Bayesian optimal design for ordinary differential equation models

Antony M. Overstall
School of Mathematics & Statistics
University of Glasgow
Glasgow

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

Benjamin M. Parker
Statistical Sciences Research Institute
University of Southampton
Southampton

Abstract

Bayesian optimal design is considered for experiments where it is hypothesised that the responses are described by the intractable solution to a system of non-linear ordinary differential equations (ODEs). Bayesian optimal design is based on the minimisation of an expected loss function where the expectation is with respect to all unknown quantities (responses and parameters). This expectation is typically intractable even for simple models before even considering the intractability of the ODE solution. New methodology is developed for this problem that involves minimising a smoothed stochastic approximation to the expected loss and using a state-of-the-art stochastic solution to the ODEs, by treating the ODE solution as an unknown quantity. The methodology is demonstrated on three illustrative examples and a real application involving estimating the properties of human placentas.

Keywords: Approximate coordinate exchange algorithm; Bayesian optimal design; Ordinary differential equations.

1 Introduction

1.1 Modelling complex physical processes

Often the dynamics behind a complex physical process can be approximately described by a system of non-linear ordinary differential equations (ODEs), where the solution to these equations provides a model predicting how a quantity of interest will behave with respect to time. It is assumed that the system of ODEs depends on some unknown physical properties (parameters) of the process in question and, potentially, may also depend on some additional controllable (design) factors.

The value of the parameters may be of direct interest or we may be interested in predicting the behaviour of the process at certain values of the design variables and/or at a certain time. In either case, we need to estimate the unknown parameters. An experiment can be conducted where observations of the quantity of interest are collected at various different times and, possibly from multiple runs of the experiment with different values of the design factors. To estimate the parameters, a statistical

Table 1: Physical parameters, $\boldsymbol{\theta}$ for the human placenta example described in Section 1.2.

Symbol	Description
θ_1	Maximum uptake
θ_2	Proportion of reaction occurring through active transport
θ_3	1st reaction rate
θ_4	2nd reaction rate

model is assumed that links the parameters to the observations via a data-generating process based on the solution to the ODEs (see, for example, Ramsay et al. 2007).

It may be possible to conduct a designed experiment where the observation times and design variables are actively selected in advance of the experiment. Optimal design of experiments refers to this selection being made optimally to minimise a loss function which reflects the ultimate goal of the experiment, e.g. estimation of the unknown parameters or prediction of a future process.

We consider Bayesian optimal design of experiments for ODE models. Bayesian optimal design has very principled foundations but can be hard to implement in practice due to the computational complexities involved. Firstly, it involves the minimisation of the expected loss function which will typically be analytically intractable. Secondly, the dimensionality of the domain of the expected loss function can sometimes be large since every quantity that can be specified for the experiment corresponds to a dimension.

For ODE models, these problems are compounded by the fact that, typically, the solution to the system of ODEs is not analytically tractable. A possible approach is to use numerical methods to find an approximate solution to the systems of ODEs (see, e.g., Iserles, 2009). However, this has two disadvantages. First, the numerical solution can be computationally expensive. Bayesian inference for computationally expensive models has begun to receive considerable attention in the Statistics literature (e.g. Kennedy and O 'Hagan 2001; Rasmussen 2003; Bliznyuk et al. 2008; Fielding et al. 2011; Overstall and Woods 2013). In each case, to some degree, evaluation of the likelihood (which depends on the computationally expensive numerical solution) is replaced by an evaluation of an approximation. The second disadvantage, which is perhaps more serious, is that the numerical error, unavoidable with numerical methods, is typically not taken account of when performing subsequent evaluations with the numerical solution. This issue was explained by Chkrebtii et al. (2015) who proposed a fully probabilistic solution to the system of ODEs. We take advantage of this methodology in our treatment of optimal design for ODE models.

Overstall and Woods (2015) proposed the approximate coordinate exchange (ACE) algorithm for Bayesian optimal design for non-ODE models. Very briefly, the ACE algorithm uses a cyclic descent algorithm (see, for example, Lange, 2013, pg 171) to minimise an approximation to the expected loss function. The approximation used is a Gaussian process (GP) emulator fitted to a Monte Carlo integration approximation of the expected loss. In this paper, we extend this algorithm to ODE models. We replace evaluation of the intractable solution to the ODEs by a value generated from the probabilistic solution as proposed by Chkrebtii et al. (2015).

1.2 Measuring human placentas

To aid exposition of the ideas and methodology introduced throughout this paper, consider the following application from biologists at the Southampton Centre for Biological Sciences (University of Southampton, UK). Interest lies in the transport of the amino acid serine within a human placenta. Specifically, we are concerned by how serine moves from the outside to the inside of a portion of placental cell membrane (called a vesicle). To investigate, an experiment is to be performed where initial amounts of radioactive and non-radioactive serine (in μl) are placed inside and outside the vesicle. The amount of radioactive serine inside the vesicle is then measured at a series of observation times. The serine transport process can be described by a system of ODEs which depend on the initial amounts of serine and four unknown physical parameters (see Table 1) which are of interest to the scientists. The solution to the system of ODEs provides theoretically predicted amounts of radioactive and non-radioactive serine inside the vesicle at a certain time. The practitioners have control over the initial amounts of non-radioactive serine inside and outside the vesicle for each experiment and the values of the observation times. Our task is to choose the initial amounts and observation times optimally with respect to the goal of estimating the physical parameters.

1.3 Organisation of the paper

The paper is organised as follows. In Section 2 we describe the background to the problem including statistical inference for ODE models, the premise of Bayesian optimal design and a brief description of the ACE algorithm. In Section 3 we describe the proposed methodology for optimal design for ODE models including a description of the probabilistic solution to ODEs of Chkrebtii et al. (2015) and how this can be embedded in the ACE algorithm. In Section 4, we apply this methodology to three illustrative examples where the goal is parameter estimation. Finally, in Section 5, the methodology is applied to the human placenta example where differing goals of parameter estimation and model selection are considered.

2 Background

2.1 Statistical inference for ordinary differential equations

Let $\mathbf{x} \in \mathcal{X}$ be a vector of k design variables, i.e. a treatment, and let $\boldsymbol{\theta} \in \Theta$ be a $p \times 1$ vector of physical parameters. Consider the following system of S ODEs which define an initial value problem

$$
\left.
\begin{aligned}
\dot{\mathbf{u}}(t) &= \mathbf{f}\left(\mathbf{u}(t), t, \boldsymbol{\theta}, \mathbf{x}\right) \\
\mathbf{u}(T_0) &= \mathbf{u}_0
\end{aligned}
\right\} \text{ for } t \in [T_0, T_1],
\tag{1}
$$

where $\dot{\mathbf{u}}(t)$ is the gradient vector of $\mathbf{u}(t)$ with respect to time t, and $\mathbf{u}_0 \in \mathbb{R}^S$ denotes the initial conditions. In (1), $\mathbf{f} : \mathbb{R}^S \times [T_0, T_1] \times \mathcal{X} \to \mathbb{R}^S$ is a suitably well-behaved function that, at the very least, we assume satisfies the Lipschitz condition (see, e.g. Iserles, 2009, pg 3). This means that (1) has a unique solution. Note that the solution actually depends on $\boldsymbol{\theta}$ and $\mathbf{x}$, i.e. $\mathbf{u}(t) = \mathbf{u}(t; \boldsymbol{\theta}, \mathbf{x})$ but we only use the longer notation when we need to be clear that there may be more than one $\boldsymbol{\theta}$ or $\mathbf{x}$.

Now we believe that the physical process in question is governed by (1). For $j = 1, \ldots, M$, we observe the process for treatment $\mathbf{x}_j$, initial conditions $\mathbf{u}_{0j}$, and at times $\mathbf{t}_j = \left(t_{j1}, \ldots, t_{jn_j}\right)$. For

$l = 1, \ldots, n_j$, we observe the $c \times 1$ vector of responses $\mathbf{y}_{jl} \in \mathcal{Y}_{jl}$, where $\mathcal{Y}_{jl}$ denotes the c-dimensional sample space for the jlth observation. Let $\mathbf{y}_j = \left(\mathbf{y}_{j1}, \ldots, \mathbf{y}_{jn_j} \right)$ be the $cn_j \times 1$ vector of responses for the jth treatment and let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_M) \in \mathcal{Y}$ be the $n \times 1$ vector of responses for the complete experiment where $n = c \sum_{j=1}^{M} n_j$ and $\mathcal{Y} = \bigcup_{j=1}^{M} \bigcup_{l=1}^{n_j} \mathcal{Y}_{jl}$ is the overall sample space. We assume that $\mathbf{y}$ are realisations according to

$$\mathbf{y}|\boldsymbol{\psi}, \mathbf{d} \sim \mathrm{F}\left(\boldsymbol{\psi}; \mathbf{d} \right), \tag{2}$$

where F is a known probability distribution, $\boldsymbol{\psi} \in \Psi$ is a $P \times 1$ vector of model parameters (with parameter space Ψ), and $\mathbf{d} \in \mathcal{D}$ is a $q \times 1$ vector specifying the design (with design space $\mathcal{D}$). The distribution in (2) essentially defines a statistical model. Note that we decompose $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma} \in \Gamma$ is a $(P - p) \times 1$ vector of nuisance parameters. The design, $\mathbf{d}$, is the set of controllable experimental conditions and can include the treatments $\mathbf{x}_1, \ldots, \mathbf{x}_M$; the initial conditions $\mathbf{u}_{01}, \ldots, \mathbf{u}_{0M}$; and the observation times $t_{j1}, \ldots, t_{jn_j}$, for $i = j, \ldots, M$. In practice, some of these may be fixed by the protocol of the experiment. Alternatively, the initial conditions may be unknown, and included in the vector of parameters, $\boldsymbol{\psi}$, either as physical or nuisance parameters.

The dependence of the distribution in (2) on $\boldsymbol{\theta}$ and $\mathbf{d}$ is through the solution of the system of ODEs given by (1). The most obvious way to do this is to assume that

$$\mathrm{E}\left(\mathbf{y}_{jl} | \boldsymbol{\theta}, \mathbf{x}_j, t_{jl} \right) = \mathcal{G}\left(\mathbf{u}(t_{jl}), \boldsymbol{\theta} \right),$$

where $\mathcal{G} : \mathbb{R}^S \times \Theta \to \mathcal{Y}_{ij}$ is a known function. If $\mathcal{G}(\mathbf{u}, \boldsymbol{\theta}) \neq \mathbf{I}_S \mathbf{u}$, then Ramsay et al. (2007) call this a distributed partial data problem.

As an example, consider the human placenta example introduced in Section 1.2. The system of $S = 2$ ODEs is given by

$$\left. \begin{aligned} \dot{u}_1(t) &= \frac{x_1(u_2(t) + \theta_2\theta_4) - u_1(t)(x_2 + \theta_2\theta_3)}{u^*(\mathbf{u}, t, \boldsymbol{\theta}, \mathbf{x})} \\ \dot{u}_2(t) &= \frac{x_2(u_1(t) + \theta_2\theta_4) - u_2(t)(x_1 + \theta_2\theta_3)}{u^*(\mathbf{u}, t, \boldsymbol{\theta}, \mathbf{x})} \\ u_1(0) &= u_{01} \\ u_2(0) &= u_{02} \end{aligned} \right\} t \in [T_0, T_1], \tag{3}$$

where

$$u^*\left(\mathbf{u}(t), t, \boldsymbol{\theta}, \mathbf{x} \right) = \frac{2(x_1 + x_2)(u_1(t) + u_2(t)) + (1 + \theta_2)\left(\theta_4(x_1 + x_2) + \theta_3(u_1(t) + u_2(t)) \right) + 2\theta_3\theta_4}{\theta_1},$$

and $T_0 = 0$ and $T_1 = 600$ seconds. The solution to this system is $\mathbf{u}(t) = (u_1(t), u_2(t))$ which are the amounts of radioactive and non-radioactive serine, respectively, inside the vesicle at time t. The values of $\mathbf{x} = (x_1, x_2) \in \mathcal{X} = [0, 1000]^2$ are the amounts of radioactive and non-radioactive serine outside the vesicle at time $t = 0$, and $\mathbf{u}_0 = (u_{01}, u_{02}) \in [0, 1000]^2$ are the corresponding amounts of serine inside the vesicle at time $t = 0$. In the experiment, we are able to control x_2 and u_{02} (i.e. x_1 and u_{01} are fixed by the experimental protocol), and the response, y_{jl}, is the amount of radioactive serine inside the vesicle at time t_{jl} for process conditions x_{2j} and u_{02j}. We assume a statistical model where

$$\mathrm{E}\left(y_{jl} | \boldsymbol{\theta}, \mathbf{x}_j, t_{jl} \right) = u_1(t_{jl}),$$

so that in this case $\mathcal{G}(\mathbf{u}, \boldsymbol{\theta}) = u_1$. The design is the collection of initial amounts of non-radioactive serine x_{2j} and u_{02j}, and the observation times $t_{j1}, \ldots, t_{jn_j}$; for $j = 1, \ldots, M$.

2.2 Decision-theoretic approach to Bayesian optimal design

We now describe the decision-theoretic approach to Bayesian optimal design. We complete the statistical model given by F in (2) by specifying a prior distribution for $\boldsymbol{\psi}$ which does not depend on the design $\mathbf{d}$. Once we observe $\mathbf{y}$, the posterior distribution of $\boldsymbol{\psi}$ is given by

$$\pi(\boldsymbol{\psi}|\mathbf{y}, \mathbf{d}) \propto \pi(\mathbf{y}|\boldsymbol{\psi}, \mathbf{d})\pi(\boldsymbol{\psi}), \tag{4}$$

where $\pi(\mathbf{y}|\boldsymbol{\psi}, \mathbf{d})$ is the mass/density function of F and $\pi(\boldsymbol{\psi})$ is the prior density function. Note that the right-hand-side of (4) also defines the joint distribution of $\boldsymbol{\psi}$ and $\mathbf{y}$. The posterior distribution of the physical parameters is found by marginalising (4) with respect to the nuisance parameters.

Bayesian optimal design relies on the specification of an appropriate (for the goal of the experiment) loss function denoted by $\lambda(\boldsymbol{\psi}, \mathbf{y}, \mathbf{d})$ which depends on the design and, potentially, the unobserved responses and unknown parameters. An optimal design, $\mathbf{d}^*$, is given by a value of $\mathbf{d}$ that minimises the expectation of $\lambda(\boldsymbol{\psi}, \mathbf{y}, \mathbf{d})$ with respect to the joint distribution of $\boldsymbol{\psi}$ and $\mathbf{y}$ (given $\mathbf{d}$), i.e.

$$\begin{aligned}
\mathbf{d}^* &= \arg\min_{\mathbf{d}\in\mathcal{D}} L(\mathbf{d}), \\
L(\mathbf{d}) &= \mathrm{E}\left(\lambda(\boldsymbol{\psi}, \mathbf{y}, \mathbf{d})|\mathbf{d}\right), \\
&= \int \lambda(\boldsymbol{\psi}, \mathbf{y}, \mathbf{d})\mathrm{dP}_{\boldsymbol{\psi},\mathbf{y}|\mathbf{d}}.
\end{aligned}$$

For example, suppose we were interested in posterior point estimation of the elements of $\boldsymbol{\theta}$, then an appropriate loss function might be the squared error loss (SEL) given by

$$\lambda_{\mathrm{SEL}}(\boldsymbol{\psi}, \mathbf{y}, \mathbf{d}) = \sum_{l=1}^{p} \left(\theta_l - \mathrm{E}\left(\theta_l|\mathbf{y}, \mathbf{d}\right)\right)^2, \tag{5}$$

which does not depend on the nuisance parameters. It can be shown that

$$\begin{aligned}
L_{\mathrm{SEL}}(\mathbf{d}) &= \mathrm{E}\left(\lambda_{\mathrm{SEL}}(\boldsymbol{\psi}, \mathbf{y}, \mathbf{d})|\mathbf{d}\right), \\
&= \int \mathrm{tr}\left(\mathrm{var}\left(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}\right)\right)\mathrm{dP}_{\mathbf{y}|\mathbf{d}},
\end{aligned}$$

and so the optimal design minimises the expected trace of the posterior variance matrix of $\boldsymbol{\theta}$.

As mentioned in Section 1, we are faced with three problems when trying to minimise the expected loss function:

- high dimensionality of the design space, $\mathcal{D}$;

- intractability of the integration required to evaluate $L(\mathbf{d})$ and $\lambda(\boldsymbol{\psi}, \mathbf{y}, \mathbf{d})$;

- both evaluation of $\lambda(\boldsymbol{\psi}, \mathbf{y}, \mathbf{d})$ and the joint distribution of $\mathbf{y}$ and $\boldsymbol{\psi}$ will typically depend on the intractable solution to the system of ODEs.

The approximate coordinate exchange (ACE) algorithm, proposed by Overstall and Woods (2015), is a solution to the first two problems and is described in the next section.

5

2.3 Approximate coordinate exchange algorithm

To overcome the problem of the high dimensionality of the design space, the coordinate exchange (CE; Meyer and Nachtsheim 1995) algorithm is used to minimise the expected loss function, $L(\mathbf{d})$. This is the same as a cyclic descent algorithm where $L(\mathbf{d})$ is minimised, sequentially, over each element (or coordinate) of the design space, where all other elements are held fixed. This process is then repeated until convergence.

However, instead of minimising the intractable $L(\mathbf{d})$, at each iteration, the ACE algorithm minimises an approximation, $\tilde{L}(\mathbf{d})$. Consider the Monte Carlo (MC) approximation to $L(\mathbf{d})$:

$$\hat{L}_B(\mathbf{d}) = \frac{1}{B} \sum_{i=1}^{B} \lambda(\boldsymbol{\psi}_i, \mathbf{y}_i, \mathbf{d}), \tag{6}$$

where $\{\boldsymbol{\psi}_i, \mathbf{y}_i\}_{i=1}^{B}$ is a sample generated from the joint distribution of $\boldsymbol{\psi}$ and $\mathbf{y}$, given $\mathbf{d}$. The $\hat{L}_B(\mathbf{d})$ is a consistent and unbiased estimator of $L(\mathbf{d})$. However, there are at least two reasons why $\hat{L}_B(\mathbf{d})$ would be a poor choice for $\tilde{L}(\mathbf{d})$. First, $\hat{L}_B(\mathbf{d})$ is a stochastic approximation and so is a non-smooth function. Secondly, $\hat{L}_B(\mathbf{d})$ is computationally expensive. This problem is aggravated by the fact that, in some cases, the loss function is, itself, an intractable function requiring MC approximation.

Instead, Overstall and Woods (2015) proposed constructing an approximation (or emulator) for $L(\mathbf{d})$ based on a "small" number of evaluations of $\hat{L}_B(\mathbf{d})$. One of the most common types of emulator is the Gaussian process (GP) emulator. The use of GP emulators in more general optimisation problems dates back to, at least, the expected improvement approach of Jones et al. (1998). The GP emulator provides a predictive distribution for $L(\mathbf{d})$ and we set $\tilde{L}(\mathbf{d})$ to be the predictive mean of this distribution. This approach of smoothing the MC approximations can be seen as an extension to the approach of Muller and Parmigiani (1996), for minimising the expected loss, to higher dimensionality design spaces, by using the CE algorithm. A similar approach was used by Gotwalt et al. (2009) who used a deterministic quadrature rule to evaluate the log determinant of the Fisher information matrix averaged over a prior distribution (a commonly-used classical objective function for optimal design) and then maximised the resulting approximation over the design space using CE.

The actual algorithm is provided in Appendix A. Note that a GP emulator, like all statistical models, can fit inadequately. Bastos and O'Hagan (2009) developed diagnostics to assess the adequacy of GP emulators. However, applying these methods automatically within the ACE algorithm is infeasible. Instead, once $\tilde{L}(\mathbf{d})$ has been minimised, we, independently of the GP emulator, decide whether to accept the change to the current design before we move onto the next element/coordinate in the ACE algorithm. This accept step is accomplished using a Bayesian hypothesis test. For more details on this feature, on the overall ACE algorithm, and on the wider issue of Bayesian optimal design, see Overstall and Woods (2015).

3 Methodology

In this section we describe how the ACE algorithm can be extended to find Bayesian optimal designs for ODE models.

In Section 3.1, we briefly describing the method of Chkrebtii et al. (2015) for finding a probabilistic

solution to a system of ODEs and then, in Section 3.2, describe how this solution can be embedded in the ACE algorithm.

3.1   Probabilistic solution to ODEs

Let R_λ denote a square integrable kernel function with length scale parameter $\lambda \in (0, \infty)$. Furthermore, let

$$
\begin{aligned}
Q_\lambda(t_1, t_2) &= \int_a^{t_1} R_\lambda(s, t_2)\mathrm{d}s, \\
S_\lambda(t_1, t_2) &= \alpha^{-1} \int_{-\infty}^{\infty} R_\lambda(s, t_1)R_\lambda(s, t_2)\mathrm{d}s, \\
W_\lambda(t_1, t_2) &= \alpha^{-1} \int_{-\infty}^{\infty} Q_\lambda(s, t_1)R_\lambda(s, t_2)\mathrm{d}s, \\
V_\lambda(t_1, t_2) &= \alpha^{-1} \int_{-\infty}^{\infty} Q_\lambda(s, t_1)Q_\lambda(s, t_2)\mathrm{d}s,
\end{aligned}
$$

where $\alpha > 0$. Let $\mathbf{u}(t) = (u_1(t), \ldots, u_S(t))$, then the central assumption of the method of Chkrebtii et al. (2015) is that $u_s(t)$ and its time derivative, $\dot{u}_s(t)$, have a joint GP prior, i.e.

$$
\begin{pmatrix} \dot{u}_s(\cdot) \\ u_s(\cdot) \end{pmatrix} \sim \mathrm{GP}\left(\begin{pmatrix} \dot{m}_s(\cdot) \\ m_s(\cdot) \end{pmatrix}, \begin{pmatrix} S_\lambda(\cdot, \cdot) & W_\lambda(\cdot, \cdot) \\ W_\lambda(\cdot, \cdot)^{\mathrm{T}} & V_\lambda(\cdot, \cdot) \end{pmatrix} \right), \tag{7}
$$

independently, for $s = 1, \ldots, S$. The probabilistic solution to the system of ODEs given by (1), conditional on $\boldsymbol{\theta}$, α and λ, is constructed by updating the joint distribution given by (7) sequentially over a discrete grid of N time points $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)$, where $T_0 = \tau_1 \le \tau_2 \le \cdots \le \tau_N = T_1$.

Let $\boldsymbol{\tau}_{1:r} = (\tau_1, \ldots, \tau_r)$ be the $r \times 1$ vector of time points up to and including the rth time point τ_r, for $r = 1, \ldots, N$. The algorithm for the sequential update is as follows:

1. Compute the $S \times 1$ row vector $\mathbf{f}_1 = \mathbf{f}(T_0, \mathbf{u}_0, \boldsymbol{\theta})$ giving the gradient at the initial time point, $T_0 = \tau_1$, and let $\Lambda_1 = 0$.

2. For $r = 1, \ldots, N - 1$ complete the following steps:

   (a) Compute the $S \times 1$ vector giving the predictive mean of $\mathbf{u}(\tau_{r+1})$ as

   $$
   \mathbf{m}(\tau_{r+1}) = \mathbf{u}_0 + W_\lambda(\tau_{r+1}, \boldsymbol{\tau}_{1:r})\left(S_\lambda(\boldsymbol{\tau}_{1:r}, \boldsymbol{\tau}_{1:r}) + \Lambda_r\right)^{-1}\mathbf{F}_r,
   $$

   where $\mathbf{F}_r$ is the $r \times S$ matrix with qth row given by $\mathbf{f}_q$ for $q = 1, \ldots, r$. Compute the common predictive variance of $u_s(\tau_{r+1})$ as

   $$
   C(\tau_{r+1}, \tau_{r+1}) = V_\lambda(\tau_{r+1}, \tau_{r+1}) - W_\lambda(\tau_{r+1}, \boldsymbol{\tau}_{1:r})\left(S_\lambda(\boldsymbol{\tau}_{1:r}, \boldsymbol{\tau}_{1:r}) + \Lambda_r\right)^{-1}W_\lambda(\tau_{r+1}, \boldsymbol{\tau}_{1:r})^{\mathrm{T}}.
   $$

   (b) For $s = 1, \ldots, S$, generate a prediction of the solution at τ_{r+1}, $u_s(\tau_{r+1})$, from the predictive distribution, i.e.
   $$
   u_s(\tau_{r+1}) \sim \mathrm{N}\left(m_s(\tau_{r+1}), C(\tau_{r+1}, \tau_{r+1})\right),
   $$
   where $m_s(\tau_{r+1})$ is the sth element of $\mathbf{m}(\tau_{r+1})$.

7

(c) Compute the true gradient vector at time τ_{r+1} and solution $\mathbf{u}(\tau_{r+1})$ as

$$\mathbf{f}_{r+1} = \mathbf{f}\left(\tau_{r+1}, \mathbf{u}(\tau_{r+1}), \boldsymbol{\theta}\right).$$

(d) Compute the common predictive variance of the time derivative

$$\dot{C}(\tau_{r+1}, \tau_{r+1}) = S_\lambda(\tau_{r+1}, \tau_{r+1}) - S_\lambda(\tau_{r+1}, \boldsymbol{\tau}_{1:r})\left(S_\lambda(\boldsymbol{\tau}_{1:r}, \boldsymbol{\tau}_{1:r}) + \Lambda_r\right)^{-1} S_\lambda(\tau_{r+1}, \boldsymbol{\tau}_{1:r})^{\mathrm{T}},$$

and let $\Lambda_{r+1} = \mathrm{diag}\left\{\Lambda_r, \dot{C}(\tau_{r+1}, \tau_{r+1})\right\}$.

Once step 2 is complete, then a probabilistic solution is given by

$$u_s(\cdot) \sim \mathrm{GP}\left(m_s(\cdot), C(\cdot, \cdot)\right),$$

for $s = 1, \ldots, S$, where $m_s(\cdot)$ is the sth element of the vector of predictive means given by

$$\mathbf{m}(\cdot) = \mathbf{u}_0 + W_\lambda(\cdot, \boldsymbol{\tau})\left(S_\lambda(\boldsymbol{\tau}, \boldsymbol{\tau}) + \Lambda_N\right)^{-1} \mathbf{F}_N,$$

and $C(\cdot, \cdot)$ is the common predictive variance given by

$$C(\cdot, \cdot) = V_\lambda(\cdot, \cdot) - W_\lambda(\cdot, \boldsymbol{\tau})\left(S_\lambda(\boldsymbol{\tau}, \boldsymbol{\tau}) + \Lambda_N\right)^{-1} W_\lambda(\cdot, \boldsymbol{\tau})^{\mathrm{T}}.$$

The methodology holds for general kernel functions. However, Chkrebtii et al. (2015) suggest two example kernel functions, the squared exponential kernel function given by

$$R_\lambda(t_1, t_2) = \exp\left[-\frac{(t_1 - t_2)^2}{2\lambda^2}\right],$$

and the uniform kernel function given by

$$R_\lambda(t_1, t_2) = I\left(t_2 \in (t_1 - \lambda, t_1 + \lambda)\right),$$

where $I(A)$ is the indicator function for the event A. Simplistically, the best choice of kernel function depends on the assumed smoothness of the $\mathbf{u}(t)$. The squared exponential kernel function is infinitely differentiable and can be used for when we expect $\mathbf{u}(t)$ to be smooth. On the other hand, if $\mathbf{u}(t)$ is non-smooth then a better choice is the uniform kernel function which is not differentiable. For closed form expressions for the functions Q_λ, S_λ, W_λ and V_λ, under both the squared exponential and uniform kernel functions, see Chkrebtii et al. (2015).

Consider the system of ODEs, given by (3), that describe the transport of serine in a human placenta. Under the squared exponential kernel function, physical parameters $\boldsymbol{\theta} = (100, 0.05, 100, 100)$, initial values $\mathbf{u}_0 = (0, 1000)$, treatment $\mathbf{x} = (7.5, 1000)$, and $\boldsymbol{\tau}$ containing $N = 501$ evenly spaced time points, Figure 1 shows 1000 draws from the probabilistic solution of $u_1(t)$ and $u_2(t)$ plotted against $t \in [T_0, T_1] = [0, 600]$. Note how the uncertainty in the solution increases as time, t, increases away from $t = T_0$ where we know, in this example, the true value of $\mathbf{u}(t)$.

Chkrebtii et al. (2015) propose several MCMC algorithms for generating a sample from the posterior distribution of the model parameters $\psi = (\boldsymbol{\theta}, \boldsymbol{\gamma})$ and the auxiliary parameters α and λ, given existing experimental responses $\mathbf{y}$. This is accomplished by embedding the probabilistic solution into standard MCMC algorithms.

Figure 1: Plots showing 1000 draws from the probabilistic solution of $u_1(t)$ and $u_2(t)$ against t for the system of ODEs, given by (3), that describe the transport of serine in a human placenta
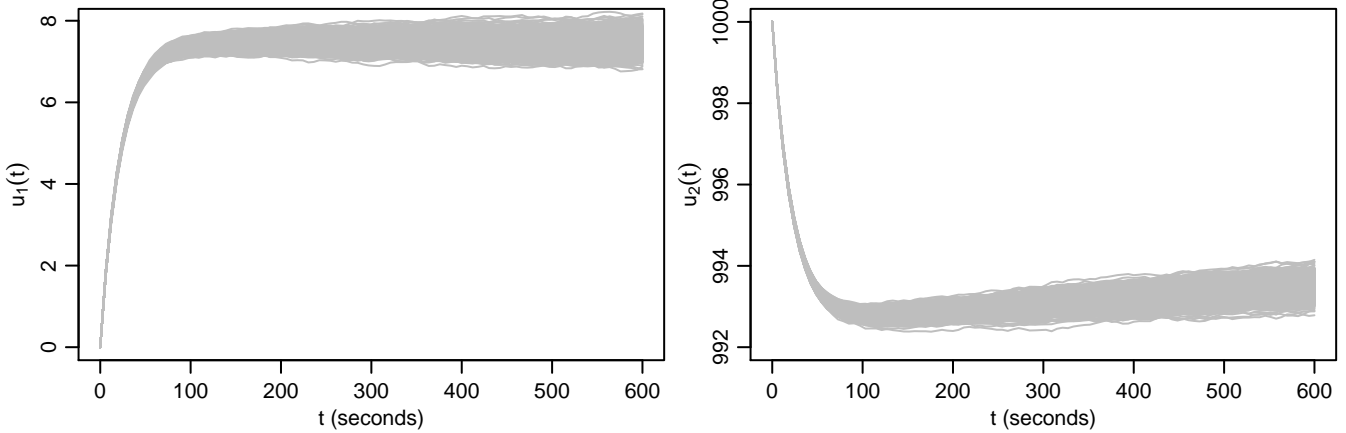
3.2 Extending the ACE algorithm to ODE models

At various points of the ACE algorithm we are required to evaluate a Monte Carlo approximation of the expected loss function, i.e. as given by (6). First, this requires a sample $\{\mathbf{y}_i, \boldsymbol{\psi}_i\}_{i=1}^B$ to be generated from the joint distribution of $\mathbf{y}$ and $\boldsymbol{\psi}$. This is accomplished by generating a value $\boldsymbol{\psi}_i$ from the prior of $\boldsymbol{\psi}$ and then generating a value $\mathbf{y}_i$ from the distribution $\mathrm{F}(\boldsymbol{\psi}_i, \mathbf{d})$ (see equation (2)). This will require B evaluations of the intractable solution of the system of ODEs, for the jth treatment and time points: $t_{j1}, \ldots, t_{jn_j}$, for $j = 1, \ldots, M$. Secondly, we need to evaluate the loss function at each value in the sample $\{\mathbf{y}_i, \boldsymbol{\psi}_i\}_{i=1}^B$. Unfortunately, the most commonly used loss functions are, themselves, intractable. For example, the squared error loss function given by (5) depends on the posterior mean, $\mathrm{E}(\theta_l|\mathbf{y})$, of each element of the vector of physical parameters, $\boldsymbol{\theta}$. Overstall and Woods (2015) used a Monte Carlo approximation of the posterior mean as follows

$$\hat{\mathrm{E}}(\theta_l|\mathbf{y}) = \frac{\sum_{j=1}^B \tilde{\theta}_{jl} \pi(\mathbf{y}|\tilde{\boldsymbol{\psi}}_j)}{\sum_{j=1}^B \pi(\mathbf{y}|\tilde{\boldsymbol{\psi}}_j)},$$

where $\left\{\tilde{\boldsymbol{\psi}}_j\right\}_{j=1}^B$ is an additional sample generated from the prior distribution of $\boldsymbol{\psi}$ where $\tilde{\boldsymbol{\psi}}_j = \left(\tilde{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{\gamma}}_j\right)$ and $\tilde{\theta}_{jl}$ is the lth element of $\tilde{\boldsymbol{\theta}}_j$ for $l = 1, \ldots, p$. Evaluation of $\mathrm{E}(\theta_l|\mathbf{y}_i)$, in the loss function, for $i = 1, \ldots, B$, is now replaced by evaluation of $\hat{\mathrm{E}}(\theta_l|\mathbf{y}_i)$ to give a nested Monte Carlo approximation to the expected loss function. A further B evaluations of the intractable solution, $\mathbf{u}(t)$, will be required for each treatment and for each time point. Therefore, in total, we need $2B$ evaluations of $\mathbf{u}(t)$ for each treatment and for each time point, one for each vector of physical parameters in the samples $\{\boldsymbol{\psi}_i\}_{i=1}^B$ and $\left\{\tilde{\boldsymbol{\psi}}_j\right\}_{j=1}^B$. Thus we are required to evaluate

$$\mathbf{u}_{ijl} = \mathbf{u}(t_{jl}; \boldsymbol{\theta}_i, \mathbf{x}_j),$$

for $i = 1, \ldots, 2B$, $j = 1, \ldots, M$ and $l = 1, \ldots, n_j$.

The basic idea is to replace evaluation of the unknown $\mathbf{u}(t)$ in the ACE algorithm by a value generated from the probabilistic solution outlined in Section 3.1. However, since B will be in the order of 1000s,

9

it will be computationally infeasible to use the full probabilistic solution where both the physical parameters, $\boldsymbol{\theta}$, and auxiliary parameters, α and λ are unknown with prior distributions. However, if we are prepared to fix the values of the auxiliary parameters, then significant computational savings can be found thus making the method feasible.

To see this note that we can rewrite the predictive mean of $\mathbf{u}(\tau_{r+1})$ in step 2(a) in Section 3.1 as follows

$$\mathbf{m}(\tau_{r+1}) = \mathbf{u}_0 + \mathbf{a}_r^{\mathrm{T}} \mathbf{F}_r,$$

where $\mathbf{a}_r = (S_\lambda(\boldsymbol{\tau}_{1:r}, \boldsymbol{\tau}_{1:r}) + \Lambda_r)^{-1} W_\lambda(\tau_{r+1}, \boldsymbol{\tau}_{1:r})$ is a $r \times 1$ vector which does not depend on $\boldsymbol{\theta}$. Also note that, in step 2(a), the common scalar predictive variance of $u_s(\tau_{r+1})$, for $s = 1, \ldots, S$, denoted as $C_r = C(\tau_{r+1}, \tau_{r+1})$, also does not depend on $\boldsymbol{\theta}$. This means we can pre-compute both $\mathbf{a}_r$ and C_r, for $r = 1, \ldots, N-1$, in advance of running the ACE algorithm.

In summary, before starting the ACE algorithm, an initial phase is completed as follows.

Initial phase

1. Set $\Lambda_1 = 0$.

2. For $r = 1, \ldots, N-1$ compute the following quantities:

$$
\begin{aligned}
\mathbf{a}_r &= \mathbf{B}_r W_\lambda(\tau_{r+1}, \boldsymbol{\tau}_{1:r})^{\mathrm{T}}, \\
C_r &= V_\lambda(\tau_{r+1}, \tau_{r+1}) - W_\lambda(\tau_{r+1}, \boldsymbol{\tau}_{1:r}) \left(S_\lambda(\boldsymbol{\tau}_{1:r}, \boldsymbol{\tau}_{1:r}) + \Lambda_r\right)^{-1} W_\lambda(\tau_{r+1}, \boldsymbol{\tau}_{1:r})^{\mathrm{T}}, \\
\Lambda_{r+1} &= \operatorname{diag}\left\{\Lambda_r, \dot{C}(\tau_{r+1}, \tau_{r+1})\right\},
\end{aligned}
$$

where

$$
\begin{aligned}
\dot{C}(\tau_{r+1}, \tau_{r+1}) &= S_\lambda(\tau_{r+1}, \tau_{r+1}) - S_\lambda(\tau_{r+1}, \boldsymbol{\tau}_{1:r}) \mathbf{B}_r S_\lambda(\tau_{r+1}, \boldsymbol{\tau}_{1:r})^{\mathrm{T}}, \\
\mathbf{B}_r &= \left(S_\lambda(\boldsymbol{\tau}_{1:r}, \boldsymbol{\tau}_{1:r}) + \Lambda_r\right)^{-1}.
\end{aligned}
$$

3. Compute $\mathbf{B}_N = (S_\lambda(\boldsymbol{\tau}, \boldsymbol{\tau}) + \Lambda_N)^{-1}$.

4. For $j = 1, \ldots, M$, compute the $n_j \times N$ matrix

$$\mathbf{D}_j = W_\lambda(\mathbf{t}_j, \boldsymbol{\tau}) \mathbf{B}_N,$$

and the $n_j \times n_j$ matrix

$$\mathbf{E}_j = V_\lambda(\mathbf{t}_j, \mathbf{t}_j) - W_\lambda(\mathbf{t}_j, \boldsymbol{\tau}) \mathbf{B}_N W_\lambda(\mathbf{t}_j, \boldsymbol{\tau})^{\mathrm{T}}.$$

Now, embedded in the ACE algorithm, we can produce a probabilistic solution for $\mathbf{u}_{ijl}$ for $i = 1, \ldots, 2B$, $j = 1, \ldots, M$ and $l = 1, \ldots, n_j$, as follows

Main phase

1. Let $\mathbf{f}_1 = \mathbf{f}(T_0, \mathbf{u}_{0j}, \boldsymbol{\theta}_i)$.

2. For $r = 1, \ldots, N-1$, complete the following steps.

(a) For $s = 1, \ldots, S$, generate

$$u_s(\tau_{r+1}) \sim \mathrm{N}\left(m_s(\tau_{r+1}), C_r\right),$$

where

$$\mathbf{m}(\tau_{r+1}) = \mathbf{u}_{0j} + \mathbf{a}_r^{\mathrm{T}} \mathbf{F}_r.$$

(b) Compute

$$\mathbf{f}_{r+1} = \mathbf{f}(\tau_{r+1}, \mathbf{u}(\tau_{r+1}), \boldsymbol{\theta}_i).$$

3. For $s = 1, \ldots, S$, generate

$$\mathbf{u}_s^* \sim \mathrm{N}\left(u_{0js} + \mathbf{D}_j \mathbf{F}_N, \mathbf{E}_j\right),$$

and let

$$\mathbf{u}_{ijl} = \left(u_{1l}^*, \ldots, u_{Sl}^*\right),$$

where u_{sl}^* is the sth element of $\mathbf{u}_s^*$.

In the above algorithm note the dependence on the initial values $\mathbf{u}_{0j}$, for the jth treatment, i.e. the initial values are known. In some situations, the initial values will be unknown and become part of the inference problem, i.e. they are given a prior distribution which we update to a posterior distribution in light of the experimental responses. If that is the case, then we can replace all occurrences of $\mathbf{u}_{0j}$ by a value, $\mathbf{u}_{0ji}$, generated from their prior distribution, as we do with the unknown parameters.

4 Illustrative Examples

4.1 Preliminaries

In this section we demonstrate the extended ACE algorithm on three illustrative examples featuring systems of ODEs:

1. Compartmental model (Section 4.2);

2. FitzHugh-Nagumo equations (Section 4.3);

3. JAK-STAT mechanism (Section 4.4).

In each example, designs are found under the three different loss functions as described below. For each loss function, let $\left\{\tilde{\boldsymbol{\psi}}_j\right\}_{j=1}^B$, where $\tilde{\boldsymbol{\psi}}_i = \left(\tilde{\boldsymbol{\theta}}_i, \tilde{\boldsymbol{\gamma}}_i\right)$, is an additional sample generated from the prior distribution of $\boldsymbol{\psi}$.

- **Squared error loss** (as described in Section 2.2).

- **Absolute error loss** (AEL) given by

$$\lambda_{AEL}(\mathbf{y}, \boldsymbol{\theta}, \mathbf{d}) = \sum_{l=1}^p |\theta_l - \mathrm{M}(\theta_l | \mathbf{y})|,$$

where $M(\theta_l|\mathbf{y})$ is the posterior median of θ_l. The posterior median is intractable and is approximated as follows. For $j = 1, \ldots, B$, let

$$w_j = \frac{\pi(\mathbf{y}|\tilde{\boldsymbol{\psi}}_j)}{\sum_{j=1}^{B} \pi(\mathbf{y}|\tilde{\boldsymbol{\psi}}_j)},$$

and let $\tilde{\theta}_{l(1)} \leq \cdots \leq \tilde{\theta}_{l(B)}$ be the ordered values of θ_l in the sample $\left\{\tilde{\boldsymbol{\theta}}_j\right\}_{j=1}^{B}$. Then an approximation to the median is given by

$$\hat{M}(\theta_l|\mathbf{y}) = \frac{1}{2}\left(\tilde{\theta}_{l(z)} + \tilde{\theta}_{l(z+1)}\right),$$

where $z = \max\{j = 1, \ldots, B | w_j \leq 1/2\}$. We then approximate the absolute error loss function by replacing $M(\theta_l|\mathbf{y})$ by $\hat{M}(\theta_l|\mathbf{y})$.

- **Self-information loss** (SIL) given by

$$
\begin{aligned}
\lambda_{\mathrm{SIL}}(\boldsymbol{\psi}, \mathbf{y}, \mathbf{d}) &= \log \pi(\boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}), \\
&= \log \pi(\mathbf{y}|\mathbf{d}) - \log \pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d}), \\
&= \log \int \pi(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{d})\pi(\boldsymbol{\gamma}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\gamma}\mathrm{d}\boldsymbol{\theta} - \log \int \pi(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{d})\pi(\boldsymbol{\gamma})\mathrm{d}\boldsymbol{\gamma}. \qquad (8)
\end{aligned}
$$

Both of the integrals in (8), denoted as

$$
\begin{aligned}
I_1 &= \int \pi(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{d})\pi(\boldsymbol{\gamma})\mathrm{d}\boldsymbol{\gamma}, \\
I_2 &= \int \pi(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{d})\pi(\boldsymbol{\gamma}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\gamma}\mathrm{d}\boldsymbol{\theta},
\end{aligned}
$$

are analytically intractable. However we can approximate them using Monte Carlo integration as follows.

$$
\begin{aligned}
\hat{I}_1 &= \frac{1}{B}\sum_{j=1}^{B} \pi(\mathbf{y}|\boldsymbol{\theta}, \tilde{\boldsymbol{\gamma}}_j), \\
\hat{I}_2 &= \frac{1}{B}\sum_{j=1}^{B} \pi(\mathbf{y}|\tilde{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{\gamma}}_j).
\end{aligned}
$$

We then replace evaluation of I_1 and I_2 in the self-information loss function by $\hat{I}_1$ and $\hat{I}_2$, respectively.

Similar to approximating the squared error loss in Section 3.2, both of the nested Monte Carlo schemes for approximating the absolute error and self-information loss require $2B$ evaluations of $\mathbf{u}(t)$ for each treatment and time point, which we now replace by a value generated from the probabilistic solution. For the probabilistic solution, the discrete grid of points denoted by $\boldsymbol{\tau}$ will be a set of N equally-spaced points where the absolute difference between any two values is denoted by h. Unless stated otherwise, for the auxiliary parameters, $\lambda = 4h$.

4.2 Compartmental model

Compartmental models are used in pharmokinetics to understand how drugs behave inside a body. The open one-compartment model with first-order absorption is described by the following system of $S = 2$ ODEs for $t \in [0, 24]$ hours

$$
\begin{aligned}
\dot{u}_1(t) &= -\theta_1 u_1(t) \\
\dot{u}_2(t) &= (\theta_2/\theta_3)u_1(t) - \theta_2 u_2(t) \\
\mathbf{u}(0) &= (D, 0)
\end{aligned}
\tag{9}
$$

where $u_1(t)$ and $u_2(t)$ are the amounts of drug outside and inside the body respectively, D is the known initial dose and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_2)$ are unknown parameters.

The system in (9) is a homogenous linear system with constant coefficients meaning it can be solved analytically as

$$
\begin{aligned}
u_1(t) &= D \exp(-\theta_1 t), \\
u_2(t) &= \frac{D\theta_2}{\theta_3(\theta_2 - \theta_1)} \left(\exp(-\theta_1 t) - \exp(-\theta_2 t) \right).
\end{aligned}
$$

For this example, we find and compare designs under the three loss functions from Section 4.1, and under both the exact and probabilistic solutions to the ODEs.

This type of compartmental model (or variants of it) is often used to demonstrate optimal experimental design methodology (see, for example, Atkinson et al. 1993; Gotwalt et al. 2009; Ryan et al. 2014; Overstall and Woods 2015). We follow the setup of Ryan et al. (2014) and Overstall and Woods (2015) where $D = 400$, $n = 15$ and

$$
\log \theta_l \sim \mathrm{N}(\mu_i, 0.05),
$$

independently, for $l = 1, 2, 3$ with $(\mu_1, \mu_2, \mu_3) = (\log 0.1, \log 1, \log 20)$. The amount of drug inside the body, y_i is observed at observation time t_i and we assume the following statistical model

$$
y_i \sim \mathrm{N}\left(u_2(t_i), \sigma^2 + \tau^2 u_2(t_i)^2\right),
\tag{10}
$$

independently, where $\sigma^2 = 0.1$ and $\tau^2 = 0.01$. Note that (10) implies that $\mathcal{G}(\mathbf{u}, \boldsymbol{\theta}) = u_2$. The design only involves the n observation times: $t_1, \ldots, t_n$. An added stipulation is that the observation times have to be at least 15 minutes apart. Such constraints are straightforward to incorporate into the ACE algorithm (see Overstall and Woods 2015).

Since we know that $\mathbf{u}(t)$ is smooth we employ the squared exponential kernel function. The discrete grid, $\boldsymbol{\tau}$, is of size $N = 501$. The auxiliary parameter α is fixed at N.

For each loss function, we compare the designs found under the exact and probabilistic solutions (using ACE) to a uniform designs of $n = 15$ equally-spaced time points in $[0, 24]$ hours. The top row of Figure 2 shows boxplots of twenty evaluations of the Monte Carlo approximation to the expected loss for the uniform design and the optimal design found under each of the loss functions. There is negligible difference between the designs found under the exact and probabilistic solutions and these designs are clearly superior to the uniform designs. In the bottom row of Figure 2 are the observation time points associated with the designs under comparison. The optimal designs appear to favour having a set of observation times near the peak of $u_2(t)$ at $t \approx 2.5$ hours and then a series of observation times at the end of the observation interval. Of the points near the peak, the optimal design under self-information loss has two distinct sets whereas the designs under squared and absolute error loss have just one set of points.

Figure 2: Plots summarising the results from the compartmental model in Section 4.2. The top row show boxplots of 20 evaluations of the Monte Carlo approximation to the expected loss for the uniform design and the optimal designs (for the exact and probabilistic solution) found under each of the loss functions. The bottom plot shows the three designs found under each of the loss functions and the uniform design. In the background to the plot in the bottom row is 100 draws from the exact solution, $u_2(t)$, giving the amount of drug at time t, for 100 values drawn from the prior distribution of $\boldsymbol{\theta}$.

14
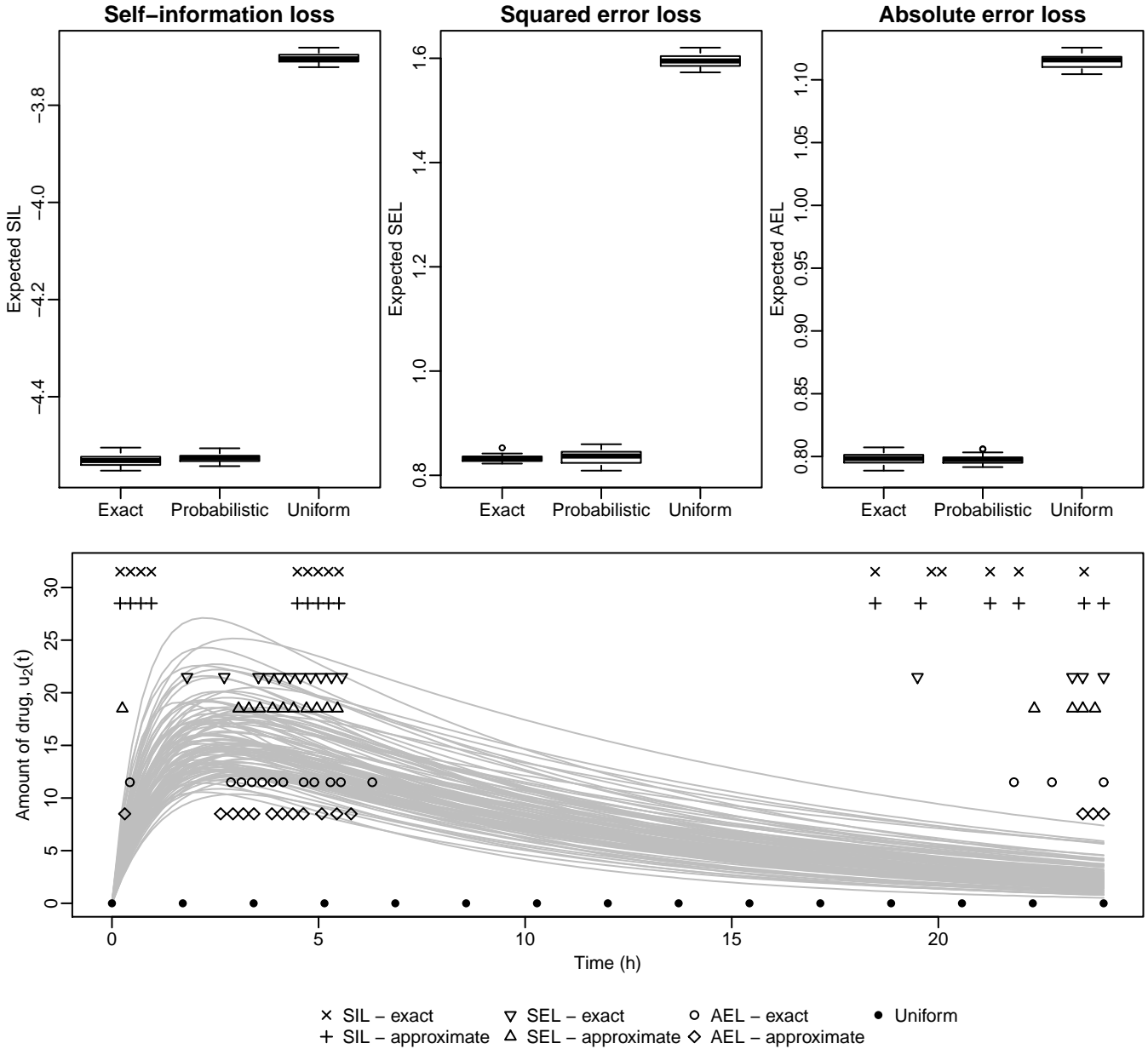
4.3 FitzHugh-Nagumo equations

The FitzHugh-Nagumo equations (FitzHugh 1961 and Nagumo et al. 1962) aim to describe the behaviour of spike potential in the giant axon of squid neurons. They are given by the following system of $S = 2$ ODEs for $t \in [0, 20]$ms

$$
\begin{aligned}
\dot{u}_1(t) &= \theta_3 \left(u_1(t) - u_1(t)^3/3 + u_2(t) \right) \\
\dot{u}_2(t) &= - \left(u_1(t) - \theta_1 + \theta_2 u_2(t) \right) / \theta_3 \\
\mathbf{u}(0) &= (-1, 1)
\end{aligned}
$$

where $u_1(t)$ is the voltage across the axon membrane, $u_2(t)$ is the recovery variable giving a summary of outward current and $\boldsymbol{\theta} = (\theta_1, \theta_1, \theta_3)$.

The experiment involves measuring the voltage, y_i, at time t_i, for $i = 1, \ldots, n = 21$. Following Ramsay et al. (2007), the following statistical model is assumed

$$
y_i \sim \mathrm{N}\left(u_1(t_i), \sigma^2 \right), \tag{11}
$$

independently, where $\sigma \sim \mathrm{U}[1/2, 1]$. Furthermore, we assume the following prior distributions for the unknown parameters: $\theta_1 \sim \mathrm{U}[0, 1]$, $\theta_2 \sim \mathrm{U}[0, 1]$ and $\theta_3 \sim \mathrm{U}[1, 5]$. In (11), $\mathcal{G}(\mathbf{u}, \boldsymbol{\theta}) = u_1$.

As noted by Ramsay et al. (2007), the solution to the FitzHugh-Nagumo equations can alternate between smooth evolution and sharp changes of direction. For this reason we employ the uniform kernel. The discrete grid is of size $N = 200$ and the auxiliary parameter α is fixed as N.

The design consists of the n observation times: $t_1, \ldots, t_n$. Similar to Section 4.2, we stipulate that the observation times have to be at least 0.25ms apart. We find designs under each of the loss functions in Section 4.1. In each case, we compare the optimal design to a uniform design of n equally spaced points in $[0, 20]$ms. Figure 3 shows boxplots of twenty evaluations of the Monte Carlo approximation to the expected loss for the uniform design and the optimal design found under each of the loss functions. In each case, there is a clear improvement to be made on using a uniform design. Also shown in Figure 3 are the four designs under comparison. Both the squared and absolute error optimal designs have a significant number of frequent observations at the beginning of the experiment. For example these designs have around a third of their observation times before 2.5ms. On the other hand, the self-information loss and uniform designs only make 3 observations before this time. A feature of all of the optimal designs is that they make no observations between about 2.5 and 6ms. The remaining observation times appear roughly evenly spaced. The background to this plot shows 100 draws from the probabilistic solution, $u_1(t)$, giving the voltage at time t, for 100 values drawn from the prior distribution of $\boldsymbol{\theta}$. It appears that the initial phase of high frequency observations is to learn about the steep increase in voltage for small t. The remaining set of evenly spaced observation times is for learning when the voltage has high noise due to parameter uncertainty.

4.4 JAK-STAT mechanism

The JAK-STAT mechanism describes a set of biochemical reactions of STAT-5 transcription factors in response to binding of the Erythropoietin hormone to cell surface receptors. The system of $S = 4$

Figure 3: Plots summarising the results from the FitzHugh-Nagumo equations in Section 4.3. The top row show boxplots of 20 evaluations of the Monte Carlo approximation to the expected loss for the uniform design and the optimal design found under each of the loss functions. The bottom plot shows the three designs found under each of the loss functions and the uniform design. In the background to the plot in the bottom row is 100 draws from the probabilistic solution, $u_1(t)$, giving the voltage at time t, for 100 values drawn from the prior distribution of $\boldsymbol{\theta}$.
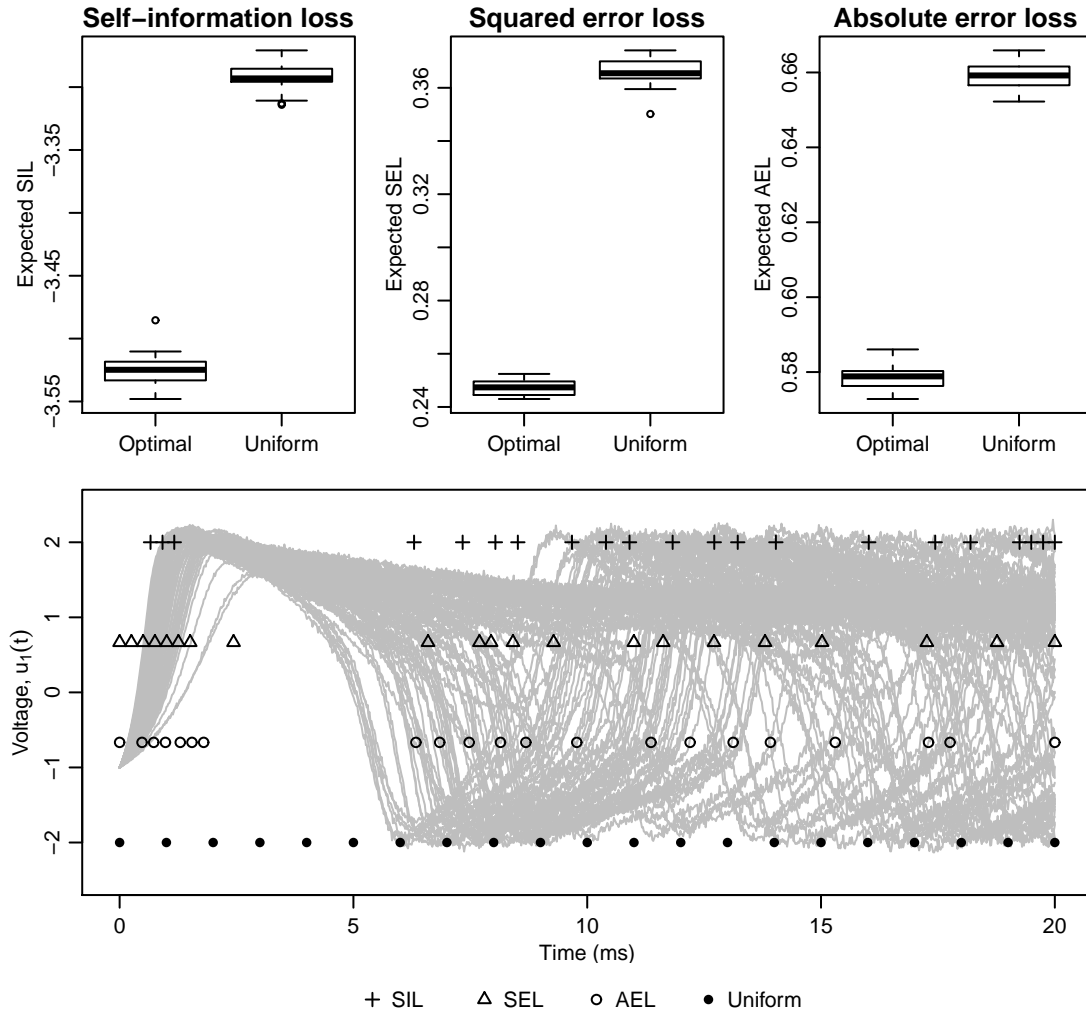
ODEs, for $t \in [0, 60]$, is

$$
\begin{aligned}
\dot{u}_1(t) &= -\theta_1 u_1(t)\kappa(t) + 2\theta_4 u_4(t - \omega) \\
\dot{u}_2(t) &= \theta_1 u_1(t)\kappa(t) - \theta_2 u_2(t)^2 \\
\dot{u}_3(t) &= -\theta_3 u_3(t) + \frac{1}{2}\theta_2 u_2(t)^2 \\
\dot{u}_4(t) &= \theta_3 u_3(t) - \theta_4 u_4(t - \omega) \\
\mathbf{u}(0) &= (u_{01}, 0, 0, 0)
\end{aligned}
$$

where u_{01} and ω are unknown, and $u_4(t) = 0$ for $t \in [-\omega, 0]$. The examples thus far have been initial value problems (IVPs) whereas the system above is an example of a delay initial function problem (DIFP). After gene activation in the cell nucleus, the transcription factors revert to their initial state, returning to the cytoplasm for the next activation cycle. This stage is explained by the unknown time delay denote by ω.

The function $\mathcal{G}$ is given by

$$
\mathcal{G}(\mathbf{u}, \boldsymbol{\theta}) = \begin{pmatrix} \theta_5(u_2 + 2u_3) \\ \theta_6(u_1 + u_2 + 2u_3) \\ u_1 \\ u_3/(u_2 + u_3) \end{pmatrix}.
$$

An experiment has already been conducted (Swameye et al., 2003) which consisted of two series of $n = 16$ measurements of the first two elements ($\mathcal{G}_1$ and $\mathcal{G}_2$) of $\mathcal{G}$ at a set of distinct observation times: $t_1, \ldots, t_n$. A further observation of $\mathcal{G}_3$ is made at time $t = 0$ and of $\mathcal{G}_4$ at time t^*. The following statistical model is assumed

$$
\begin{aligned}
(y_{1i}, y_{2i}) &\sim \mathrm{N}\left(\mathcal{G}(\mathbf{u}(t_i), \boldsymbol{\theta})_{1:2}, \mathbf{C}_i\right), \\
y_3 &\sim \mathrm{N}\left(\mathcal{G}(\mathbf{u}(t_i), \boldsymbol{\theta})_3, \sigma_3^2\right), \\
y_4 &\sim \mathrm{N}\left(\mathcal{G}(\mathbf{u}(t_i), \boldsymbol{\theta})_4, \sigma_4^2\right)
\end{aligned}
$$

independently, for $i = 1, \ldots, n$, where $\mathbf{C}_i = \mathrm{diag}\{\sigma_{1i}^2, \sigma_{2i}^2\}$. See Raue et al. (2009) and Chkrebtii et al. (2015) for analyses of these data. In these papers, σ_{1i}^2, σ_{2i}^2, σ_3^2 and σ_4^2 (for $i = 1, \ldots, n$) are experimentally determined, i.e. they are known.

The design task considered here will be to optimally find a follow-up design based on information on the parameters from the existing data. We assume the same statistical model as above and find $t_1, \ldots, t_n$ and t^*. In the terminology of Sections 2 and 3, the prior distribution for $\boldsymbol{\theta}$, ω and u_{01} is the posterior distribution for the existing data as per the analysis of Chkrebtii et al. (2015). Instead of the variance parameters being fixed, we assume that $\sigma_{1i}^2 = \sigma_1^2$, $\sigma_{2i}^2 = \sigma_2^2$, for all $i = 1, \ldots, n$ and

$$
\begin{aligned}
\sigma_1 &\sim \mathrm{U}[0, 0.1] \\
\sigma_2 &\sim \mathrm{U}[0, 0.1] \\
\sigma_3 &\sim \mathrm{U}[0, 20] \\
\sigma_4 &\sim \mathrm{U}[0, 0.1].
\end{aligned}
$$

These prior distributions are consistent with the experimentally determined values from the original analysis.

Note that the system of ODEs depends on the forcing function $\kappa(t)$. This is unknown, but has been measured at 16 time points. Following Chkrebtii et al. (2015), we assume these measurements are without error and use a GP to give a probabilistic approximation to $\kappa(t)$ at any value of $t \in [0, 60]$.

The nature of the DIFP does introduce an added complexity to our implementation of the probabilistic solution. In step 2(b) of the main phase of the algorithm in Section 3.2 we are required to compute

$$\mathbf{f}_{r+1} = \mathbf{f}(\tau_{r+1}, \mathbf{u}(\tau_{r+1}), \boldsymbol{\theta}_i).$$

In an IVP, this is dependent on $\mathbf{u}(\tau_{r+1})$ which is generated from a normal distribution in step 2(a). However in the DIFP, to compute $\mathbf{f}_{r+1}$, we also need $u_4(\tau_{r+1} - \omega_i)$, where ω_i is a value generated from the prior (posterior from original analysis) distribution of ω. If $\tau_{r+1} - \omega_i \leq 0$, then $u_4(\tau_{r+1} - \omega_i) = 0$ by the initial conditions of the system of ODEs. For $\tau_{r+1} - \omega_i > 0$, in the probabilistic solution of Chkrebtii et al. (2015), the conditional distribution of $u_4(\tau_{r+1} - \omega_i)$ can be derived and a value generated. However, this will be computationally expensive to incorporate in the implementation of the probabilistic solution described in Section 3.2. Instead, if $\tau_{r+1} - \omega_i > 0$, then we replace $u_4(\tau_{r+1} - \omega_i)$ by $u_4(\tau_{\bar{r}})$ where

$$\bar{r} = \arg \min_{r'=1,\ldots,r+1} |\tau_{r+1} - \omega_i - \tau_{r'}|,$$

i.e. from the series of $u_4(\tau_1), \ldots, u_4(\tau_{r+1})$ generated in step 2(a), thus far, we choose the value at time $\tau_{\bar{r}}$ that is "closest" to $\tau_{r+1} - \omega_i$.
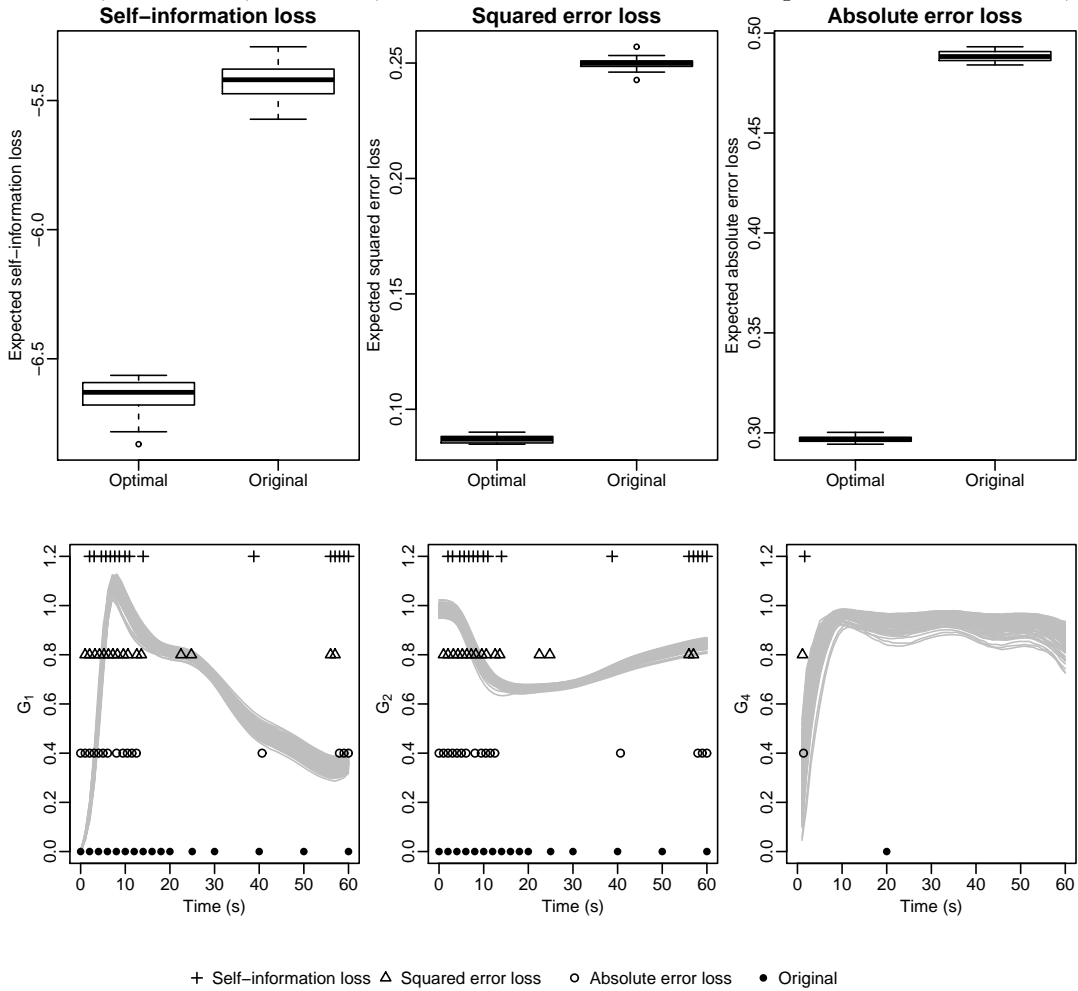
As noted by Chkrebtii et al. (2015), the time delay can cause discontinuities in the derivative meaning we emply the uniform kernel function. The discrete grid, $\boldsymbol{\tau}$, is of size $N = 500$, and the auxiliary parameters are $\lambda = 0.085$ and $\alpha = 8000$ which are consistent with their posterior distribution from the original analysis.

We use the extended ACE algorithm to generate designs under the three loss functions from Section 4.1 and compare these designs to the original design used in the experiment of Swameye et al. (2003). As in the previous examples, the observation times need to be at least 1 second apart. Figure 4 shows boxplots of twenty evaluations of the Monte Carlo approximation to the expected loss for the original design and the optimal design found under each of the loss functions. In each case, there is a clear improvement to be made on repeating the experiment with the same set of observation times. Also shown in Figure 4 are the four designs under comparison. Clearly the optimal designs favour having a dense set of points early in the observation window and then a smaller set of times at the end of the window. This is especially true for the designs under the squared and absolute error loss where 75% of the observation times are less than 15 seconds, compared to about 60% for self-information loss and 50% for the original design. It appears that the early observation times can learn about the peak in $\mathcal{G}_1$ and the sharp decrease in $\mathcal{G}_2$ at and up to 10 seconds, respectively. For the single observation time, t^*, of $\mathcal{G}_4$, the optimal designs clearly favour making a very early observation. Note that t^* for each of the optimal designs is between 1 and 2 seconds.

5   Application: Transport of serine across human placenta

Now consider the human placenta example introduced in Sections 1.2 and 2.1. By the protocol of the experiment, the initial amounts of radioactive serine inside (u_{01}) and outside (x_1) are fixed as 0 and 7.5, respectively.

Figure 4: Plots summarising the results from the JAK-STAT example in Section 4.4. The top row show boxplots of 20 evaluations of the Monte Carlo approximation to the expected loss for the original design and the optimal design found under each of the loss functions. The bottom plot shows the three designs found under each of the loss functions and the original design. In the background to the plot in the bottom row is 100 draws from $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_4$, at time t, for 100 values drawn from the prior distribution of $\boldsymbol{\theta}$, u_{01} and ω.

+ Self−information loss   △ Squared error loss   ○ Absolute error loss   • Original

19

We consider $M = 2, \ldots, 7$ placentas and let $n_j = n^* = 8$, for all $j = 1, \ldots, M$. We also fix $t_{jl} = t_l$ for $l = 1, \ldots, n_j$ and $j = 1, \ldots, M$. The following statistical model is assumed for the experimental responses

$$y_{jl} = u_1(t_l; \boldsymbol{\theta}_j, \mathbf{x}_j) + \epsilon_{jl}, \tag{12}$$

for $j = 1, \ldots, M$ and $l = 1, \ldots, n^*$, where $\mathbf{x}_j = (x_1, x_{2j})$,

$$\epsilon_{jl} \overset{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma^2\right),$$

and $\boldsymbol{\theta}_j$ are the physical parameters for the jth placenta. For $d = 1, \ldots, p$, we specify a multiplicative hierarchical prior structure for $\boldsymbol{\theta}_j$ as

$$\theta_{jd} \overset{\text{iid}}{\sim} \mathrm{U}\left[\theta_d\left(1 - c_d\right), \theta_d\left(1 + c_d\right)\right],$$

where $c_d \overset{\text{iid}}{\sim} \mathrm{U}[0, 0.05]$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ are the population physical parameters. For the latter, we assume the following independent prior distributions for each element

$$\theta_d \sim \mathrm{Tri}[a_l, b_l],$$

where $\mathrm{Tri}[a, b]$ denotes the symmetric triangle distribution on the interval $[a, b]$. We assume $a_1 = a_2 = a_4 = 80$, $b_1 = b_2 = b_4 = 120$, $a_2 = 0.02$ and $b_2 = 0.08$. This reflects actual prior knowledge on the value of these parameters from past experiments. For the response variance, we assume

$$\sigma^2 \sim \mathrm{U}[0, 1].$$

We expect the solution to be smooth so use the squared exponential kernel function. The discrete grid, $\boldsymbol{\tau}$, is of size $N = 601$ and the auxiliary parameter $\alpha = 10N$.

Specifying a design corresponds to specifying the $M = 7$ experimental conditions $x_{21}, \ldots, x_{2M}$, and initial values $u_{021}, \ldots, u_{02M}$, as well as the common $n^* = 8$ observation times $t_1, \ldots, t_{n^*}$. Therefore, the design space has 22 dimensions.

For each value of M, we use the ACE algorithm to find designs under the three loss functions from Section 4.1. In addition, suppose a question of interest concerns whether the reaction rates are symmetric, i.e. is $\theta_3 = \theta_4$? To answer this question, we define two models: m_1 (where $\theta_3 = \theta_4$) and m_2 (where $\theta_3 \neq \theta_4$). An appropriate loss function, termed the Model selection loss (MSL), is

$$\lambda(\mathbf{y}, m, \mathbf{d}) = 1 - I(\hat{m} = m),$$

where $m \in \mathcal{M} = \{m_1, m_2\}$ denotes the model and $\hat{m} = \mathrm{argmax}_{m \in \mathcal{M}} \pi(m|\mathbf{y})$ is the model that maximises the posterior model probability. The posterior model probability of model m is given by

$$\pi(m|\mathbf{y}) = \frac{\pi(\mathbf{y}|m)\pi(m)}{\sum_{j \in \mathcal{M}} \pi(\mathbf{y}|m_j)\pi(m_j)},$$

where $\pi(m)$ is the prior model probability of model m and

$$\pi(\mathbf{y}|m_j) = \int \pi(\mathbf{y}|\boldsymbol{\theta}^j, \boldsymbol{\gamma})\pi(\boldsymbol{\theta}^j, \boldsymbol{\gamma})\mathrm{d}\boldsymbol{\theta}^j \mathrm{d}\boldsymbol{\gamma} \tag{13}$$

for $j = 1, 2$, with $\boldsymbol{\theta}^j$ being the parameters under m_j. Under $j = 1, 2$, the integration in (13) is intractable but can be approximated using a Monte Carlo approximation in a similar way to I_2

Figure 5: Plots summarising the results from the placenta example in Section 5. Shown are boxplots of 20 evaluations of the Monte Carlo approximation to the expected loss for the proposed design and the optimal design found under each of the loss functions and each value of M.
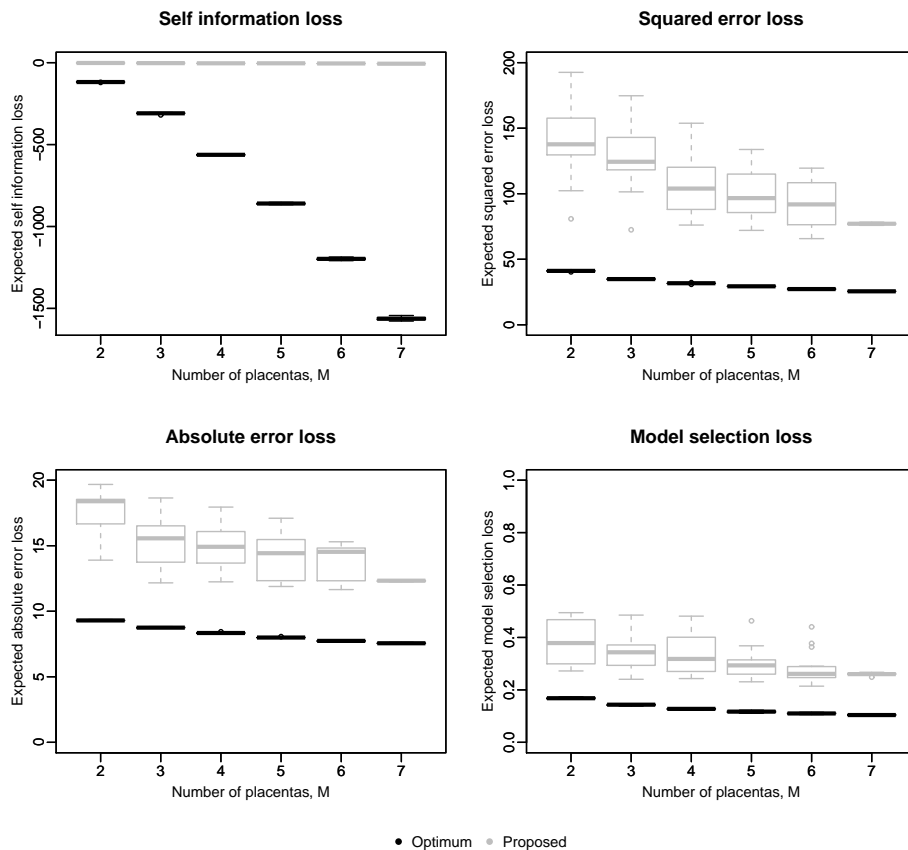
Table 2: Initial concentrations (to nearest integer) of non-radioactive serine inside ($u_{02} = u_2(0)$) and outside (x_2) each of the $M = 7$ placentas for the optimal designs found under the four loss functions and the proposed design.

Placenta	Self-information		Squared error		Absolute error		Model selection		Proposed	
	x_2	u_{02}	x_2	u_{02}	x_2	u_{02}	x_2	u_{02}	x_2	u_{02}
1	0	1000	0	0	0	0	0	0	0	0
2	0	1000	0	0	0	0	0	105	250	0
3	0	1000	0	58	0	40	0	169	250	250
4	0	1000	0	56	0	67	302	0	250	1000
5	0	1000	217	952	188	932	306	457	1000	0
6	0	1000	215	1000	184	1000	755	1000	1000	250
7	0	1000	194	1000	207	1000	570	117	1000	1000

under the self-information loss. In each case, as in the previous examples, the observation times need to be at least 5 seconds apart.

Figure 5 shows boxplots of twenty evaluations of the Monte Carlo approximation to the expected loss for the optimal design found under each loss function plotted against $M = 2, \ldots, 7$. For $M = 7$, also shown is a boxplot of twenty evaluations of the Monte Carlo approximation for a design (see Table 2) proposed by the biologists at the Southampton Centre for Biological Sciences. For $M = 2, \ldots, 6$, the corresponding boxplots show evaluations of the Monte Carlo approximation for twenty designs formed by randomly choosing M rows from the proposed design. As expected, the expected loss decreases with the number of placentas, M. Also, the optimal designs are clearly superior to the proposed designs. However, it should be noted that, for each loss function, the optimal design is expected to gain more information from $M = 2$ placentas than the proposed design is expected to do so from $M = 7$ placentas.

We now look at the case of $M = 7$ placentas in more detail. The initial concentrations of non-radioactive serine inside (u_{02}) and outside (x_2) each of the $M = 7$ placentas for the optimal designs found under each of the four loss functions are shown in Table 2. Also shown are the conditions for the design proposed by the biologists. Figure 6 show the optimal (for each loss function) and proposed observation times. Each row shows 100 draws from $u_1(t)$ plotted against t for 100 draws from the prior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, for each placenta and initial concentrations shown in Table 2.

Under the self-information loss, Table 2 shows that the optimal design has replicates of the same initial concentrations of non-radioactive serine. Figure 6 shows that, compared to the other designs, these initial concentrations lead to a larger maximum value of $u_1(t)$. Note that the between-placenta variability of $u_1(t)$ in Figure 6 is due to the placenta-specific physical parameters, $\boldsymbol{\theta}_j$. The optimal observation times under the self-information loss occur in two clusters, just before and after the peak in $u_1(t)$.

The designs under the squared and absolute error loss functions appear similar. In both cases they are superior to the proposed design. The initial concentrations in Table 2 lead to three distinct profiles of $u_1(t)$ (Placentas 1 and 2; 3 and 4; 5, 6 and 7). The profile for placentas 1 and 2 has a slow steady increase in $u_1(t)$ with respect to t. Placentas 3 and 4 have a steep initial increase and subsequent decrease in $u_1(t)$ with respect to t. Finally, placentas 5 to 7 has a steep initial increase in $u_1(t)$ with respect to t followed by a slow decrease. The optimal observation times are predominantly at the beginning of the observation window.

Figure 6: Plots summarising the results from the placenta example in Section 5. Shown are 100 draws from $u_1(t)$ plotted against t for 100 values drawn from the the prior distribution of $\boldsymbol{\theta}$, for each of the $M = 7$ placentas and experimental conditions given by Table 2.

23
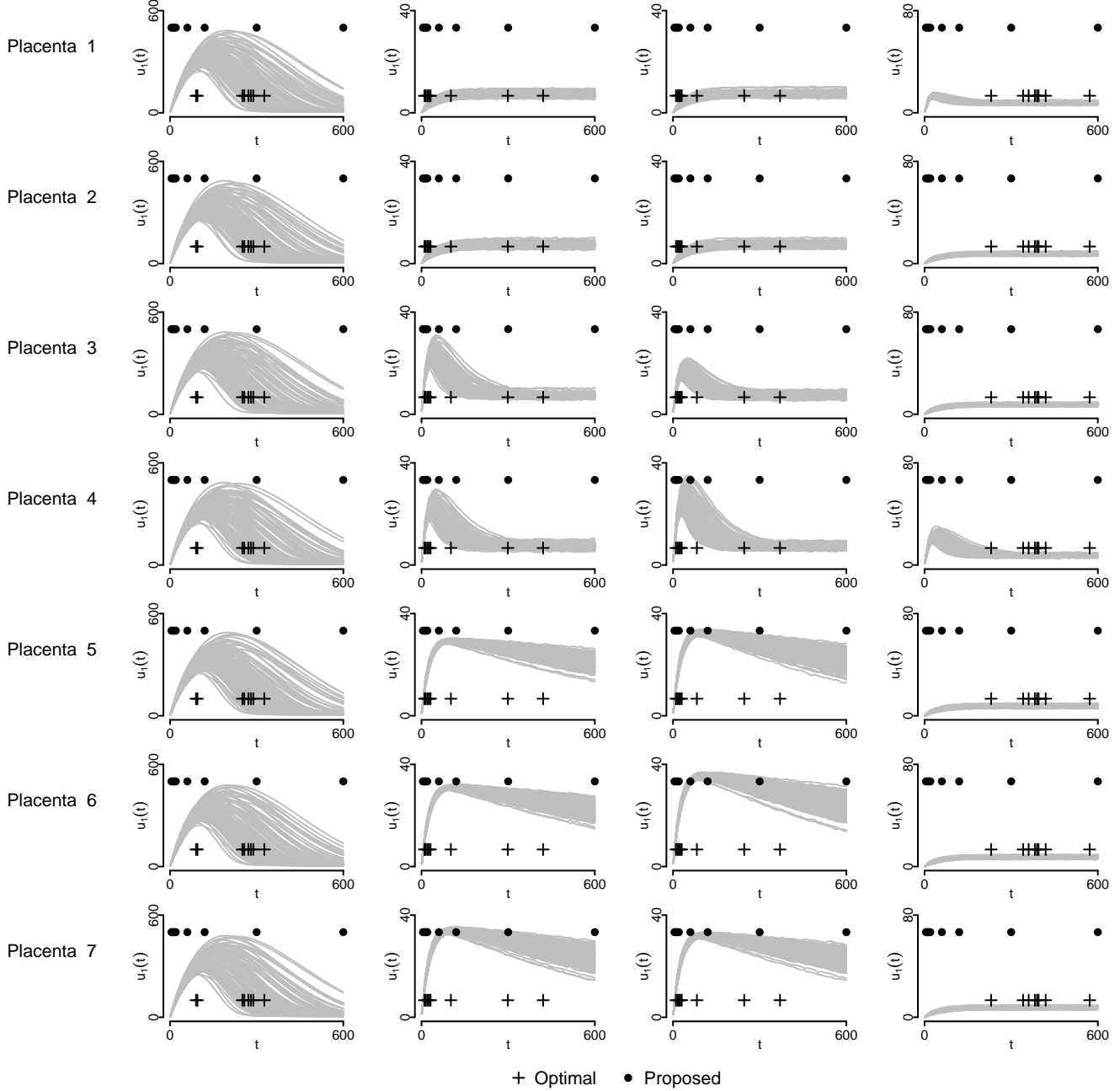
The initial concentrations of the optimal design under the model selection loss result in two distinct profiles of $u_1(t)$. For placentas 2, 3 and 5 to 7, $u_1(t)$ has a slow steady increase in $u_1(t)$ with respect to t. Placentas 1 and 4 have a steep initial increase and subsequent decrease in $u_1(t)$ with respect to t. Opposite to the other loss functions, the optimal observation times are predominantly towards the end of the observation window.

6 Concluding Remarks

This paper introduces an extension of the ACE algorithm for Bayesian optimal design so that the challenging problem of experimental design for ODE models can now be attempted.

The method relies on the probabilistic solution to a system of ODEs as recently proposed by Chkrebtii et al. (2015) and is demonstrated on four examples where the goal of the experiment is estimating unknown physical parameters.

One key issue not addressed is model discrepancy (Kennedy and O 'Hagan, 2001). This is a systematic mis-match between the true physical process and the solution to the ODEs. Not taking account of this bias can lead to significant bias in the posterior estimates of the physical parameters (Brynjarsdottir and O 'Hagan, 2014). Future work will focus on Bayesian optimal design for physical models including model discrepancy.

A The ACE algorithm

1. Choose an initial design $\mathbf{d}^0 = \left(d_1^0, \ldots, d_q^0\right)$ and set the current design to be $\mathbf{d}^C = \left(d_1^C, \ldots, d_q^C\right) = \mathbf{d}^0$.

2. For each element $i = 1, \ldots, q$ of $\mathbf{d}$:

   (a) Let $L^i(d) = L(d_1^C, \ldots, d_{i-1}^C, d, d_{i+1}^C, \ldots, d_q^C)$ be the function given by the expected loss function which only varies over the design space, $\mathcal{D}_i$, for the ith element.

   (b) For $j = 1, \ldots, Q$, evaluate
   $$z_j = \hat{L}_B^i(d_j),$$
   for $\{d_1, \ldots, d_Q\} \in \mathcal{D}_i$. Fit a GP emulator to $\{z_j, d_j\}_{j=1}^Q$ and set $\tilde{L}^i(d)$ to be the resulting predictive mean.

   (c) Find
   $$d_i^* = \operatorname{argmin}_{d \in \mathcal{D}_i} \tilde{L}^i(d),$$
   and let $\mathbf{d}^* = \left(d_1^C, \ldots, d_{i-1}^C, d^*, d_{i+1}^C, \ldots \ldots, d_q^C\right)$ be the proposed design.

   (d) For $i = j, \ldots, B$, set
   $$\begin{aligned} \lambda_j^C &= \lambda(\boldsymbol{\psi}_j^C, \mathbf{y}_j^C, \mathbf{d}^C), \\ \lambda_j^* &= \lambda(\boldsymbol{\psi}_j^*, \mathbf{y}_j^*, \mathbf{d}^*), \end{aligned}$$
   where $\{\boldsymbol{\psi}_j^C, \mathbf{y}_j^C\}_{i=1}^B$ and $\{\boldsymbol{\psi}_j^*, \mathbf{y}_j^*\}_{j=1}^B$ are samples generated from $\boldsymbol{\psi}, \mathbf{y} | \mathbf{d}^C$ and $\boldsymbol{\psi}, \mathbf{y} | \mathbf{d}^*$, respectively.

(e) Calculate

$$p^* = 1 - F\left(-\frac{\sum_{i=j}^{B} \lambda_i^C - \sum_{i=1}^{B} \lambda_j^*}{\sqrt{2B\hat{v}}}\right),$$

where $F(\cdot)$ is the distribution function of the t-distribution with $2B - 2$ degrees of freedom,

$$\hat{v} = \frac{\sum_{j=1}^{B}(\lambda_j^C - \bar{\lambda}^C)^2 + \sum_{j=1}^{B}(\lambda_j^* - \bar{\lambda}^*)^2}{2B - 2},$$

and $\bar{\lambda}^C$ and $\bar{\lambda}^*$ are the sample means of the λ_j^C's and λ_j^*'s, respectively.

(f) Set $\mathbf{d}^C = \mathbf{d}^*$ with probability p^*.

3. Return to step 2, until convergence.

Convergence can be assessed informally using trace plots of the evaluations of either $\bar{\lambda}^*$ or $\bar{\lambda}^C$ at step 2(e), dependent on whether the proposed design was accepted or not, respectively.

The ACE algorithm should be started from multiple different starting designs $\mathbf{d}^0$. Out of the resulting designs, the one with the lowest value of $\hat{L}_B(\mathbf{d})$ is returned.

References

Atkinson, A., Chaloner, K., Herzberg, A., and Juritz, J. (1993), "Experimental Designs for Properties of a Compartmental Model," *Biometrics*, 49, 325–337.

Bastos, L. and O'Hagan, A. (2009), "Diagnostics for Gaussian Process Emulators," *Technometrics*, 51, 425–438.

Bliznyuk, N., Ruppert, D., Shoemaker, C., Regis, R., Wild, S., and Mugunthan, P. (2008), "Bayesian calibration of computationally expensive models using optimization and radial basis function approximations." *Journal of Computational and Graphical Statistics*, 17, 270–294.

Brynjarsdottir, J. and O 'Hagan, A. (2014), "Learning about physical parameters: The importance of model discrepancy," *Inverse Problems*, 30, 114007.

Chkrebtii, O., Campbell, D., Girolami, M., and Calderhead, B. (2015), "Bayesian Uncertainty Quantification for Differential Equations," Tech. rep., Ohio State University, USA.

Fielding, M., Nott, D., and Liong, S. (2011), "Efficient MCMC schemes for computationally expensive posterior distributions," *Technometrics*, 53, 16–28.

FitzHugh, R. (1961), "Impulses and physiological states in models of nerve membrane," *Biophysical Journal*, 1, 445–466.

Gotwalt, C., Jones, B., and Steinberg, D. (2009), "Fast Computation of Designs Robust to Parameter Uncertainty for Nonlinear Settings," *Technometrics*, 51, 88–95.

Iserles, A. (2009), *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press.

Jones, D., Schonlau, M., and Welch, W. (1998), "Efficient global optimization of expensive black-box functions," *Journal of Global Optimization*, 13, 455–492.

Kennedy, M. C. and O 'Hagan, A. (2001), "Bayesian calibration of computer models (with discussion)," *Journal of the Royal Statistical Society, Series B*, 63, 425–464.

Lange, K. (2013), *Optimization*, Spinger, 2nd ed.

Meyer, R. and Nachtsheim, C. (1995), "The Coordinate Exchange Algorithm for Constructing Exact Optimal Experimental Designs," *Technometrics*, 37, 60–69.

Muller, P. and Parmigiani, G. (1996), "Optimal Design via Curve Fitting of Monte Carlo Experiments," *Journal of the American Statistical Association*, 90, 1322–1330.

Nagumo, J., Arimoto, S., and Joshizawa, S. (1962), "An active pulse transmission line simulating a nerve axon," *Proceedings of the Institute of Radio Engineers*, 50, 2061–2070.

Overstall, A. and Woods, D. (2013), "A Strategy for Bayesian Inference for Computationally Expensive Models with Application to the Estimation of Stem Cell Properties," *Biometrics*, 69, 458–468.

— (2015), "The approximate coordinate exchange algorithm for Bayesian optimal design of experiments," *arXiv:1501.00264*.

Ramsay, J., Hooker, G., Campbell, D., and Cao, J. (2007), "Parameter estimation for differential equations: a generalised smoothing approach (with discussion)," *Journal of the Royal Statistical Society, Series B*, 69, 741–796.

Rasmussen, C. (2003), "Gaussian processes to speed up hybrid Monte Carlo for expensive Bayesian integrals," in *Bayesian Statistics 7*, eds. Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., Oxford.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmuller, U., and Timmer, J. (2009), "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood," *Bioinformatics*, 25, 1923–1929.

Ryan, E., Drovandi, C., Thompson, M., and Pettitt, A. (2014), "Towards Bayesian experimental design for nonlinear models that require a large number of sampling times," *Computational Statistics and Data Analysis*, 70, 45–60.

Swameye, I., Muller, T., Timmer, J., Sandra, O., and Klingmuller, U. (2003), "Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling," *Proceedings of the National Academy of Sciences*, 100, 1028–1033.