Bayesian Optimal Experimental Design of Clinical Studies

Tobias Freidling
DPMMS, University of Cambridge
taf40@cam.ac.uk

supervised by Thomas Moxon, Unilever

1 Introduction

Starting with Smith's early work (1918) on finding a good choice for circumstances of an experiment a researcher can influence, the field of optimal experimental design (OED) developed a rich literature from the 1960s onwards. It rigorously assesses how factors, such as the length of an experiment, the strata of the participants or the time points of measurements, can be set in an optimal way to achieve a certain goal. Hence, various optimality criteria were proposed and their relationships investigated. For a brief review, we refer to Fedorov (2010) and references therein. More recently, Bayesian methods became more popular due to both the increase in computational power and the possibility to incorporate prior knowledge. Likewise, different optimal criteria emerged among which the expected information gain (EIG) is the most popular.
Parallel to this development, physiologically based pharmacokinetic (PBPK) models became the go-to method to describe the propagation of chemicals through the human body, see Kuepfer et al. (2016). Originating in pharmaceutical research and drug development, they were more recently also applied to risk assessment of cosmetic products, e.g. Baltazar et al. (2020), Hatherell et al. (2020) and Moxon et al. (2020). Determining the correct values of the model parameters is key to obtain exact predictions. Due to the mechanistic nature of PBPK models some of these parameters can be inferred through in vitro experiments. Nevertheless, clinical studies are still necessary to obtain reliable in vivo values. Against this backdrop, OED methods seem promising in order to design experiments that yield a maximum amount of information about the parameters in question. Since there are only few examples, noteably Bois et al. (1999) and Gueorguieva et al. (2007), we hope that this report contributes to the exploration of Bayesian optimal experimental design methods for PBPK models.
We use the three-compartment model of Remifentanil according to Cascone et al. (2013) as a case study and concentrate on assessing the feasibility of different Bayesian OED approaches rather than on tuning prior distributions or hyper-parameters. We implement our experiments in Python using the SciPy library and the Pyro package, cf. Bingham et al. (2019), which provides many different Bayesian OED algorithms. We particularly focus on methods that do not require large computational resources and can be run in reasonable time on a usual PC or Mac.

2 Physiologically Based Pharmacokinetic Models

This section introduces physiologically based pharmacokinetic (PBPK) models from a mathematician's point of view and describes the three-compartment model that we analyse in the following sections.

2.1 Background

Physiologically based pharmacokinetic models describe the absorption, distribution, metabolism and excretion of certain chemicals in the human body over time. Tissues, such as liver, stomach and lung, are modelled as compartments which are typically connected by arterial and venous blood flow. We assign a concentration value of the investigated chemical to each compartment and describe the propagation of the drug in the human body by a system of ordinary differential equations (ODEs). Common parameters of PBPK models include the volume of the tissues, inter-compartment flow rates and clearance rates among others.

Unlike the more empirical pharmacokinetic models, PBPK models highly utilise knowledge of the human physiology and principles of fluid dynamics as well as physiochemical properties of the drug to describe the propagation of the investigated chemical. This mechanistic point of view allows to get a more granular understanding of the involved parameters but entails complex models in many cases. These can be used to extrapolate findings to diseased or pediatric populations, conduct sensitivity analysis, scale between different treatment scenarios and explore drug-drug interactions without the need for additional clinical studies. For an in-depth introduction we refer to Jones and Rowland-Yeo (2013) and Kuepfer et al. (2016) and included references.

While some parameters of the in vivo model can be inferred with certain confidence from in vitro experiments, others still require calibration through clinical studies. To this end, a group of patients is injected with the drug to be tested and the concentration in different tissues is measured at certain time points. It goes without saying that we would like to keep the number of participating patients and measurements as small as possible while still gathering enough information about the parameters. The fact that the time points can be chosen gives rise to the question how to determine them in an optimal way given constraints on the number of measurements and the distance between each two of them. We elaborate on this problem in Section 3.1 and present a Bayesian perspective.

2.2 Three-Compartment Model

This report uses the three-compartment model, which was proposed by Cascone et al. (2013), as a case study to better understand the challenges arising in Bayesian OED with PBPK models. This model describes the propagation of Remifentanil, an ultra-short acting opioid, in the human body after a bolus injection or a continuous infusion administration.

We specify three different compartments: C_1 which describes the concentration in the plasma, C_2 which models the concentration in highly perfused tissues and organs, and C_3 which represents the concentration in scarcely perfused tissues and organs. A schematic is shown in Figure 1. Taking absorption, distribution and excretion processes into account, we arrive at the following system of ODEs

$$\begin{aligned} V_1 \frac{\mathrm{d}C_1}{\mathrm{d}t} &= -\big[c_1 + V_1(k_{12} + k_{13} + k_{10})\big]C_1 + k_{21}V_2C_2 + k_{31}V_3C_3 + I(t), \\ V_2 \frac{\mathrm{d}C_2}{\mathrm{d}t} &= -\big[c_2 + k_{21}V_2\big]C_2 + k_{12}V_1C_1, \\ V_3 \frac{\mathrm{d}C_3}{\mathrm{d}t} &= -\big[c_3 + k_{31}V_3\big]C_3 + k_{13}V_1C_1. \end{aligned} \tag{1}$$

The V-terms refer to the volumes of the respective compartments, the c-parameters denote clearance rates, the k-coefficients describe the flow rates between compartments and we use the initial condition $C_1(0) = C_2(0) = C_3(0) = 0$. The function I models the injection of the chemical into the body and can accommodate for all kinds of infusion schemes. In this study, however, we concentrate on the case of intravenous constant-rate infusion of 20 minutes, i.e. $I(t) = 1\text{µg}\,\text{kg}^{-1}\,\text{min}^{-1}\mathbf{1}\{t \leq 20\}$. The values of the model parameters that Cascone et al. estimated are given in Table 1.

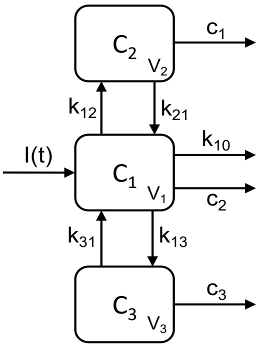

Figure 1: Schematic of the three-compartment model.

Parameter	Value	Parameter	Value
V_1	7.88 mL	k_{10}	0.172 min^{-1}
V_2	23.9 mL	k_{12}	0.373 min^{-1}
V_3	13.8 mL	k_{21}	0.103 min^{-1}
c_1	2.08 mL·min^{-1}	k_{13}	0.0367 min^{-1}
c_2	0.828 mL·min^{-1}	k_{31}	0.0124 min^{-1}
c_3	0.0784 mL·min^{-1}		

Table 1: Estimated model parameters from Cascone et al. (2013).

The mass balance differential equations which are typically used in PBPK models are of the form of an inhomogeneous, linear system of ODEs with time-independent matrix. Although this problem class has a general solution, the involved quantities such as integrals and matrix exponentials cannot be evaluated exactly in many cases and require a numerical approximation, see for example Markley (2004). In the considered three-compartment model an explicitly computable solution might be possible; nonetheless, this work relies on a numerical solution of the ODE system as we aim to demonstrate a procedure which is also applicable for more involved models. To this end, we use the `SciPy` implementation of the LSODA algorithm, which was proposed by Petzold (1983), due to its fast execution time. Figure 2 shows the evolution of the concentrations of Remifentanil in the three compartments when using the estimated paramters from Table 1. We observe that the concentrations range over more than two orders of magnitude during the first two hours. This already gives an indication that C_1 and C_2 become too small for acurate measurements after roughly 70 minutes. Hence, we do not consider time points after this time for optimal experimental design.

On a more general note, due to the complexity of the ODE system we only have limited insight into the relationship between time, the concentrations in the compartments and the parameters of the model. This renders many standard approaches to Bayesian optimal experimental design infeasible, as explained in Section 3.3.

3 Bayesian Experimental Design

In this section we state the optimal experimental design problem, embed the three-compartment model in a Bayesian framework and present different methods of assessing the expected information gain for a specified design.

3.1 Background

The field of optimal experimental design is concerned with setting up an experiment such that a maximum amount of information about the model parameters θ can be obtained from the observations of the variables y. We assume that the model is given by the probability density function $p(y|\theta, d)$, where d specifies the design. This refers to conditions that a researcher can set when conducting an experiment, e.g. the proportion of patients in different strata, the randomisation scheme or the length of the study. In our problem the design refers to the time

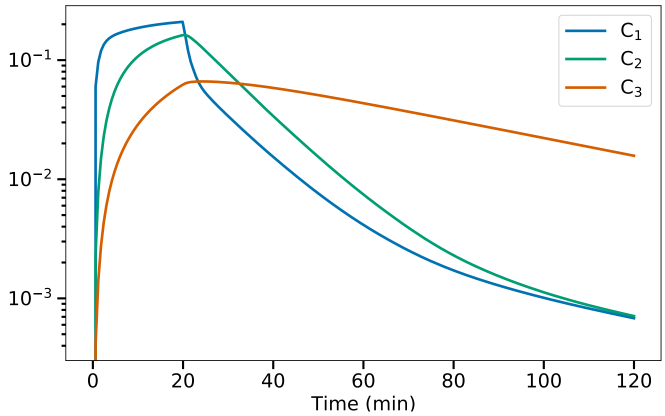

Figure 2: Concentrations [µg mL^{-1}] of Remifentanil in the three compartments under constant-rate infusion during the initial 20 minutes.

points when the concentrations are measured.

In order to assess the quality of a design, we consider a utility function $U(\theta, d)$. Common choices are the log-determinant of the inverse Fisher information matrix M (D-optimality) or the largest eigenvalue of M (E-optimality). For a review of different optimality criteria we refer to Fedorov (2010). In simple models, such as linear regression, U is indeed independent of θ and we can obtain a universally optimal design by maximising $U(d)$. In general, however, this is not the case and we can only find locally optimal designs by guessing a value θ^* for the true θ and considering $U(\theta^*, d)$. In many applications this is not satisfactory as the very reason to conduct an experiment is finding a value close to the true θ.

Bayesian optimal experimental design overcomes this weakness by integrating $U(d, \theta)$ over a prior $p(\theta)$ which encapsulates the pre-experiment knowledge about the parameters and can be interpreted as averaging over possible values of θ. Moreover, new criteria for optimality can be defined. This report concentrates on the expected information gain (EIG) which is given by the expected Kullback-Leibler divergence between the posterior distribution $p(\theta|y, d)$ and the prior $p(\theta)$, cf. Chaloner and Verdinelli (1995). For a fixed design d, EIG is defined as

$$\mathrm{EIG}(d) := \mathbb{E}_{p(y|d)}\Big[\mathrm{D}_{\mathrm{KL}}\big(p(\theta|y,d)\,\|\,p(\theta)\big)\Big] = \mathbb{E}_{p(y,\theta|d)}\left[\log\frac{p(\theta|y,d)}{p(\theta)}\right] = \mathbb{E}_{p(y,\theta|d)}\left[\log\frac{p(y|\theta,d)}{p(y|d)}\right]. \tag{2}$$

The expected information gain is an elegant way of describing the average amount of information that we can get when conducting an experiment with design d. However, as for many Bayesian methods, its computation is delicate and will be explored in the Sections 3.3 and following.

3.2 Bayesian Three-Compartment Model

In order to apply Bayesian OED methods, we define a Bayesian model based on the ODE system (1). We denote its solution evaluated at the timepoints $d = (t_1, \ldots, t_p)$ by $C(d, \bar{\theta})$, where $\bar{\theta}$ contains all parameters of the three-compartment model; hence $C(d, \bar{\theta}) \in \mathbb{R}^{3p}$. As Figure 2 shows, the concentrations that we expect to measure range over multiple orders of magnitude. For this reason, we apply a log-transformation, similar to Bois et al. (1999), and model measurement errors as Gaussian noise with a Gamma-prior for the variance

$$p(y|\theta, d) \sim \mathcal{N}\big(\log C(d, \bar{\theta}), \sigma^2 \mathrm{Id}\big), \qquad p(\sigma^2) \sim \mathrm{Gamma}(2, 31),$$

where $\theta = (\bar{\theta}, \sigma^2)$.

The parameters of the Gamma-prior are chosen according to the following heuristic. Experience of practitioners and other articles, such as White et al. (2016), suggest that the standard deviation of measurements of C is roughly 25% of the mean value. If $X \sim \mathcal{N}(\log C, \sigma^2)$, then e^X follows a log-normal distribution and its mean and variance are given by

$$\mathbb{E}\big[e^X\big] = e^{\log C + \sigma^2/2} \approx C, \qquad \mathrm{Var}\big(e^X\big) = \big(e^{\sigma^2} - 1\big)e^{\sigma^2} C^2.$$

Applying $\mathrm{Var}(e^X)^{1/2} = 0.25\,\mathbb{E}[e^X]$, we arrive at

$$[(e^{\sigma^2} - 1)e^{\sigma^2}]^{1/2} = 0.25\, e^{\sigma^2} \quad \Leftrightarrow \quad \sigma^2 \approx 0.065.$$

Consequently, we choose the prior on σ^2 such that its mean is close to 0.065.

The prior distributions for the parameters $\bar{\theta}$ take a bearing on the estimated values of Cascone et al.'s study. The volume and flow rate parameters use a Gaussian prior with mean according to the values of Table 1 and variance as 20% of the mean. For the clearance rates we apply a log-normal prior with mean-parameter corresponding to the values of Table 1 and variance-parameter set to 40% of the mean-parameter. We denote the prior distribution for all involved model parameters by $p(\theta)$.

We would like to highlight that this report concentrates on finding the best experimental design for only one patient and sets up the Bayesian model accordingly. Nevertheless, if the design shall be the same for all participants of a study and external factors, such as age, gender or preexisting conditions, are neglibile or only stratified subgroups are considered, the described Bayesian model can be used for multiple patients as well.

3.3 Infeasible methods

In some applications it may be desirable to find an optimal design that maximises EIG for a subset of model parameters; for example, we might only be interested in pinning down the value of c_1. However, this problem is much more intricate as the implicit likelihood $p(y|c_1, d) = \mathbb{E}_{p(c_1|\tilde{\theta})}[p(y|\theta, d)]$ is intractable. (Here, we split $\theta = (\tilde{\theta}, c_1)$.) For this reason, we focus on estimating the expected information gain for all model parameters.

Nested Monte Carlo The fundamental challenge of estimating EIG is the fact that neither $p(\theta|y, d)$ nor $p(y|d)$ are known in closed form. Therefore, conventional MC methods fail to compute the integrals in (2) and we have to resort to a nested Monte Carlo (NMC) approach. We create iid samples $\theta_{n,m} \sim p(\theta)$ and $y_n \sim p(y|\theta_{n,0}, d)$, where $n \in \{1, \ldots, N\}, m \in \{0, \ldots, M\}$, and compute the estimate as follows

$$\hat{\mu}_{\mathrm{NMC}}(d) = \frac{1}{N} \sum_{n=1}^{N} \log \frac{p(y_n|\theta_{n,0}, d)}{\frac{1}{M} \sum_{m=1}^{M} p(y_n|\theta_{n,m}, d)}.$$

The computational costs are of order $\mathcal{O}(NM)$ and Rainforth et al. (2018) showed that the RMSE decreases with rate $\mathcal{O}(N^{-1/2} + M^{-1})$ as $N, M \to \infty$; hence, we set $M \propto \sqrt{N}$. Moreover, $\hat{\mu}_{\mathrm{NMC}}(d)$ is an upper bound for $\mathrm{EIG}(d)$. Although the nested Monte Carlo estimator is asymptotically consistent, we found that for a complex system like the Bayesian three-compartment model the computational costs on a single core are prohibitively large to achieve convergence. This complicates the assessment of other estimation techniques as we cannot compute the ground-truth value of EIG for a given design.

Laplace Approximation Since non-parametric methods seem to fail for complex models, we might consider Laplace approximations to the posterior probability density function $p(\theta|y,d)$, as proposed by Long et al. (2013). This approach promises an accelerated estimation of EIG by using a Gaussian approximation $q(\theta|y,d)$. The Laplace estimator is given by

$$\hat{\mu}_{\text{laplace}}(d) = \frac{1}{N}\sum_{n=1}^{N} \hat{H}(q(\theta|y_n,d), p(\theta)) - \hat{H}(q(\theta|y_n,d)),$$

where $\hat{H}$ denotes an estimate of the (cross) entropy and $q(\theta|y_n,d)$ is a Laplace approximation to $p(\theta|y,d)$ computed once for each $y_n \sim p(y|d), n \in \{1,\ldots,N\}$. Since there are no theoretical guarantees for the performance of this approach and it is prone to large biases, as Foster et al. (2020) point out, we focus on other methods and leave the assessment of Laplace approximations for future work.

LFIRE Kleinegesse and Gutmann (2019) suggested to apply the liklihood-free inference by ratio estimation (LFIRE) method to Bayesian optimal experimental design. Since we can generate samples of the ratio

$$r(y,\theta,d) = \frac{p(y|\theta,d)}{p(y|d)}$$

in (2) but do not known an analytical expression, Thomas et al. (2021) proposed to train a logistic regression model that distinguishes whether y was sampled from $p(y|d)$ or $p(y|\theta,d)$. More precisely, for a fixed value of θ^* we define the model

$$\mathbb{P}(y \text{ was sampled from } p(y|\theta^*,d)) = \frac{1}{1+\exp(-h(y,\theta^*,d))},$$

where h is a function that is estimated on samples generated from $p(y|d)$ and $p(y|\theta^*,d)$. We denote the estimate $\hat{h}$. It can be shown that for large sample sizes $\hat{h}(y,\theta,d) \approx \log r(y,\theta,d)$. Hence, we can define the LFIRE estimator as

$$\hat{\mu}_{\text{LFIRE}}(d) = \frac{1}{n}\sum_{n=1}^{N} \hat{h}(y_n,\theta_n,d),$$

where $\hat{h}$ is computed separately for each sample pair $y_n,\theta_n \sim p(y,\theta|d), n \in \{1,\ldots,N\}$. In our experiments we found that the major difficulty of employing this method is the definition of a good regression model, i.e. a class of functions h that is capable of capturing the differences of $p(y|d)$ and $p(y|\theta,d)$ well. Moreover, the computational costs are considerable as for every sample y_n,θ_n a new logistic regression model has to be fitted.

Donsker-Varadhan The Kullback-Leibler divergence of two measures $\mathbb{P},\mathbb{Q}$ on a domain Ω admits the dual presentation $\mathrm{D}_{\mathrm{KL}}(\mathbb{P}\,\|\,\mathbb{Q}) = \sup_{T\colon\Omega\to\mathbb{R}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T])$, which was introduced by Donsker and Varadhan (1975). (The supremum is taken over all functions such that the expectations are finite.) By restricting the functions T to a parametrised class $\mathcal{F}_\Phi$, we can learn the optimal $T_\phi \in \mathcal{F}_\Phi$ which yields a lower approximation of the KL divergence, cf. Belghazi et al. (2018). Applying the Donsker-Varadhan (DV) representation to our problem, we obtain

$$\mathrm{EIG}(d) = \mathbb{E}_{p(y|d)}\left[\sup_{T:\Theta\to\mathbb{R}} \mathbb{E}_{p(\theta|y,d)}[T] - \log(\mathbb{E}_{p(\theta)}[e^T])\right].$$

Introducing a parametrised approximation $T(y,\theta|d,\phi)$ to $\log\frac{p(y,\theta|d)}{p(y|d)p(\theta)}$, we define the estimator

$$\hat{\mu}_{\mathrm{DV}}(d)=\frac{1}{N}\sum_{n=1}^{N}T(y_n,\theta_n|d,\phi)-\log\left(\frac{1}{N}\sum_{n=1}^{N}e^{T(y_n,\theta'_n|d,\phi)}\right),$$

where y_n,θ_n are iid samples from $p(y,\theta|d)$ and $\theta'_n\sim p(\theta)$. Since the (DV) representation yields a lower bound, we maximise the estimate over the parameters ϕ. Similar to LFIRE, finding a good approximation T is key to obtaining a decent estimate of EIG. However, as the three-compartment model is complex and allows little insight, we could not find a suitable class of approximations.

Variational methods Intractable likelihoods arise in many Bayesian problems. A common approach to tackle this issue are variational methods which define a parametric approximation to the likelihoods in question which can be optimised in some sense. Foster et al. (2020) introduced this set of techniques to Bayesian optimal experimental design proposing a variational posterior and variational nested Monte Carlo (VNMC) estimator.
For the former, we specify an approximation $q(\theta|y,d,\phi)$, parametrised by ϕ, to the posterior probability density $p(\theta|y,d)$. It is proven that using q instead of the true posterior in 2 entails a lower value which is equal to EIG if and only if the approximation and the true posterior agree. Hence, we can find the best approximation and thus the EIG-estimate closest to the truth by maximising over ϕ. The variational posterior estimator is given by

$$\hat{\mu}_{\mathrm{post}}(d)=\frac{1}{N}\sum_{n=1}^{N}\log\frac{q(\theta_n|y_n,d,\phi)}{p(\theta_n)},$$

where $y_n,\theta_n\sim p(y,\theta|d)$ are iid samples.
The VNMC approach aims to improve on the original NMC estimator by constructing an importance sampling estimate of $p(y|d)$. To this end, a proposal $q(\theta|y,d,\phi)$ approximating $p(\theta|y,d)$ is learnt adjusting ϕ. The estimate is computed as

$$\hat{\mu}_{\mathrm{VNMC}}(d)=\frac{1}{N}\sum_{n=1}^{N}\left(\log p(y_n|\theta_{n,0},d)-\log\frac{1}{M}\sum_{m=1}^{M}\frac{p(y_n,\theta_{n,m}|d)}{q(\theta_{n,m}|y_n,d,\phi)}\right),$$

with the iid sampels $\theta_{n,0}\sim p(\theta)$, $y_n\sim p(y|\theta_{n,0},d)$ and $\theta_{n,m}\sim q(\theta|y_n,d,\phi)$. Foster et al. showed that VNMC is an upper bound to EIG which can be used to find an optimal ϕ. Additionally, unlike the variational posterior, the VNMC estimator is asymptotically consistent, even when $q(\theta|y,d,\phi)\neq p(\theta|y,d)$ for all $\phi\in\Phi$.
Both approaches rely on finding good approximations to the posterior probability density. This, however, is highly challenging due to the ODE-system at the core of our model and gave rise to an entire field of mathematics, namely Bayesian Inverse Problems. Therefore, these two variational methods are not suited for our problem. However, a third can be successfully applied as explained in the following section.

3.4 Variational Marginal

Similar to the variational posterior and VNMC, the variational marginal approach estimates EIG by finding an approximation to one of the unknown likelihoods in (2). However, we now approximate the marginal likelihood $p(y|d)$ by $q(y|d,\phi)$ and compute the estimator as follows

$$\hat{\mu}_{\mathrm{marg}}(d)=\frac{1}{N}\sum_{n=1}^{N}\log\frac{p(y_n|\theta_n,d)}{q(y_n|d,\phi)},$$

where $y_n, \theta_n \sim p(y, \theta|d)$ is an iid sample. Foster et al. proved an upper variational bound which can be used to optimise the parameter ϕ.

This approach is favourable for our problem as we only need to have a grasp of $p(y|d)$ which we can observe simply by generating samples of the Bayesian three-compartment model. Figure 3 exhibits that the marginal densities of $p(y|d)$ resemble a Gaussian distribution for all considered time points and concentrations.

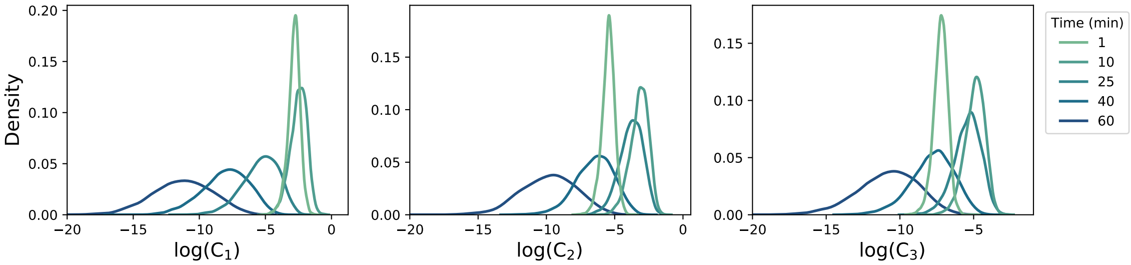

Figure 3: Distribution of the log-concentrations in the three compartments at five different time points. The probability densities were estimated with Gaussian kernel density estimation on 5000 samples each.

We further investigate our presumption by considering P-P plots, see Figure 4.

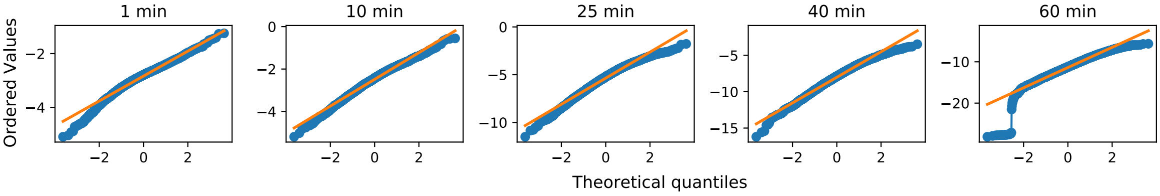

Figure 4: P-P plot of a Gaussian distribution against the log-concentrations in the central compartment at different time points. We simulated 5000 values each.

They show that the assumption of a Gaussian $p(y|d)$ is not exactly, but reasonably well fulfilled. Lastly, we investigate the correlation structure among different measurements and time points, see Figure 5. The correlation between different measurements is, in general, strong, in particular along the time dimension, i.e. two measurements of the same concentration at neighbouring time points are stronger correlated than two measurements of different concentrations at the same time point.

Based on these investigations we consider three different Gaussian approximations to $p(y|d)$, i.e. $q(y|d, \phi) \sim \mathcal{N}(\mu, \Sigma)$ where $\phi = (\mu, \Sigma)$: We either restrict Σ to be diagonal, which resembles a mean field approximation, allow only for correlation along time or lift any constraints on the covariance matrix. In the following we refer to the first variant as factorised and to the last variant as non-factorised approximation. In order to generate an EIG-estimate, we proceed as follows: First, we sample N_1 data points from the model and compute the empirical mean and covariance matrix, $\hat{\mu}$ and $\hat{\Sigma}$; then, we apply the variational marginal method with $q(y|d, \phi) \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma})$.

Besides parametric approximations to the marginal distribution, we also consider the non-parametric and thus more flexible kernel density estimates of $p(y|d)$. For d-dimensional iid random variables $X_1, \ldots, X_n$ the kernel density estimator is given by

$$\hat{f}_n(x) := \frac{1}{n \det(H)} \sum_{i=1}^{n} K\left(H^{-1}(x - X_i)\right),$$

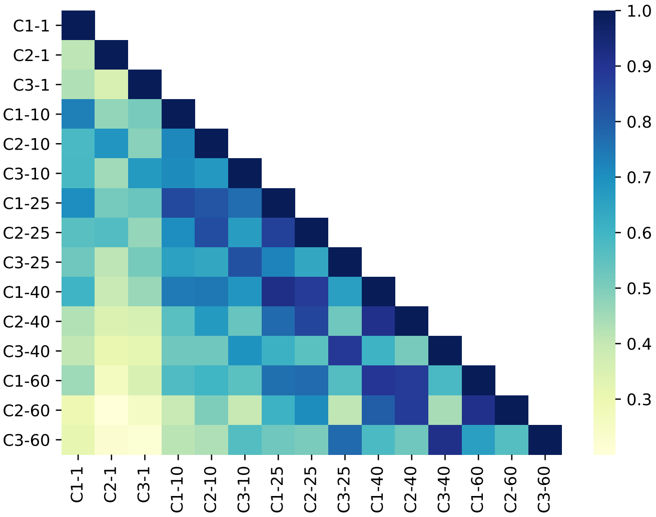

Figure 5: Correlation between the measurements of log-concentrations in different compartments at five different time points. The sample size is 5000 and "Cj-t" denotes the log-concentration in the j-th compartment at time t.

where K is a non-negative function that integrates to 1, the so-called kernel, and H denotes the bandwith parameter matrix. Common choices for K include the Gaussian or Epanechnikov kernel and the bandwith is often chosen according to Scott's rule, that is $H \approx 1.06\,\hat{\Sigma}\,n^{-1/(d+4)}$, which exhibits good performance for data that does not deviate too far from a Gaussian distribution. For more details, we refer to Scott (2015). Kernel density estimation is usually applied in low-dimensional settings as the convergence rate suffers from the curse of dimensionality if the true probability density function is not sufficiently smooth. For this reason, we compare three different KDE approximations, similiar to the Gaussian case: One which restricts H to be diagonal, one that only allows for interactions along time but not between different concentrations and one which does not put any restrictions.

We under take some experiments to evaluate the empirical behaviour of the discussed estimators. The simulations depicted in Figure 6 show that nested Monte Carlo exhibits the weakest performance as it is more volatile than the variational marginal estimators and has a higher bias. The latter can be inferred from the fact that all considered estimators yield upper bounds to the true EIG-value. Following the same rationale, we see that the non-factorised approximations are less biased compared to their factorised counterparts. Lastly, the estimates obtained from the normal approximation are less instable, supposedly due to their parametric structure. Hence, we conclude that the variational marginal estimator with non-factorised Gaussian approximation showed the best performance in our experiments.

4 Optimisation

Having found methods to get estimates of the expected information gain for one particular design, we now turn to discussing optimisation over the design space $\mathcal{D}$. In many applications constraints on $\mathcal{D}$ have to be taken into account: In our case study, for example, we demand that each two measurements have to be at least 5 minutes apart.
First, we note that we cannot take derivatives of EIG with respect to d. Consequently, the

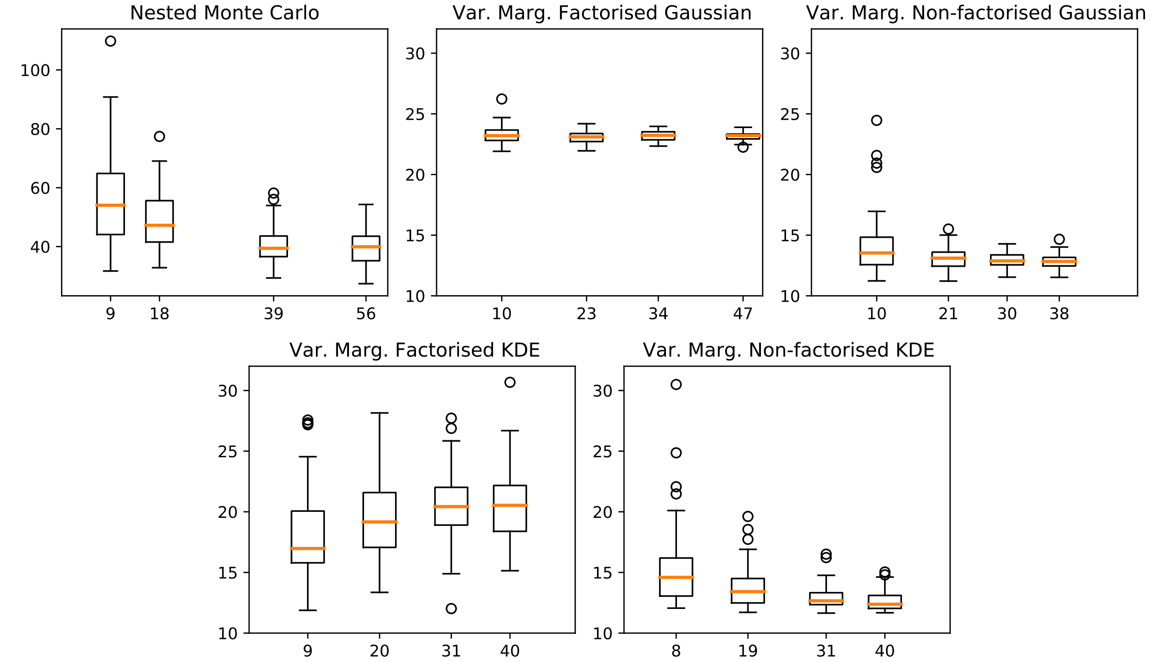

Figure 6: Box plot of the EIG-estimate of different methods for the design $d = (1, 10, 25, 40, 60)$. We used four different sets of hyper-parameters for each estimator, which correspond to the average run-times [s] on a single core noted on the x-axis, and ran 50 simulations each. For the nested Monte Carlo estimate, we used $M \in \{20, 27, 37, 45\}$ and $N = M^2$; for the variational marginal approach, we set $N_1 \in \{200, 450, 700, 900\}$ and $N = N_1$.

majority of off-the-shelf optimisation algorithms, such as gradient descent and Newton methods as well as variants thereof, cannot be applied to our problem as they require knowledge of the gradient. For this reason, we consider two popular optimisation schemes that only rely on function evaluations: simulated annealing, proposed by Kirkpatrick et al. (1983), and differential evolution, which was introduced by Storn and Price (1997). It is crucial that both of them are suitable for global optimisation as we have little knowledge of the smoothness of EIG as a function of d. Setting the hyper-parameters appropriately, we can strike a balance between exploring the design space and exploiting already gained information. Thus, we trade-off convergence speed to a maximum and certainty of it being global. In our implementations we found that the differential evolution algorithm can more easily accomodate for constraints than simulated annealing.

Yet, applying one of these algorithms requires a large amount of function evaluations. Since estimates that can be computed on a single core in reasonable time are instable, see Figure 6, any optimisation algorithm sooner or later finds a design which has a seemingly high EIG. This, however, is most often not the result of an optimal design but the instability of the EIG function evaluation. In our experiments, the optimisation algorithm was repeatedly trapped by erroneous EIG estimates.

Allocating more computational resources to the estimation method to produce more reliable EIG-values can be a remedy for this issue. Nonetheless, if we aim for low computational complexity, performing a grid search or testing a given set of designs, proposed by practitioners, seem to be the only straightforward solutions.

Utilising grid search for three time points, we obtain the optimal designs according to Table 2. We notice that later time points are favoured and EIG seems to be a fairly smooth function due

EIG	Design	EIG	Design
13.69	(30, 55, 60)	11.93	(5, 45, 60)
13.02	(35, 55, 60)	11.87	(20, 50, 60)
12.75	(35, 45, 60)	11.86	(20, 55, 60)
12.04	(30, 50, 60)	11.85	(40, 55, 60)
11.93	(50, 55, 60)	11.84	(30, 50, 55)

Table 2: The 10 best designs for three measurements according to grid search. We consider time points within the first 60 minutes under the constraint that each two measurements have to be at least 5 minutes apart and set the grid length to 5. EIG was estimated applying the variational marginal method with non-factorised normal approximation where $N_1 = N = 1500$.

to the similarity of close to optimal designs.

5 Outlook

In the previous sections we explored different ways of leveraging Bayesian optimal experimental design methods for PBPK models. Nonetheless, there are many aspects that we did not touch on.

For application to clinical studies tuning hyper-parameters of the estimation methods as well as clarifying the sensitivity of the results with respect to the choice of the prior distributions is certainly important. Moreover, studying other PBPK models, especially those with a higher number of compartments, can foster a better understanding of the behaviour of the discussed methods and provide more examples as reference for new experiments. In addition, we may also consider multiple patients and allow for measurement schemes that differ among participants.

From a theoretical point of view, there are interesting questions and ideas for future research as well. For example, Foster et al. propose in Appendix F of their paper (2020) an approach using control variates which is asymptotically consistent and promises a reduced variance. Furthermore, we conjecture that we can adapt VNMC using an importance sampling estimate of $p(\theta|y, d)$ and proof asymptotic consistency. If these ideas prove to be successful, we may contribute them to `Pyro`'s OED module. A certainly challenging problem in Bayesian optimal experimental design is the treatment of implicit likelihoods. These arise when we want to estimate EIG only for a subset of parameters as we can sample from an implicit likelihood in a straightforward way but cannot evaluate it. Although some estimators, such as NMC, can accommodate such a situation, for many others a solution is not known. Finally, the development of optimisation algorithms dedicated to functions that cannot be evaluated precisely, or rather are subject to uncertainty coming from the estimation procedure, would facilitate the detection of good designs.

References

Baltazar, Maria T, Sophie Cable, Paul L Carmichael, Richard Cubberley, Tom Cull, Mona Delagrange, Matthew P Dent, Sarah Hatherell, Jade Houghton, Predrag Kukic, Hequn Li, Mi-Young Lee, Sophie Malcomber, Alistair M Middleton, Thomas E Moxon, Alexi V Nathanail, Beate Nicol, Ruth Pendlington, Georgia Reynolds, Joe Reynolds, Andrew White, and Carl Westmoreland (2020). "A Next-Generation Risk Assessment Case Study for Coumarin in Cosmetic Products". In: *Toxicological Sciences* 176.1, pp. 236–252.

Belghazi, Mohamed Ishmael, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm (2018). "Mutual Information Neural Estimation". In: *Proceedings of the 35th International Conference on Machine Learning.* Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 531–540.

Bingham, Eli, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman (2019). "Pyro: Deep Universal Probabilistic Programming". In: *J. Mach. Learn. Res.* 20, 28:1–28:6.

Bois, F Y, T J Smith, A Gelman, H Y Chang, and A E Smith (1999). "Optimal design for a study of butadiene toxicokinetics in humans." In: *Toxicological Sciences* 49.2, pp. 213–224.

Cascone, Sara, Gaetano Lamberti, Giuseppe Titomanlio, and Ornella Piazza (2013). "Pharmacokinetics of Remifentanil: A three-compartmental modeling approach". In: *Translational medicine @ UniSa* 7, pp. 18–22.

Chaloner, Kathryn and Isabella Verdinelli (1995). "Bayesian experimental design: A review". In: *Statistical Science* 10.3, pp. 273–304.

Donsker, M.D. and S. R. S Varadhan (1975). "Asymptotic evaluation of certain markov process expectations for large time, I". In: *Communications on Pure and Applied Mathematics* 28.1, pp. 1–47.

Fedorov, Valerii (2010). "Optimal experimental design". In: *WIREs Computational Statistics* 2.5, pp. 581–589.

Foster, Adam, Martin Jankowiak, Eli Bingham, Paul Horsfall, Yee Whye Teh, Tom Rainforth, and Noah Goodman (2020). "Variational Bayesian Optimal Experimental Design". In: *arXiv preprint.*

Gueorguieva, Ivelina, Kayode Ogungbenro, Gordon Graham, Sophie Glatt, and Leon Aarons (2007). "A program for individual and population optimal design for univariate and multivariate response pharmacokinetic–pharmacodynamic models". In: *Computer Methods and Programs in Biomedicine* 86.1, pp. 51–61.

Hatherell, Sarah, Maria T Baltazar, Joe Reynolds, Paul L Carmichael, Matthew Dent, Hequn Li, Stephanie Ryder, Andrew White, Paul Walker, and Alistair M Middleton (2020). "Identifying and Characterizing Stress Pathways of Concern for Consumer Safety in Next-Generation Risk Assessment". In: *Toxicological Sciences* 176.1, pp. 11–33.

Jones, HM and K Rowland-Yeo (2013). "Basic Concepts in Physiologically Based Pharmacokinetic Modeling in Drug Discovery and Development". In: *CPT: Pharmacometrics & Systems Pharmacology* 2.8, p. 63.

Kirkpatrick, S., C. D. jun. Gelatt, and M. P. Vecchi (1983). "Optimization by simulated annealing". In: *Science* 220.4598, pp. 671–680.

Kleinegesse, Steven and Michael U. Gutmann (2019). "Efficient Bayesian Experimental Design for Implicit Models". In: *Proceedings of Machine Learning Research.* Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 476–485.

Kuepfer, L, C Niederalt, T Wendl, J-F Schlender, S Willmann, J Lippert, M Block, T Eissing, and D Teutonico (2016). "Applied Concepts in PBPK Modeling: How to Build a PBPK/PD Model". In: *CPT: Pharmacometrics & Systems Pharmacology* 5.10, pp. 516–531.

Long, Quan, Marco Scavino, Raúl Tempone, and Suojin Wang (2013). "Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations". In: *Computer Methods in Applied Mechanics and Engineering* 259, pp. 24–39.

Markley, Nelson G. (2004). *Principles of differential equations.* Hoboken, NJ: Wiley-Interscience, pp. x + 339.

Moxon, Thomas E., Hequn Li, Mi-Young Lee, Przemyslaw Piechota, Beate Nicol, Juliette Pickles, Ruth Pendlington, Ian Sorrell, and Maria Teresa Baltazar (2020). "Application of physiologically based kinetic (PBK) modelling in the next generation risk assessment of dermally applied consumer products". In: *Toxicology in Vitro* 63, p. 104746.

Petzold, Linda (1983). "Automatic Selection of Methods for Solving Stiff and Nonstiff Systems of Ordinary Differential Equations". In: *SIAM Journal on Scientific and Statistical Computing* 4.1, pp. 136–148.

Rainforth, Tom, Rob Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood (2018). "On Nesting Monte Carlo Estimators". In: *Proceedings of the 35th International Conference on Machine Learning.* Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 4267–4276.

Scott, David W. (2015). *Multivariate density estimation. Theory, practice, and visualization. 2nd ed.* 2nd ed. Hoboken, NJ: John Wiley & Sons, pp. xviii + 350.

Smith, Kirstine (1918). "On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and its Constants and the Guidance they give Towards a Proper Choice of the Distribution of Observations". In: *Biometrika* 12.1/2, pp. 1–85.

Storn, Rainer and Kenneth Price (1997). "Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces". In: *Journal of Global Optimization* 11.4, pp. 341–359.

Thomas, Owen, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U. Gutmann (2021). "Likelihood-Free Inference by Ratio Estimation". In: *Bayesian Analysis* -1.-1, pp. 1–31.

White, John R., Jeannie M. Padowski, Yili Zhong, Gang Chen, Shaman Luo, Philip Lazarus, Matthew E. Layton, and Sterling McPherson (2016). "Pharmacokinetic analysis and comparison of caffeine administered rapidly or slowly in coffee chilled or hot versus chilled energy drink in healthy young adults". In: *Clinical Toxicology* 54.4, pp. 308–312.