Statistics and Computing (2025) 35:12
https://doi.org/10.1007/s11222-024-10544-z

ORIGINAL PAPER

Laplace-based strategies for Bayesian optimal experimental design with nuisance uncertainty

Arved Bartuska[1] · **Luis Espath**[2] · **Raúl Tempone**[1,3,4]

Received: 1 April 2024 / Accepted: 26 November 2024

Abstract

Finding the optimal design of experiments in the Bayesian setting typically requires estimation and optimization of the expected information gain functional. This functional consists of one outer and one inner integral, separated by the logarithm function applied to the inner integral. When the mathematical model of the experiment contains uncertainty about the parameters of interest and nuisance uncertainty, (i.e., uncertainty about parameters that affect the model but are not themselves of interest to the experimenter), two inner integrals must be estimated. Thus, the already considerable computational effort required to determine good approximations of the expected information gain is increased further. The Laplace approximation has been applied successfully in the context of experimental design in various ways, and we propose two novel estimators featuring the Laplace approximation to alleviate the computational burden of both inner integrals considerably. The first estimator applies Laplace's method followed by a Laplace approximation, introducing a bias. The second estimator uses two Laplace approximations as importance sampling measures for Monte Carlo approximations of the inner integrals. Both estimators use Monte Carlo approximation for the remaining outer integral estimation. We provide four numerical examples demonstrating the applicability and effectiveness of our proposed estimators.

Keywords Bayesian experimental design · Nuisance uncertainty · Nested integration · Monte Carlo · Laplace approximation · Importance sampling

Mathematics Subject Classification 65K05 · 65C05

Contents

✉ Arved Bartuska
arved.bartuska@gmail.com

Luis Espath
espath@gmail.com

Raúl Tempone
rtempone@gmail.com

[1] Department of Mathematics, RWTH Aachen University, Gebäude-1953 1.OG, Pontdriesch 14-16, 161, 52062 Aachen, Germany

[2] School of Mathematical Sciences, University of Nottingham, Nottingham NG7 2RD, UK

[3] King Abdullah University of Science & Technology (KAUST), Computer, Electrical and Mathematical Sciences & Engineering Division (CEMSE), 23955-6900 Thuwal, Saudi Arabia

[4] RWTH Aachen University, Aachen, Germany

Published online: 13 December 2024

1 Introduction

The goal of optimal experimental design (OED) (Chaloner and Verdinelli 1995) is to provide designs for a given scientific experiment for which the expected information gain (EIG) (Lindley 1956) is optimal. In the Bayesian setting, the knowledge of the experimenter about the parameters of interest before conducting the experiment is expressed as the prior probability density (pdf) of the parameters of interest, whereas the knowledge after the experiment is expressed as the posterior pdf, conditioned on the data. The increase in knowledge (or the reduction in uncertainty) is given by the Kullback–Leibler divergence (Kullback 1959; Kullback and Leibler 1951) between the posterior and prior. The EIG can be expressed as the expected Kullback–Leibler divergence over the data. A higher EIG indicates that the data obtained from the experiment are expected to provide more information on the parameters of interest. Therefore, optimizing the EIG reduces the uncertainty about the parameters of interest.

If additional uncertainty is present in the model via parameters not of interest to the experimenter, we refer to them as nuisance parameters and marginalize them before optimizing the EIG. Thus, we find a design that only reduces the uncertainty about the parameters of interest, not the nuisance parameters. This approach ultimately maximizes the amount of information the experiment provides about the parameters of interest while keeping track of the uncertainty introduced by the nuisance parameters. A similar setting was considered in Alexanderian et al. (2022); Bartuska et al. (2022); Feng and Marzouk (2019). For a more general background on nuisance uncertainty, see Bernardo (1979); Levine (2014); Liepe et al. (2013); Polson (1988).

There is often no closed-form expression for the EIG; thus, it must be estimated numerically (Ryan 2003), which involves estimating nested integrals separated by the logarithm function and typically entails high computational costs. Marginalizing nuisance parameters introduces a second inner integral, increasing the computational cost further. The Laplace approximation (Stigler 1986; Tierney and Kadane 1986; Tierney et al. 1989; Kass et al. 1990; Friston et al. 2007) of the inner integral in combination with the Monte Carlo (MC) method is an effective tool for estimating the EIG (Beck et al. 2018; Long et al. 2013). In this work, we extend this approach to the case with additional nuisance uncertainty and derive Laplace approximations for two distinct estimators. The first estimator uses Laplace's method followed by a Laplace approximation in a nested fashion to approximate the posterior pdf of the parameters of interest directly, and then uses the MC method for the outer integral. This method introduces a bias relative to the number of experiments performed, which is considered fixed from a modeling perspective. The second estimator uses two separate Laplace approximations as importance sampling measures for the two inner integrals, which are approximated using the MC method.

Our approach differs from that used in Feng and Marzouk (2019), as the Laplace approximation is incorporated centrally in both estimators, whereas (Feng and Marzouk 2019) developed a gradient-free importance sampling scheme to reduce variance in their proposed MC estimator. Similar to Feng and Marzouk (2019), we consider reducible nuisance uncertainty, which is accessible to Bayesian updates. On the other hand, Alexanderian et al. (2022) considers irreducible nuisance uncertainty. The studies (Alexanderian et al. 2022) and Bartuska et al. (2022) assumed Gaussianity for the distribution of the nuisance parameters. Such assumptions are not present in Feng and Marzouk (2019) and the present work. The small-noise approximation derived in Bartuska et al. (2022) also further restricts the size of the nuisance uncertainty to be considered.

In Englezou et al. (2022), an approximate Laplace importance sampling for EIG estimation was proposed, where outer samples were used to compute the Hessian for the Laplace approximation rather than using the maximum a posteriori (MAP). A double-importance sampling scheme for EIG with nuisance uncertainty based on the Laplace approximation was first proposed in Englezou et al. (2022). For the importance sampling density of the nuisance parameters conditioned on the data and the parameters of interest, the joint density of nuisance parameters and parameters of interest was first approximated by the standard Laplace approximation. The conditional density was then computed based on the formula for the conditional normal density. Instead, in this work we propose a novel Laplace approximation that is directly applied to the conditional density of the nuisance parameters. This approach requires additional numerical approximation methods, namely, the MAP and related Hessian for this conditional density; however, we also expect to obtain a more robust importance sampling density from our approach. Moreover, we demonstrate that the cost to obtain importance sampling densities does not affect the asymptotic cost of the overall estimator.

The work (Overstall et al. 2017) considers normal-based Laplace approximations also for utility functions different from the EIG, called therein the expected Shannon information gain or expected self-information loss. Moreover, model uncertainty is also discussed in Overstall et al. (2017), which can be interpreted as an application of nuisance parameters. However, the setting discussed in Overstall et al. (2017) is of a finite and countable number of possible models. Inte-

gral approximations, as required in the current manuscript for nuisance parameters modeled by a continuous random variable, are not discussed in their setting.

For asymptotic error bounds on the Laplace approximation, see Schillings et al. (2020); Wacker (2017). Pre-asymptotic error bounds and the effect of the nonlinearity of the experiment model on the data likelihood were investigated in Helin and Kretschmann (2022). For additional errors from estimating the mean and covariance in the Laplace approximation, see Spokoiny (2022).

The rest of this work is structured as follows. In Sect. 2, we introduce the Bayesian formulation and the standard double-loop MC (DLMC) estimator with nuisance uncertainty, which serves as a reference method for the novel estimators. Next, in Sect. 3, we provide an error analysis, computational work analysis, and a method to obtain the required number of outer and inner samples to achieve a certain error tolerance for the DLMC estimator with nuisance uncertainty. As the main contribution of this work, we derive the Laplace approximations necessary for the MC double-Laplace (MC2LA) estimator in Sect. 4. Subsequently, in Sect. 5, we develop the MC2LA estimator and a probabilistic error bound. Next, we develop the DLMC double importance sampling (DLMC2IS) estimator. Finally, in Sect. 6, we demonstrate the effectiveness of the proposed estimators on four numerical examples. The first example provides insight into the effect of nuisance parameters on the optimal design of an experiment. An example from pharmacokinetics indicates the possibility for optimization in a larger design space of 15 dimensions. The third and fourth examples reveal the applicability of the estimators to electrical impedance tomography (EIT) experiments involving a finite element approximation of the underlying partial differential equation (PDE).

2 Bayesian formulation

Before we discuss the Bayesian setting, we must first describe the experiment in mathematical terms. For this purpose, we consider an additive data model:

$$\boldsymbol{y}_i(\boldsymbol{\xi}) = \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}_t, \boldsymbol{\phi}_t) + \boldsymbol{\epsilon}_i, \tag{1}$$

where

- $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$ is the observed data vector, $i = 1, \cdots, N_e$;
- $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{N_e}) \in \mathbb{R}^{d_y \times N_e}$ is the data for each experiment;
- N_e is the number of observations for a specific experiment setup;
- $\boldsymbol{\theta}_t \in \mathbb{R}^{d_\theta}$ is the true vector of the parameters of interest;
- $\boldsymbol{\phi}_t \in \mathbb{R}^{d_\phi}$ is the true vector of the nuisance parameters;
- $\boldsymbol{\xi} \in \mathbb{R}^{d_\xi}$ is the vector of the design parameters;
- $\boldsymbol{g} \colon \mathbb{R}^{d_\xi} \times \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_\phi} \to \mathbb{R}^{d_y}$ is the deterministic model of the experiment;
- $\boldsymbol{\epsilon}_i \in \mathbb{R}^{d_y}$ is the error vector assumed to be Gaussian $\boldsymbol{\epsilon}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}), i = 1, \cdots, N_e$,

where iid refers to independent and identically distributed. The knowledge about the parameters of interest $\boldsymbol{\theta}$ before the experiment is encompassed by the prior pdf $\pi(\boldsymbol{\theta})$, whereas the knowledge after the experiment is given by the posterior pdf $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$. We make no general assumptions on the independence of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$; therefore, we also consider the joint distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\phi})$. To complete the Bayesian setup, we consider the likelihood $p(\boldsymbol{Y}|\boldsymbol{\theta})$ and evidence $p(\boldsymbol{Y})$. The latter two pdfs are denoted by p rather than π to distinguish between pdfs for model parameters and data. Moreover, the posterior, likelihood, and the evidence are conditioned on the design $\boldsymbol{\xi}$, but we omit this dependence for concision. In this context, the Bayes formula for the posterior reads

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{Y}) &= \frac{\pi(\boldsymbol{\theta})\, p(\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{Y})}, \\ &= \frac{\pi(\boldsymbol{\theta}) \int_{\boldsymbol{\Phi}} p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\phi}) \pi(\boldsymbol{\phi}|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\phi}}{\int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Phi}} p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\phi}) \pi(\boldsymbol{\theta}, \boldsymbol{\phi}) \, \mathrm{d}\boldsymbol{\phi} \, \mathrm{d}\boldsymbol{\theta}},\end{aligned} \tag{2}$$

where the likelihood terms are obtained through marginalization. The likelihood conditioned on ϕ results from the data assumption (1), as

$$\begin{aligned}p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\phi}) &:= \det(2\pi \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})^{-\frac{N_e}{2}} \\ &\exp\left(-\frac{1}{2} \sum_{i=1}^{N_e} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{\phi})\right),\end{aligned} \tag{3}$$

where

$$\boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) := \boldsymbol{y}_i - \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \boldsymbol{g}(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) + \boldsymbol{\epsilon}_i - \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}) \tag{4}$$

is the data residual. The amount of information gained about the parameters of interest $\boldsymbol{\theta}$ from the experiment is expressed as the Kullback–Leibler divergence (Kullback 1959; Shannon 1948) given by

$$D_{\mathrm{KL}} = \int_{\boldsymbol{\Theta}} [\log(\pi(\boldsymbol{\theta}|\boldsymbol{Y})) - \log(\pi(\boldsymbol{\theta}))] \pi(\boldsymbol{\theta}|\boldsymbol{Y}) \, \mathrm{d}\boldsymbol{\theta}. \tag{5}$$

The goal of OED is to determine a design $\boldsymbol{\xi}$ for which the experiment provides informative data; therefore, we consider the expectation of (5) over the data $\boldsymbol{Y}$, yielding the EIG

$$I = \int_{\mathcal{Y}} \int_{\boldsymbol{\Theta}} [\log(\pi(\boldsymbol{\theta}|\boldsymbol{Y})) - \log(\pi(\boldsymbol{\theta}))] \pi(\boldsymbol{\theta}|\boldsymbol{Y}) \, \mathrm{d}\boldsymbol{\theta} \, p(\boldsymbol{Y}) \, \mathrm{d}\boldsymbol{Y}. \tag{6}$$

This quantity only depends on the design $\boldsymbol{\xi}$, as all other dependencies have been marginalized. We rewrite this expression

in terms of the known likelihood function (3) conditioned on ϕ using the Bayes formula (2)

$$
\begin{aligned}
I &= \int_{\boldsymbol{\Theta}} \int_{\mathcal{Y}} [\log(p(\boldsymbol{Y}|\boldsymbol{\theta})) - \log(p(\boldsymbol{Y}))] p(\boldsymbol{Y}|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{Y} \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}, \\
&= \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Phi}} \int_{\mathcal{Y}} \left[\log \left(\int_{\boldsymbol{\Phi}} p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\varphi}) \pi(\boldsymbol{\varphi}|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\varphi} \right) \right. \\
&\quad \left. - \log \left(\int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Phi}} p(\boldsymbol{Y}|\boldsymbol{\vartheta}, \boldsymbol{\varphi}) \pi(\boldsymbol{\vartheta}, \boldsymbol{\varphi}) \, \mathrm{d}\boldsymbol{\varphi} \, \mathrm{d}\boldsymbol{\vartheta} \right) \right] \\
&\quad p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\phi}) \, \mathrm{d}\boldsymbol{Y} \pi(\boldsymbol{\theta}, \boldsymbol{\phi}) \, \mathrm{d}\boldsymbol{\phi} \, \mathrm{d}\boldsymbol{\theta}, \qquad (7)
\end{aligned}
$$

where $\boldsymbol{\vartheta}$ and $\boldsymbol{\varphi}$ are dummy variables to distinguish them from $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ from the outer integrals. As a reference to approximate expression (7), we introduce the DLMC estimator with two inner loops:

$$
\begin{aligned}
I_{\mathrm{DL}} = \frac{1}{N} \sum_{n=1}^{N} \left[\log \left(\frac{1}{M_1} \sum_{m=1}^{M_1} p(\boldsymbol{Y}^{(n)}|\boldsymbol{\theta}^{(n)}, \boldsymbol{\varphi}^{(n,m)}) \right) \right. \\
\left. - \log \left(\frac{1}{M_2} \sum_{k=1}^{M_2} p(\boldsymbol{Y}^{(n)}|\boldsymbol{\vartheta}^{(n,k)}, \boldsymbol{\varphi}^{(n,k)}) \right) \right]. \qquad (8)
\end{aligned}
$$

The samples are drawn as follows. First, we sample $(\boldsymbol{\theta}^{(n)}, \boldsymbol{\phi}^{(n)}) \overset{\text{iid}}{\sim} \pi(\boldsymbol{\theta}, \boldsymbol{\phi})$. This allows us to sample $\boldsymbol{Y}^{(n)} \overset{\text{iid}}{\sim} p(\boldsymbol{Y}|\boldsymbol{\theta}^{(n)}, \boldsymbol{\phi}^{(n)})$, $1 \leq n \leq N$. Next, we sample $\boldsymbol{\varphi}^{(n,m)} \overset{\text{iid}}{\sim} \pi(\boldsymbol{\phi}|\boldsymbol{\theta}^{(n)})$, $1 \leq m \leq M_1$, and $(\boldsymbol{\vartheta}^{(n,k)}, \boldsymbol{\varphi}^{(n,k)}) \overset{\text{iid}}{\sim} \pi(\boldsymbol{\theta}, \boldsymbol{\phi})$, $1 \leq k \leq M_2$.

3 Bias, statistical error, and optimal number of samples

To obtain the optimal number of samples for the outer and inner loops, we aim to minimize the total work to achieve a certain prescribed tolerance in the EIG estimator. Thus, we first stipulate that the total work W of computing the estimator I_{DL} in expression (8) is given by

$$
W(I_{\mathrm{DL}}) \propto N(M_1 + M_2). \qquad (9)
$$

Moreover, we split the error of the estimator into the bias and statistical errors:

$$
|I_{\mathrm{DL}} - I| \leq \underbrace{|\mathbb{E}[I_{\mathrm{DL}}] - I|}_{\text{bias error}} + \underbrace{|I_{\mathrm{DL}} - \mathbb{E}[I_{\mathrm{DL}}]|}_{\text{statistical error}}. \qquad (10)
$$

Emulating the reckoning by Feng and Marzouk (2019), with a second-order Taylor approximation, the bias component reads as

$$
|\mathbb{E}[I_{\mathrm{DL}}] - I| \approx \left| \frac{C_2}{M_2} - \frac{C_1}{M_1} \right|. \qquad (11)
$$

In (11), we have the difference between two positive terms; thus, these terms could cancel each other. In this scenario, the error estimation depends on the ignored higher-order terms. Thus, we lose all knowledge of the error estimates. To avoid this, we use the triangular inequality to arrive at a slightly more conservative error estimate:

$$
\begin{aligned}
|\mathbb{E}[I_{\mathrm{DL}}] - I| &\approx \left| \frac{C_2}{M_2} - \frac{C_1}{M_1} \right|, \\
&\leq \frac{C_2}{M_2} + \frac{C_1}{M_1}. \qquad (12)
\end{aligned}
$$

For the statistical error, we also recall (Feng and Marzouk 2019) and express the variance of the EIG estimator as follows:

$$
\mathbb{V}[I_{\mathrm{DL}}] \approx \frac{D_3}{N} + \frac{D_1}{NM_1} + \frac{D_2}{NM_2}, \qquad (13)
$$

where the constants $C_i > 0$ and $D_i > 0$ depend on the design parameter $\boldsymbol{\xi}$. Although we base this part of the work on Feng and Marzouk (2019), the error estimates differ because we use (12) instead of (11). Furthermore, Feng and Marzouk (2019) consider that $M_1 = M_2$ for simplicity, whereas we derive an allocation of inner samples that is proportional to the constants C_1 and C_2, following (Bartuska et al. 2022 Appendix A).

The constants appearing in (12) and (13) are given as follows (see Appendix A):

$$
C_1 = \frac{1}{2} \mathbb{E} \left[\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\varphi})|\boldsymbol{Y}, \boldsymbol{\theta}]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})} \right], \qquad (14)
$$

$$
C_2 = \frac{1}{2} \mathbb{E} \left[\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\vartheta}, \boldsymbol{\varphi})|\boldsymbol{Y}]}{p^2(\boldsymbol{Y})} \right], \qquad (15)
$$

$$
D_1 = \mathbb{E} \left[\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\varphi})|\boldsymbol{Y}, \boldsymbol{\theta}]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})} \right], \qquad (16)
$$

$$
D_2 = \mathbb{E} \left[\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\vartheta}, \boldsymbol{\varphi})|\boldsymbol{Y}]}{p^2(\boldsymbol{Y})} \right], \qquad (17)
$$

and

$$
D_3 = \mathbb{V} \left[\log \left(\frac{p(\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{Y})} \right) \right]. \qquad (18)
$$

For a given error tolerance $TOL > 0$, we can introduce a splitting parameter $\kappa \in (0, 1)$ to distribute the error between the bias and statistical error. Thus, aided by the central limit theorem, we arrive at the following constraints:

$$
|\mathbb{E}[I_{\mathrm{DL}}] - I| \leq (1 - \kappa) TOL, \qquad (19)
$$

$$
\mathbb{V}[I_{\mathrm{DL}}] \leq \left(\frac{\kappa TOL}{C_\alpha} \right)^2, \qquad (20)
$$

where $C_\alpha = \Phi^{-1}(1 - \alpha/2)$ is the constant depending on the chosen confidence level α, and Φ^{-1} is the inverse cumulative distribution function of a standard normal random variable.

To determine the optimal number of samples, we formulate the following subproblem. We find N, M_1, M_2, and κ arising from the minimization of the total work (9) subject to the bias and statistical error constraints (19) and (20):

$$\begin{aligned} N^*, M_1^*, M_2^*, \kappa^* := \underset{N, M_1, M_2, \kappa}{\arg\min} \; & N(M_1 + M_2) \\ \text{s.t.} \; & \frac{C_1}{M_1} + \frac{C_2}{M_2} \leq (1-\kappa) TOL, \\ & \frac{1}{N}\left(D_3 + \frac{D_1}{M_1} + \frac{D_2}{M_2}\right) \leq \left(\frac{\kappa TOL}{C_\alpha}\right)^2, \\ & N, M_1, M_2 > 0, \\ & 1 > \kappa > 0. \end{aligned} \tag{21}$$

In Appendix B, we solve this problem analytically. We present the optimal sample choices as follows:

$$\begin{cases} \kappa^* = \dfrac{8TOL + 3D_3 - \sqrt{16TOLD_3 + 9D_3^2}}{8TOL}, \\ N^* = \dfrac{C_\alpha^2 \left(D_3 + 2(1-\kappa^*)TOL\right)}{(\kappa^*)^2 TOL^2}, \\ M_1^* = \dfrac{C_1 + \sqrt{C_1 C_2}}{(1-\kappa^*)TOL}, \\ M_2^* = \dfrac{C_2 + \sqrt{C_1 C_2}}{(1-\kappa^*)TOL}, \end{cases} \tag{22}$$

whereas the total optimal work reads as

$$\begin{aligned} W^* \; & \propto \left(\frac{C_\alpha^2 \left(D_3 + 2(1-\kappa^*)TOL\right)}{(\kappa^*)^2 TOL^2}\right)\left(\frac{(\sqrt{C_1} + \sqrt{C_2})^2}{(1-\kappa^*)TOL}\right) \\ & \propto TOL^{-3}. \end{aligned} \tag{23}$$

As TOL approaches infinity, κ approaches 1. Thus, for large tolerances, only the statistical error is relevant. As TOL approaches 0, κ approaches 2/3, which can be observed from the fact that the statistical error (13) can be rewritten to include the bias error using (14)–(17) as follows:

$$\mathbb{V}[I_{\mathrm{DL}}] \approx \frac{1}{N}\left(D_3 + 2|\mathbb{E}[I_{\mathrm{DL}}] - I|\right). \tag{24}$$

Thus, the statistical error always takes priority over the bias error.

When the bias error due to the numerical discretization of the forward problem is also considered, we obtain the following term for the average computational work:

$$W(I_{\mathrm{DL}}^h) \propto N(M_1 + M_2)h^{-\gamma}, \tag{25}$$

where $h^{-\gamma}$ is proportional to the average work of evaluating the forward model $\boldsymbol{g}$ with discretization parameter h. The subproblem (21) becomes

$$\begin{aligned} N^*, M_1^*, M_2^*, h^*, \kappa^* := \underset{N, M_1, M_2, h, \kappa}{\arg\min} \; & N(M_1 + M_2)h^{-\gamma} \\ \text{s.t.} \; & C_3 h^\eta + \frac{C_1}{M_1} + \frac{C_2}{M_2} \leq (1-\kappa) TOL, \\ & \frac{1}{N}\left(D_3 + \frac{D_1}{M_1} + \frac{D_2}{M_2}\right) \\ & \leq \left(\frac{\kappa TOL}{C_\alpha}\right)^2, \\ & N, M_1, M_2, h > 0, \\ & 1 > \kappa > 0, \end{aligned} \tag{26}$$

where η is the weak convergence rate of the discretization method. We obtain the solution

$$\begin{cases} \kappa^* = \eta \dfrac{8TOL\eta + 3D_3\eta + D_3\gamma + 4TOL\gamma - \sqrt{D_3\left(9D_3\eta^2 + 6D_3\eta\gamma + D_3\gamma^2 + 16\eta^2 TOL + 8TOL\eta\gamma\right)}}{2TOL\left(4\eta^2 + 4\eta\gamma + \gamma^2\right)}, \\ N^* = \dfrac{C_\alpha^2}{\kappa^2 TOL}\left(\dfrac{D_3}{TOL} + 2\left(1 - \kappa\left(1 + \dfrac{\gamma}{2\eta}\right)\right)\right), \\ M_1^* = \dfrac{C_1\left(1 + \sqrt{\frac{C_2}{C_1}}\right)}{\left(1 - \kappa\left(1 + \frac{\gamma}{2\eta}\right)\right)TOL}, \\ M_2^* = \dfrac{C_2\left(1 + \sqrt{\frac{C_1}{C_2}}\right)}{\left(1 - \kappa\left(1 + \frac{\gamma}{2\eta}\right)\right)TOL}, \\ h^* = \left(\dfrac{\gamma\kappa TOL}{2\eta C_3}\right)^{\frac{1}{\eta}}. \end{cases} \tag{27}$$

For the derivation of this result, see Appendix C and Bartuska et al. (2022); Beck et al. (2018).

4 Laplace approximation

4.1 Laplace's method

The DLMC estimator is expensive and often suffers from numerical underflow (Beck et al. 2018). Some inexpensive and accurate estimators have been proposed for the EIG using importance sampling (Beck et al. 2018) or multilevel techniques combined with importance sampling (Beck et al. 2020) to reduce the computational cost. In addition, the inner integral may be estimated directly using the Laplace approximation. In this work, we enhance the computational performance of the DLMC estimator with nuisance uncertainty by deriving suitable Laplace approximations to marginalize the likelihoods arising in the Bayes update and integrate the inner loop. The Laplace approximation for the Bayes update in this setting is novel. Through what follows, we demonstrate that the Laplace approximation for the inner loop differs from that in Long et al. (2013); Beck et al. (2018).

Next, using (4), the posterior (2) may be expressed as follows:

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{Y}) &= \frac{\pi(\boldsymbol{\theta})}{p(\boldsymbol{Y})\det(2\pi\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})^{\frac{N_e}{2}}} \\ &\quad\times\int_{\boldsymbol{\Phi}}\exp\left(-\frac{1}{2}\sum_{i=1}^{N_e}\boldsymbol{r}(\boldsymbol{y}_i,\boldsymbol{\theta},\boldsymbol{\phi})\cdot\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{r}(\boldsymbol{y}_i,\boldsymbol{\theta},\boldsymbol{\phi})\right)\pi(\boldsymbol{\phi}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\phi}, \\ &= \frac{\pi(\boldsymbol{\theta})}{p(\boldsymbol{Y})\det(2\pi\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})^{\frac{N_e}{2}}} \\ &\quad\int_{\boldsymbol{\Phi}}\exp\underbrace{\left(-\frac{1}{2}\sum_{i=1}^{N_e}\boldsymbol{r}(\boldsymbol{y}_i,\boldsymbol{\theta},\boldsymbol{\phi})\cdot\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{r}(\boldsymbol{y}_i,\boldsymbol{\theta},\boldsymbol{\phi})+\log(\pi(\boldsymbol{\phi}|\boldsymbol{\theta}))\right)}_{:=-f(\boldsymbol{\theta},\boldsymbol{\phi})}\mathrm{d}\boldsymbol{\phi}, \\ &= \frac{\pi(\boldsymbol{\theta})}{p(\boldsymbol{Y})\det(2\pi\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})^{\frac{N_e}{2}}}\int_{\boldsymbol{\Phi}}e^{-f(\boldsymbol{\theta},\boldsymbol{\phi})}\mathrm{d}\boldsymbol{\phi}. \end{aligned} \tag{28}$$

We approximate the integral in (28) using Laplace's method. Assuming that $f(\boldsymbol{\theta},\boldsymbol{\phi})$ has a unique minimum in $\boldsymbol{\phi}$, and that its Hessian in $\boldsymbol{\phi}$ is negative definite for almost all $\boldsymbol{\theta}$, we write the second-order Taylor approximation of $f(\boldsymbol{\theta},\boldsymbol{\phi})$ in $\boldsymbol{\phi}$ around its minimum in $\boldsymbol{\phi}$, denoted as $\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})$ and given by

$$\begin{aligned}\hat{\boldsymbol{\phi}}(\boldsymbol{\theta}) &= \arg\min_{\boldsymbol{\phi}} f(\boldsymbol{\theta},\boldsymbol{\phi}) \\ &= \arg\min_{\boldsymbol{\phi}}\left(\frac{1}{2}\sum_{i=1}^{N_e}\boldsymbol{r}(\boldsymbol{y}_i,\boldsymbol{\theta},\boldsymbol{\phi})\cdot\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{r}(\boldsymbol{y}_i,\boldsymbol{\theta},\boldsymbol{\phi})\right. \\ &\quad\left.-\log(\pi(\boldsymbol{\phi}|\boldsymbol{\theta}))\right), \\ &= \arg\min_{\boldsymbol{\phi}}\left(\frac{1}{2}N_e(\boldsymbol{g}(\boldsymbol{\theta}_t,\boldsymbol{\phi}_t)-\boldsymbol{g}(\boldsymbol{\theta},\boldsymbol{\phi}))\cdot\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_t,\boldsymbol{\phi}_t)\right. \\ &\quad-\boldsymbol{g}(\boldsymbol{\theta},\boldsymbol{\phi})), \\ &\quad\left.+\sum_{i=1}^{N_e}\boldsymbol{\epsilon}_i\cdot\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\boldsymbol{g}(\boldsymbol{\theta}_t,\boldsymbol{\phi}_t)-\boldsymbol{g}(\boldsymbol{\theta},\boldsymbol{\phi}))-\log(\pi(\boldsymbol{\phi}|\boldsymbol{\theta}))\right). \end{aligned} \tag{29}$$

The Taylor expansion in $\boldsymbol{\phi}$ reads as

$$\begin{aligned}\tilde{f}(\boldsymbol{\theta},\boldsymbol{\phi}) &= \underbrace{f(\boldsymbol{\theta},\hat{\boldsymbol{\phi}}(\boldsymbol{\theta}))}_{\text{const. in }\boldsymbol{\phi}}+\underbrace{\nabla_{\boldsymbol{\phi}}f(\boldsymbol{\theta},\hat{\boldsymbol{\phi}}(\boldsymbol{\theta}))}_{=\mathbf{0}}\cdot(\boldsymbol{\phi}-\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) \\ &\quad+\frac{1}{2}(\boldsymbol{\phi}-\hat{\boldsymbol{\phi}}(\boldsymbol{\theta}))\cdot\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\phi}}f(\boldsymbol{\theta},\hat{\boldsymbol{\phi}}(\boldsymbol{\theta}))(\boldsymbol{\phi}-\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})), \end{aligned} \tag{30}$$

with the following terms:

$$\begin{aligned}f(\boldsymbol{\theta},\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) &= \frac{1}{2}\sum_{i=1}^{N_e}\boldsymbol{r}(\boldsymbol{y}_i,\boldsymbol{\theta},\hat{\boldsymbol{\phi}}(\boldsymbol{\theta}))\cdot\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{r}(\boldsymbol{y}_i,\boldsymbol{\theta},\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) \\ &\quad-\log(\pi(\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})|\boldsymbol{\theta})), \end{aligned} \tag{31}$$

$$\begin{aligned}\nabla_{\boldsymbol{\phi}}f(\boldsymbol{\theta},\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) &= -\sum_{i=1}^{N_e}\nabla_{\boldsymbol{\phi}}\boldsymbol{g}(\boldsymbol{\theta},\hat{\boldsymbol{\phi}}(\boldsymbol{\theta}))^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{r}(\boldsymbol{y}_i,\boldsymbol{\theta},\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) \\ &\quad-\nabla_{\boldsymbol{\phi}}\log(\pi(\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})|\boldsymbol{\theta})), \end{aligned} \tag{32}$$

and

$$\begin{aligned}\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\phi}}f(\boldsymbol{\theta},\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) &= -\sum_{i=1}^{N_e}\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\phi}}\boldsymbol{g}(\boldsymbol{\theta},\hat{\boldsymbol{\phi}}(\boldsymbol{\theta}))^{\top} \\ &\quad\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{r}(\boldsymbol{y}_i,\boldsymbol{\theta},\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) \end{aligned}$$

$$+ N_e \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) - \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \log(\pi(\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})|\boldsymbol{\theta})), \tag{33}$$

where $\nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) \in \mathbb{R}^{d_y \times d_\phi}$ is the Jacobian of $\boldsymbol{g}$ with respect to $\boldsymbol{\phi}$, evaluated at $\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})$. In addition, $\nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})))$ $\in \mathbb{R}^{d_y \times d_\phi \times d_\phi}$ is the Hessian of $\boldsymbol{g}$ with respect to $\boldsymbol{\phi}$, evaluated at $\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})$. Because $\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})$ is the minimizer of $f(\boldsymbol{\theta}, \boldsymbol{\phi})$, $\nabla_{\boldsymbol{\phi}} f(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) = \mathbf{0}$.

In index notation, using Einstein's convention for summation with Greek letters, we obtain

$$f = \frac{1}{2} \sum_{i=1}^{N_e} \boldsymbol{r}_\alpha (\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1})_{\alpha\beta} \boldsymbol{r}_\beta - \log(\pi(\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})|\boldsymbol{\theta})), \tag{34}$$

$$(\nabla_{\boldsymbol{\phi}} f)_\gamma = - \sum_{i=1}^{N_e} (\partial_\gamma \boldsymbol{g}_\alpha)(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1})_{\alpha\beta} \boldsymbol{r}_\beta - \partial_\gamma (\log(\pi(\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})|\boldsymbol{\theta}))), \tag{35}$$

and

$$(\nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} f)_{\gamma\zeta} = - \sum_{i=1}^{N_e} (\partial_\zeta \partial_\gamma \boldsymbol{g}_\alpha)(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1})_{\alpha\beta} \boldsymbol{r}_\beta + N_e (\partial_\gamma \boldsymbol{g}_\alpha)(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1})_{\alpha\beta} (\partial_\zeta \boldsymbol{g}_\beta) - \partial_\zeta \partial_\gamma (\log(\pi(\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})|\boldsymbol{\theta}))). \tag{36}$$

Laplace's method reads as

$$\int_{\boldsymbol{\Phi}} e^{-f(\boldsymbol{\theta}, \boldsymbol{\phi})} \, \mathrm{d}\boldsymbol{\phi} \approx \frac{(2\pi)^{\frac{d_\phi}{2}}}{\det(\nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} f(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})))^{\frac{1}{2}}} e^{-f(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta}))}, \tag{37}$$

yielding the following approximation of the posterior:

$$\pi_{\hat{\boldsymbol{\phi}}}(\boldsymbol{\theta}|\boldsymbol{Y}) := \frac{\pi(\boldsymbol{\theta}) \pi(\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})|\boldsymbol{\theta}) (2\pi)^{\frac{d_\phi}{2}}}{p(\boldsymbol{Y}) \det(2\pi \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})^{\frac{N_e}{2}} \det(\nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} f(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})))^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \sum_{i=1}^{N_e} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) \right). \tag{38}$$

Moreover, (38) is not Gaussian unless the prior is Gaussian, and the model $\boldsymbol{g}$ is linear on $\boldsymbol{\theta}$.

4.2 Laplace approximation

Assuming that (38) has a unique minimum and negative definite Hessian as a function of $\boldsymbol{\theta}$, we follow the approach in Long et al. (2013) to obtain a Gaussian approximation of the posterior distribution, taking the negative logarithm of (38) and evolving it around the maximum a posteriori estimate $\hat{\boldsymbol{\theta}}$ up to second order:

$$F(\boldsymbol{\theta}) := - \log(\pi_{\hat{\boldsymbol{\phi}}}(\boldsymbol{\theta}|\boldsymbol{Y})), = \frac{1}{2} \sum_{i=1}^{N_e} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) - h(\boldsymbol{\theta}) + k(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}) + C, \tag{39}$$

where

$$h(\boldsymbol{\theta}) := \log(\pi(\boldsymbol{\theta})),$$
$$k(\boldsymbol{\theta}) := \frac{1}{2} \log(\det(\nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} f(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})))),$$

and

$$\ell(\boldsymbol{\theta}) := \log(\pi(\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})|\boldsymbol{\theta})).$$

The last two terms are new compared to their approach, and C is a constant.

The maximum a posteriori estimate $\hat{\boldsymbol{\theta}}$ of (38) is given by

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \pi_{\hat{\boldsymbol{\phi}}}(\boldsymbol{\theta}|\boldsymbol{Y}), \\ &= \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}), \\ &= \arg\min_{\boldsymbol{\theta}} \left[\frac{1}{2} \sum_{i=1}^{N_e} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) - h(\boldsymbol{\theta}) + k(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}) \right]. \end{aligned} \tag{40}$$

The Taylor expansion is given by

$$\tilde{F}(\boldsymbol{\theta}) = \underbrace{F(\hat{\boldsymbol{\theta}})}_{\text{const. in } \boldsymbol{\theta}} + \underbrace{\nabla_{\boldsymbol{\theta}} F(\hat{\boldsymbol{\theta}})}_{=\mathbf{0}} \cdot (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \cdot \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} F(\hat{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}). \tag{41}$$

From (39), we obtain

$$\nabla_{\boldsymbol{\theta}} F(\hat{\boldsymbol{\theta}}) = - \sum_{i=1}^{N_e} \left(\nabla_z \boldsymbol{g}(z(\hat{\boldsymbol{\theta}})) \nabla_{\boldsymbol{\theta}} z(\hat{\boldsymbol{\theta}}) \right)^\top \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{r}(\boldsymbol{y}_i, z(\hat{\boldsymbol{\theta}})) - \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}) + \nabla_{\boldsymbol{\theta}} k(\hat{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}), \tag{42}$$

by the chain rule, where we write

$$z(\hat{\boldsymbol{\theta}}) := (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}(\hat{\boldsymbol{\theta}})) \in \mathbb{R}^{d_\theta + d_\phi}, \tag{43}$$

and

$$\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} F(\hat{\boldsymbol{\theta}}) = - \sum_{i=1}^{N_e} \Big(\nabla_{\boldsymbol{\theta}} z(\hat{\boldsymbol{\theta}})^\top \nabla_z \nabla_z \boldsymbol{g}(z(\hat{\boldsymbol{\theta}})) \nabla_{\boldsymbol{\theta}} z(\hat{\boldsymbol{\theta}})$$

$$+\nabla_z \boldsymbol{g}(z(\hat{\boldsymbol{\theta}}))\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} z(\hat{\boldsymbol{\theta}})\Big)^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{r}(\boldsymbol{y}_i, z(\hat{\boldsymbol{\theta}})) + N_e \left(\nabla_z \boldsymbol{g}(z(\hat{\boldsymbol{\theta}}))\nabla_{\boldsymbol{\theta}} z(\hat{\boldsymbol{\theta}})\right)^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \left(\nabla_z \boldsymbol{g}(z(\hat{\boldsymbol{\theta}}))\nabla_{\boldsymbol{\theta}} z(\hat{\boldsymbol{\theta}})\right) - \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}) + \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} k(\hat{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}). \tag{44}$$

Given (39), (40), (41), and (44) we find the Gaussian approximation of the posterior at $\hat{\boldsymbol{\theta}}$:

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y}) := \frac{1}{\det(2\pi\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\cdot\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}{2}\right), \tag{45}$$

where $\boldsymbol{\Sigma}^{-1} = \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} F(\hat{\boldsymbol{\theta}})$. The first term in (44) is $\mathcal{O}_{\mathbb{P}}(\sqrt{N_e})$,[1] and the second term is $\mathcal{O}_{\mathbb{P}}(N_e)$ (see Long et al. 2013 and Appendix D). The last two terms are of order $\mathcal{O}_{\mathbb{P}}(1)$.

Remark 1 (Differences and additional expenses in accounting for nuisance uncertainty in the Laplace approximation of the posterior) In this setting, we obtain terms relating to $z(\hat{\boldsymbol{\theta}})$, k, and ℓ, requiring additional Jacobian and Hessian evaluations to adequately account for nuisance uncertainty.

Remark 2 (Uniqueness of the minimum) The Laplace approximation and Laplace's method can be applied to functions without a unique minimum, see Bornkamp (2011) and Long (2022). However, the examples considered in Sect. 6 displayed behavior consistent with the assumptions stated at the beginning of Sect. 4.1 and Sect. 4.2.

5 Expected information gain estimators

5.1 Double Laplace approximation: Monte Carlo double Laplace

The first estimator to compute the EIG uses Laplace's method and a Laplace approximation. We begin by rewriting the log ratio between posterior and prior:

$$\log\left(\frac{\pi(\boldsymbol{\theta}|\boldsymbol{Y})}{\pi(\boldsymbol{\theta})}\right) = \underbrace{\log\left(\frac{\pi(\boldsymbol{\theta}|\boldsymbol{Y})}{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})}\right)}_{:=\epsilon_{La}} + \log\left(\frac{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})}{\pi(\boldsymbol{\theta})}\right). \tag{46}$$

[1] The notation $X_M = \mathcal{O}_{\mathbb{P}}(a_M)$ for a sequence of random variables X_M and constants a_M is as follows. For any $\epsilon > 0$, there exists a finite $K(\epsilon) > 0$ and finite $M_0 > 0$ such that $\mathbb{P}(|X_M| > K(\epsilon)|a_M|) < \epsilon$ holds for all $M \geq M_0$.

Using the second Laplace approximation given in (45), we obtain

$$\log\left(\frac{\pi(\boldsymbol{\theta}|\boldsymbol{Y})}{\pi(\boldsymbol{\theta})}\right) = \epsilon_{La} - \frac{1}{2}\log(\det(2\pi\boldsymbol{\Sigma})) - \left(\frac{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\cdot\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}{2}\right) - \log(\pi(\boldsymbol{\theta})). \tag{47}$$

This formulation is identical to that Long et al. (2013), and the only difference in the formulation without nuisance uncertainty is encoded in the covariance matrix $\boldsymbol{\Sigma}$.

Next, we write the Kullback–Leibler divergence (5) as follows:

$$\begin{aligned} D_{\mathrm{KL}} &= \int_{\boldsymbol{\Theta}} \log\left(\frac{\pi(\boldsymbol{\theta}|\boldsymbol{Y})}{\pi(\boldsymbol{\theta})}\right)\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})\,\mathrm{d}\boldsymbol{\theta} \\ &\quad + \underbrace{\int_{\boldsymbol{\Theta}} \log\left(\frac{\pi(\boldsymbol{\theta}|\boldsymbol{Y})}{\pi(\boldsymbol{\theta})}\right)(\pi(\boldsymbol{\theta}|\boldsymbol{Y}) - \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y}))\,\mathrm{d}\boldsymbol{\theta}}_{:=\epsilon_{int}}, \\ &= \int_{\boldsymbol{\Theta}} \epsilon_{La}\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})\,\mathrm{d}\boldsymbol{\theta} + \int_{\boldsymbol{\Theta}} \Bigg[-\frac{1}{2}\log(\det(2\pi\boldsymbol{\Sigma})) \\ &\quad - \left(\frac{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\cdot\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}{2}\right) \\ &\quad - \log(\pi(\boldsymbol{\theta}))\Bigg]\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})\,\mathrm{d}\boldsymbol{\theta} + \epsilon_{int}, \\ &= -\frac{1}{2}\log(\det(2\pi\boldsymbol{\Sigma})) - \frac{d_\theta}{2} - \log(\pi(\hat{\boldsymbol{\theta}})) \\ &\quad - \frac{\mathrm{tr}(\boldsymbol{\Sigma}\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\log(\pi(\hat{\boldsymbol{\theta}})))}{2} + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{N_e^2}\right), \end{aligned} \tag{48}$$

where ϵ_{La}, $\epsilon_{int} = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{N_e^2}\right)$ (see Long et al. 2013 Appendices A, B, and C) and $\log(\pi(\boldsymbol{\theta})) = \log(\pi(\hat{\boldsymbol{\theta}})) + \frac{\mathrm{tr}(\boldsymbol{\Sigma}\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\log(\pi(\hat{\boldsymbol{\theta}})))}{2} + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{N_e^2}\right)$ (see Long et al. 2013 Appendix A).

Taking the expected value over all $\boldsymbol{Y}$ and using the law of total probability results in the following EIG:

$$\begin{aligned} I &= \int_{\boldsymbol{\Theta}}\int_{\boldsymbol{\Phi}}\int_{\mathcal{Y}} \Bigg[-\frac{1}{2}\log(\det(2\pi\boldsymbol{\Sigma})) - \frac{d_\theta}{2} \\ &\quad - \log(\pi(\hat{\boldsymbol{\theta}})) - \frac{\mathrm{tr}(\boldsymbol{\Sigma}\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\log(\pi(\hat{\boldsymbol{\theta}})))}{2}\Bigg] \\ &\quad \times p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{Y}\pi(\boldsymbol{\theta},\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\phi}\,\mathrm{d}\boldsymbol{\theta} + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{N_e^2}\right), \end{aligned} \tag{49}$$

whereas its sample-based version reads as

$$\begin{aligned} \hat{I} &= \frac{1}{N}\sum_{n=1}^{N}\Bigg[-\frac{1}{2}\log(\det(2\pi\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}(\boldsymbol{Y}^{(n)})))) - \frac{d_\theta}{2} \\ &\quad - \log(\pi(\hat{\boldsymbol{\theta}}(\boldsymbol{Y}^{(n)}))) - \frac{\mathrm{tr}(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}(\boldsymbol{Y}^{(n)}))\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\log(\pi(\hat{\boldsymbol{\theta}}(\boldsymbol{Y}^{(n)}))))}{2}\Bigg], \end{aligned} \tag{50}$$

where $\boldsymbol{\Sigma}$ is the inverse Hessian (44) evaluated at $\hat{\boldsymbol{\theta}}$. Both $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\theta}}$ depend on $\boldsymbol{Y}^{(n)} \overset{\text{iid}}{\sim} p(\boldsymbol{Y}|\boldsymbol{\theta}^{(n)}, \boldsymbol{\phi}^{(n)})$, where $(\boldsymbol{\theta}^{(n)}, \boldsymbol{\phi}^{(n)}) \overset{\text{iid}}{\sim} \pi(\boldsymbol{\theta}, \boldsymbol{\phi})$, $1 \leq n \leq N$, are sampled from the prior. The optimal setting for this estimator was derived in Beck et al. (2018).

5.2 Double Laplace-based importance sampling: double loop Monte Carlo double importance sampling

Rather than directly approximating the inner integrals, we can also use the Laplace approximation as a change of measure in the EIG (7). This method is known as importance sampling and reduces the variances (14), (15), (16), and (17) in the inner MC loops, reducing the total work (9). We aim to estimate the following:

$$I = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Phi}} \int_{\mathcal{Y}} \left[\log \left(\int_{\boldsymbol{\Phi}} p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\varphi}) \frac{\pi(\boldsymbol{\varphi}|\boldsymbol{\theta})}{\tilde{\pi}(\boldsymbol{\varphi}|\boldsymbol{Y}, \boldsymbol{\theta})} \tilde{\pi}(\boldsymbol{\varphi}|\boldsymbol{Y}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\varphi} \right) \right.$$
$$\left. - \log \left(\int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Phi}} p(\boldsymbol{Y}|\boldsymbol{\vartheta}, \boldsymbol{\varphi}) \frac{\pi(\boldsymbol{\vartheta}, \boldsymbol{\varphi})}{\tilde{\pi}(\boldsymbol{\vartheta}, \boldsymbol{\varphi}|\boldsymbol{Y})} \tilde{\pi}(\boldsymbol{\vartheta}, \boldsymbol{\varphi}|\boldsymbol{Y}) \, \mathrm{d}\boldsymbol{\varphi} \, \mathrm{d}\boldsymbol{\vartheta} \right) \right]$$
$$p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\phi}) \, \mathrm{d}\boldsymbol{Y} \pi(\boldsymbol{\theta}, \boldsymbol{\phi}) \, \mathrm{d}\boldsymbol{\phi} \, \mathrm{d}\boldsymbol{\theta}, \tag{51}$$

where $\tilde{\pi}(\boldsymbol{\varphi}|\boldsymbol{Y}, \boldsymbol{\theta})$ is the Laplace approximation of $\pi(\boldsymbol{\varphi}|\boldsymbol{Y}, \boldsymbol{\theta})$ and $\tilde{\pi}(\boldsymbol{\vartheta}, \boldsymbol{\varphi}|\boldsymbol{Y})$ is the Laplace approximation of $\pi(\boldsymbol{\vartheta}, \boldsymbol{\varphi}|\boldsymbol{Y})$. These Laplace approximations are different from the one developed in the previous section. The data $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{N_e})$ in (51) can be decomposed into $\boldsymbol{y}_i = \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}) + \boldsymbol{\epsilon}_i$, $1 \leq i \leq N_e$; therefore, we write the distribution of $\boldsymbol{\varphi}$ as follows:

$$\pi(\boldsymbol{\varphi}|\boldsymbol{Y}, \boldsymbol{\theta}) \propto \pi(\boldsymbol{\varphi}|\boldsymbol{\theta}) p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\varphi}), \tag{52}$$

where

$$p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\varphi}) \propto \exp\left(-\frac{1}{2} \sum_{i=1}^{N_e} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{\varphi}) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{\varphi}) \right), \tag{53}$$

and

$$\begin{aligned} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{\varphi}) &= \boldsymbol{y}_i - \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\varphi}), \\ &= \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}) + \boldsymbol{\epsilon}_i - \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\varphi}), \quad 1 \leq i \leq N_e. \end{aligned} \tag{54}$$

Following the steps in Long et al. (2013), we arrive at

$$\tilde{\pi}(\boldsymbol{\varphi}|\boldsymbol{Y}, \boldsymbol{\theta}) := \frac{1}{\det(2\pi \boldsymbol{\Sigma}_{\boldsymbol{\phi}})^{\frac{1}{2}}} \exp \times \left(-\frac{(\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1} (\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}})}{2} \right), \tag{55}$$

where

$$\hat{\boldsymbol{\varphi}} = \arg\min_{\boldsymbol{\varphi}} \left[\frac{1}{2} \sum_{i=1}^{N_e} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{\varphi}) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{\varphi}) - \log(\pi(\boldsymbol{\varphi}|\boldsymbol{\theta})) \right], \tag{56}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1} = -\sum_{i=1}^{N_e} \nabla_{\boldsymbol{\varphi}} \nabla_{\boldsymbol{\varphi}} \boldsymbol{g}(\boldsymbol{\theta}, \hat{\boldsymbol{\varphi}})^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}})$$
$$+ N_e \nabla_{\boldsymbol{\varphi}} \boldsymbol{g}(\boldsymbol{\theta}, \hat{\boldsymbol{\varphi}})^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \nabla_{\boldsymbol{\varphi}} \boldsymbol{g}(\boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}) - \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \log(\pi(\hat{\boldsymbol{\varphi}}|\boldsymbol{\theta})). \tag{57}$$

We drop the first term in (57), as it is $\mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right)$. As for the joint density $\pi(\boldsymbol{\vartheta}, \boldsymbol{\varphi}|\boldsymbol{Y})$, we write $\boldsymbol{z} := (\boldsymbol{\vartheta}, \boldsymbol{\varphi}) \in \mathbb{R}^{d_\theta + d_\phi}$ to obtain the Laplace approximation:

$$\tilde{\pi}(\boldsymbol{z}|\boldsymbol{Y}) := \frac{1}{\det(2\pi \boldsymbol{\Sigma}_z)^{\frac{1}{2}}} \exp\left(-\frac{(\boldsymbol{z} - \hat{\boldsymbol{z}}) \cdot \boldsymbol{\Sigma}_z^{-1} (\boldsymbol{z} - \hat{\boldsymbol{z}})}{2} \right), \tag{58}$$

with

$$\hat{\boldsymbol{z}} = \arg\min_{\boldsymbol{z}} \left[\frac{1}{2} \sum_{i=1}^{N_e} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{z}) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{r}(\boldsymbol{y}_i, \boldsymbol{z}) - \log(\pi(\boldsymbol{z})) \right], \tag{59}$$

and

$$\boldsymbol{\Sigma}_z^{-1} = -\sum_{i=1}^{N_e} \nabla_{\boldsymbol{z}} \nabla_{\boldsymbol{z}} \boldsymbol{g}(\hat{\boldsymbol{z}})^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{r}(\boldsymbol{y}_i, \hat{\boldsymbol{z}})$$
$$+ N_e \nabla_{\boldsymbol{z}} \boldsymbol{g}(\hat{\boldsymbol{z}})^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \nabla_{\boldsymbol{z}} \boldsymbol{g}(\hat{\boldsymbol{z}}) - \nabla_{\boldsymbol{z}} \nabla_{\boldsymbol{z}} \log(\pi(\hat{\boldsymbol{z}})). \tag{60}$$

This last Laplace approximation (58) has the same shape as that derived in Long et al. (2013). We again drop the first term in (60), as it is $\mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right)$. This method provides the DLMC2IS estimator, as follows:

$$\hat{I} = \frac{1}{N} \sum_{n=1}^{N} \left[\log \left(\frac{1}{M_1} \sum_{m=1}^{M_1} p(\boldsymbol{Y}^{(n)}|\boldsymbol{\theta}^{(n)}, \boldsymbol{\varphi}^{(n,m)}) \right.\right.$$
$$\left. \frac{\pi(\boldsymbol{\varphi}^{(n,m)})}{\tilde{\pi}(\boldsymbol{\varphi}^{(n,m)}|\boldsymbol{Y}^{(n)}, \boldsymbol{\theta}^{(n)})} \right)$$
$$- \log \left(\frac{1}{M_2} \sum_{k=1}^{M_2} p(\boldsymbol{Y}^{(n)}|\boldsymbol{\vartheta}^{(n,k)}, \boldsymbol{\varphi}^{(n,k)}) \right.$$
$$\left.\left. \frac{\pi(\boldsymbol{\vartheta}^{(n,k)}, \boldsymbol{\varphi}^{(n,k)})}{\tilde{\pi}(\boldsymbol{\vartheta}^{(n,k)}, \boldsymbol{\varphi}^{(n,k)}|\boldsymbol{Y}^{(n)})} \right) \right]. \tag{61}$$

First, we sample $(\boldsymbol{\theta}^{(n)}, \boldsymbol{\phi}^{(n)}) \overset{\text{iid}}{\sim} \pi(\boldsymbol{\theta}, \boldsymbol{\phi})$, then we sample $\boldsymbol{Y}^{(n)} \overset{\text{iid}}{\sim} p(\boldsymbol{Y}|\boldsymbol{\theta}^{(n)}, \boldsymbol{\phi}^{(n)})$, $\boldsymbol{\varphi}^{(n,m)} \overset{\text{iid}}{\sim} \tilde{\pi}(\boldsymbol{\varphi}|\boldsymbol{Y}^{(n)}, \boldsymbol{\theta}^{(n)})$ and $(\boldsymbol{\vartheta}^{(n,k)}, \boldsymbol{\varphi}^{(n,k)}) \overset{\text{iid}}{\sim} \tilde{\pi}(\boldsymbol{\vartheta}, \boldsymbol{\varphi}|\boldsymbol{Y}^{(n)})$. The derivation of the optimal setting for the DLMC2IS estimator follows the same basic structure as for the DLMC estimator. Laplace-based importance sampling helps reduce the variance of the inner MC approximation, and thus affects the constants appearing in the bias and variance bounds presented in Sect. 3. Moreover, the total work for the DLMC2IS estimator includes the cost of numerically estimating the MAP and Hessian for each outer sample (Beck et al. 2018). The asymptotic computational work of the DLMC estimator is

$$W_{DLMC} = \mathcal{O}\left(N \times (M_1 + M_2) \times h^{-\gamma}\right), \tag{62}$$

where N, M_1, M_2, and h are expressed as functions of the error tolerance TOL. Preasymptotically, this work may be expressed as

$$W_{DLMC} = c_1 N \times (c_2 M_1 + c_3 M_2 + 1) \times c_4 h^{-\gamma}, \tag{63}$$

for appropriately chosen constants c_1, c_2, c_3, c_4 independent of TOL. The factor 1 enters as the cost of evaluating the experiment model once per outer sample to generate the data $\boldsymbol{Y}$, which can be ignored asymptotically. For the DLMC2IS estimator, it holds that

$$W_{DLIS} = c_1 N \times (\tilde{c}_2 M_1 + \tilde{c}_3 M_2 + J) \times c_4 h^{-\gamma}, \tag{64}$$

where typically $\tilde{c}_2 \ll c_2$, $\tilde{c}_3 \ll c_3$ due to importance sampling, and $J \geq 2$ represents the cost to numerically estimate MAP and Hessian for each outer sample. This factor J, as well as $\tilde{c}_1$ and $\tilde{c}_2$, are independent of TOL, thus, it follows that

$$\begin{aligned} &c_1 N \times (\tilde{c}_2 M_1 + \tilde{c}_3 M_2 + J) \times c_4 h^{-\gamma} \\ &= c_1 N \times (\tilde{c}_2 M_1 + \tilde{c}_3 M_2) \\ &\times c_4 h^{-\gamma} + c_1 N \times J \times c_4 h^{-\gamma}, \\ &\geq (1+\epsilon) \times c_1 N \times (\tilde{c}_2 M_1 + \tilde{c}_3 M_2) \times c_4 h^{-\gamma}, \end{aligned} \tag{65}$$

for any $\epsilon > 0$ as $TOL \to 0$, where the last term in the first line is of higher order in TOL. The asymptotic work of the estimator is therefore influenced by a reduction in a multiplicative error term proportional to the combined importance sampling effects. The cost for finding MAPs and Hessians only enters as an arbitrary $\epsilon > 0$. Numerical results in Fig. 3 (Panel (A)) show that our estimator is conservative, implying the number of samples used was larger than necessary for a given tolerance.

Remark 3 (Importance sampling for a prior distribution with compact support) If the support of the prior distribution is compact, using a Gaussian distribution for importance sampling can result in samples outside this domain (see Bisetti et al. 2016). However, this problem is rarely observed in cases where the covariance of the importance sampling distribution is highly concentrated.

Remark 4 (Numerical estimation of MAPs and Hessians) All MAPs and Hessians in the following section were estimated numerically. Specific experiment models allow for closed-form expressions or the use of automatic differentiation; however, this was beyond the scope of this work.

6 Numerical results

6.1 Linear Gaussian example

This example is based on a similar formulation used by Feng and Marzouk (2019) to demonstrate the effects of nuisance parameters on the optimal design $\boldsymbol{\xi}$. We assume a linear model,

$$\boldsymbol{y}(\xi) = \boldsymbol{g}(\xi, \theta, \phi) + \boldsymbol{\epsilon}, \tag{66}$$

where

$$\begin{aligned} \boldsymbol{g}(\xi, \theta, \phi) &= \begin{pmatrix} \xi & 0 \\ 0 & (1-\xi) \end{pmatrix} \begin{pmatrix} \theta \\ \phi \end{pmatrix}, \\ &= \begin{pmatrix} \xi\theta \\ (1-\xi)\phi \end{pmatrix}, \end{aligned} \tag{67}$$

and the parameters are sampled from the following distributions:

$$\begin{aligned} &\theta \sim \mathcal{N}(0,1), \quad \phi \sim \mathcal{N}(0, 10^{-2}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}), \\ &\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \begin{pmatrix} 10^{-2} & 0 \\ 0 & 10^{-2} \end{pmatrix}. \end{aligned} \tag{68}$$

The vector $(0,0)^\top$ is denoted by $\mathbf{0}$. The design ξ is chosen from the interval $(0,1)$, and we only consider one experiment; thus, $N_e = 1$. First, we run a pilot to estimate the constants (14) to (18) required for the optimal setting of the DLMC and the DLMC2IS estimators using $N = 1000$ outer samples and $M_1 = M_2 = 200$ inner samples. The optimal number of samples and the splitting parameter for a certain tolerance $TOL > 0$ are given by (22). For the optimal setting of the MC2LA estimator, we run another pilot using $N = 1000$ outer samples to estimate the variance. The results are displayed in Fig. 1 as a function of the tolerance TOL. For the DLMC estimator, the number of required inner samples M_1^* is smaller than M_2^* because the constant C_1 only encompasses the variance from the nuisance parameter ϕ, whereas the constant C_2 encompasses the variance from θ and ϕ.

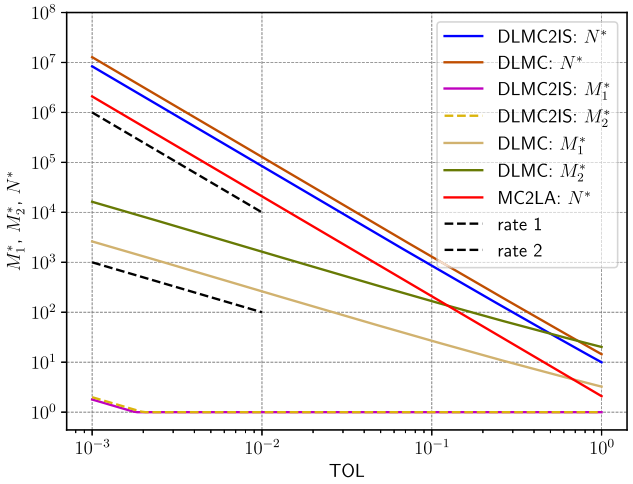


Fig. 1 Example 1: Optimal number of outer (N^*) and inner (M_1^*, M_2^*) samples vs. tolerance TOL for the DLMC, DLMC2IS, and MC2LA estimators

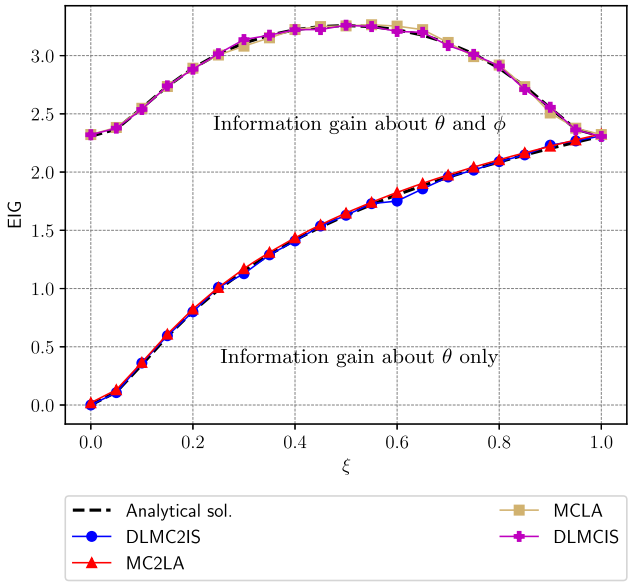


Fig. 2 Example 1: EIG vs. design parameter ξ. Analytical solution (dashed black), DLMC2IS estimator (solid blue), and MC2LA estimator (solid red) for the case with nuisance uncertainty and DLMCIS estimator (solid magenta) and MCLA estimator (solid tan) for the case without nuisance uncertainty

Importance sampling reduces the number of required inner samples to one, even for small tolerances for the DLMC2IS estimator.

We run the estimators for various designs ξ between 0 and 1. The results are presented in Fig. 2. The DLMC2IS and MC2LA estimators indicate that $\xi = 1$ is optimal. The model (67) is linear; thus, the EIG can be computed analytically. For comparison, we also present the optimal design with no nuisance uncertainty (i.e., both θ and ϕ are considered parameters of interest). For this scenario, we estimate the EIG using the DLMCIS and MCLA estimators developed in Beck et al. (2018), which only use one Laplace approximation. The optimal design is found at $\xi = 1/2$. Although a higher overall information gain occurs in this scenario, the information we gain solely about θ is less at $\xi = 1/2$ than at $\xi > 1/2$.

Figure 3 presents 100 runs of the DLMC2IS and MC2LA estimator for various tolerances and design $\xi = 1/2$. For every tolerance, the probabilistic error bounds specified by the central limit theorem with confidence constant $C_\alpha = 1.96$ ($\alpha = 0.05$) predict five runs resulting in an error greater than that tolerance. The probabilistic error bound for the DLMC2IS estimator was overly conservative, whereas the probabilistic error bound for the MC2LA estimator was overly optimistic for small tolerances. The reason for the unexpectedly small error of the DLMC2IS estimator is likely due to the fact that even using just one inner sample results in an inner variance that is much smaller than required, thanks to Laplace-based importance sampling.

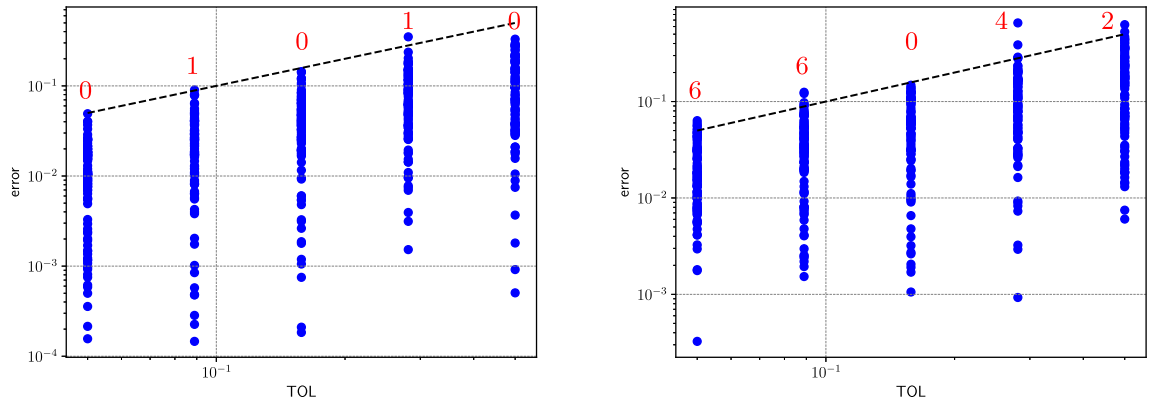
6.2 Pharmacokinetics example

The aim of pharmacokinetics is to learn patient-specific parameters in a medical setting. The following example is based on the setting introduced in Ryan et al. (2014) and modified in Goda et al. (2020). After a drug is administered to a patient at time $t_0 = 0$, a total of 15 blood samples are taken over the next 24 h to determine how fast the drug is absorbed and subsequently eliminated. We follow the simplified approach in Goda et al. (2020), where only additive noise is present. The design space is $(0, 24]^{d_\xi}$, where $d_\xi = 15$. The data model is as follows:

$$g_j(\boldsymbol{\theta}, \phi, \boldsymbol{\xi}) := \frac{D}{\phi} \frac{\theta_1}{\theta_1 - \theta_2} \left(e^{-\theta_2 \xi_j} - e^{-\theta_1 \xi_j} \right), \quad 1 \le j \le 15, \tag{69}$$

where $D = 400$ indicates the administered dose, $\log(\theta_1) \sim \mathcal{N}(0, 0.05)$ indicates the first-order absorption constant, $\log(\theta_2) \sim \mathcal{N}(\log(0.1), 0.05)$ indicates the first-order elimination constant, and $\log(\phi) \sim \mathcal{N}(\log(20), 0.05)$ indicates the volume of distribution. This last parameter is considered a nuisance parameter in the present work, deviating from the original setting in Goda et al. (2020). The design $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{15}) := (t_1, \ldots, t_{15})$ of the experiment signifies the sample times. Moreover, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$, where

$$\boldsymbol{\Sigma}_\epsilon = \begin{pmatrix} 10^{-2} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 10^{-2} \end{pmatrix} \tag{70}$$



(A) DLMC2IS (B) MC2LA

Fig. 3 Example 1: Error vs. tolerance consistency plot for various tolerances TOL with a predefined confidence parameter $C_\alpha = 1.96$ ($\alpha = 0.05$). Panel **A** DLMC2IS estimator. Panel **B** MC2LA estimator

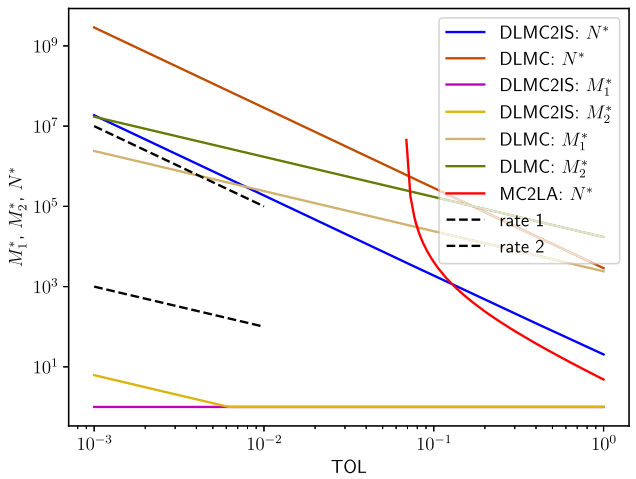


Fig. 4 Example 2: Optimal number of outer (N^*) and inner (M_1^*, M_2^*) samples vs. tolerance TOL for the DLMC, DLMC2IS, and MC2LA estimators for geometrically spaced sampling times

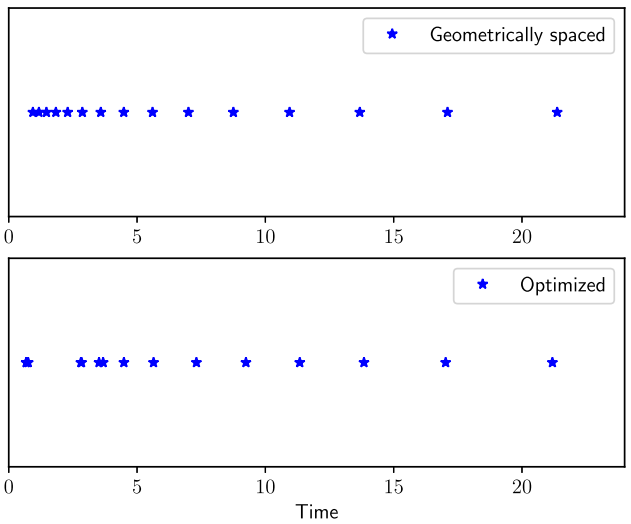


Fig. 5 Example 2: Geometrically spaced design (Goda et al. 2020) yielding expected information gain (EIG) of 6.12 vs. optimized design yielding EIG of 6.25. Clustering of measurement times appeared to have a positive impact on the EIG of the experiment. Later measurement times had little effect on the EIG and were mostly unaffected by the optimization

signifies the observation noise. The work (Goda et al. 2020) found that a geometrically spaced design $\xi_j = 0.94 \times 1.25^{j-1}$, where $1 \leq j \leq 15$, performed best for their experiments, which did not consider nuisance uncertainty. We adopt this choice as a starting point for our optimization and present the optimal number of samples for the DLMC, DLMC2IS, and MC2LA estimators in Fig. 4. For the pilot of the DLMC estimator, we used $N = 300$ outer samples and $M_1 = M_2 = 13000$ inner samples. For the pilot of the DLMC2IS estimator, we used $N = 2000$ outer samples and $M_1 = M_2 = 200$ inner samples. Finally, for the pilot of the MC2LA estimator, we used $N = 2000$ outer samples and the DLMC2IS estimator to ascertain the bias resulting from the Laplace approximation.

To improve upon this design, we used a greedy minibatch stochastic gradient descent algorithm, where $N = 300$ outer samples of the MC2LA estimator are used to optimize the EIG with respect to ξ_1 with $(\xi_2, \ldots, \xi_{15})$ fixed. Next, the EIG is optimized with respect to ξ_2 with $(\xi_1, \xi_3, \ldots, \xi_{15})$ fixed, and so on and so forth. That is, we employ a combination of stochastic gradient descent and coordinate descent methods to find a local minimum of the EIG. The geometrically spaced design is used as a starting point. A comparison between both design choices is presented in Fig. 5. Optimization of measurement times indicates that a certain clustering is beneficial for the EIG, whereas measurements towards the end of the 24 h appeared to have little effect on the EIG. A clustering effect of sampling times was also observed in Ryan et al. (2014).

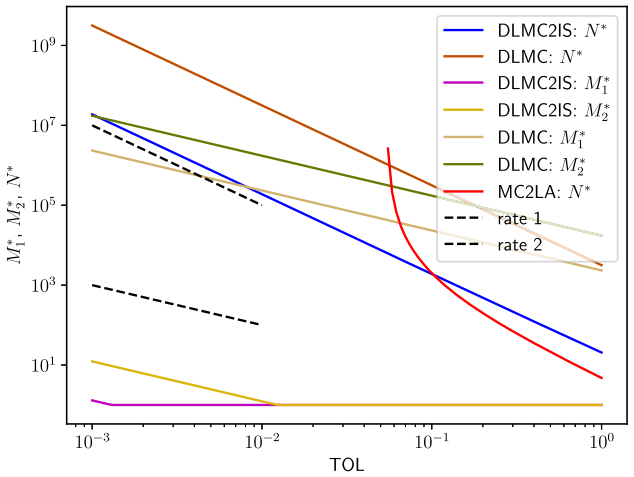


Fig. 6 Example 2: Optimal number of outer (N^*) and inner (M_1^*, M_2^*) samples vs. tolerance TOL for the DLMC, DLMC2IS, and MC2LA estimators for optimized sampling times

The optimal number of inner and outer samples for the optimized design is displayed in Fig. 6. The design had minimal effect on the optimal number of samples; however, the EIG was improved for the optimized design, as the DLMC2IS estimator with a tolerance of $TOL = 10^{-2}$ yielded an EIG of 6.25 for the optimized design and an EIG of 6.12 for the geometrically spaced design.

6.3 Electrical impedance tomography example I

For this example, we consider a more challenging model based on the solution operator of a PDE. Solving the PDE for EIT is generally not possible in closed form; therefore, we employ a finite element method (FEM) approximation instead. This example demonstrates the practical applicability of the derived estimators and their ability to incorporate approximate models $\boldsymbol{g}_h$, where h is the mesh-discretization parameter. We examine a two-dimensional model of a composite laminate material with a fiber structure. The material consists of two plies, both conducting an electric current more easily along the direction of their fibers than transversal to them. The experimental setup involves attaching five electrodes to the top and five to the bottom boundaries, injecting, and measuring electric current. The experimenter can learn about the fiber angles in each ply; hence, these angles are considered parameters of interest. The electrode positions can be chosen freely in principle but affect the meaningfulness of the measurements. Hence, this is the design the experimenter aims to optimize. We further assume that the exact conductivities of the plies are known only up to a concentrated normal distribution. Therefore, they are considered nuisance parameters. We consider a rectangular domain $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 = [0, 20] \times [0, 1] \cup [0, 20] \times [1, 2]$, consisting of two subdomains. For the quasi-static potential field u, current flux $\boldsymbol{j}$, and conductivity field $\bar{\boldsymbol{\sigma}}$, we solve the PDE

$$\nabla \cdot \boldsymbol{j}(\boldsymbol{x}, \omega) = 0 \quad \text{in } \mathcal{D} \text{ and} \tag{71}$$

$$\boldsymbol{j}(\boldsymbol{x}, \omega) = \bar{\boldsymbol{\sigma}}(\omega) \cdot \nabla u(\boldsymbol{x}, \omega), \tag{72}$$

where $\boldsymbol{x} \in \mathcal{D}$,

$$\bar{\boldsymbol{\sigma}}(\omega) = \boldsymbol{Q}(\theta_i(\omega))^\top \cdot \boldsymbol{\sigma}(\phi_i(\omega)) \cdot \boldsymbol{Q}(\theta_i(\omega)), \quad i = 1, 2, \tag{73}$$

$$\boldsymbol{Q}(\theta_i) = \begin{pmatrix} \cos(\theta_i) & 0 & -\sin(\theta_i) \\ 0 & 1 & 0 \\ \sin(\theta_i) & 0 & \cos(\theta_i) \end{pmatrix}, \quad i = 1, 2, \tag{74}$$

and

$$\boldsymbol{\sigma}(\phi_i) = \begin{pmatrix} \sigma_1(\phi_i) & 0 & 0 \\ 0 & \sigma_2(\phi_i) & 0 \\ 0 & 0 & \sigma_3(\phi_i) \end{pmatrix}, \quad i = 1, 2. \tag{75}$$

For the fiber angle in $\mathcal{D}_1$, we assume the following prior distribution:

$$\theta_1 \sim \pi(\theta_1) = \mathcal{U}\left(-\frac{\pi}{4} - 0.05, -\frac{\pi}{4} + 0.05\right) \tag{76}$$

and for the fiber angle in $\mathcal{D}_2$, we assume the prior distribution:

$$\theta_2 \sim \pi(\theta_2) = \mathcal{U}\left(\frac{\pi}{4} - 0.05, \frac{\pi}{4} + 0.05\right). \tag{77}$$

For the conductivities, we consider the prior distribution:

$$\sigma_j(\phi_i) = \exp(\mu_j + \phi_i), \quad j = 1, 2, 3, \quad i = 1, 2, \tag{78}$$

where

$$\phi_i = \sigma_\phi z_i, \quad z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \quad i = 1, 2. \tag{79}$$

Moreover, $\mu_1 = \log(0.1)$ and $\mu_2 = \mu_3 = \log(0.02)$ and σ_ϕ^2 is the covariance of ϕ_i for all $i = 1, 2$. The design $\boldsymbol{\xi} = (\xi_1, \xi_2) = [0, 2] \times [0, 2]$ signifies a shift between top and bottom electrodes and the distance between electrodes, respectively. For a detailed description of the problem setting, including boundary conditions, and a finite element formulation, see Bartuska et al. (2022); Beck et al. (2018); Somersalo et al. (1992).

As in the previous example, we start by running a pilot with $N = 50$ outer samples for the MC2LA estimator and $N = 50$ outer samples, with $M_1 = M_2 = 10$ inner samples for the DLMC2IS and DLMC estimators. The bias resulting from the Laplace approximations in the MC2LA estimator is measured by comparison with the DLMC2IS estimator results. In addition, we estimate the FEM constants η, C_{disc}, and γ. The optimal number of samples for these estimators is depicted in Fig. 7 as a function of the error tolerance TOL. Next, we compare the computational work required for the

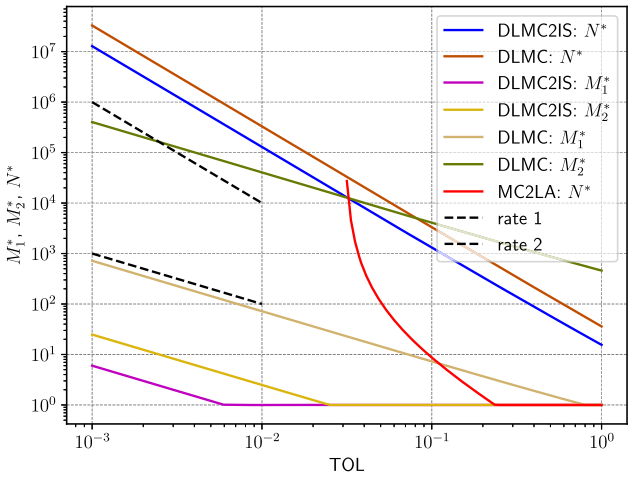


Fig. 7 Example 3: Optimal number of outer (N^*) and inner (M_1^*, M_2^*) samples vs. tolerance TOL for the DLMC, DLMC2IS, and MC2LA estimators

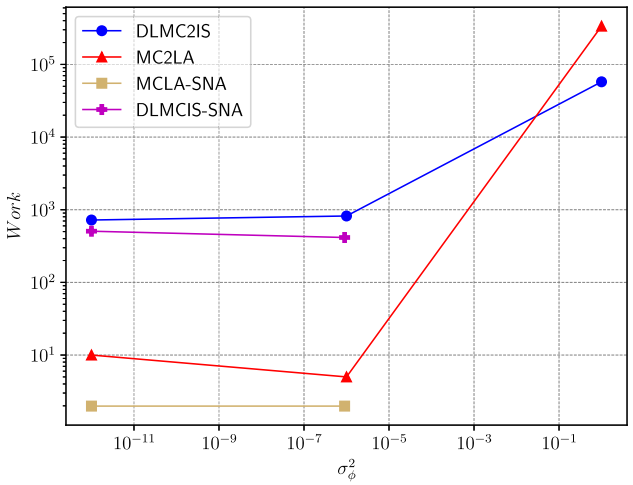


Fig. 8 Example 3: Computational work vs. nuisance covariance σ_ϕ^2 for the MCLA with small-noise approximation, DLMCIS with small-noise approximation, DLMC2IS, and MC2LA estimators

MC2LA and DLMC2IS estimators and for the MCLA and DLMCIS estimators with small-noise approximation developed in Bartuska et al. (2022) as a function of the covariance of the nuisance parameters in Fig. 8. The computational work for the MC2LA and the MCLA with small-noise approximation estimators is given by $N \times h^{-\gamma}$. The computational work for the DLMC2IS estimator is given by $N \times (M_1 + M_2) \times h^{-\gamma}$. Finally, the computational work for the DLMCIS estimator with small-noise approximation is given by $N \times M \times h^{-\gamma}$. The small-noise approximation is only applicable up to a small nuisance covariance $\sigma_\phi^2 \approx 9.03 \times 10^{-7}$. However, the cost for the MC2LA and the DLMC2IS estimators increase sharply for a larger nuisance covariance.

6.4 Electrical impedance tomography example II

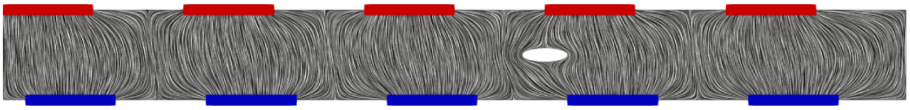
Next, we consider a slightly different setup from the previous example. The electric conductivity is now fixed at $\sigma_1 = 0.1$, $\sigma_2 = \sigma_3 = 0.02$, and no longer a nuisance parameter. Furthermore, we assume that a small ellipsoid exclusion exists between the two plies. The vertical axis of this ellipsoid is the new nuisance parameter, with the following distribution:

$$\phi \sim \mathcal{U}(0.2, 0.4). \tag{80}$$

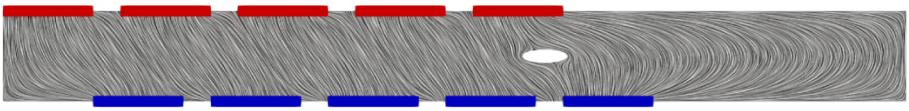
The horizontal axis is considered fixed at 1 and the center is fixed at (12,1). The parameters of interest are still the angles of the fibers in each ply. The FEM formulation remains the same as in the previous example, except for the different mesh that now incorporates the ellipsoid hole (Fig. 9). Figure 10 demonstrates that although the design $\boldsymbol{\xi} = (2.0, 1.5)$ is optimal regardless of whether $\boldsymbol{\theta}$ and ϕ or only $\boldsymbol{\theta}$ are considered parameters of interest, the overall response surface changes depending on the parameters that are considered to be of interest. In particular, ξ_2, the distance between electrodes, is almost irrelevant when recovering the fiber angles, whereas it has a much more prominent effect when recovering the height of the exclusion as well. The design choices $\boldsymbol{\xi} = (2.0, 1.0)$ and $\boldsymbol{\xi} = (2.0, 0.5)$ yield EIG that is within the tolerance $TOL = 0.2$ of the optimal choice when considering nuisance parameters. The MCLA estimator with optimal sampling was used to estimate the EIG without nuisance uncertainty. The pilot run for this estimator was performed using $N = 150$ samples. The DLMC2IS estimator was used with optimal sampling to estimate the EIG with nuisance uncertainty. For the pilot run, we used $N = M_1 = M_2 = 200$ samples. The bias introduced by the Laplace approximation (45) rendered the MC2LA estimator ineffective for this example for a tolerance of $TOL = 0.2$; thus the DLMC2IS estimator was used instead.

7 Conclusion

We propose two estimators for the expected information gain under nuisance uncertainty: the Monte Carlo double-Laplace estimator based on Laplace's method for the nuisance parameters and the Laplace approximation for the parameters of interest, and the double-loop Monte Carlo double importance sampling estimator based on two Laplace approximations. We demonstrate the applicability of these estimators in four numerical examples, showcasing their computational efficiency provided by Laplace-based integral approximations.

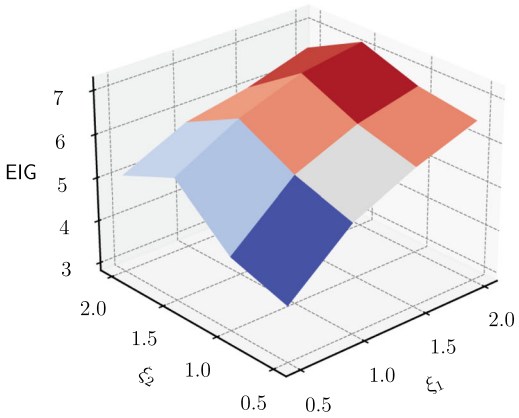


(A) Current streamlines for design $\boldsymbol{\xi} = (0.5, 2.0)$.

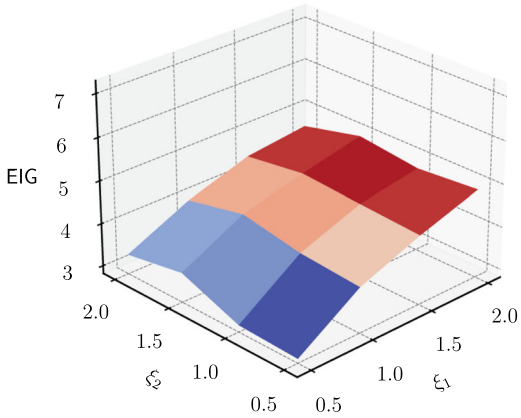


(B) Current streamlines for design $\boldsymbol{\xi} = (2.0, 0.5)$.

Fig. 9 Example 4: Inlet: red electrodes. Outlet: blue electrodes. The current flow is affected by the electrode positions



(A) Expected information gain regarding the fiber angles and the vertical axis of the ellipsoid exclusion.



(B) Expected information gain regarding the fiber angles only.

Fig. 10 Example 4: Expected information gain (EIG) as a function of the design $\boldsymbol{\xi}$. The response surface changes when considering $\boldsymbol{\theta}$ and ϕ as parameters of interest (Panel **a**), or $\boldsymbol{\theta}$ only (Panel **b**). The EIG regarding the fiber angles is virtually unaffected by the distance between the electrodes (ξ_2) in this case

Appendix A. Error estimates for the double-loop Monte Carlo estimator with two inner loops

Derivation of the bias error approximation

Following (Beck et al. 2018), we derive an estimate for the bias, given as follows:

$$|\mathbb{E}[I_{\mathrm{DL}}] - I|. \tag{81}$$

We recall that I_{DL} is given as follows:

$$I_{\mathrm{DL}} = \frac{1}{N}\sum_{n=1}^{N} \log\left(\frac{1}{M_1}\sum_{m=1}^{M_1} p(\boldsymbol{Y}^{(n)}|\boldsymbol{\theta}^{(n)}, \boldsymbol{\varphi}^{(n,m)})\right) - \log\left(\frac{1}{M_2}\sum_{k=1}^{M_2} p(\boldsymbol{Y}^{(n)}|\boldsymbol{\vartheta}^{(n,k)}, \boldsymbol{\varphi}^{(n,k)})\right),$$

$$:= \frac{1}{N}\sum_{n=1}^{N} \log(\hat{p}_{M_1}(\boldsymbol{Y}^{(n)}|\boldsymbol{\theta}^{(n)})) - \log(\hat{p}_{M_2}(\boldsymbol{Y}^{(n)})). \tag{82}$$

Replacing (82) in (81) yields

$$\begin{aligned}|\mathbb{E}[I_{\mathrm{DL}}] - I| &= \left|\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{\hat{p}_{M_1}(\boldsymbol{Y}^{(n)}|\boldsymbol{\theta}^{(n)})}{\hat{p}_{M_2}(\boldsymbol{Y}^{(n)})}\right)\right] - \mathbb{E}\left[\log\left(\frac{p(\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{Y})}\right)\right]\right|, \\ &= |\underbrace{\mathbb{E}[\log(\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta}))]}_{I} - \underbrace{\mathbb{E}[\log(p(\boldsymbol{Y}|\boldsymbol{\theta}))]}_{II} - \underbrace{\mathbb{E}[\log(\hat{p}_{M_2}(\boldsymbol{Y}))]}_{III} + \underbrace{\mathbb{E}[\log(p(\boldsymbol{Y}))]}_{IV}|. \end{aligned} \tag{83}$$

The estimation centers on the following Taylor expansion of $\log(X)$ around $\mathbb{E}[X]$:

$$\log(X) = \log(\mathbb{E}[X]) + \frac{1}{\mathbb{E}[X]}(X - \mathbb{E}[X])$$

$$-\frac{1}{2}\frac{1}{\mathbb{E}[X]^2}(X-\mathbb{E}[X])^2 + \int_0^1 \frac{1-s}{(\mathbb{E}[X]+s\,(X-\mathbb{E}[X]))^3}\,\mathrm{d}s(X-\mathbb{E}[X])^3. \quad (84)$$

As $\hat{p}_{M_2}(\boldsymbol{Y})$ is an unbiased estimator, we have

$$\mathbb{E}[\hat{p}_{M_2}(\boldsymbol{Y})|\boldsymbol{Y}] = p(\boldsymbol{Y}).$$

By the Taylor expansion and taking the expected value, we obtain

$$\mathbb{E}[\log(\hat{p}_{M_2}(\boldsymbol{Y}))|\boldsymbol{Y}] = \mathbb{E}[\log(p(\boldsymbol{Y}))|\boldsymbol{Y}] - \frac{1}{2}\mathbb{E}\left[\frac{1}{p^2(\boldsymbol{Y})}(\hat{p}_{M_2}(\boldsymbol{Y})-p(\boldsymbol{Y}))^2|\boldsymbol{Y}\right] + \mathbb{E}\left[\int_0^1 \frac{1-s}{\left(p(\boldsymbol{Y})+s\left(\hat{p}_{M_2}(\boldsymbol{Y})-p(\boldsymbol{Y})\right)\right)^3}\,\mathrm{d}s(\hat{p}_{M_2}(\boldsymbol{Y}) - p(\boldsymbol{Y}))^3\middle|\boldsymbol{Y}\right], \quad (85)$$

where the linear term vanishes because of the expected value. The quadratic term can be rewritten as follows:

$$\mathbb{E}\left[\frac{1}{p^2(\boldsymbol{Y})}(\hat{p}_{M_2}(\boldsymbol{Y})-p(\boldsymbol{Y}))^2|\boldsymbol{Y}\right] = \frac{\mathbb{V}[\hat{p}_{M_2}(\boldsymbol{Y})|\boldsymbol{Y}]}{p^2(\boldsymbol{Y})}, \quad (86)$$

$$= \frac{\mathbb{V}\left[\frac{1}{M_2}\sum_{k=1}^{M_2} p(\boldsymbol{Y}|\boldsymbol{\vartheta}^{(k)},\boldsymbol{\varphi}^{(k)})|\boldsymbol{Y}\right]}{p^2(\boldsymbol{Y})},$$

$$= \frac{1}{M_2}\frac{\mathbb{V}\left[p(\boldsymbol{Y}|\boldsymbol{\vartheta},\boldsymbol{\varphi})\middle|\boldsymbol{Y}\right]}{p^2(\boldsymbol{Y})}. \quad (87)$$

The third-order term has the following bound:

$$\mathbb{E}\left[\int_0^1 \frac{1-s}{\left(p(\boldsymbol{Y})+s\left(\hat{p}_{M_2}(\boldsymbol{Y})-p(\boldsymbol{Y})\right)\right)^3}\,\mathrm{d}s(\hat{p}_{M_2}(\boldsymbol{Y}) - p(\boldsymbol{Y}))^3\middle|\boldsymbol{Y}\right]$$

$$\leq \mathbb{E}\left[\left(\int_0^1 \frac{1-s}{\left(p(\boldsymbol{Y})+s\left(\hat{p}_{M_2}(\boldsymbol{Y})-p(\boldsymbol{Y})\right)\right)^3}\,\mathrm{d}s\right)^p\middle|\boldsymbol{Y}\right]^{\frac{1}{p}} \mathbb{E}\left[(\hat{p}_{M_2}(\boldsymbol{Y})-p(\boldsymbol{Y}))^{3q}\middle|\boldsymbol{Y}\right]^{\frac{1}{q}} \quad (88)$$

by Hölder's inequality for some $1 \leq p, q \leq \infty$, where $1/p + 1/q = 1$. For the last term in (88), we have by the discrete Burkholder–Davis–Gundy inequality (Burkholder et al. 1972; Giles and Goda 2019) that:

$$\mathbb{E}\left[(\hat{p}_{M_2}(\boldsymbol{Y})-p(\boldsymbol{Y}))^{3q}\middle|\boldsymbol{Y}\right]^{\frac{1}{q}}$$

$$= \mathbb{E}\left[\left(\frac{1}{M_2}\sum_{k=1}^{M_2} p(\boldsymbol{Y}|\boldsymbol{\vartheta}^{(k)},\boldsymbol{\varphi}^{(k)}) - p(\boldsymbol{Y})\right)^{3q}\middle|\boldsymbol{Y}\right]^{\frac{1}{q}},$$

$$\leq \frac{C_{\mathrm{BDG}}}{M_2^{\frac{3}{2}}}\mathbb{E}\left[(p(\boldsymbol{Y}|\boldsymbol{\vartheta},\boldsymbol{\varphi})-p(\boldsymbol{Y}))^{3q}\middle|\boldsymbol{Y}\right]^{\frac{1}{q}} \quad (89)$$

for some constant C_{BDG} for almost all $\boldsymbol{Y}$. From (85), we obtain that

$$\underbrace{\mathbb{E}[\log(\hat{p}_{M_2}(\boldsymbol{Y}))]}_{III} = \underbrace{\mathbb{E}[\log(p(\boldsymbol{Y}))]}_{IV} - \frac{1}{2}\frac{1}{M_2}\mathbb{E}\left[\frac{\mathbb{V}\left[p(\boldsymbol{Y}|\boldsymbol{\vartheta},\boldsymbol{\varphi})\middle|\boldsymbol{Y}\right]}{p^2(\boldsymbol{Y})}\right] + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{M_2^{\frac{3}{2}}}\right). \quad (90)$$

Similarly, we obtain

$$\mathbb{E}[\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta})|\boldsymbol{Y},\boldsymbol{\theta}] = p(\boldsymbol{Y}|\boldsymbol{\theta}), \quad (91)$$

and

$$\mathbb{E}[\log(\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta}))|\boldsymbol{Y},\boldsymbol{\theta}] = \mathbb{E}[\log(p(\boldsymbol{Y}|\boldsymbol{\theta}))|\boldsymbol{Y},\boldsymbol{\theta}] - \frac{1}{2}\mathbb{E}\left[\frac{1}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})}(\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta})-p(\boldsymbol{Y}|\boldsymbol{\theta}))^2|\boldsymbol{Y},\boldsymbol{\theta}\right] + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{M_1^{\frac{3}{2}}}\right), \quad (92)$$

where the quadratic term can be written as follows:

$$\mathbb{E}\left[\frac{1}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})}(\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta})-p(\boldsymbol{Y}|\boldsymbol{\theta}))^2|\boldsymbol{Y},\boldsymbol{\theta}\right] = \frac{\mathbb{V}[\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta})|\boldsymbol{Y},\boldsymbol{\theta}]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})},$$

$$= \frac{\mathbb{V}\left[\frac{1}{M_1}\sum_{m=1}^{M_1} p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\varphi}^{(m)})|\boldsymbol{Y},\boldsymbol{\theta}\right]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})},$$

$$= \frac{1}{M_1}\frac{\mathbb{V}\left[p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\varphi})\middle|\boldsymbol{Y},\boldsymbol{\theta}\right]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})}. \quad (93)$$

This calculation results in

$$\underbrace{\mathbb{E}[\log(\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta}))]}_{I} = \underbrace{\mathbb{E}[\log(p(\boldsymbol{Y}|\boldsymbol{\theta}))]}_{II} - \frac{1}{2}\frac{1}{M_1}\mathbb{E}\left[\frac{\mathbb{V}\left[p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\varphi})\middle|\boldsymbol{Y},\boldsymbol{\theta}\right]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})}\right]$$

$$+\mathcal{O}_{\mathbb{P}}\left(\frac{1}{M_1^{\frac{3}{2}}}\right). \tag{94}$$

After combining everything, we obtain

$$|\mathbb{E}[I_{\mathrm{DL}}] - I| \approx \left|\frac{1}{2M_2}\mathbb{E}\left[\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\vartheta},\boldsymbol{\varphi})|\boldsymbol{Y}]}{p^2(\boldsymbol{Y})}\right] - \frac{1}{2M_1}\mathbb{E}\left[\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\varphi})|\boldsymbol{Y},\boldsymbol{\theta}]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})}\right]\right|. \tag{95}$$

Derivation of the statistical error approximation

For the variance estimation, we obtain

$$\mathbb{V}[I_{\mathrm{DL}}] = \mathbb{V}\left[\frac{1}{N}\sum_{n=1}^{N}\log(\hat{p}_{M_1}(\boldsymbol{Y}^{(n)}|\boldsymbol{\theta}^{(n)})) - \log(\hat{p}_{M_2}(\boldsymbol{Y}^{(n)}))\right]. \tag{96}$$

By the law of total variance, we can rewrite this as

$$\mathbb{V}[I_{\mathrm{DL}}] = \frac{1}{N}\mathbb{V}\left[\mathbb{E}\left[\log(\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta})) - \log(\hat{p}_{M_2}(\boldsymbol{Y}))|\boldsymbol{Y},\boldsymbol{\theta}\right]\right] \tag{97}$$

$$+ \frac{1}{N}\mathbb{E}\left[\mathbb{V}\left[\log(\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta})) - \log(\hat{p}_{M_2}(\boldsymbol{Y}))|\boldsymbol{Y},\boldsymbol{\theta}\right]\right]. \tag{98}$$

Using (90) and (94), we can rewrite (97) as

$$\frac{1}{N}\mathbb{V}\left[\mathbb{E}\left[\log(\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta}))|\boldsymbol{Y},\boldsymbol{\theta}\right] - \mathbb{E}\left[\log(\hat{p}_{M_2}(\boldsymbol{Y}))|\boldsymbol{Y}\right]\right] = \frac{1}{N}\mathbb{V}\left[\log\left(\frac{p(\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{Y})}\right) + \frac{1}{2M_2}\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\vartheta},\boldsymbol{\varphi})|\boldsymbol{Y}]}{p^2(\boldsymbol{Y})} - \frac{1}{2M_1}\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\varphi})|\boldsymbol{Y},\boldsymbol{\theta}]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})}\right] + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{NM_1^2}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{NM_2^2}\right),$$

which results in

$$\frac{1}{N}\mathbb{V}\left[\log\left(\frac{p(\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{Y})}\right)\right] + \frac{1}{4NM_2^2}\mathbb{V}\left[\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\vartheta},\boldsymbol{\varphi})|\boldsymbol{Y}]}{p^2(\boldsymbol{Y})}\right] + \frac{1}{4NM_1^2}\mathbb{V}\left[\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\varphi})|\boldsymbol{Y},\boldsymbol{\theta}]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})}\right] + \frac{1}{NM_2}Cov\left[\log\left(\frac{p(\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{Y})}\right), \frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\vartheta},\boldsymbol{\varphi})|\boldsymbol{Y}]}{p^2(\boldsymbol{Y})}\right] - \frac{1}{NM_1}Cov\left[\log\left(\frac{p(\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{Y})}\right), \frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\varphi})|\boldsymbol{Y},\boldsymbol{\theta}]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})}\right] - \frac{1}{4NM_1M_2}Cov\left[\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\vartheta},\boldsymbol{\varphi})|\boldsymbol{Y}]}{p^2(\boldsymbol{Y})}, \frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\varphi})|\boldsymbol{Y},\boldsymbol{\theta}]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})}\right]$$

$$+ \mathcal{O}_{\mathbb{P}}\left(\frac{1}{NM_1^2}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{NM_2^2}\right). \tag{99}$$

This yields

$$\frac{1}{N}\mathbb{V}\left[\log\left(\frac{p(\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{Y})}\right)\right] + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{NM_1}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{NM_2}\right). \tag{100}$$

From (98), we obtain

$$\frac{1}{N}\mathbb{E}\left[\mathbb{V}\left[\log(\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta})) - \log(\hat{p}_{M_2}(\boldsymbol{Y}))|\boldsymbol{Y},\boldsymbol{\theta}\right]\right] = \frac{1}{N}\mathbb{E}[\mathbb{V}[\log(\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta}))|\boldsymbol{Y},\boldsymbol{\theta}]] + \frac{1}{N}\mathbb{E}[\mathbb{V}[\log(\hat{p}_{M_2}(\boldsymbol{Y}))|\boldsymbol{Y}]]$$

as the inner samples are independent. For the first term, by the first-order Taylor expansion (84), we obtain

$$\mathbb{V}[\log(\hat{p}_{M_1}(\boldsymbol{Y}|\boldsymbol{\theta}))|\boldsymbol{Y},\boldsymbol{\theta}] \tag{101}$$

$$= \mathbb{V}\left[\log(p(\boldsymbol{Y}|\boldsymbol{\theta})) + \frac{1}{p(\boldsymbol{Y}|\boldsymbol{\theta})}\left(\frac{1}{M_1}\sum_{m=1}^{M_1}p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\varphi}^{(m)}) - p(\boldsymbol{Y}|\boldsymbol{\theta})\right)\middle|\boldsymbol{Y},\boldsymbol{\theta}\right] + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{M_1^2}\right) \tag{102}$$

$$= \frac{1}{M_1}\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\varphi})|\boldsymbol{Y},\boldsymbol{\theta}]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})} + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{M_1^2}\right). \tag{103}$$

Similarly, for the second term, we obtain

$$\mathbb{V}[\log(\hat{p}_{M_2}(\boldsymbol{Y}))|\boldsymbol{Y}] = \frac{1}{M_2}\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\vartheta},\boldsymbol{\varphi})|\boldsymbol{Y}]}{p^2(\boldsymbol{Y})} + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{M_2^2}\right), \tag{104}$$

resulting in

$$(98) = \frac{1}{NM_1}\mathbb{E}\left[\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\varphi})|\boldsymbol{Y},\boldsymbol{\theta}]}{p^2(\boldsymbol{Y}|\boldsymbol{\theta})}\right] + \frac{1}{NM_2}\mathbb{E}\left[\frac{\mathbb{V}[p(\boldsymbol{Y}|\boldsymbol{\vartheta},\boldsymbol{\varphi})|\boldsymbol{Y}]}{p^2(\boldsymbol{Y})}\right] + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{NM_1^2}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{NM_2^2}\right),$$

completing the derivation.

Remark 5 (Covariance terms in the statistical error approximation) Applying the inequality

$$\mathbb{V}[A+B] \leq 2\mathbb{V}[A] + 2\mathbb{V}[B] \tag{105}$$

demonstrates that the covariance terms in (99) are of a similar magnitude as the remaining variance terms. Moreover, the covariance terms are challenging to estimate numerically and were thus neglected in the optimization of the number of samples (see Beck et al. 2018). Figure 3 (Panel A) demonstrates that this simplification did not significantly impact the consistency of the DLMC2IS estimator.

Appendix B. Optimal setting for the double-loop Monte Carlo estimator with two inner loops

We solve the minimization problem stated in (21) by introducing a Lagrange function, which we can take the derivative of with respect to N, M_1, M_2, the error splitting parameter κ, and the Lagrange multipliers λ and μ and set the resulting equations to 0[2]

$$\begin{aligned}\mathcal{L}(N, M_1, M_2, \kappa, \lambda, \mu) &= N(M_1 + M_2) \\ &-\lambda\left(\frac{C_1}{M_1} + \frac{C_2}{M_2} - (1-\kappa)TOL\right) \\ &-\mu\left(\frac{1}{N}\left(D_3 + \frac{D_1}{M_1} + \frac{D_2}{M_2}\right)\right. \\ &\left.-\left(\frac{\kappa TOL}{C_\alpha}\right)^2\right),\end{aligned} \tag{106}$$

with derivatives

$$\frac{\partial \mathcal{L}}{\partial N} = M_1 + M_2 + \frac{\mu}{N^2}\left(D_3 + \frac{D_1}{M_1} + \frac{D_2}{M_2}\right) \stackrel{!}{=} 0, \tag{107}$$

$$\frac{\partial \mathcal{L}}{\partial M_1} = N + \lambda\frac{C_1}{M_1^2} + \mu\frac{D_1}{NM_1^2} \stackrel{!}{=} 0, \tag{108}$$

$$\frac{\partial \mathcal{L}}{\partial M_2} = N + \lambda\frac{C_2}{M_2^2} + \mu\frac{D_2}{NM_2^2} \stackrel{!}{=} 0, \tag{109}$$

$$\frac{\partial \mathcal{L}}{\partial \kappa} = 2\mu\kappa\frac{TOL^2}{C_\alpha^2} - \lambda TOL \stackrel{!}{=} 0, \tag{110}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = (1-\kappa)TOL - \frac{C_1}{M_1} - \frac{C_2}{M_2} \stackrel{!}{=} 0, \tag{111}$$

and

$$\frac{\partial \mathcal{L}}{\partial \mu} = \left(\frac{\kappa TOL}{C_\alpha}\right)^2 - \frac{1}{N}\left(D_3 + \frac{D_1}{M_1} + \frac{D_2}{M_2}\right) \stackrel{!}{=} 0. \tag{112}$$

From (14)–(17), we have that $D_1 = 2C_1$ and that $D_2 = 2C_2$. Thus, it follows from (108) that

$$\frac{M_1^2}{C_1} = -\frac{1}{N}\left(\lambda + \frac{2\mu}{N}\right), \tag{113}$$

[2] Denoted by $\stackrel{!}{=} 0$.

and from (109) that

$$\frac{M_2^2}{C_2} = -\frac{1}{N}\left(\lambda + \frac{2\mu}{N}\right). \tag{114}$$

Thus, it follows that

$$M_2 = M_1\frac{\sqrt{C_2}}{\sqrt{C_1}}. \tag{115}$$

Next, from (111), we have that

$$\begin{aligned}(1-\kappa)TOL &= \frac{C_1}{M_1} + \frac{C_2}{M_2}, \\ &= \frac{C_1}{M_1} + \frac{\sqrt{C_1}\sqrt{C_2}}{M_1},\end{aligned} \tag{116}$$

and thus that

$$M_1^* = \frac{\sqrt{C_1}(\sqrt{C_1} + \sqrt{C_2})}{(1-\kappa)TOL}, \tag{117}$$

and also that

$$M_2^* = \frac{\sqrt{C_2}(\sqrt{C_1} + \sqrt{C_2})}{(1-\kappa)TOL}. \tag{118}$$

Moreover, from (110), we have that

$$\lambda = \frac{2\mu\kappa TOL}{C_\alpha^2}. \tag{119}$$

Substituting (117) and (118) into (107), we find that

$$\begin{aligned}&M_1^* + M_2^* + \frac{2\mu}{N^2}\left(\frac{D_3}{2} + \frac{C_1}{M_1} + \frac{C_2}{M_2}\right) \\ &= \frac{(\sqrt{C_1} + \sqrt{C_2})^2}{(1-\kappa)TOL} \\ &+ \frac{2\mu}{N^2}\left(\frac{D_3}{2} + (1-\kappa)TOL\right), \\ &= 0,\end{aligned} \tag{120}$$

yielding

$$\mu = -\frac{N^2}{2}\frac{(\sqrt{C_1} + \sqrt{C_2})^2}{(1-\kappa)TOL}\left(\frac{D_3}{2} + (1-\kappa)TOL\right)^{-1}. \tag{121}$$

substituting into (119) leads to

$$\lambda = -\frac{N^2\kappa}{C_\alpha^2}\frac{(\sqrt{C_1} + \sqrt{C_2})^2}{(1-\kappa)}\left(\frac{D_3}{2} + (1-\kappa)TOL\right)^{-1}. \tag{122}$$

From (112), we obtain that

$$N^* = 2\left(\frac{\kappa TOL}{C_\alpha}\right)^{-2}\left(\frac{D_3}{2} + (1-\kappa)TOL\right). \tag{123}$$

Finally, substituting (117), (119) and (121) into (113), it follows that

$$\begin{aligned}\frac{\left(\sqrt{C_1}+\sqrt{C_2}\right)^2}{(1-\kappa)^2 TOL^2} &= -\frac{1}{N^*}\left(\lambda + \frac{2\mu}{N^*}\right),\\ &= \frac{1}{N^*}\left(\frac{N^{*2}\kappa}{C_\alpha^2}\frac{(\sqrt{C_1}+\sqrt{C_2})^2}{(1-\kappa)}\right.\\ &\left.+\frac{N^*(\sqrt{C_1}+\sqrt{C_2})^2}{(1-\kappa)TOL}\right)\\ &\times\left(\frac{D_3}{2}+(1-\kappa)TOL\right)^{-1},\\ &= \left(\frac{N^*\kappa}{C_\alpha^2}\frac{(\sqrt{C_1}+\sqrt{C_2})^2}{(1-\kappa)}\right.\\ &\left.+\frac{(\sqrt{C_1}+\sqrt{C_2})^2}{(1-\kappa)TOL}\right)\\ &\times\left(\frac{D_3}{2}+(1-\kappa)TOL\right)^{-1}.\end{aligned} \tag{124}$$

Cancellations lead to the following quadratic equation for κ^*:

$$\begin{aligned}1 &= \left(\frac{N^*\kappa(1-\kappa)TOL^2}{C_\alpha^2}+(1-\kappa)TOL\right)\\ &\left(\frac{D_3}{2}+(1-\kappa)TOL\right)^{-1},\\ &= \left(\frac{2(1-\kappa)}{\kappa}\left(\frac{D_3}{2}+(1-\kappa)TOL\right)\right.\\ &\left.+(1-\kappa)TOL\right)\left(\frac{D_3}{2}+(1-\kappa)TOL\right)^{-1}.\end{aligned} \tag{125}$$

Rearranging terms yields

$$\begin{aligned}\frac{D_3}{2}+(1-\kappa)TOL &= \frac{2(1-\kappa)}{\kappa}\left(\frac{D_3}{2}\right.\\ &\left.+(1-\kappa)TOL\right)+(1-\kappa)TOL,\end{aligned} \tag{126}$$

and ultimately

$$D_3\kappa = 2D_3(1-\kappa)+4(1-\kappa)^2 TOL, \tag{127}$$

or, in the standard form

$$\left(\frac{4TOL}{D_3}\right)\kappa^2 - \left(3+\frac{8TOL}{D_3}\right)\kappa + 2 + \frac{4TOL}{D_3} = 0 \tag{128}$$

with the solutions

$$\kappa_{1,2} = 1+\frac{3D_3}{8TOL} \pm \sqrt{\frac{D_3}{4TOL}+\frac{9D_3^2}{64TOL^2}}. \tag{129}$$

From the requirement that $0 < \kappa < 1$, it follows that only the solution

$$\kappa^* = 1+\frac{3D_3}{8TOL} - \sqrt{\frac{D_3}{4TOL}+\frac{9D_3^2}{64TOL^2}} \tag{130}$$

is permissible. It follows immediately that

$$\frac{3D_3}{8TOL} < \sqrt{\frac{D_3}{4TOL}+\frac{9D_3^2}{64TOL^2}} \tag{131}$$

for any $D_3, TOL > 0$, and thus that $\kappa^* < 1$. Moreover, it follows that

$$\begin{aligned}\frac{D_3}{4TOL}+\frac{9D_3^2}{64TOL^2} &< \left(1+\frac{3D_3}{8TOL}\right)^2,\\ &= 1+\frac{3D_3}{4TOL}+\frac{9D_3^2}{64TOL^2}\end{aligned} \tag{132}$$

for any $D_3, TOL > 0$, and thus that $0 < \kappa^*$. Applying L'Hôpital's rule, we observe that

$$\begin{aligned}\lim_{TOL\to 0}\kappa^* &= \lim_{TOL\to 0} 1+\frac{3D_3}{8TOL} - \sqrt{\frac{D_3}{4TOL}+\frac{9D_3^2}{64TOL^2}}\\ &= \lim_{TOL\to 0}\frac{\left(8TOL+3D_3-\sqrt{16TOLD_3+9D_3^2}\right)}{8TOL},\\ &= \lim_{TOL\to 0}\frac{\frac{\mathrm{d}}{\mathrm{d}TOL}\left(8TOL+3D_3-\sqrt{16TOLD_3+9D_3^2}\right)}{\frac{\mathrm{d}}{\mathrm{d}TOL}8TOL},\\ &= \lim_{TOL\to 0} 1-\frac{D_3}{\sqrt{16TOLD_3+9D_3^2}} = \frac{2}{3},\end{aligned} \tag{133}$$

and that

$$\lim_{TOL\to\infty}\kappa^* = 1. \tag{134}$$

Appendix C. Optimal setting with additional discretization bias

We solve the minimization problem in (26):

$$\begin{aligned}\mathcal{L}(N, M_1, M_2, h, \kappa, \lambda, \mu) &= N(M_1+M_2)h^{-\gamma}\\ &-\lambda\left(\frac{D_3}{N}+\frac{D_1}{NM_1}+\frac{D_2}{NM_2}-\left(\frac{\kappa TOL}{C_\alpha}\right)^2\right)\\ &-\mu\left(C_3h^\eta+\frac{C_1}{M_1}+\frac{C_2}{M_2}-(1-\kappa)TOL\right),\end{aligned} \tag{135}$$

with the following derivatives, which we set to 0:

$$\frac{\partial \mathcal{L}}{\partial N} = (M_1 + M_2)h^{-\gamma} + \frac{\lambda D_3}{N^2} + \frac{\lambda D_1}{N^2 M_1} + \frac{\lambda D_2}{N^2 M_2} \stackrel{!}{=} 0, \tag{136}$$

$$\frac{\partial \mathcal{L}}{\partial M_1} = Nh^{-\gamma} + \frac{\lambda D_1}{N M_1^2} + \frac{\mu C_1}{M_1^2} \stackrel{!}{=} 0, \tag{137}$$

$$\frac{\partial \mathcal{L}}{\partial M_2} = Nh^{-\gamma} + \frac{\lambda D_2}{N M_2^2} + \frac{\mu C_2}{M_2^2} \stackrel{!}{=} 0, \tag{138}$$

$$\frac{\partial \mathcal{L}}{\partial h} = -\gamma N h^{-\gamma-1}(M_1 + M_2) - \mu C_3 \eta h^{\eta-1} \stackrel{!}{=} 0, \tag{139}$$

$$\frac{\partial \mathcal{L}}{\partial \kappa} = 2\lambda\kappa \left(\frac{TOL}{C_\alpha}\right)^2 - \mu TOL \stackrel{!}{=} 0, \tag{140}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \left(\frac{\kappa TOL}{C_\alpha}\right)^2 - \frac{D_3}{N} - \frac{D_1}{N M_1} - \frac{D_2}{N M_2} \stackrel{!}{=} 0, \tag{141}$$

and

$$\frac{\partial \mathcal{L}}{\partial \mu} = (1-\kappa)TOL - C_3 h^\eta - \frac{C_1}{M_1} - \frac{C_2}{M_2} \stackrel{!}{=} 0. \tag{142}$$

The idea is to express every parameter of the Lagrangian (135) as a function of the splitting parameter κ and solve the remaining quadratic equation for κ. Subtracting (137) from (138) results in

$$M_2 = \sqrt{\frac{C_2}{C_1}} M_1, \tag{143}$$

which we use in (141) to obtain

$$\begin{aligned} M_1 &= \frac{\left(D_1 + \sqrt{\frac{C_1}{C_2}} D_2\right)}{N\left(\frac{\kappa TOL}{C_\alpha}\right)^2 - D_3}, \\ &= \frac{2C_1\left(1 + \sqrt{\frac{C_2}{C_1}}\right)}{N\left(\frac{\kappa TOL}{C_\alpha}\right)^2 - D_3}. \end{aligned} \tag{144}$$

The last equation follows from the definitions of C_1, C_2, D_1, and D_2. From (142) we obtain

$$\begin{aligned} h &= \left(\frac{(1-\kappa)TOL - \frac{C_1}{M_1}\left(1 + \sqrt{\frac{C_2}{C_1}}\right)}{C_3}\right)^{\frac{1}{\eta}}, \\ &= \left(\frac{(1-\kappa)TOL - \frac{1}{2}\left(N\left(\frac{\kappa TOL}{C_\alpha}\right)^2 - D_3\right)}{C_3}\right)^{\frac{1}{\eta}}. \end{aligned} \tag{145}$$

From (139), we obtain

$$\mu = -\frac{\gamma N M_1 \left(1 + \sqrt{\frac{C_2}{C_1}}\right)}{\eta C_3 h^{\gamma+\eta}}. \tag{146}$$

Inserting this into (140) results in

$$\lambda = -\frac{\gamma N M_1 \left(1 + \sqrt{\frac{C_2}{C_1}}\right) C_\alpha^2}{\eta C_3 h^{\gamma+\eta} 2\kappa TOL}. \tag{147}$$

Inserting (143) to (147) into (136), after some reordering, yields

$$N^* = \frac{C_\alpha^2}{\kappa^2 TOL}\left(\frac{D_3}{TOL} + 2\left(1 - \kappa\left(1 + \frac{\gamma}{2\eta}\right)\right)\right). \tag{148}$$

This result is the same optimal number of outer samples N^* as in the case for only one inner loop. We can insert this into (144) to obtain

$$M_1^* = \frac{C_1\left(1 + \sqrt{\frac{C_2}{C_1}}\right)}{\left(1 - \kappa\left(1 + \frac{\gamma}{2\eta}\right)\right)TOL} \tag{149}$$

and

$$M_2^* = \frac{C_2\left(1 + \sqrt{\frac{C_1}{C_2}}\right)}{\left(1 - \kappa\left(1 + \frac{\gamma}{2\eta}\right)\right)TOL}. \tag{150}$$

Equation (145) results in

$$h^* = \left(\frac{\gamma\kappa TOL}{2\eta C_3}\right)^{\frac{1}{\eta}}. \tag{151}$$

This quantity also remains unchanged compared to the case with only one inner loop.

Finally, (137) provides the following quadratic equation:

$$\begin{aligned} &\left[\frac{1}{D_3}\left(1 + \frac{\gamma}{2\eta}\right)^2 TOL\right]\kappa^{*2} - \left[\frac{1}{4} + \left(\frac{1}{2} + \frac{2}{D_3}TOL\right)\right. \\ &\left.\left(1 + \frac{\gamma}{2\eta}\right)\right]\kappa^* + \left[\frac{1}{2} + \frac{1}{D_3}TOL\right], \end{aligned} \tag{152}$$

coinciding with the equation in the case with only one inner loop.

Appendix D. Derivation of the order of the additional terms when accounting for nuisance uncertainty

We demonstrate the order of $\nabla_{\boldsymbol{\theta}} z(\hat{\boldsymbol{\theta}})$, $z(\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}(\hat{\boldsymbol{\theta}}))$. We let

$$\begin{aligned} S(\boldsymbol{\theta}, \boldsymbol{\phi}) = &\frac{1}{2} N_e (\boldsymbol{g}(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi})) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} (\boldsymbol{g}(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \\ &- \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi})) \\ &+ \sum_{i=1}^{N_e} \boldsymbol{\epsilon}_i \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} (\boldsymbol{g}(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi})) \\ &- \log(\pi(\boldsymbol{\phi}|\boldsymbol{\theta})), \end{aligned} \tag{153}$$

with the gradient

$$\begin{aligned} \nabla_{\boldsymbol{\phi}} S(\boldsymbol{\theta}, \boldsymbol{\phi}) = &-N_e \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi})^\top \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} (\boldsymbol{g}(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi})) \\ &- \sum_{i=1}^{N_e} \boldsymbol{\epsilon}_i \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}) - \nabla_{\boldsymbol{\phi}} \log(\pi(\boldsymbol{\phi}|\boldsymbol{\theta})). \end{aligned} \tag{154}$$

The first-order approximation of $\nabla_{\boldsymbol{\phi}} S(\boldsymbol{\theta}, \boldsymbol{\phi})$ around $\boldsymbol{\phi}_t$, given by

$$\nabla_{\boldsymbol{\phi}} S(\boldsymbol{\theta}, \boldsymbol{\phi}) \approx \nabla_{\boldsymbol{\phi}} S(\boldsymbol{\theta}, \boldsymbol{\phi}_t) + \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} S(\boldsymbol{\theta}, \boldsymbol{\phi}_t)(\boldsymbol{\phi} - \boldsymbol{\phi}_t), \tag{155}$$

yields

$$\begin{aligned} \nabla_{\boldsymbol{\phi}} S(\boldsymbol{\theta}, \boldsymbol{\phi}) \approx &\Big(-N_e \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t)^\top \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} (\boldsymbol{g}(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t)) \\ &- \sum_{i=1}^{N_e} \boldsymbol{\epsilon}_i \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t) - \nabla_{\boldsymbol{\phi}} \log(\pi(\boldsymbol{\phi}_t|\boldsymbol{\theta})) \Big) \\ &+ \Big(- N_e \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t)^\top \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} (\boldsymbol{g}(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t)) \\ &+ N_e \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t)^\top \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t) \\ &- \sum_{i=1}^{N_e} \boldsymbol{\epsilon}_i \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t) \\ &- \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \log(\pi(\boldsymbol{\phi}_t|\boldsymbol{\theta})) \Big) (\boldsymbol{\phi} - \boldsymbol{\phi}_t). \end{aligned} \tag{156}$$

Applying Newton's method and $\nabla_{\boldsymbol{\phi}} S(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}(\boldsymbol{\theta})) = \mathbf{0}$ implies that $\nabla_{\boldsymbol{\phi}} S(\boldsymbol{\theta}, \boldsymbol{\phi}_t) + \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} S(\boldsymbol{\theta}, \boldsymbol{\phi}_t)(\hat{\boldsymbol{\phi}}(\boldsymbol{\theta}) - \boldsymbol{\phi}_t) \approx \mathbf{0}$; therefore

$$\hat{\boldsymbol{\phi}}(\boldsymbol{\theta}) \approx \boldsymbol{\phi}_t - \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} S(\boldsymbol{\theta}, \boldsymbol{\phi}_t)^{-1} \nabla_{\boldsymbol{\phi}} S(\boldsymbol{\theta}, \boldsymbol{\phi}_t), \tag{157}$$

resulting in

$$\begin{aligned} \hat{\boldsymbol{\phi}}(\boldsymbol{\theta}) \approx \boldsymbol{\phi}_t &- \Big(- N_e \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t)^\top \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} (\boldsymbol{g}(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t)) \\ &+ N_e \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t)^\top \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t) \\ &- \sum_{i=1}^{N_e} \boldsymbol{\epsilon}_i \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t) - \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \log(\pi(\boldsymbol{\phi}_t|\boldsymbol{\theta})) \Big)^{-1} \\ &\times \Big(-N_e \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t)^\top \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} (\boldsymbol{g}(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t)) \\ &- \sum_{i=1}^{N_e} \boldsymbol{\epsilon}_i \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \nabla_{\boldsymbol{\phi}} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_t) - \nabla_{\boldsymbol{\phi}} \log(\pi(\boldsymbol{\phi}_t|\boldsymbol{\theta})) \Big). \end{aligned} \tag{158}$$

Both terms in the product have leading order N_e; thus, assuming that Newton's method converges, the Jacobian

$$\nabla_{\boldsymbol{\theta}} z(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} I_{d_\theta \times d_\theta} \\ \nabla_{\boldsymbol{\theta}} \hat{\boldsymbol{\phi}}(\hat{\boldsymbol{\theta}}) \end{pmatrix} \tag{159}$$

has leading order terms that are constant in N_e by the quotient rule. The term $k(\boldsymbol{\theta})$ is of order $\mathcal{O}_{\mathbb{P}}(\log(N_e))$ and the term $\ell(\boldsymbol{\theta})$ is constant in N_e. Evaluating $\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_t$ rather than at $\hat{\boldsymbol{\theta}}$ reduces the leading order of the numerator in (158) to $\sqrt{N_e}$ (see Long et al. 2013). This would suggest approximating $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ by $(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)$, however, numerical experiments show that the influence from the nuisance uncertainty would not be captured accurately by such an approximation.

Acknowledgements This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. OSR-2019-CRG8-4033, the Alexander von Humboldt Foundation, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 338 49990/GRK2379 (IRTG Hierarchical and Hybrid Approaches in Modern Inverse Problems), and was partially supported by the Flexible Interdisciplinary Research Collaboration Fund at the University of Nottingham Project ID 7466664.

Author contributions All authors conceived the theory and analysis. A.B. and L.E. wrote the main manuscript text and prepared the python code. All authors reviewed the manuscript..

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest Authors Luis Espath and Raúl Tempone are associate editors of the Statistics and Computing Journal.

References

Alexanderian, A., Nicholson, R., Petra, N.: Optimal design of large-scale nonlinear Bayesian inverse problems under model uncertainty. arXiv preprint arXiv:2211.03952, (2022)

Bartuska, A., Espath, L.F., Tempone, R.: Small-noise approximation for Bayesian optimal experimental design with nuisance uncertainty. Comput. Methods Appl. Mech. Eng. **399**, 115320 (2022)

Beck, J., Dia, B.M., Espath, L.F., Long, Q., Tempone, R.: Fast Bayesian experimental design: Laplace-based importance sampling for the expected information gain. Comput. Methods Appl. Mech. Eng. **334**, 523–553 (2018)

Beck, J., Dia, B.M., Espath, L.F., Tempone, R.: Multilevel double loop Monte Carlo and stochastic collocation methods with importance sampling for Bayesian optimal experimental design. Int. J. Numer. Meth. Eng. **121**, 3482–3503 (2020)

Bernardo, J.M.: Reference posterior distributions for Bayesian inference. Journal of the Royal Statistical Society, Series B 41, (1979)

Bisetti, F., Kim, D., Knio, O., Long, Q., Tempone, R.: Optimal Bayesian experimental design for priors of compact support with application to shock-tube experiments for combustion kinetics. Int. J. Numer. Meth. Eng. **108**, 136–155 (2016)

Bornkamp, B.: Approximating probability densities by iterated Laplace approximations. J. Comput. Graph. Stat. **20**(3), 656–669 (2011)

Burkholder, D.L., Davis, B.J., Gundy, R.F.: Integral inequalities for convex functions of operators on martingales. Proceed. Sixth Berkeley Symp. Math. Stat. Probab. **2**, 223–240 (1972)

Chaloner, K., Verdinelli, I.: Bayesian experimental design: a review. Stat. Sci. **10**, 273–304 (1995)

Englezou, Y., Waite, T.W., Woods, D.C.: Approximate Laplace importance sampling for the estimation of expected Shannon information gain in high-dimensional Bayesian design for nonlinear models. Stat. Comput. **32**, 82 (2022)

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

Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W (Eds.).: Statistical parametric mapping: The analysis of functional brain images. Academic Press, (2007)

Giles, M.B., Goda, T.: Decision-making under uncertainty: using MLMC for efficient estimation of EVPPI. Stat. Comput. **29**(4), 739–751 (2019)

Goda, T., Hironaka, T., Iwamoto, T.: Multilevel Monte Carlo estimation of expected information gains. Stoch. Anal. Appl. **38**(4), 581–600 (2020)

Helin, T., Kretschmann, R.: Non-asymptotic error estimates for the Laplace approximation in Bayesian inverse problems. Numer. Math. **150**, 521–549 (2022)

Kass, R.E., Tierney, L., Kadane, J.B.: The validity of posterior expansions based on Laplace's method. in: Geisser S, Hodges JS, Press SJ, and Zellner A (Eds.), Essays in Honor of George Barnard, 473–488, (1990)

Kullback, S.: Information theory and statistics. Wiley, (1959)

Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)

Levine, D.S.: Focused active inference. Ph.D. thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, (2014)

Liepe, J., Filippi, S., Komorowski, M., Stumpf, M.P.H.: Maximizing the information content of experiments in systems biology. PLoS Comput. Biol. **9**, 1–13 (2013)

Lindley, D.V.: On a measure of information provided by an experiment. Ann. Math. Stat. **27**, 986–1005 (1956)

Long, Q.: Multimodal information gain in Bayesian design of experiments. Comput. Stat. **37**(2), 865–885 (2022)

Long, Q., Scavino, M., Tempone, R., Wang, S.: Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. Comput. Methods Appl. Mech. Eng. **259**, 24–39 (2013)

Overstall, A.M., McGree, J.M., Drovandi, C.C.: An approach for finding fully Bayesian optimal designs using normal-based approximations to loss functions. Stat. Comput. **28**, 343–358 (2017)

Polson, N.G.: Bayesian Perspectives on Statistical Modelling. Ph.D. thesis, Department of Mathematics, University of Nottingham, (1988)

Ryan, E.G., Drovandi, C.C., Thompson, M.H., Pettitt, A.N.: Towards Bayesian experimental design for nonlinear models that require a large number of sampling times. Comput. Stat. Data Anal. **70**, 45–60 (2014)

Ryan, K.J.: Estimating expected information gains for experimental designs with application to the random fatigue-limit model. J. Comput. Graph. Stat. **12**, 585–603 (2003)

Schillings, C., Sprungk, B., Wacker, P.: On the convergence of the Laplace approximation and noise-level-robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems. Numer. Math. **145**, 915–971 (2020)

Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)

Somersalo, E., Cheney, M., Isaacson, D.: Existence and uniqueness for electrode models for electric current computed tomography. SIAM J. Appl. Math. **52**, 1023–1040 (1992)

Spokoiny, V.: Dimension free non-asymptotic bounds on the accuracy of high dimensional Laplace approximation. arXiv preprint arXiv:2204.11038, (2022)

Stigler, S.M.: Laplace's 1774 memoir on inverse probability. Stat. Sci. **1**, 359–378 (1986)

Tierney, L., Kadane, J.B.: Accurate approximations for posterior moments and marginal densities. J. Am. Stat. Assoc. **81**, 82–86 (1986)

Tierney, L., Kass, R.E., Kadane, J.B.: Fully exponential Laplace approximations to expectations and variances of nonpositive functions. J. Am. Stat. Assoc. **84**, 710–716 (1989)

Wacker, P.: Laplace's method in Bayesian inverse problems. arXiv preprint arXiv:1701.07989, (2017)