performance of the designs from different methods are seen for the lubricant model, which is of substantially higher dimension. In particular, the ALIS and LIS designs substantially outperform the designs from all other methods when $n = 20$ (see Figure 3), and all methods except LA1 when $n = 53$. This improved performance for LA1 for large experiment sizes is expected due to the asymptotic nature of the Laplace approximation.

We did not record computational times for finding (near-)optimal designs. However, ACE is usually performed with a fixed number of iterations, and the dominant computational cost is that of the expected utility evaluations. Thus the relative cost for the different

utility approximation methods will be similar to Section 4.2.

5 Nuisance parameters

In this section we discuss the case where the model contains nuisance parameters, meaning parameters that are not of direct interest but which must nonetheless be considered when making inference about the parameters of interest. Laplace approximations for this case have been developed by Overstall et al. (2018), and a Layered Multiple Importance Sampling approximation has been developed by Feng & Marzouk (2019), who refer to the resulting optimal designs as 'focused'. Both of these approaches used the idea of conditioning a multivariate normal approximation to the posterior. Similar ideas can be applied in the ALIS/LIS context, as follows.

First we partition the overall parameter vector as $\psi = (\theta^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}}$, where $\theta \in \Theta \subseteq \mathbb{R}^{p_\theta}$ is the vector of interest parameters, and $\gamma \in \Gamma \subseteq \mathbb{R}^{p_\gamma}$ is the vector of nuisance parameters. The expected Shannon information gain for the interest parameters now takes the form

$$U(\boldsymbol{\xi}) = \int_{\Theta}\int_{\mathbb{R}^n} \log \frac{f_M(\mathbf{y}|\theta,\boldsymbol{\xi})}{f_E(\mathbf{y}|\boldsymbol{\xi})} f(\mathbf{y},\theta) d\mathbf{y} d\theta\,, \tag{6}$$

where $f_M(\mathbf{y}|\theta,\boldsymbol{\xi}) = \int_\Gamma f_R(\mathbf{y}|\theta,\gamma,\boldsymbol{\xi}) f_B(\gamma|\theta) d\gamma$ denotes the marginal density of the response after integrating out the nuisance parameters. The expected utility (6) can be estimated via

$$\tilde{U}_{\text{(A)LIS}}(\boldsymbol{\xi}) = \frac{1}{M_1}\sum_{h=1}^{M_1} \log \frac{\tilde{f}_M^h}{\tilde{f}_E^h}\,,$$

where as before $\tilde{f}_E^h = \frac{1}{M_2}\sum_{k=1}^{M_2} f_R(\mathbf{y}_h|\tilde{\psi}_{hk},\boldsymbol{\xi}) \frac{f_B(\tilde{\psi}_{hk})}{q_h(\tilde{\psi}_{hk})}$ is a LIS/ALIS estimate of the evidence. In addition, now we also require a second importance sampling approximation, $\tilde{f}_M^h$, to estimate the marginal likelihood, $f_M(\mathbf{y}|\theta,\boldsymbol{\xi})$, of the interest parameters after integrating out the nuisance parameters.

In particular we suggest using the following approximation for the marginal likelihood:

$$\tilde{f}_M^h = \frac{1}{M_3}\sum_{s=1}^{M_3} f_R(\mathbf{y}_h|\theta_h,\tilde{\tilde{\gamma}}_{hs},\boldsymbol{\xi}) \frac{f_B(\tilde{\tilde{\gamma}}_{hs}|\theta_h)}{q_{\gamma|\theta_h}(\tilde{\tilde{\gamma}}_{hs})}\,, \tag{7}$$

where $\{\tilde{\tilde{\gamma}}_{hs}\}_{s=1}^{M_3}$ is an i.i.d. sample from the importance density $q_{\gamma|\theta_h}$. To minimize the variance of the estimator, the importance distribution $q_{\gamma|\theta_h}$ should approximate the conditional posterior $f_A(\gamma|\mathbf{y}_h,\boldsymbol{\theta}_h,\boldsymbol{\xi})$ for the nuisance parameters. To obtain such an approximation we suggest closed-form conditioning of the ALIS/LIS approximation to the joint posterior, $\psi|\mathbf{y}_h \overset{\text{approx}}{\sim} N(\hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Sigma}}_h)$, giving

$$q_{\gamma|\theta_h} \sim N\Big[\hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^h + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma\theta}}^h (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta\theta}}^h)^{-1}(\boldsymbol{\theta}_h - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^h)\,, \\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma\gamma}}^h - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma\theta}}^h (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta\theta}}^h)^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta\gamma}}^h\Big]\,,$$

where $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^h$, $\hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}^h$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta\theta}}^h$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta\gamma}}^h$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma\theta}}^h$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma\gamma}}^h$ denote the appropriate subcomponents of $\hat{\boldsymbol{\mu}}_h$ and $\hat{\boldsymbol{\Sigma}}_h$.

Note that we already gave some examples of models with nuisance parameters in Section 4. Approximation (7) was not needed in these cases, as the nuisance parameters could be integrated out analytically. Approximation (7) will be more useful when the nuisance parameters are analytically intractable.

6 Discussion

Given the results here, our overall recommendation would be to use ALIS or LIS when finding (near)-optimal designs if the computational budget allows. If a smaller cost is required then LA1 may give a competitive design if the experiment size is sufficiently large relative to the number of parameters in the model. We would discourage the use of other methods, especially the reuse estimator, due to their potential for poor performance.

Uptake of the ALIS and LIS methods would likely be enhanced by their inclusion in software such as the `acebayes` package. A major barrier to this is that to obtain acceptable computation times we found it necessary to hard-code various model-specific functions in C++, including the mean function $\eta(x,\theta)$, the likelihood and prior, and their derivatives. A non-specialist user is unlikely to have the time or inclination to implement such functions in C++ for their models, even with the benefit of high level linear algebra packages. One potential solution to this quandry may be to leverage recent probabilistic programming languages such as STAN (Stan Development Team 2021) or Turing.jl (Ge et al. 2018), a package for the Julia language (Bezanson et al. 2017). Both of these frameworks allow user-friendly high-level specification of Bayesian models but achieve performance comparable to compiled code. In addition these frameworks allow automatic differentiation, avoiding the need for detailed manual calculation of derivatives.

The results in this paper are limited to the expected Shannon information gain criterion. This is the most common choice in the literature, and it is a good one when the goal is inference and uncertainty quantification about the parameters using the full posterior distribution for reasons discussed in Section 1. In other situations, such as point estimation, a different utility function may be preferable. We believe that similar numerical approximations to LIS/ALIS could be developed for other utility functions. The idea of approximating the optimal importance distribution could again be used, though this would no longer be the posterior distribution. However, such methods are outside the scope of the present paper. While of interest, comparisons with other recent approaches such as amortized variational inference (Foster et al. 2019) and layered

multiple importance sampling (Feng & Marzouk 2019) are also outside of scope.

Acknowledgments

This work was supported by the UK Engineering and Physical Sciences Research Council through a PhD studentship (YE) and a Fellowship (DCW, EP/J018317/1). Some calculations were performed using the Iridis computational facility at the University of Southampton. Part of the work was completed while the authors were visiting the Isaac Newton Institute, Cambridge, UK, as part of the programme 'Uncertainty quantification for complex systems: theory and methodologies' during 2018.

References

Bates, D. M. & Watts, D. G. (1988), *Nonlinear regression analysis and its applications*, Wiley, New York.

Beck, J., Dia, B. M., Espath, L. F., Long, Q. & Tempone, R. (2018), 'Fast Bayesian experimental design: Laplace-based importance sampling for the expected information gain', *Computer Methods in Applied Mechanics and Engineering* **334**, 523–553.

Bernardo, J. M. (1979), 'Expected information as expected utility', *Annals of Statistics* **7**, 686–690.

Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. (2017), 'Julia: A fresh approach to numerical computing', *SIAM Review* **59**, 65–98.

Chaloner, K. & Verdinelli, I. (1995), 'Bayesian experimental design: a review', *Statistical Science* **10**, 273–304.

Chernoff, H. (1953), 'Locally optimal designs for estimating parameters', *The Annals of Mathematical Statistics* **24**, 586–602.

DiCiccio, T. J., Kass, R. E., Raftery, A. & Wasserman, L. (1997), 'Computing Bayes factors by combining simulation and asymptotic approximations', *Journal of the American Statistical Association* **92**, 903–915.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J. & Bates, D. (2011), 'Rcpp: Seamless R and C++ integration', *Journal of Statistical Software* **40**, 1–18.

Eddelbuettel, D. & Sanderson, C. (2014), 'RcppArmadillo: Accelerating R with high-performance C++ linear algebra', *Computational Statistics & Data Analysis* **71**, 1054–1063.

Englezou, Y. (2018), Bayesian design for calibration of physical models, PhD thesis, University of Southampton.
URL: *https://eprints.soton.ac.uk/427145/*

Feng, C. (2015), Optimal Bayesian experimental design in the presence of model error, Master's thesis, Center for Computational Engineering, Massachussets Institute of Technology.

Feng, C. & Marzouk, Y. M. (2019), 'A layered multiple importance sampling scheme for focused optimal bayesian experimental design', *arXiv preprint arXiv:1903.11187* .

Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y. W., Rainforth, T. & Goodman, N. (2019), 'Variational Bayesian optimal experimental design', *arXiv preprint arXiv:1903.05480* .

Ge, H., Xu, K. & Ghahramani, Z. (2018), Turing: a language for flexible probabilistic inference, *in* 'International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain', pp. 1682–1690.
URL: *http://proceedings.mlr.press/v84/ge18b.html*

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian data analysis*, 3rd edn, Chapman and Hall/CRC.

Huan, X. & Marzouk, Y. M. (2013), 'Simulation-based optimal Bayesian experimental design for nonlinear systems', *Journal of Computational Physics* **232**, 288–317.

Lemieux, C. (2009), *Monte Carlo and Quasi-Monte Carlo Sampling*, Springer, New York.

Lindley, D. V. et al. (1956), 'On a measure of the information provided by an experiment', *The Annals of Mathematical Statistics* **27**, 986–1005.

Long, Q., Scavino, M., Tempone, R. & 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.

Müller, P., Sansó, B. & De Iorio, M. (2004), 'Optimal Bayesian design by inhomogeneous Markov chain simulation', *Journal of the American Statistical Association* **99**, 788–798.

Overstall, A. M., McGree, J. M. & Drovandi, C. C. (2018), 'An approach for finding fully Bayesian optimal designs using normal-based approximations to loss functions', *Statistics and Computing* **28**, 343–358.

Overstall, A. M. & Woods, D. C. (2017), 'Bayesian design of experiments using approximate coordinate exchange', *Technometrics* **59**, 458–470.

Overstall, A., Woods, D. & Adamou, M. (2019), 'acebayes - An R package for Bayesian optimal design of experiments via approximate coordinate exchange', *Journal of Statistical Software* **95**(13), 1–33.

Ryan, E., Drovandi, C. & Pettitt, A. (2015), 'Fully Bayesian experimental design for pharmacokinetic studies', *Entropy* **17**, 1063–1089.

Ryan, K. J. (2003), 'Estimating expected information gains for experimental designs with application to the random fatigue-limit model', *Journal of Compu-*

tational and Graphical Statistics **12**, 585–603.

Senarathne, S., Drovandi, C. C. & McGree, J. M. (2020), 'A Laplace-based algorithm for Bayesian adaptive design', *Statistics and Computing* **30**, 1183–1208.

Stan Development Team (2021), *Stan Modeling Language Users Guide and Reference Manual, 2.27*. **URL:** *https://mc-stan.org*

A Appendix: derivative calculations

The ALIS and LIS methods require the (marginal) likelihood and the gradient and Hessian of the log unnormalized posterior. In this appendix we report the details of these calculations for the models in Section 4.

A.1 Nonlinear models with σ^2 treated as nuisance

Here we assume that the model is a general nonlinear regression (5) where the noise variance is a nuisance parameter and has the conjugate prior $\sigma_\varepsilon^2 \sim \text{IG}(a,b)$, where a and b denote the shape and scale hyperparameters. In this case the nuisance parameter can be integrated out analytically. We work on the scale $\vartheta_j = \log\theta_j$ on which the parameters are assumed to follow independent normal priors, $\vartheta_j \sim N(\bar\vartheta_j, v_j)$. The Michaelis-Menten and BOD examples from Section 4 both fit into this framework.

The marginal likelihood is

$$\begin{aligned} f_R(\mathbf{y}|\boldsymbol\vartheta,\boldsymbol\xi) &= \int_0^\infty f_R(\mathbf{y}|\boldsymbol\vartheta,\sigma_\varepsilon^2,\boldsymbol\xi) f_B(\boldsymbol\vartheta) d\sigma_\varepsilon^2 \\ &= \text{const.} \times \left\{ 1 + \frac{1}{2b}\sum_{i=1}^n (y_i - \eta(x_i,\boldsymbol\vartheta))^2 \right\}^{-(a+n/2)} . \end{aligned}$$

The unnormalized log-posterior for $\boldsymbol\vartheta$ is

$$\begin{aligned} &\log f_J(\boldsymbol\vartheta,\mathbf{y}|\boldsymbol\xi) \\ &= \text{const.} - \left(a+\frac{n}{2}\right)\log\left(1+\frac{1}{2b}\sum_{i=1}^n (y_i-\eta(x_i;\boldsymbol\vartheta))^2\right) \\ &\quad - \frac{1}{2}\sum_{j=1}^2 \log(2\pi v_j) - \sum_{j=1}^2 \frac{(\vartheta_j-\bar\vartheta_j)^2}{2v_j} . \end{aligned}$$

Its gradient and Hessian are given by

$$\begin{aligned} &\frac{\partial}{\partial\vartheta_k}\log f_J(\boldsymbol\vartheta,\mathbf{y}|\boldsymbol\xi) \\ &= \frac{(2a+n)\sum_{i=1}^n (y_i-\eta(x_i,\boldsymbol\vartheta))\frac{\partial\eta}{\partial\vartheta_k}(x_i,\boldsymbol\vartheta)}{2b+\sum_{i=1}^n (y_i-\eta(x_i,\boldsymbol\vartheta))^2} - \frac{\vartheta_k-\bar\vartheta_k}{v_k} \\ &\frac{\partial^2}{\partial\vartheta_k\partial\vartheta_l}\log f_J(\boldsymbol\vartheta,\mathbf{y}|\boldsymbol\xi) \\ &= -\frac{\delta_{kl}}{v_k} + (2a+n)\left(\frac{\sum_{i=1}^n (y_i-\eta(x_i,\boldsymbol\vartheta))\frac{\partial^2\eta(x_i,\boldsymbol\vartheta)}{\partial\vartheta_k\partial\vartheta_l}}{2b+\sum_{i=1}^n (y_i-\eta(x_i,\boldsymbol\vartheta))^2} \right. \\ &\quad - \frac{\sum_{i=1}^n \frac{\partial\eta}{\partial\vartheta_k}(x_i,\boldsymbol\vartheta)\frac{\partial\eta}{\partial\vartheta_l}(x_i,\boldsymbol\vartheta)}{2b+\sum_{i=1}^n (y_i-\eta(x_i,\boldsymbol\vartheta))^2} \\ &\quad \left. + 2\frac{\sum_{i=1}^n (y_i-\eta(x_i,\boldsymbol\vartheta))\frac{\partial\eta}{\partial\vartheta_k}(x_i,\boldsymbol\vartheta)\sum_{i=1}^n (y_i-\eta(x_i,\boldsymbol\vartheta))\frac{\partial\eta}{\partial\vartheta_l}(x_i,\boldsymbol\vartheta)}{(2b+\sum_{i=1}^n (y_i-\eta(x_i,\boldsymbol\vartheta))^2)^2} \right), \end{aligned}$$

where δ_{kl} denotes the Kronecker delta, i.e. $\delta_{kl}=1$ if $k=l$, and 0 otherwise. These expressions can be evaluated for a particular nonlinear model by substituting in appropriate expressions for the mean function and its partial derivatives.

For the Michaelis-Menten the appropriate formulae are

$$\begin{aligned} &\eta(x,\boldsymbol\vartheta) = \frac{xe^{\vartheta_1}}{e^{\vartheta_2}+x}, \\ &\frac{\partial\eta}{\partial\vartheta_1}(x,\boldsymbol\vartheta) = \frac{xe^{\vartheta_1}}{e^{\vartheta_2}+x}, \qquad \frac{\partial\eta}{\partial\vartheta_2}(x,\boldsymbol\vartheta) = -\frac{xe^{\vartheta_1}e^{\vartheta_2}}{(e^{\vartheta_2}+x)^2}, \\ &\frac{\partial^2\eta}{\partial\theta_1^2}(x,\boldsymbol\vartheta) = \frac{xe^{\vartheta_1}}{e^{\vartheta_2}+x}, \qquad \frac{\partial^2\eta(x,\boldsymbol\vartheta)}{\partial\theta_1\partial\theta_2} = -\frac{xe^{\vartheta_1}e^{\vartheta_2}}{(e^{\vartheta_2}+x)^2}, \\ &\frac{\partial^2\eta}{\partial\vartheta_2^2}(x,\boldsymbol\vartheta) = \frac{2xe^{\vartheta_1}e^{2\vartheta_2}}{(e^{\vartheta_2}+x)^3} - \frac{xe^{\theta_1}e^{\theta_2}}{(e^{\vartheta_2}+x)^2}, \end{aligned}$$

while for BOD they are

$$\begin{aligned} \eta(x,\boldsymbol\vartheta) &= e^{\vartheta_1}(1-\exp(-xe^{\vartheta_2})), \\ \frac{\partial\eta}{\partial\vartheta_1}(x,\boldsymbol\vartheta) &= e^{\vartheta_1}(1-\exp(-xe^{\vartheta_2})), \\ \frac{\partial\eta}{\partial\vartheta_2}(x,\boldsymbol\vartheta) &= xe^{\vartheta_1}\exp(-xe^{\vartheta_2})e^{\vartheta_2}, \\ \frac{\partial^2\eta(x,\boldsymbol\vartheta)}{\partial\vartheta_1^2} &= e^{\vartheta_1}(1-\exp(-xe^{\vartheta_2})), \\ \frac{\partial^2\eta(x,\boldsymbol\vartheta)}{\partial\vartheta_1\partial\vartheta_2} &= xe^{\vartheta_1}\exp(-xe^{\vartheta_2})e^{\vartheta_2}, \\ \frac{\partial^2\eta(x,\boldsymbol\vartheta)}{\partial\vartheta_2^2} &= xe^{\vartheta_1}\exp(-xe^{\vartheta_2})(-xe^{2\vartheta_2}+e^{\vartheta_2}). \end{aligned}$$

Note that the approach of Englezou (2018) was to reparameterize only when finding the importance distribution inside the ALIS/LIS algorithm, which requires the introduction of Jacobian terms. Here we instead take the more direct approach of reparameterizing the model before finding any derivatives. This leads to expressions that are more general and easier to check, but the two approaches ultimately yield equivalent answers after appropriate simplification.

A.2 Nonlinear models with σ_ε^2 treated as an interest parameter

In this section the interest parameter is $\boldsymbol\psi = (\theta_1,\ldots,\theta_p,\varsigma)^{\mathrm{T}}$ where the nonlinear model parameters have prior distributions $\theta_j \sim N(\bar\theta_j, v_j)$ and $\varsigma = \log(\sigma_\varepsilon) \sim N(\bar\varsigma, v_\varsigma)$. The unnormalized log posterior is

$$\begin{aligned} \log f_J(\boldsymbol\psi,\mathbf{y}|\boldsymbol\xi) = &-\frac{n}{2}\log 2\pi - n\varsigma - \frac{1}{2e^{2\varsigma}}\sum_{i=1}^n (y_i-\eta(x_i,\boldsymbol\theta))^2 \\ &- \frac{1}{2}\sum_{j=1}^p \log(2\pi v_j) - \sum_{j=1}^p \frac{(\theta_j-\bar\theta_j)^2}{2v_j} \\ &- \frac{1}{2}\log(2\pi v_\varsigma) - \frac{(\varsigma-\bar\varsigma)^2}{2v_\varsigma}, \end{aligned}$$

with gradient given by

$$\begin{aligned} &\frac{\partial}{\partial\theta_k}\log f_J(\boldsymbol\psi,\mathbf{y}|\boldsymbol\xi) \\ &= e^{-2\varsigma}\sum_{i=1}^n (y_i-\eta(x_i,\boldsymbol\theta))\frac{\partial\eta}{\partial\theta_k}(x_i,\boldsymbol\theta) - \frac{\theta_k-\bar\theta_k}{v_k} \\ &\frac{\partial}{\partial\varsigma}\log f_J(\boldsymbol\psi,\mathbf{y}|\boldsymbol\xi) \\ &= -n + e^{-2\varsigma}\sum_{i=1}^n (y_i-\eta(x_i,\boldsymbol\theta))^2 - \frac{\varsigma-\bar\varsigma}{v_\varsigma}, \end{aligned}$$

and Hessian given by

$$\frac{\partial^2 \log f_J(\psi,\mathbf{y}|\boldsymbol{\xi})}{\partial\theta_k\partial\theta_l} = e^{-2\varsigma}\sum_{i=1}^{n}(y_i-\eta(x_i,\boldsymbol{\theta}))\frac{\partial^2\eta(x_i,\boldsymbol{\theta})}{\partial\theta_k\partial\theta_l} - e^{-2\varsigma}\sum_{i=1}^{n}\frac{\partial\eta}{\partial\theta_k}(x_i,\boldsymbol{\theta})\frac{\partial\eta}{\partial\theta_l}(x_i,\boldsymbol{\theta}) - \frac{\delta_{kl}}{v_k}$$

$$\frac{\partial^2 \log f_J(\psi,\mathbf{y}|\boldsymbol{\xi})}{\partial\theta_k\partial\varsigma} = -2e^{-2\varsigma}\sum_{i=1}^{n}(y_i-\eta(x_i,\boldsymbol{\theta}))\frac{\partial\eta}{\partial\theta_k}(x_i,\boldsymbol{\theta})$$

$$\frac{\partial^2 \log f_J(\psi,\mathbf{y}|\boldsymbol{\xi})}{\partial\varsigma^2} = -2e^{-2\varsigma}\sum_{i=1}^{n}(y_i-\eta(x_i,\boldsymbol{\theta}))^2 - \frac{1}{v_\varsigma}.$$

For a given nonlinear model the above can be evaluated by substituting in appropriate expressions for the mean function η and its partial derivatives. For the lubricant model we have

$$\eta(\mathbf{x},\boldsymbol{\theta}) = \frac{\theta_1}{\theta_2+x_1} + \theta_3 x_2 + \theta_4 x_2^2 + \theta_5 x_2^3 + (\theta_6+\theta_7 x_2^2)x_2\exp\left(-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right),$$

$$\frac{\partial\eta}{\partial\theta_1} = \frac{1}{\theta_2+x_1}, \qquad \frac{\partial\eta}{\partial\theta_2} = -\frac{\theta_1}{(\theta_2+x_1)^2},$$

$$\frac{\partial\eta}{\partial\theta_3} = x_2, \qquad \frac{\partial\eta}{\partial\theta_4} = x_2^2, \qquad \frac{\partial\eta}{\partial\theta_5} = x_2^3,$$

$$\frac{\partial\eta}{\partial\theta_6} = x_2\exp\left(-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right),$$

$$\frac{\partial\eta}{\partial\theta_7} = x_2^3\exp\left(-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right),$$

$$\frac{\partial\eta}{\partial\theta_8} = x_1x_2(\theta_6+\theta_7x_2^2)\exp\left(-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right)\frac{1}{(\theta_8+\theta_9x_2^2)^2},$$

$$\frac{\partial\eta}{\partial\theta_9} = x_1x_2^3(\theta_6+\theta_7x_2^2)\exp\left(-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right)\frac{1}{(\theta_8+\theta_9x_2^2)^2}.$$

Among the $\frac{\partial^2\eta}{\partial\theta_k\partial\theta_l}$, $(k\le l)$, the non-zero terms are

$$\frac{\partial^2\eta}{\partial\theta_1\partial\theta_2} = -\frac{1}{(\theta_2+x_1)^2}, \qquad \frac{\partial^2\eta}{\partial\theta_2\partial\theta_2} = \frac{2\theta_1}{(\theta_2+x_1)^3},$$

$$\frac{\partial^2\eta}{\partial\theta_6\partial\theta_8} = x_1x_2\exp\left(-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right)\frac{1}{(\theta_8+\theta_9x_2^2)^2},$$

$$\frac{\partial^2\eta}{\partial\theta_6\partial\theta_9} = x_1x_2^3\exp\left(-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right)\frac{1}{(\theta_8+\theta_9x_2^2)^2},$$

$$\frac{\partial^2\eta}{\partial\theta_7\partial\theta_8} = x_1x_2^3\exp\left(-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right)\frac{1}{(\theta_8+\theta_9x_2^2)^2},$$

$$\frac{\partial^2\eta}{\partial\theta_7\partial\theta_9} = x_1x_2^5\exp\left(-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right)\frac{1}{(\theta_8+\theta_9x_2^2)^2},$$

$$\frac{\partial^2\eta}{\partial\theta_8\partial\theta_8} = x_1x_2(\theta_6+\theta_7x_2^2) \times\exp\left(-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right)\left\{\frac{x_1}{(\theta_8+\theta_9x_2^2)^4}-\frac{2}{(\theta_8+\theta_9x_2^2)^3}\right\},$$

$$\frac{\partial^2\eta}{\partial\theta_8\partial\theta_9} = x_1x_2^3(\theta_6+\theta_7x_2^2) \times\exp\left(-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right)\left\{\frac{x_1}{(\theta_8+\theta_9x_2^2)^4}-\frac{2}{(\theta_8+\theta_9x_2^2)^3}\right\},$$

$$\frac{\partial^2\eta}{\partial\theta_9\partial\theta_9} = x_1x_2^5(\theta_6+\theta_7x_2^2) \times\exp\left(-\frac{x_1}{\theta_8+\theta_9 x_2^2}\right)\left\{\frac{x_1}{(\theta_8+\theta_9x_2^2)^4}-\frac{2}{(\theta_8+\theta_9x_2^2)^3}\right\}.$$

All other second-order derivatives are either zero or can be obtained from the above by symmetry.