See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/320321933>

Fast Bayesian experimental design: Laplace-based importance sampling for the expected information gain

Article in *Computer Methods in Applied Mechanics and Engineering* · October 2017

DOI: 10.1016/j.cma.2018.01.053

CITATIONS

0

READS

49

5 authors, including:

[Joakim Beck](#)

King Abdullah University of Science and Tech...

16 PUBLICATIONS 218 CITATIONS

[SEE PROFILE](#)

[Ben Mansour Dia](#)

King Abdullah University of Science and Tech...

10 PUBLICATIONS 20 CITATIONS

[SEE PROFILE](#)

[Luis Felipe da Rosa Espath](#)

King Abdullah University of Science and Tech...

14 PUBLICATIONS 56 CITATIONS

[SEE PROFILE](#)

[Quan Long](#)

United Technologies Research Center

26 PUBLICATIONS 122 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:

collaboration_work [View project](#)

Sparse approximation [View project](#)

Fast Bayesian experimental design: Laplace-based importance sampling for the expected information gain

Joakim Beck^a, Ben Mansour Dia^b, Luis FR Espath^a, Quan Long^c, Raul Tempone^a

^a*Computer, Electrical and Mathematical Science and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia*

^b*CIPR, College of Petroleum Engineering and Geosciences, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia*

^c*United Technologies Research Center, East Hartford, CT, 06108, United States*

Abstract

In calculating expected information gain in optimal Bayesian experimental design, the computation of the inner loop in the classical double-loop Monte Carlo requires a large number of samples and suffers from underflow if the number of samples is small. These drawbacks can be avoided by using an importance sampling approach. We present a computationally efficient method for optimal Bayesian experimental design that introduces importance sampling based on the Laplace method to the inner loop. We derive the optimal values for the method parameters in which the average computational cost is minimized according to the desired error tolerance. We use three numerical examples to demonstrate the computational efficiency of our method compared with the classical double-loop Monte Carlo, and a more recent single-loop Monte Carlo method that uses the Laplace method as an approximation of the return value of the inner loop. The first example is a scalar problem that is linear in the uncertain parameter. The second example is a nonlinear scalar problem. The third example deals with the optimal sensor placement for an electrical impedance tomography experiment to recover the fiber orientation in laminate composites.

Keywords: Bayesian experimental design, Expected information gain, Monte Carlo, Laplace approximation, Importance sampling, Composite materials.

AMS 2010 subject classification: 62K05, 65N21, 65C60, 65C05

1. Introduction

This work proposes an efficient method for the computation of expected information gain [1–3] in optimal Bayesian experimental design. The expected information gain, also known as the expected Kullback-Leibler divergence, is an information metric commonly used to estimate the information provided by the proposed experiment. Its straightforward estimator,

Email addresses: joakim.beck@kaust.edu.sa (Joakim Beck), mansourben2002@yahoo.fr (Ben Mansour Dia), espath@gmail.com (Luis FR Espath), longq@utrc.utc.com (Quan Long), raul.tempone@kaust.edu.sa (Raul Tempone)

the double-loop Monte Carlo (DLMC), typically requires a large number of samples [4, 5] because it embeds a nested sample-average structure. The computation of the inner loop can suffer from arithmetic underflow, especially for small sample sizes.

Another approach is to employ the Laplace method (MCLA) [6] to approximate the inner integral of the expected information gain with a second-order Taylor expansion around the mode and then compute the inner integral analytically [7–9], which leads to a single integral. Other studies have introduced the Laplace method to the design of experiments for partial differential equation (PDE) models and achieved a substantial improvement in the efficiency compared to DLMC [10, 11]. The posterior distribution was expanded at the maximum a posteriori (MAP) estimate; therefore, the associated integrals of the inner loop could be approximated analytically by Gaussian integration. However, this leads to an extra bias, except when the Laplace approximation is exact. Furthermore, in [12], the Laplace method was used to approximate the entropy of the posterior distribution for optimal sensor placement, and in [13], the method was extended to handle under-determined experiments. A truncated Gaussian has also been used to approximate the posterior distribution, see [14]. Related work in optimal experimental design are the Bayesian A-optimality for PDE models [15], and the use of Gaussian process models for the approximation of the expected information gain in the context of sequential design of computer experiments [16].

A self-normalized importance sampling approach for the computation of the inner loop of DLMC has been proposed by Feng [17], where a multivariate normal distribution is used as the inner sampling distribution with weighted sample mean and weighted sample covariance based on the outer samples. This leads to a substantial reduction in the number of inner samples. However, the approximation of the normalization constant can suffer from underflow when the posterior distributions are concentrated, in particular when the measurement error is small or the number of repetitive experiments is large.

In this work, we instead propose a Laplace-based importance sampling where the inner sampling distribution is the Laplace approximation on the MAP estimates. These MAP estimates are obtained by solving an optimization problem. The method does not introduce any extra bias, in contrast to the Laplace method, and shows a reduction of orders of magnitude in the number of inner samples. This approach also mitigates the risk of underflow.

We also devise a strategy to estimate optimal values for the method parameters for the desired error tolerance. The method parameters are the outer and inner number of samples, as well as the discretization parameter in the context of PDEs.

The outline of the paper is as follows: Section 2 deals with optimal Bayesian experimental design where the data model is composed of a deterministic computer model and an additive observational noise component. A brief introduction of the expected information gain criteria is also presented. In Section 3, we provide an error analysis of the DLMC estimator, and derive optimal values for the method parameters for a given error tolerance. In Section 4, we perform the minimization of the average computational work of the MCLA estimator. In Section 5, we propose our double-loop Monte Carlo importance sampling (DLMCIS) method, which is proven to substantially reduce the average computational work, mitigate the risk of underflow, and in contrast to the Laplace method it does not introduce any additional bias. In Section 6, three numerical examples are used to compare the methods in terms of

robustness and computational efficiency. The first example is a linear model, the second example is a nonlinear model, and the third example is a sensor placement design problem, where the goal is to maximize signal information during electrical impedance tomography (EIT) in order to inversely obtain the parameters of the inter-ply delaminations.

2. Optimal Bayesian experimental design

2.1. Problem setting

We consider the data model given by

$$\mathbf{y}_i(\boldsymbol{\xi}) = \mathbf{g}(\boldsymbol{\theta}_t, \boldsymbol{\xi}) + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, N_e, \quad (1)$$

where $\mathbf{y}_i \in \mathbb{R}^q$ is a vector of q observed responses, $\mathbf{g}(\boldsymbol{\theta}_t, \boldsymbol{\xi}) \in \mathbb{R}^q$ is the deterministic model responses, $\boldsymbol{\theta}_t \in \mathbb{R}^d$ is the true parameter vector, $\boldsymbol{\xi} \in \Xi$ is the design parameter vector, Ξ is the experimental design space, $\boldsymbol{\epsilon}_i$ are independent and identically distributed (i.i.d.) zero-mean Gaussian measurement errors with covariance matrix $\boldsymbol{\Sigma}_\epsilon$, and N_e is the number of repetitive experiments. The observed dataset is denoted by $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N_e}$. In our problem setting, we assume the true value of $\boldsymbol{\theta}_t$ is unknown. We let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space where $\mathcal{F}$ is the σ -field of events, $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ is a probability measure, and Ω is the set of outcomes. We consider a vector of random variables, $\boldsymbol{\theta} : \Theta \mapsto \mathbb{R}^d$, with prior space Θ and prior distribution $\pi(\boldsymbol{\theta})$ in lieu of the unknown vector $\boldsymbol{\theta}_t$, i.e.,

$$\mathbf{y}_i(\omega_1, \omega_2, \boldsymbol{\xi}) = \mathbf{g}(\boldsymbol{\theta}(\omega_1), \boldsymbol{\xi}) + \boldsymbol{\epsilon}_i(\omega_2), \quad i = 1, \dots, N_e. \quad (2)$$

We note that the data model accounts for parametric but not structural uncertainty; for example, we do not account for the model error in $\mathbf{g}$ that cannot be eliminated by evaluating $\mathbf{g}$ at $\boldsymbol{\theta}_t$.

We denote the resulting approximation of the forward model $\mathbf{g}$ using mesh discretization parameter h by $\mathbf{g}_h$. As $h \rightarrow 0$ asymptotically, the convergence order of $\mathbf{g}_h$ is given by

$$\mathbb{E} [\|\mathbf{g}(\boldsymbol{\theta}) - \mathbf{g}_h(\boldsymbol{\theta})\|_2] = \mathcal{O}(h^\eta),$$

where $\eta > 0$ is the h -convergence rate. The work of $\mathbf{g}_h$ is assumed to be $\mathcal{O}(h^{-\gamma})$, for some $\gamma > 0$. We also assume that $\mathbf{g}$ is twice differentiable with respect to $\boldsymbol{\theta}$. The matrix norm given by $\|\mathbf{x}\|_{\boldsymbol{\Sigma}^{-1}}^2 = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}$ for a vector $\mathbf{x}$ and covariance matrix $\boldsymbol{\Sigma}$ is used throughout.

The objective of optimal Bayesian experimental design is to determine the most informative experimental design setup about $\boldsymbol{\theta}_t$, denoted by $\boldsymbol{\xi}^* \in \Xi$. The utility function employed is the expectation of the Kullback-Leibler divergence, see, e.g., [4, 5, 10]. For the sake of conciseness, $\boldsymbol{\xi}$ is omitted until Section 6, as the expected information gain for each $\boldsymbol{\xi}$ is computed separately.

2.2. Expected information gain

The Kullback-Leibler divergence, D_{kl} , also known as the information gain [18, 19], is an entropic function that can quantify our uncertainty about $\boldsymbol{\theta}_t$ through the distance between the prior $\pi(\boldsymbol{\theta})$ and the posterior $\pi(\boldsymbol{\theta}|\mathbf{Y})$ as

$$D_{kl}(\mathbf{Y}) = \int_{\Theta} \pi(\boldsymbol{\theta}|\mathbf{Y}) \log \left(\frac{\pi(\boldsymbol{\theta}|\mathbf{Y})}{\pi(\boldsymbol{\theta})} \right) d\boldsymbol{\theta}, \quad (3)$$

where $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_{N_e})$ is the data, and $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\mathbf{Y})$ are the prior and posterior probability density functions (pdfs), respectively. The larger the value of D_{kl} , the more informative the dataset is about the unknown $\boldsymbol{\theta}_t$.

The value of $D_{kl}(\mathbf{Y})$ can be derived exactly when the prior and posterior pdfs are both Gaussian. For the one-dimensional case ($d=1$), the Gaussian integral formula yields

$$D_{kl}(\mathbf{Y}) = \log \left(\frac{\sigma_{prior}}{\sigma_{post}(\mathbf{Y})} \right) + \frac{1}{2} \left[\frac{\sigma_{post}^2(\mathbf{Y})}{\sigma_{prior}^2} - 1 + \frac{(\mu_{post}(\mathbf{Y}) - \mu_{prior})^2}{\sigma_{prior}^2} \right], \quad (4)$$

where the first and second moments of the prior distribution $\pi(\boldsymbol{\theta})$ are denoted by μ_{prior} and σ_{prior} , and the moments of the posterior pdf $\pi(\boldsymbol{\theta}|\mathbf{Y})$ by $\mu_{post}(\mathbf{Y})$ and $\sigma_{post}(\mathbf{Y})$.

In the first stage of experimental design, there are no available observations. Thus, we take the expectation of D_{kl} , denoted by I , over the sample space, $\mathcal{Y} \subseteq \mathbb{R}^q$, i.e.,

$$\begin{aligned} I &= \int_{\mathcal{Y}} \int_{\Theta} \log \left(\frac{\pi(\boldsymbol{\theta}|\mathbf{Y})}{\pi(\boldsymbol{\theta})} \right) \pi(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} p(\mathbf{Y}) d\mathbf{Y} \\ &= \int_{\Theta} \int_{\mathcal{Y}} \log \left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \right) p(\mathbf{Y}|\boldsymbol{\theta}) d\mathbf{Y} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \end{aligned} \quad (5)$$

The latter equality in (5) follows from Bayes' rule. For clarity, we adopt the notation $\pi(\cdot)$ for the pdf of the parameters and $p(\cdot)$ for the pdf of the data sample. In the rest of the paper, we use the approximation of the likelihood with respect to $\mathbf{g}_h$, given by

$$p(\mathbf{Y}|\boldsymbol{\theta}) = (2\pi|\boldsymbol{\Sigma}_{\epsilon}|)^{-\frac{N_e}{2}} \exp \left(-\frac{1}{2} \sum_{i=1}^{N_e} \|\mathbf{g}_h(\boldsymbol{\theta}_t) + \boldsymbol{\epsilon}_i - \mathbf{g}_h(\boldsymbol{\theta})\|_{\boldsymbol{\Sigma}_{\epsilon}^{-1}}^2 \right). \quad (6)$$

2.3. Fast numerical estimators for expected information gain

Here we propose a strategy for designing a computationally-efficient numerical estimator $\mathcal{I}$ of $I = \mathbb{E}[D_{kl}]$ that satisfies the tolerance $\text{TOL} > 0$ at a confidence level given by $0 < \alpha \ll 1$:

$$\mathbb{P}(|I - \mathcal{I}| \leq \text{TOL}) \geq 1 - \alpha. \quad (7)$$

So the absolute difference $|\mathcal{I} - I|$ should be less or equal to TOL , with a probability of $1 - \alpha$.

Similar to [20, 21], we optimize the numerical estimator by minimizing its average computational cost based on the h -convergence rate, η , and the work rate, γ , of the underlying

forward problem. Before defining the cost minimization problem, we introduce some notation. The total error is split into a bias component and a statistical error,

$$|I - \mathcal{I}| \leq |I - \mathbb{E}[\mathcal{I}]| + |\mathbb{E}[\mathcal{I}] - \mathcal{I}|, \quad (8)$$

and we introduce a balance parameter $\kappa \in]0, 1[$ such that

$$|I - \mathbb{E}[\mathcal{I}]| \leq (1 - \kappa)\text{TOL}, \text{ and} \quad (9)$$

$$|\mathbb{E}[\mathcal{I}] - \mathcal{I}| \leq \kappa\text{TOL}, \quad (10)$$

where $(1 - \kappa)\text{TOL}$ is the bias tolerance and κTOL is the statistical error tolerance. As in a previous study [21], we recast the statistical error constraint by using the central theorem limit (CTL) as follows:

$$\mathbb{V}[\mathcal{I}] \leq \left(\frac{\kappa\text{TOL}}{C_\alpha} \right)^2, \quad (11)$$

where $C_\alpha = \Phi^{-1}(1 - \frac{\alpha}{2})$ and $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution. Furthermore, W denotes the average computational work of a single evaluation of $\mathcal{I}$, and the method parameters of the estimator are denoted by $\boldsymbol{\zeta}$. For instance, the classical Monte Carlo (MC) estimator has $\boldsymbol{\zeta} = \{N, h\}$, where N is the number of samples and h is the discretization parameter for $\mathbf{g}_h$.

Within this framework, the optimal setting for the estimator $\mathcal{I}_{\boldsymbol{\zeta}}$ is the solution of the cost minimization problem,

$$(\boldsymbol{\zeta}^*, \kappa^*) = \arg \min_{(\boldsymbol{\zeta}, \kappa)} W(\boldsymbol{\zeta}) \quad \text{subject to} \quad \begin{cases} \mathbb{V}[\mathcal{I}_{\boldsymbol{\zeta}}] \leq (\kappa\text{TOL}/C_\alpha)^2 \\ |I - \mathbb{E}[\mathcal{I}_{\boldsymbol{\zeta}}]| \leq (1 - \kappa)\text{TOL}, \end{cases} \quad (12)$$

for the specified tolerance $\text{TOL} > 0$ at a confidence level given by $1 - \alpha$. We note that the balance parameter κ is chosen in conjunction with the method parameters. In practice, we need to provide estimates for W , $\mathbb{V}[\mathcal{I}]$, and $|I - \mathbb{E}[\mathcal{I}]|$, which depend on $\boldsymbol{\zeta}$. Two optimal settings can be derived, with and without a mesh discretization, which lead to different optimal $\boldsymbol{\zeta}^*$. In this work, we consider both settings, but only the derivation with the mesh discretization is presented.

Below we review the DLMC, MCLA, and DLMCIS estimators, and derive their optimal parameter settings.

3. Double-loop Monte Carlo

3.1. Double-loop Monte Carlo (DLMC) estimator

The DLMC estimator, $\mathcal{I}_{dl}$, of the expected information gain (5) is given by

$$\mathcal{I}_{dl} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^N \log \left(\frac{p(\mathbf{Y}_n | \boldsymbol{\theta}_n)}{\hat{p}_M(\mathbf{Y}_n)} \right), \quad (13)$$

where $\hat{p}_M(\mathbf{Y}_n)$ denotes the inner averaging,

$$\hat{p}_M(\mathbf{Y}_n) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^M p(\mathbf{Y}_n | \tilde{\boldsymbol{\theta}}_{n,m}), \quad (14)$$

and N and M are the respective numbers of samples. The tilde ($\sim$) on $\boldsymbol{\theta}$ differentiates the inner loop samples from the outer loop samples, as they are independent of each other. The average computational work of $\mathcal{I}_{dl}$ is assumed to follow

$$W_{dl} \propto NMh^{-\gamma}, \quad (15)$$

where $h^{-\gamma}$ is proportional to the average work of a single evaluation of $\mathbf{g}_h$. The next section shows the optimal setting (i.e., N , M , and h) with respect to the choice of TOL.

3.2. Optimal setting for the DLMC estimator

The optimal setting for the DLMC estimator is derived by exploiting Proposition 1:

Proposition 1. *The bias and variance of DLMC estimator $\mathcal{I}_{dl}$ can be estimated by:*

$$|I - \mathbb{E}[\mathcal{I}_{dl}]| \leq C_{dl,3}h^\eta + \frac{C_{dl,4}}{M} + o(h^\eta) + \mathcal{O}\left(\frac{1}{M^2}\right), \quad (16)$$

$$\mathbb{V}[\mathcal{I}_{dl}] = \frac{C_{dl,1}}{N} + \frac{C_{dl,2}}{NM} + \mathcal{O}\left(\frac{1}{NM^2}\right), \quad (17)$$

respectively, where

$$\begin{aligned} C_{dl,1} &= \mathbb{V}\left[\log\left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})}\right)\right], \\ C_{dl,2} &= \left(1 + \mathbb{E}\left[\log\left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})}\right)\right]\right) \mathbb{E}\left[\mathbb{V}\left[\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \middle| \mathbf{Y}\right]\right] - \mathbb{E}\left[\log\left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})}\right) \mathbb{V}\left[\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \middle| \mathbf{Y}\right]\right], \\ C_{dl,4} &= \frac{1}{2} \mathbb{E}\left[\mathbb{V}\left[\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \middle| \mathbf{Y}\right]\right], \end{aligned}$$

and $C_{dl,3}$ is the constant of the h -convergence of $\mathcal{I}_{dl}$.

The proof of Proposition 1 is given in Appendix A. We remark that a second-order Taylor expansion is used in the proof, instead of a first order as was used elsewhere [22], which results in a better estimate of $C_{dl,2}$. The terms $C_{dl,3}h^\eta$ and $\frac{C_{dl,4}}{M}$ respectively account for the biases due to the numerical discretization of the forward problem and due to the error of the inner averaging that arise through the nonlinearity of the logarithmic function. We insert (16), (17), and (15) into (12), and the optimal parameter setting of the estimator is obtained by

$$(N^*, M^*, h^*, \kappa^*) = \arg \min_{(N, M, h, \kappa)} W_{dl} \quad \text{subject to} \quad \begin{cases} \frac{C_{dl,1}}{N} + \frac{C_{dl,2}}{NM} \leq (\kappa \text{TOL}/C_\alpha)^2 \\ C_{dl,3}h^\eta + \frac{C_{dl,4}}{M} \leq (1 - \kappa)\text{TOL}, \end{cases}$$

where superscript $*$ is used to denote the optimal solution for the method parameters. The solution to the above problem using the Pontryagin's principle of minimization is given by

$$\frac{1}{2C_{dl,1}} \left(1 + \frac{\gamma}{2\eta}\right)^2 \kappa^{*2} \text{TOL} - \left[\frac{1}{2} + \left(1 - \frac{1}{C_{dl,1}} \text{TOL}\right) \left(1 + \frac{\gamma}{2\eta}\right)\right] \kappa^* + \left[1 + \frac{1}{2C_{dl,1}} \text{TOL}\right] = 0,$$

for $\kappa^* \in]0, 1[$, and

$$\begin{aligned} N^* &= \frac{C_\alpha^2}{2\kappa^*} \frac{C_{dl,1}}{1 - \kappa^* \left(1 + \frac{\gamma}{2\eta}\right)} \text{TOL}^{-2}, \\ M^* &= \frac{C_{dl,2}}{2 \left[1 - \kappa^* \left(1 + \frac{\gamma}{2\eta}\right)\right]} \text{TOL}^{-1}, \\ h^* &= \left(\frac{\gamma}{\eta} \frac{\kappa^*}{2C_{dl,3}}\right)^{1/\eta} \text{TOL}^{1/\eta}. \end{aligned}$$

In practice, we take the ceilings of N and M to obtain the optimal parameter values, N^* and M^* . The minimal average work is then given by

$$W_{dl}^* \propto \text{TOL}^{-(3+\frac{\gamma}{\eta})}. \quad (18)$$

We note that the optimal setting without discretization, which is not presented herein but is applied in the numerical section (Section 6), is different from the one presented above with $\gamma/\eta = 0$; however, the same rates are attained with respect to TOL. The constants C_{dl} are independent of h , but for the numerical estimation we approximate these constants using $\mathbf{g}_h$.

It has been proposed that the same samples can be used for the inner and outer loops as a means to reduce the number of evaluations of $\mathbf{g}_h$ [5]; more specifically, the computational cost is reduced from $\mathcal{O}(NM)$ to $\mathcal{O}(N)$ whenever the cost is dominated by the forward problem solver. Generally, such an approach leads to an extra bias [13].

3.3. Arithmetic underflow

Arithmetic underflow can cause the return value of $\hat{p}_M(\mathbf{Y}_n)$, which is the denominator of its logarithm term in the DLMC estimator (13), to be numerically zero. This occurs if all of the likelihoods, $p(\mathbf{Y}_n | \tilde{\boldsymbol{\theta}}_{n,m})$, $m = 1, 2, \dots, M$, return values below the machine precision. For example, this is likely to happen if the prior distribution π is not sufficiently concentrated on the posterior of $\boldsymbol{\theta}_n$. We also note that the posterior distribution itself becomes more concentrated for lower measurement noise, and the likelihood decreases exponentially with the size of $\mathbf{Y}$ ($= qN_e$). This can be seen from looking at the following leading-order terms of the log-likelihood:

$$\begin{aligned}
\log \left(p(\mathbf{Y}_n | \tilde{\boldsymbol{\theta}}_{n,m}) \right) &= -\frac{N_e}{2} \log(2\pi |\boldsymbol{\Sigma}_\epsilon|) - \frac{N_e}{2} \left\| \mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2 \\
&\quad - \frac{1}{2} \mathbf{v}_\epsilon^T \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right) \mathcal{O}_{\mathbb{P}} \left(\sqrt{N_e} \right) - \frac{1}{2} \sum_{j=1}^q \sigma_{\epsilon_j}^{-1} \left(N_e + \mathcal{O}_{\mathbb{P}} \left(\sqrt{N_e} \right) \right),
\end{aligned} \tag{19}$$

where $\mathbf{v}_\epsilon$ is the vector of the diagonal elements of $\boldsymbol{\Sigma}_\epsilon^{-1/2}$. The derivation of the above expression is given in Appendix B. We remark that the term $-\frac{N_e}{2} \log(2\pi |\boldsymbol{\Sigma}_\epsilon|)$ is canceled out by the ratio in the DLMC estimator. The notation $X_M = \mathcal{O}_{\mathbb{P}}(a_M)$ for a sequence of random variables X_M with corresponding constants a_M , indexed by M , is defined as follows: for any $\epsilon > 0$, there exists a finite $K > 0$ and a finite $M_0 > 0$, such that for all $M \geq M_0$, $\mathbb{P}(|X_M| > K|a_M|) < \epsilon$.

The top of Figure 1 presents a normal prior $\pi(\boldsymbol{\theta})$ (depicted in red) and a particular likelihood $p(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\xi})$ (depicted in blue), both centered at $\tilde{\boldsymbol{\theta}}$. The shaded region over $\pi(\boldsymbol{\theta})$ represents the non-overlapping region between the two distributions. The lower illustration in Figure 1 presents the distance, multiplied by N_e , between the model $\mathbf{g}$ evaluated at $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$.

Figure 1: Illustration of the underflow that can occur when sampling the inner integral using the conventional DLMC.

4. Monte Carlo with the Laplace method

In this section, we consider MCLA for the estimation of the expected information gain [10]. MCLA approximates the inner part of the nested integral analytically as a normal

approximation of the posterior pdf. Thus, the estimation of the expected information gain (5) is reduced to a single integral over the prior space Θ , typically at the price of some additional bias. The size of the bias decreases with the increasing number of measurements, i.e., the number of independent and repetitive experiments, N_e . MCLA has been extended previously to the approximation of posterior pdfs characterized by a non-informative manifold [13].

In this section, we first show the main steps of the construction of MCLA, which can also be found elsewhere [10, 11], and then present our optimized MCLA estimator.

4.1. Laplace approximation

The posterior pdf of $\boldsymbol{\theta}$ is given by Bayes' rule,

$$\pi(\boldsymbol{\theta}|\mathbf{Y}) = \frac{\prod_{i=1}^{N_e} \exp\left(-\frac{1}{2}\mathbf{r}_i^T(\boldsymbol{\theta})\boldsymbol{\Sigma}_\epsilon^{-1}\mathbf{r}_i(\boldsymbol{\theta})\right)\pi(\boldsymbol{\theta})}{p(\mathbf{Y})}, \quad (20)$$

where $\mathbf{r}_i(\boldsymbol{\theta}) = \mathbf{g}(\boldsymbol{\theta}_t) + \boldsymbol{\epsilon}_i - \mathbf{g}(\boldsymbol{\theta})$ is the residual of the i -th experimental data.

The Gaussian approximation of the posterior pdf (20) can be written as

$$\pi_g(\boldsymbol{\theta}|\mathbf{Y}) = (2\pi)^{-\frac{d}{2}}|\hat{\boldsymbol{\Sigma}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_{\hat{\boldsymbol{\Sigma}}^{-1}}^2\right), \quad (21)$$

where $\hat{\boldsymbol{\theta}}$ is the MAP estimate, i.e.,

$$\hat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \arg \min_{\boldsymbol{\theta} \in \Theta} \left[\sum_{i=1}^{N_e} \|\mathbf{y}_i - \mathbf{g}(\boldsymbol{\theta})\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2 + h(\boldsymbol{\theta}) \right], \text{ and} \quad (22)$$

$$\hat{\boldsymbol{\Sigma}}^{-1} = N_e \mathbf{J}(\hat{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{J}(\hat{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}) + \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right) \quad (23)$$

is the inverse Hessian matrix of the negative logarithm of the posterior pdf evaluated at $\hat{\boldsymbol{\theta}}$, and $h(\boldsymbol{\theta}) = \log(\pi(\boldsymbol{\theta}))$. It has been shown [10] that

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_t - (N_e \mathbf{J}^T(\boldsymbol{\theta}_t) \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{J}(\boldsymbol{\theta}_t) + \mathbf{H}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{E}_\epsilon - \nabla \nabla h(\boldsymbol{\theta}_t))^{-1} \mathbf{J}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{E}_\epsilon + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{N_e}\right), \quad (24)$$

where

$$\mathbf{E}_\epsilon(\boldsymbol{\theta}) = \sum_{i=1}^{N_e} \mathbf{r}_i^T(\boldsymbol{\theta}), \quad \mathbf{J}(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \mathbf{g}(\boldsymbol{\theta}) \quad \text{and} \quad \mathbf{H}(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \mathbf{g}(\boldsymbol{\theta}).$$

For a sufficiently large N_e , we have

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_t + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{N_e}}\right), \quad (25)$$

which is the approximation used in the MCLA estimator, i.e., $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}_t$.

4.2. Monte Carlo with the Laplace approximation (MCLA) estimator

The Gaussian approximation (21) with $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$, given by (24) and (23), respectively, leads to an analytical expression of the Kullback-Leibler divergence, which subsequently yields the following approximation of the expected information gain:

$$I = \int_{\Theta} \left[-\frac{1}{2} \log((2\pi)^d |\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})|) - \frac{d}{2} - h(\hat{\boldsymbol{\theta}}) \right] p(\boldsymbol{\theta}) d\boldsymbol{\theta} + \mathcal{O}\left(\frac{1}{N_e}\right). \quad (26)$$

The detailed proof of (26) is given in Appendix D. Then we can use MC sampling on the integral in (26). The additional bias (Laplace error) is of the order $\mathcal{O}\left(\frac{1}{N_e}\right)$. Many other schemes could also be applied, e.g., Gaussian quadrature [10, 11]. The covariance matrix is proportional to N_e^{-1} , which can be seen by expanding (23) using the Sherman-Morrison formula; hence, $|\hat{\boldsymbol{\Sigma}}| = \mathcal{O}(N_e^{-d})$ and $I = \mathcal{O}(\log(N_e))$, where $d = \dim(\boldsymbol{\theta})$.

The MCLA estimator of I is defined as

$$\mathcal{I}_{la} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^N \left(-\frac{1}{2} \log((2\pi)^d |\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_n)|) - \frac{d}{2} - h(\boldsymbol{\theta}_n) \right), \quad (27)$$

where N is the number of MC samples. The average computational work of MCLA can be estimated by

$$W_{la} \propto N N_{jac} h^{-\gamma},$$

where N_{jac} is the number of forward model evaluations required in the Jacobian matrix $\mathbf{J}(\boldsymbol{\theta})$ in (24) and (23). For the finite difference,

$$N_{jac} = \begin{cases} d+1 & \text{if forward or backward differentiation,} \\ 2d & \text{if central differentiation.} \end{cases}$$

To derive the optimal setting for the MCLA estimator $\mathcal{I}_{la}$, we use Proposition 2.

Proposition 2. *The bias and variance of the MCLA estimator $\mathcal{I}_{la}$ can be estimated by:*

$$|I - \mathbb{E}[\mathcal{I}_{la}]| \leq \frac{C_{la,2}}{N_e} + C_{la,3} h^\eta + o(h^\eta), \quad (28)$$

$$\mathbb{V}[\mathcal{I}_{la}] = \frac{C_{la,1}}{N}, \quad (29)$$

where $C_{la,1}$, $C_{la,2}$ and $C_{la,3}$ are constants.

The total bias is composed of the bias $C_{la,2}/N_e$ introduced by the Laplace approximation in (26), and the bias $C_{la,3}h^\eta$ from the numerical discretization. The constant $C_{la,1}$ is given by $C_{la,1} = \mathbb{V}[D_{kl}]$, since

$$\mathbb{V}[\mathcal{I}_{la}] = \mathbb{V}\left[\frac{1}{N} \sum_{n=1}^N D_{kl}(\boldsymbol{\theta}_n)\right] = \frac{\mathbb{V}[D_{kl}]}{N}. \quad (30)$$

4.3. Optimal setting for the MCLA estimator

The cost minimization problem for the MCLA estimator of a given $\text{TOL} > 0$ is

$$(N^*, h^*, \kappa^*) = \arg \min_{(N, h, \kappa)} W_{la} \quad \text{subject to} \quad \begin{cases} \frac{C_{la,1}}{N} \leq (\kappa \text{TOL}/C_\alpha)^2, \\ \frac{C_{la,2}}{N_e} + C_{la,3} h^\eta \leq (1 - \kappa) \text{TOL}. \end{cases} \quad (31)$$

We note that the Laplace bias, $C_{la,2}/N_e$, is related to the number of repetitive experiments, N_e , and is therefore not a method parameter. Also, there exists no solution to the above problem if the constraint,

$$(1 - \kappa) \text{TOL} \geq \frac{C_{la,2}}{N_e}, \quad (32)$$

does not hold. Thus, for $\text{TOL} \geq \frac{C_{la,2}}{N_e}$, the solution to the Pontryagin's principle of minimization is given by

$$\begin{aligned} \kappa^* &= \frac{1 - \frac{C_{la,2}}{N_e} \text{TOL}^{-1}}{1 + \frac{\gamma}{2\eta}}, \\ h^* &= \left(\frac{\gamma}{\eta} \frac{\kappa^*}{2C_{la,3}} \right)^{1/\eta} \text{TOL}^{1/\eta}, \\ N^* &= \frac{2\eta}{\gamma} C_{la,1} \frac{1 - \kappa^*}{\kappa^{*3}} C_\alpha \text{TOL}^{-2} - \frac{2\eta}{\gamma} \frac{C_{la,1} C_{la,2}}{N_e \kappa^{*3}} C_\alpha^2 \text{TOL}^{-3}, \end{aligned}$$

with an average work of order $\mathcal{O}(\text{TOL}^{-(2+\frac{\gamma}{\eta})})$. This means that the work rate of MCLA is better than that of DLMC, although an accuracy lower than $C_{la,2}/N_e$ cannot be achieved.

5. Double-loop Monte Carlo with Laplace-based importance sampling

In this section, we devise a new approach that entails a measure change based on the Laplace approximation for the inner loop of DLMC. The approach does not introduce any extra bias, avoids the occurrence of underflow, and substantially reduces the number of samples of the inner loop. In other words, we use the Laplace approximation to achieve a more efficient sampling on the posterior of θ_n .

5.1. Double-loop Monte Carlo with importance sampling (DLMCIS) estimator

To compute the evidence term in (13), we change the sampling pdf, π , to a new pdf, $\tilde{\pi}_n$, given by $\tilde{\pi}_n(\theta) \sim \mathcal{N}(\hat{\theta}_n, \hat{\Sigma}(\hat{\theta}_n))$, where

$$\hat{\theta}_n \stackrel{\text{def}}{=} \arg \min_{\theta \in \Theta} \left[\frac{1}{2} \sum_{i=1}^{N_e} \|\mathbf{y}_{n,i} - \mathbf{g}_h(\theta)\|_{\Sigma_\epsilon^{-1}}^2 + \log(\pi(\theta)) \right], \quad (33)$$

and

$$\hat{\Sigma}(\hat{\theta}_n) = \left(N_e \mathbf{J}(\hat{\theta}_n)^T \Sigma_\epsilon^{-1} \mathbf{J}(\hat{\theta}_n) - \nabla_{\theta} \nabla_{\theta} h(\hat{\theta}_n) \right)^{-1} + \mathcal{O}_{\mathbb{P}} \left(\frac{1}{\sqrt{N_e}} \right). \quad (34)$$

The importance sampling based on the Laplace approximation leads to the DLMCIS estimator,

$$\mathcal{I}_{dis} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^N \log \left(\frac{p(\mathbf{Y}_n | \theta_n)}{\frac{1}{M} \sum_{m=1}^M L(\mathbf{Y}_n, \tilde{\theta}_{n,m})} \right), \quad (35)$$

where

$$L(\mathbf{Y}_n, \tilde{\theta}_{n,m}) = \frac{p(\mathbf{Y}_n | \tilde{\theta}_{n,m}) \pi(\tilde{\theta}_{n,m})}{\tilde{\pi}_n(\tilde{\theta}_{n,m})}. \quad (36)$$

Here θ_n and $\tilde{\theta}_{n,m}$ are sampled from the prior π and the Laplace-based pdf $\tilde{\pi}_n$ respectively.

It would be computationally less costly to construct the Laplace approximation around θ_n rather than $\hat{\theta}_n$. Unfortunately, we have verified numerically that it would be inefficient as the discrepancy between θ_n and $\hat{\theta}_n$ can be too large, and thus not enough concentration on the posterior; see Figure 2.

5.2. Optimal setting for the DLMCIS estimator

The average computational work of the DLMCIS estimator (35) is assumed to be

$$W_{dis} = C_1 N M h^{-\gamma} + C_2 N N_{jac} h^{-\gamma}, \quad (37)$$

with constants $C_1, C_2 > 0$. The cost minimization problem can be written as

$$(N^*, M^*, h^*, \kappa^*) = \arg \min_{(N, M, h, \kappa)} W_{dis} \quad \text{subject to} \quad \begin{cases} \frac{C_{dis,1}}{N} + \frac{C_{dis,2}}{NM} \leq (\kappa \text{TOL} / C_\alpha)^2, \\ C_{dis,3} h^\eta + \frac{C_{dis,4}}{M} \leq (1 - \kappa) \text{TOL}, \end{cases}$$

with constants $C_{dis,k}$, $k = 1, \dots, 4$. The optimal κ , denoted by κ^* , is the solution of a second order polynomial equation that belongs to $]0, 1[$, and the optimal N , M , and h are

$$\begin{aligned} N^* &= \frac{C_{dis,1} C_\alpha^2}{\kappa^{*2}} \text{TOL}^{-2} + \left(1 - \kappa^* \left(1 + \frac{\gamma}{2\eta} \right) \right) \frac{C_{dis,1} C_\alpha^2}{\kappa^{*2}} \text{TOL}^{-1}, \\ M^* &= \frac{C_{dis,4}}{1 - \kappa^* \left(1 + \frac{\gamma}{2\eta} \right)} \text{TOL}^{-1}, \\ h^* &= \left(\frac{\gamma}{\eta} \frac{\kappa^*}{2C_{dis,3}} \right)^{1/\eta} \text{TOL}^{1/\eta}, \text{ respectively.} \end{aligned}$$

The average computational work is $W_{dis}^* \propto \text{TOL}^{-(3+\frac{\gamma}{\eta})}$, i.e., the method preserves the order as DLMC, but the constants $C_{dis,1}$, $C_{dis,2}$, and $C_{dis,3}$, are typically several magnitudes smaller than their corresponding counterparts C_{dl} , due to the variance reduction achieved by the importance sampling; $C_{dis,3}$ is still equal to $C_{dl,3}$.

Figure 2: Posterior pdfs by DLMCIS: top with θ_n and bottom with $\hat{\theta}_n$.

5.3. Effect of the change of measure on the arithmetic underflow

The change of measure mitigates the risk of underflow in DLMC as discussed in Section 3.3. The advantage of changing measure in this regard can be seen by observing that the constants in front of the leading-order terms of the following expansion of the likelihood

become much smaller:

$$\begin{aligned}
\log \left(L(\mathbf{Y}_n; \tilde{\boldsymbol{\theta}}_{n,m}) \right) &= \frac{N_e}{2} \left(-\log(2\pi|\boldsymbol{\Sigma}_\epsilon|) + \log \left(2\pi|\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}_n)| \right) \right) - \frac{N_e}{2} \left\| \mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2 \\
&\quad + \frac{N_e}{2} \left\| \mathbf{J}(\hat{\boldsymbol{\theta}}_n) \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right) \right\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2 - \frac{1}{2} \left\| \tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right\|_{\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}_n)}^2 \\
&\quad - \frac{1}{2} \mathbf{v}_\epsilon^T \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right) \mathcal{O}_{\mathbb{P}} \left(\sqrt{N_e} \right) - \frac{1}{2} \sum_{j=1}^q \sigma_{\epsilon_j}^{-1} \left(N_e + \mathcal{O}_{\mathbb{P}} \left(\sqrt{N_e} \right) \right) \Bigg).
\end{aligned} \tag{38}$$

The detailed proof can be found in Appendix E.

6. Numerical examples

We present three examples to demonstrate the numerical performance of our proposed method, DLMCIS, in comparison with DLMC and MCLA. In Example 1, we consider a scalar model whose parameter follows a normal prior distribution. In Example 2, we apply the methods to a nonlinear benchmark test, which was studied in [5]. In the third example, we design the placement of electrode sensors in order to maximize signal information about unknown parameters, in this case the orientation of fibers, in a laminate composite material during EIT experiments.

6.1. Example 1: Linear scalar model

In this example, we consider an algebraic linear model given by

$$y_i(\theta, \xi) = \theta(1 + \xi)^2 + \epsilon_i, \quad \text{for } i = 1, 2, \dots, N_e, \tag{39}$$

with prior $\pi(\theta) \sim \mathcal{N}(1, 0.01)$ and observational noise $\epsilon_i \sim \mathcal{N}(0, (2 + (\xi - 10)/10)^2)$. Two repetitive experiments ($N_e = 2$) are conducted.

6.1.1. Optimality and error convergence

We study the problem in two parts. First, we analyze the optimality, that is, N^* , M^* , and κ^* for the tolerance range $[10^{-5}, 1]$. Second, we analyze the error, computational time, and work in relation to TOL; due to computational constraints, we shrink the tolerance range to $[10^{-2}, 1]$.

We compute the expected information gain for a fixed experimental setup, $\xi = 10$, at a given confidence level of 97.5%, i.e., $\alpha = 0.05$. For DLMCIS, the constants $C_{dis,1}$ and $C_{dis,2}$ are estimated using DLMC with $N = M = 100$; for MCLA, the constant $C_{la,1}$ is estimated by MC with $N = 100$. Because the model is linear on θ and the prior is normal, the Laplace method is exact in this case and does not induce any bias. We estimate $C_{la,2}$ by the absolute difference between MCLA and DLMCIS in D_{kl} for a small, fixed number of outer samples with identical realizations.

Figure 3 shows N^* and M^* for the tolerance range. The optimal value M^* for DLMCIS follows the asymptotic rate of the order $\mathcal{O}(\text{TOL}^{-1})$, except for tolerances larger than 10^{-2} ,

where it is small (less than 5). The number of samples needed with DLMCIS can be several magnitudes lower than the number required with DLMC, which shows the efficiency of the importance sampling with the Laplace approximation. The optimal value κ^* is constant over the considered tolerance window: roughly 0.64, 0.67, and 1, for DLMC, DLMCIS, and MCLA, respectively. This means that the statistical error contributes more than the bias to the total error in this example, especially for MCLA.

Figure 3: Optimal setting (outer N^* and inner M^* number of samples) vs. tolerance for linear scalar model with normal prior (Example 1).

We now look at the accuracies achieved by MCLA and DLMCIS using their optimal settings with respect to the specified tolerance TOL. In Figure 4, we provide a consistency test between the actual computed absolute error, $|\mathcal{I} - I|$, and TOL. The absolute error and TOL are in agreement for both MCLA and DLMCIS, which numerically validates our bias estimation in the optimal setting derivation. The absolute error of MCLA is consistently a bit higher than the tolerance, but this is expected since the error is below TOL with 97.5% probability. Here DLMCIS is here a bit more conservative than expected.

The average computational work required for MCLA is consistent with the theoretical asymptotic result $\mathcal{O}(\text{TOL}^{-2})$, whereas the work for DLMCIS, interestingly enough, has the same rate of $\mathcal{O}(\text{TOL}^{-2})$, which is better than the theoretical result, $\mathcal{O}(\text{TOL}^{-3})$, for this method; see Figures 5a and 5b for the running time and the average computational work. The DLMCIS work rate is a pre-asymptotic rate since the optimal value M^* is unaffected over the tolerance range $[0.01, 1]$ (see Figure 3). The computation of the inner loop using DLMCIS is very efficient in this example. We observe that MCLA has a lower error constant

Figure 4: Error vs. tolerance for linear model with normal prior (Example 1).

than DLMCIS; this is due to the fact that the Laplace approximation is exact owing to as the Gaussian response of the linear model with normal prior.

Figure 5: Average computational work and running time vs. tolerance for linear scalar model with normal prior (Example 1).

6.1.2. Expected information gain

In this section, we show the expected information gain estimated by the methods at $TOL = 0.01$ over a set of experiment setups, $\xi \in [10, 30]$. As seen in Figure 6, DLMCIS and MCLA show agreement, though the variance of the MCLA estimator is larger than that of DLMCIS.

Figure 6: Expected information gain for linear scalar model with normal prior (Example 1).

6.2. Example 2: Nonlinear scalar model

We consider the algebraic nonlinear model found in Huan and Marzouk [5]:

$$y_i = \theta^3 \xi^2 + \theta \exp(-|0.2 - \xi|) + \epsilon_i, \quad \text{for } i = 1, 2, \dots, N_e, \quad (40)$$

where $N_e = 1$ and $N_e = 10$. The prior is $\pi(\theta) \sim \mathcal{U}(0, 1)$, and the observational noise is $\epsilon_i \sim \mathcal{N}(0, 10^{-3})$.

6.2.1. Optimality and error convergence

Same as in Example 1, we provide a comparison between MCLA and DLMCIS; here for the experiment $\xi = 1$. Example 2 is more challenging for MCLA than Example 1, since the use of the Laplace approximation leads to an inherent bias due to the uniform prior and the nonlinearity in the forward model with respect to the unknown parameter, θ .

We start with a tolerance range $[10^{-4}, 1]$ to analyze the optimal method parameters, i.e., N^* , M^* , and κ^* . Figure 7 shows the optimal setting for $N_e = 1$ (top) and $N_e = 10$ (bottom). MCLA can reach lower tolerances with $N_e = 10$ than with $N_e = 1$, which is a constraint given by the Laplace bias $\mathcal{O}(1/N_e)$. DLMCIS can achieve an accuracy of 10^{-3} with $M^* = 5$, compared to $M^* = 10^5$ with DLMLC. However, there is an additional cost associated with the change of measure in DLMCIS; namely, the method requires about 30 extra forward model solves per outer sample θ_n in order to find $\hat{\theta}_n$ by solving an optimization problem (33).

Figure 7: Optimal setting (outer N^* and inner M^* number of samples) vs. tolerance for nonlinear scalar model with uniform prior (Example 2).

The optimal balance parameter κ^* is shown against TOL in Figure 15; in contrast to the previous example, here κ^* is not constant. For MCLA, the optimal value κ^* is quadratically decreasing to zero as TOL approaches the size of the Laplace bias. MCLA is unable to achieve an accuracy better than the Laplace bias; accordingly, the Laplace bias decreases linearly with N_e .

Figure 8: Optimal split factor κ^* vs. tolerance for nonlinear scalar model with uniform prior (Example 2).

The absolute error for all three methods is below the specified tolerance TOL in the range $[10^{-2}, 1]$ with at least 97.5% probability (as specified); see Figure 9. We follow the same procedure as in Example 1 for the estimation of the constants in the optimal parameter

setting derivation.

Figure 9: Error vs. tolerance for nonlinear model with uniform prior (Example 2).

The running time against TOL has rate 2 for MCLA and DLMCIS, and rate 3 for DLMC; see Figure 10. Again, as in Example 1, DLMCIS exhibits a pre-asymptotic rate that is one order less than the theoretical asymptotic rate, since M^* is constant for tolerances within $[10^{-2}, 1]$. In this case, DLMCIS is about hundred times faster than DLMC for $TOL = 0.01$. MCLA performs similarly to DLMCIS; however, MCLA could not be computed for tolerances lower than its bias.

Figure 10: Average computational work and running time vs. tolerance for nonlinear scalar model with uniform prior (Example 2).

6.2.2. Expected information gain

In Figure 11, we present the estimation of the expected information gain using MCLA and DLMCIS for the experiment setups $\xi \in [0, 1]$. In Figure 11a, MCLA is applied for $\text{TOL} = 3 \times 10^{-2}$, and DLMCIS for $\text{TOL} = 10^{-3}$. The confidence bars of the MCLA curve show the 97.5% confidence intervals. Different tolerances are specified for each of the two methods due to the Laplace bias constraint. However, in Figure 11b, we omit the bias constraint by enforcing $\kappa^* = 1$, and see that the MCLA curve matches well with the DLMCIS curve for $\text{TOL} = 10^{-3}$. The resulting expected information gain curve is in agreement with the one reported previously, see [5].

(a) MCLA for $\text{TOL} = 3 \times 10^{-2}$, and DLMCIS for $\text{TOL} = 10^{-3}$. (b) MCLA when enforcing $\kappa^* = 1$ for $\text{TOL} = 10^{-3}$.

Figure 11: Expected information gain for nonlinear scalar model with uniform prior (Example 2).

6.3. Example 3: Electrical impedance tomography

Here, we consider the optimal design of EIT experiments, an imaging technique in which the conductivity is inferred of a closed body from the measurements of electrodes placed on its boundary surface. In the forward problem, low-frequency electrical currents are injected through the electrodes attached to a composite laminate material, where each ply is orthotropic. The potential field in the body of the material is considered quasi-static for a given conductivity. The mathematical model is a second-order partial differential equation with an electrode boundary model.

6.3.1. Mathematical model

We use the complete electrode model (CEM) to formulate the EIT problem for a composite laminate material, see [23]. The composite body, denoted by D , is composed of N_p orthotropic laminated plies, which yields a macroscale, anisotropic material. The configuration is such that the plies overlap with their fibers at different orientation angles.

The equations governing the potential field $\mathbf{u}$ are

$$\nabla \cdot \mathbf{j}(\omega, \mathbf{x}) = 0, \quad \text{in } D, \text{ and} \quad (41)$$

$$\mathbf{j}(\omega, \mathbf{x}) = \bar{\boldsymbol{\sigma}}(\omega, \mathbf{x}) \cdot \nabla \mathbf{u}(\omega, \mathbf{x}), \quad (42)$$

where $\mathbf{j}$ is the flux of electric current, the conductivity $\bar{\boldsymbol{\sigma}}$ is given by

$$\bar{\boldsymbol{\sigma}}(\omega, \mathbf{x}) = \mathbf{Q}^T(\theta_k(\omega)) \cdot \boldsymbol{\sigma} \cdot \mathbf{Q}(\theta_k(\omega)), \quad \text{for } \mathbf{x} \in D_k, \quad k = 1, \dots, N_p,$$

and the boundary conditions are specified in equations (43)-(45). The domain of a ply k is denoted by D_k ; thus, $D = \bigcup_{k=1}^{N_p} D_k$. The orthogonal matrix $\mathbf{Q}(\theta_k)$ is the rotational matrix that defines the orientation of the fibres, in ply k for a given angle θ_k , while $\boldsymbol{\sigma}$ stands for the orthotropic conductivity:

$$\mathbf{Q}(\theta_k) = \begin{bmatrix} \cos(\theta_k) & 0 & -\sin(\theta_k) \\ 0 & 1 & 0 \\ \sin(\theta_k) & 0 & \cos(\theta_k) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\sigma} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix}.$$

The upper and lower surfaces of boundary ∂D are equipped with N_{el} square-shaped electrodes E_l , $l = 1, \dots, N_{\text{el}}$, with dimensions e_{el} . On the free surface of the boundary, $\partial D \setminus (\cup E_l)$, we assume a no-flux condition, i.e., no current flow in the out-of-surface direction:

$$\mathbf{j} \cdot \mathbf{n} = 0, \quad (\bar{\boldsymbol{\sigma}}(\omega, \mathbf{x}) \cdot \nabla \mathbf{u}(\omega, \mathbf{x})) \cdot \mathbf{n} = 0, \quad (43)$$

where $\mathbf{n}$ represents the outward normal unit vector. CEM [23] is applied in the electrodes E_l . This means that we adopt (43) along with the assumption that the total injected current I_l through each electrode is known and given by

$$\int_{E_l} \mathbf{j} \cdot \mathbf{n} \, d\mathbf{x} = I_l \quad \text{on } E_l, \quad l = 1, \dots, N_{\text{el}}, \quad (44)$$

and that the shared interface of the electrode and the material has an infinitesimally thin layer with a surface impedance of z_l :

$$\frac{1}{E_l} \int_{E_l} \mathbf{u} \, d\mathbf{x} + z_l \int_{E_l} \mathbf{j} \cdot \mathbf{n} \, d\mathbf{x} = U_l \quad \text{on } E_l, \quad l = 1, \dots, N_{\text{el}}. \quad (45)$$

For sake of well-posedness, the Kirchhoff law of charge conservation and the ground potential condition are set as constraints to complete the boundary model,

$$\sum_{l=1}^{N_{\text{el}}} I_l = 0 \quad \text{and} \quad \sum_{l=1}^{N_{\text{el}}} U_l = 0. \quad (46)$$

In the rest of the paper, the EIT model refers to (41), (42), (43), (44), (45), and (46). Due to the randomness of θ_k , the conductivity field $\bar{\sigma}$ is random and assumed to be a uniformly and strictly positive element of $L^\infty(\Omega \times D)$ in order to guarantee ellipticity. The vectors $\mathbf{I} = (I_1, I_2, \dots, I_{N_{\text{el}}})^T$, and $\mathbf{U} = (U_1, U_2, \dots, U_{N_{\text{el}}})^T$ respectively design the vector of the injected (deterministic) current and the vector measurement of the (random) potential at the electrodes. According to the constraints, (46), $\mathbf{I}$ belongs to the mean-free subspace $\mathbb{R}_{\text{free}}^{N_{\text{el}}}$ of $\mathbb{R}^{N_{\text{el}}}$ and $\mathbf{U}$ is a random element of $\mathbb{R}_{\text{free}}^{N_{\text{el}}}$.

6.3.2. Finite element formulation

We let $H^1(D)$ be the Hilbert space of L^2 -integrable functions with L^2 -integrable derivatives over the physical domain, D . Moreover, $\mathcal{H} \stackrel{\text{def}}{=} H^1(D) \times \mathbb{R}_{\text{free}}^{N_{\text{el}}}$ denotes the space of the solution $(\mathbf{u}(\omega), \mathbf{U}(\omega))$ for a given random event $\omega \in \Omega$, and we introduce the Bochner space,

$$L_{\mathbb{P}}^2(\Omega; \mathcal{H}) \stackrel{\text{def}}{=} \left\{ (\mathbf{u}, \mathbf{U}) : \Omega \rightarrow \mathcal{H} \quad \text{s.t.} \quad \int_{\Omega} \|(\mathbf{u}(\omega), \mathbf{U}(\omega))\|_{\mathcal{H}}^2 d\mathbb{P}(\omega) < \infty \right\}.$$

The variational form associated with the EIT problem finds $(\mathbf{u}, \mathbf{U}) \in L_{\mathbb{P}}^2(\Omega; \mathcal{H})$ such that

$$\mathbb{E}[B((\mathbf{u}, \mathbf{U}), (\mathbf{v}, \mathbf{V}))] = \mathbf{I} \cdot \mathbb{E}[\mathbf{U}], \quad \text{for all } (\mathbf{v}, \mathbf{V}) \in L_{\mathbb{P}}^2(\Omega; \mathcal{H}), \quad (47)$$

where for any event $\omega \in \Omega$, the bilinear form $B : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ is

$$B((\mathbf{u}, \mathbf{U}), (\mathbf{v}, \mathbf{V})) = \int_D \mathbf{j} \cdot \nabla \mathbf{v} dD + \sum_{l=1}^{N_{\text{el}}} \frac{1}{z_l} \int_{E_l} (U_m - \mathbf{u})(V_m - \mathbf{v}) dE_l. \quad (48)$$

We let K be an element in the triangulation $\mathcal{T}_h$ and $\bar{D} \stackrel{\text{def}}{=} \cup_{K \in \mathcal{T}_h} K$. Then we define the subspace $H_h^1(D_k) \stackrel{\text{def}}{=} \{\mathbf{u}_h|_K \subset C^0(\bar{D}), \forall K \in \mathcal{T}_h\}$. Also, we let $\mathcal{H}_h \stackrel{\text{def}}{=} H_h^1(D) \times \mathbb{R}_{\text{free}}^{N_{\text{el}}} \subset \mathcal{H}$. Next we consider the finite-dimensional problem in the Bubnov-Galerkin sense. The trial $(\mathbf{u}_h, \mathbf{U}_h)$ (test $(\mathbf{v}_h, \mathbf{V}_h)$) function pair is denoted such that $(\mathbf{u}_h, \mathbf{U}_h) \in \mathcal{H}_h$ ($(\mathbf{v}_h, \mathbf{V}_h) \in \mathcal{H}_h$). Finally, by rephrasing $L_{\mathbb{P}}^2$ on $(\Omega; \mathcal{H}_h)$, we conclude the finite element formulation.

6.3.3. Bayesian experimental design formulation

We perform the experiments using the potentials measured at the electrodes. Hence, the Bayesian formulation of the EIT model is given by

$$\mathbf{y}_i = \mathbf{g}_h(\boldsymbol{\theta}_t) + \boldsymbol{\epsilon}_i \stackrel{\text{def}}{=} \mathbf{U}_h(\boldsymbol{\theta}_t) + \boldsymbol{\epsilon}_i, \quad \text{for } i = 1, \dots, N_e, \quad (49)$$

where $\mathbf{y}_i \in \mathbb{R}^{N_{\text{el}}-1}$ (i.e., $q = N_{\text{el}} - 1$), $\boldsymbol{\theta}_t = (\theta_{t,1}, \theta_{t,2})$, and the error distribution is Gaussian, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 0.25)$. We consider a uniform distribution to describe our prior knowledge of $\boldsymbol{\theta}_t$, i.e.,

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

and

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

We consider a body consisting of two plies whose parameters are $\sigma_{11} = 0.05$, $\sigma_{22} = \sigma_{33} = 10^{-3}$, and $z_l = 0.1$. A total of 10 electrodes are placed on the surface of the plies to measure the potential at the electrodes. The orientations of the angles θ_1 and θ_2 of the fibers are the uncertain parameters. Figure 12 presents the physical configuration of the experiment.

Figure 12: Description of the domain for the EIT problem (Example 3).

Figure 13a presents the experiment for which we compute the respective optimal settings for MCLA, DLMC and DLMCIS. The domain is $D = [0, 20] \times [0, 2]$, and we set $e_{el} = e_{sh} = e_{sp} = 2$. We depict the potential field with its equipotential lines. Figure 13b illustrates the current streamlines, $\mathbf{j}$.

Figure 13: Experimental setup for the EIT problem (Example 3).

To understand how the potential field may behave at the surface boundaries, we show the potential profiles at $x_2 = 0$ (Figure 14a) and $x_2 = 2$ (Figure 14b). Sharp changes in the potential field are appreciated at the boundaries of the electrodes, meaning that the meshes utilized must be very fine. The solutions depicted in Figures 13 and 17 are computed using 1600×160 linear elements.

Figure 14: Potential profiles for the EIT problem (Example 3).

6.3.4. Optimality and error convergence

Figure 15a shows the optimal number of outer and inner samples with respect to tolerance. Given the Laplace bias, MCLA can be performed to $\text{TOL} = 1$, whereas DLMCIS, which is not limited by any uncontrolled bias, is carried out to $\text{TOL} = 0.1$. The constants are estimated using $N = M = 10$ for DLMCIS, instead of $N = M = 100$ as in the previous examples, due to the expensive nature of this model. Even though the constants are roughly estimated, our results show a good agreement in tolerance versus absolute error for DLMCIS when using a confidence level of 97.5%, as seen in Figure 16a.

(a) N^* and M^*

(b) κ^*

Figure 15: Optimal setting (outer N^* and inner M^* number of samples, and balance κ^*) vs. tolerance for the EIT problem (Example 3).

The optimal discretization is roughly given by $h^* \approx 20\text{TOL}^{1/\eta}$. We constrain the maximum allowed h to the mesh size that corresponds to 50 elements along the plies, and to 2 elements throughout the thickness of each ply; therefore, the same discretization size is used for $\text{TOL} = 10$ and $\text{TOL} = 1$. The discontinuity in the material and the discontinuity between the electrode and the no-flux boundary condition lead to discontinuity in the gra-

dient of the potential field and, as a result, the h -convergence rate of $\mathbf{U}_h$ is $\eta \approx 1.15$ and the work rate is $\gamma \approx 3.5$.

To obtain an accurate reference solution for the computational error, we substitute the MC sampling for the outer integral by a Gauss-Legendre quadrature with 20×20 points, as the inner integral is low dimensional (in our case, $\dim(\boldsymbol{\theta}) = 2$). Figure 16a shows the computational error in relation to tolerance. The relationship between computational work and tolerance is given in Figure 16b, where the theoretical asymptotic rates are numerically verified.

Figure 16: To the left is error vs. tolerance for DLMCIS only, and to the right is running time vs. tolerance for MCLA and DLMCIS (Example 3).

To reach $\text{TOL} \sim 0.01$, the number of samples needed for the inner loop in DLMCIS is $M^* = 1$, and $M^* \sim 10^3$ for DLMC; for MCLA we are unable to reach accuracies better than $\text{TOL} \sim 1$. For all of the methods, the bias is larger than the statistical error; more specifically, κ^* is less 0.5 (see Figure 15b). As discussed in Example 2, the Laplace approximation for the importance sampling requires about 30 forward model evaluations per outer sample $\boldsymbol{\theta}_n$ to estimate $\hat{\boldsymbol{\theta}}_n$.

For MCLA, the rate of the computational work cannot be observed, as the work required is the same for $\text{TOL} = 1$ and 10. This is because the Laplace bias is the dominant error component and N^* is equal to 1 for both tolerance levels. For DLMCIS, the observed computational work is $\mathcal{O}(\text{TOL}^{-2})$, which is a pre-asymptotic rate that is two orders better than the asymptotic result. The reason for this is that the spatial discretization and the number of inner samples are constant for the range of tolerances considered.

6.3.5. Expected information gain

For this demonstration, we assign the design variables $\boldsymbol{\xi} = (\xi_1, \xi_2)$, where $\xi_1 = e_{sh}$ and $\xi_2 = e_{sp}$. We compute the expected information gain over the range $\boldsymbol{\xi} \in [0, 2] \times$

$[0.05, 0.5]$. The reference solution is computed using the Gauss-Legendre quadrature for the outer integral to achieve a relative error 0.01. The measurement error follows the distribution $\epsilon \sim \mathcal{N}(0, 0.01)$.

The experiment is different from the previously considered experimental setup in Figure 13, and is depicted in Figure 17a. All of the electrodes at the top boundary inject current, and the electrodes at the bottom act as outlets, and the domain is $D = [0, 20] \times [0, 2]$. The current streamlines, $\mathbf{j}$, are shown in Figure 17b.

Figure 17: The experimental setup for the expected information gain computation (Example 3).

Figure 18a shows the potential profile at the bottom of the domain, $x_2 = 0$, whereas Figure 18b shows the potential profile at the top of the domain $x_2 = 2$. The potential field is smoother than seen in the experimental setup shown in Figure 13, due to all of the current being injected at the top and leaving at the bottom.

Figure 18: The potential profile for the EIT problem (Example 3).

The expected information gain is presented in Figure 19 together with the posterior distributions for three experimental designs. In the figure, we see that there is oscillatory behavior at the surface; when the set of electrodes at the bottom is moved by ξ_1 , the main current flow changes from one pair of electrodes to a different pair. Similar behavior occurs for the other design component, ξ_2 . Moreover, the expected information gain is lower when the distance between the electrodes is larger.

Figure 19: Expected information gain for the EIT problem (Example 3).

The maximum expected information gain occurs at $\xi = (2.0, 0.6)$, and the corresponding posterior distribution is shown at the top of Figure 19; the worst experiment occurs at $\xi = (0.5, 2.0)$, and its posterior distribution is depicted in the bottom-left of the figure. We show the electrode placement and current flux for the best and worst experiment in Figure 20. The best experiment suggests that even better results can be achieved by further increasing the shift of the electrodes, e_{sh} . From a design perspective, it is interesting that the optimal distance between the electrodes is about a quarter of the electrode size.

Figure 20: The electrode placement and current flux for the worst (bottom) and best (top) experiments according to in accordance with the expected information gain presented in Figure 19 (Example 3).

Conclusion

We presented a computationally efficient method, DLMCIS, for computing the expected information gain in optimal Bayesian experimental design. The method uses importance sampling in the inner loop of the classical DLMC method through a change of measure based on the Laplace method. We demonstrated that the use of importance sampling can substantially reduce the number of samples in the inner loop, leading to an improvement in the running time of DLMC by several orders of magnitude. Further benefits include preservation of the asymptotic unbiasedness and mitigates the risk of underflow in the computation of the inner loop. We derived optimal settings for DLMCIS, DLMC and MCLA. The methods were optimized by minimizing the average computational work for a given tolerance. We showed that DLMCIS achieves the same work rate as MCLA for higher tolerances since only a few inner samples were needed, and for lower tolerances MCLA is not applicable due to its inherent bias. The efficiency of DLMCIS was demonstrated for optimal sensor placement in an EIT problem, where the composite material model was solved using high-fidelity finite elements.

Acknowledgements

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST); KAUST CRG3 Award Ref:2281 and the KAUST CRG4 Award Ref:2584.

References

References

- [1] S. Kullback, R. A. Leibler, On information and sufficiency, *Annals of Mathematical Statistics* 22 (1951) 79–86.
- [2] S. Kullback, *Information Theory and Statistics*, Wiley, 1959.
- [3] J. Ginebra, On the measure of the information in a statistical experiment, *Bayesian Analysis* 2 (2007) 167–211.

- [4] K. J. Ryan, Estimating expected information gains for experimental designs with application to the random fatigue-limit model, *Journal of Computational and Graphical Statistics* 12 (2003) 585–603.
- [5] X. Huan, Y. M. Marzouk, Simulation-based optimal Bayesian experimental design for nonlinear systems, *Journal of Computational Physics* 232 (1) (2013) 288–317.
- [6] S. M. Stigler, Laplace’s 1774 memoir on inverse probability, *Statistical Science* 1 (1986) 359–363.
- [7] L. Tierney, J. B. Kadane, Accurate approximations for posterior moments and marginal densities, *Journal of American Statistical Association* 81 (1986) 82–86.
- [8] L. Tierney, R.E.Kass, J.B.Kadane, Fully exponential Laplace approximations to expectations and variances of nonpositive functions, *Journal of American Statistical Association* 710-716 (84) (1989) 710–716.
- [9] R.E.Kass, L.Tierney, J.B.Kadane, *Essays in Honor of George Barnard* (eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner), North-Holland, 1990, Ch. The Validity of Posterior Expansions Based on Laplace’s Method, pp. 473–488.
- [10] Q. Long, M. Scavino, R. Tempone, S. Wang, Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations, *Computer Methods in Applied Mechanics and Engineering* 259 (2013) 24–39.
- [11] Q. Long, M. Motamed, R. Tempone, Fast Bayesian optimal experimental design for seismic source inversion, *Computer Methods in Applied Mechanics and Engineering* 155 (2015) 123–145.
- [12] C. Papadimitriou, Optimal sensor placement methodology for parametric identification of structural systems, *Journal of Sound and Vibration* 278 (2004) 923–947.
- [13] Q. Long, M. Scavino, R. Tempone, S. Wang, A Laplace method for under-determined Bayesian optimal experimental design, *Computer Methods in Applied Mechanics and Engineering* 285 (2015) 849–876.
- [14] F. Bisetti, D. Kim, O. Knio, Q. Long, R. Tempone, Optimal Bayesian experimental design for priors of compact support with application to shock-tube experiments for combustion kinetics, *International Journal for Numerical Methods in Engineering* 108 (2) (2016) 136–155.
- [15] A. Alexanderian, N. Petra, G. Stadler, O. Ghattas, A fast and scalable method for a-optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems, *SIAM Journal on Scientific Computing* 38 (1) (2016) A243–A272.
- [16] J. Beck, S. Guillas, Sequential design with mutual information for computer experiments (mice): Emulation of a tsunami model, *SIAM/ASA Journal on Uncertainty Quantification* 4 (1) (2016) 739–766.
- [17] C. Feng, Optimal bayesian experimental design in the presence of model error, Ph.D. thesis, Massachusetts Institute of Technology (2015).
- [18] D. V. Lindley, On a measure of information provided by an experiment, *The Annals of Mathematical Statistics* 27 (1956) 986–1005.
- [19] C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal* 27 (1948) 379–423.
- [20] M. B. Giles, Multilevel Monte Carlo path simulation, *Operations Research* 56 (3) (2008) 607–617.
- [21] N. Collier, A.-L. Haji-Ali, F. Nobile, E. von Schwerin, R. Tempone, A continuation multilevel monte carlo algorithm, *BIT Numerical Mathematics* (2014) 1–34.
- [22] K. J. Ryan, Estimating expected information gains for experimental designs with application to the random fatigue-limit model, *Journal of Computational and Graphical Statistics* 12 (3) (2003) 585–603.
- [23] E. Somersalo, M. Cheney, D. Isaacson., Existence and uniqueness for electrode models for electric current computed tomography, *SIAM J. Appl. Math.*, 52 (1992) 1023–1040.

Appendix A. Estimation of the bias and variance of the DLMC estimator

Proof. This is the proof of Proposition 1. First, we show that the bias of the DLMC estimator, $\mathcal{I}_{dl,h}$, for the expected information gain can be upper-bounded by

$$|I - \mathbb{E}[\mathcal{I}_{dl,h}]| \leq C_{dl,3}h^\eta + \frac{C_{dl,4}}{M} + o(h^\eta) + \mathcal{O}\left(\frac{1}{M^2}\right),$$

where I is the expected information gain, and the subscript h of the estimator shows its dependence on the forward model, $\mathbf{g}_h$.

We decompose the bias as follows:

$$|I - \mathbb{E}[\mathcal{I}_{dl,h}]| \leq |I - \mathbb{E}[\mathcal{I}_{dl}]| + |\mathbb{E}[\mathcal{I}_{dl} - \mathcal{I}_{dl,h}]|, \quad (\text{A.1})$$

where $\mathcal{I}_{dl} = \lim_{h \rightarrow 0^+} \mathcal{I}_{dl,h}$. The numerical bias of the estimator, due to the mesh discretization of $\mathbf{g}_h$, follows as

$$|\mathbb{E}[\mathcal{I}_{dl} - \mathcal{I}_{dl,h}]| = Ch^\eta + o(h^\eta), \quad (\text{A.2})$$

for a constant $C > 0$, where $\mathcal{O}(h^\eta)$ is of the same order as for $\mathbf{g}_h$. This can be seen by tracking $\mathcal{O}(h^\eta)$ through the computation of $\log p(\mathbf{Y}|\boldsymbol{\theta})$ in $\mathcal{I}_{dl,h}$.

The other bias component, $|I - \mathbb{E}[\mathcal{I}_{dl}]|$, resulting from the inner averaging $\hat{p}_M(\mathbf{Y}) = \frac{1}{M} \sum_{m=1}^M p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_m)$ within the logarithmic function of the estimator, is proportional to the number of samples, M . This bias term can be recast as:

$$\begin{aligned} |I - \mathbb{E}[\mathcal{I}_{dl}]| &= \mathbb{E} \left[\frac{1}{N} \sum_{n=1}^N \log \left(\frac{p(\mathbf{Y}_n|\boldsymbol{\theta}_n)}{p(\mathbf{Y}_n)} \right) \right] - \mathbb{E} \left[\frac{1}{N} \sum_{n=1}^N \log \left(\frac{p(\mathbf{Y}_n|\boldsymbol{\theta}_n)}{\hat{p}_M(\mathbf{Y}_n)} \right) \right] \\ &= \mathbb{E}[\log(\hat{p}_M(\mathbf{Y}))] - \mathbb{E}[\log(p(\mathbf{Y}))]. \end{aligned} \quad (\text{A.3})$$

We apply the second-order Taylor expansion of $\log(X)$ around $\mathbb{E}[X]$ for both terms in (A.3):

$$\log(X) = \log(\mathbb{E}[X]) + \frac{1}{\mathbb{E}[X]} (X - \mathbb{E}[X]) - \frac{1}{2\mathbb{E}[X]^2} (X - \mathbb{E}[X])^2 + \mathcal{O}([X - \mathbb{E}[X]]^3).$$

That is, the expectation of $\log(\hat{p}_M(\mathbf{Y}))$ in (A.3) can be written as

$$\begin{aligned} \mathbb{E}[\log(\hat{p}_M(\mathbf{Y}))] &= \mathbb{E}[\log(\mathbb{E}[\hat{p}_M(\mathbf{Y})])] - \frac{1}{2(\mathbb{E}[\hat{p}_M(\mathbf{Y})])^2} \mathbb{E}[(\hat{p}_M(\mathbf{Y}) - \mathbb{E}[\hat{p}_M(\mathbf{Y})])^2] + \mathcal{O}\left(\frac{1}{M^2}\right) \\ &= \log(p(\mathbf{Y})) - \frac{1}{2p^2(\mathbf{Y})} \mathbb{E}[(\hat{p}_M(\mathbf{Y}) - p(\mathbf{Y}))^2] + \mathcal{O}\left(\frac{1}{M^2}\right). \end{aligned}$$

Here, the order $\mathcal{O}(\frac{1}{M^2})$ comes from the fourth-order term of the Taylor expansion. Using

the above results, we derive the result for (A.3) as

$$\begin{aligned}
|I - \mathbb{E}[\mathcal{I}_{dl}]| &= \frac{1}{2} \mathbb{E} \left[\frac{1}{p^2(\mathbf{Y})} \mathbb{E} \left[(\hat{p}_M(\mathbf{Y}) - p(\mathbf{Y}))^2 \middle| \mathbf{Y} \right] \right] + \mathcal{O} \left(\frac{1}{M^2} \right) \\
&= \frac{1}{2} \mathbb{E} \left[\frac{1}{p^2(\mathbf{Y})} \mathbb{V} \left[\hat{p}_M(\mathbf{Y}) \middle| \mathbf{Y} \right] \right] + \mathcal{O} \left(\frac{1}{M^2} \right) \\
&= \frac{1}{2} \mathbb{E} \left[\frac{1}{p^2(\mathbf{Y})} \frac{1}{M^2} \mathbb{V} \left[\sum_{m=1}^M p(\mathbf{Y} | \tilde{\boldsymbol{\theta}}_m) \middle| \mathbf{Y} \right] \right] + \mathcal{O} \left(\frac{1}{M^2} \right) \\
&= \frac{1}{2} \mathbb{E} \left[\frac{1}{p^2(\mathbf{Y})} \frac{1}{M^2} \sum_{m=1}^M \mathbb{V} \left[p(\mathbf{Y} | \tilde{\boldsymbol{\theta}}) \middle| \mathbf{Y} \right] \right] + \mathcal{O} \left(\frac{1}{M^2} \right) \\
&= \frac{1}{2M} \mathbb{E} \left[\frac{\mathbb{V} \left[p(\mathbf{Y} | \tilde{\boldsymbol{\theta}}) \middle| \mathbf{Y} \right]}{p^2(\mathbf{Y})} \right] + \mathcal{O} \left(\frac{1}{M^2} \right).
\end{aligned}$$

The above result completes the derivation of the upper bound of the total bias (A.1) with respect to M and h . Throughout the proof, we drop the subscripts when they are not required since they are identically distributed variables.

Next, we derive the expression for the variance of the estimator,

$$\mathbb{V}[\mathcal{I}_{dl}] = \frac{C_{dl,1}}{N} + \frac{C_{dl,2}}{NM} + \mathcal{O} \left(\frac{1}{NM^2} \right).$$

By the law of total variance, we attain the following result:

$$\begin{aligned}
\mathbb{V}[\mathcal{I}_{dl}] &= \mathbb{V} \left[\frac{1}{N} \sum_{n=1}^N \log \left(\frac{p(\mathbf{Y}_n | \boldsymbol{\theta}_n)}{\hat{p}_M(\mathbf{Y})} \right) \right] \\
&= \frac{1}{N} \mathbb{V} \left[\mathbb{E} \left[\log(p(\mathbf{Y} | \boldsymbol{\theta})) - \log(\hat{p}_M(\mathbf{Y})) \middle| \boldsymbol{\theta}, \mathbf{Y} \right] \right] \\
&\quad + \frac{1}{N} \mathbb{E} \left[\mathbb{V} \left[\log(p(\mathbf{Y} | \boldsymbol{\theta})) - \log(\hat{p}_M(\mathbf{Y})) \middle| \boldsymbol{\theta}, \mathbf{Y} \right] \right] \\
&= \frac{1}{N} \mathbb{V} \left[\mathbb{E} \left[\log(p(\mathbf{Y} | \boldsymbol{\theta})) - \log(\hat{p}_M(\mathbf{Y})) \middle| \boldsymbol{\theta}, \mathbf{Y} \right] \right] + \frac{1}{N} \mathbb{E} \left[\mathbb{V} \left[\log(\hat{p}_M(\mathbf{Y})) \middle| \mathbf{Y} \right] \right].
\end{aligned} \tag{A.4}$$

Here, we apply the second-order Taylor expansion around $\mathbb{E}[\hat{p}_M(\mathbf{Y})]$ for both terms. The

first term of (A.4) can be rewritten as follows:

$$\begin{aligned}
& \frac{1}{N} \mathbb{V} \left[\mathbb{E} \left[\log(p(\mathbf{Y}|\boldsymbol{\theta})) - \log(\hat{p}_M(\mathbf{Y})) \middle| \boldsymbol{\theta}, \mathbf{Y} \right] \right] \\
&= \frac{1}{N} \mathbb{V} \left[\mathbb{E} \left[\log \left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \right) - \frac{1}{2p^2(\mathbf{Y})} \left(\frac{1}{M} \sum_{m=1}^M p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_m) - p(\mathbf{Y}) \right)^2 \middle| \boldsymbol{\theta}, \mathbf{Y} \right] \right] + \mathcal{O} \left(\frac{1}{NM^2} \right) \\
&= \frac{1}{N} \mathbb{V} \left[\log \left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \right) - \frac{1}{2} \frac{1}{M} \mathbb{V} \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \middle| \mathbf{Y} \right] \right] + \mathcal{O} \left(\frac{1}{NM^2} \right) \\
&= \frac{1}{N} \left(\mathbb{V} \left[\log \left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \right) \right] + \frac{1}{4M^2} \mathbb{V} \left[\mathbb{V} \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \middle| \mathbf{Y} \right] \right] \right. \\
&\quad \left. - \frac{1}{M} \text{Cov} \left[\log \left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \right), \mathbb{V} \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \middle| \mathbf{Y} \right] \right] \right) + \mathcal{O} \left(\frac{1}{NM^2} \right) \\
&= \frac{1}{N} \left(\mathbb{V} \left[\log \left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \right) \right] + \frac{1}{4M^2} \mathbb{V} \left[\mathbb{V} \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \middle| \mathbf{Y} \right] \right] \right. \\
&\quad \left. - \frac{1}{M} \left(\mathbb{E} \left[\log \left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \right) \mathbb{V} \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \middle| \mathbf{Y} \right] \right] \right. \right. \\
&\quad \left. \left. - \mathbb{E} \left[\log \left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \right) \right] \mathbb{E} \left[\mathbb{V} \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \middle| \mathbf{Y} \right] \right] \right) \right) + \mathcal{O} \left(\frac{1}{NM^2} \right) \\
&= \frac{1}{N} \mathbb{V} \left[\log \left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \right) \right] - \frac{1}{NM} \left(\mathbb{E} \left[\log \left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \right) \mathbb{V} \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \middle| \mathbf{Y} \right] \right] \right. \\
&\quad \left. - \mathbb{E} \left[\log \left(\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \right) \right] \mathbb{E} \left[\mathbb{V} \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \middle| \mathbf{Y} \right] \right] \right) + \mathcal{O} \left(\frac{1}{NM^2} \right) \tag{A.5}
\end{aligned}$$

Similarly, for the second term of (A.4), we obtain:

$$\begin{aligned}
& \frac{1}{N} \mathbb{E} \left[\mathbb{V} \left[\log(\hat{p}_M(\mathbf{Y})) \middle| \mathbf{Y} \right] \right] \\
&= \frac{1}{N} \mathbb{E} \left[\mathbb{V} \left[-\log(p(\mathbf{Y})) + \frac{1}{p(\mathbf{Y})} \left(\frac{1}{M} \sum_{m=1}^M p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_m) - p(\mathbf{Y}) \right) \middle| \mathbf{Y} \right] \right] \\
&= \frac{1}{N} \mathbb{E} \left[\mathbb{V} \left[\frac{1}{p(\mathbf{Y})} \left(\frac{1}{M} \sum_{m=1}^M p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_m) - p(\mathbf{Y}) \right) \middle| \mathbf{Y} \right] \right] \\
&= \frac{1}{N} \mathbb{E} \left[\mathbb{V} \left[\frac{1}{M} \frac{\sum_{m=1}^M p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_m)}{p(\mathbf{Y})} \middle| \mathbf{Y} \right] \right] \\
&= \frac{1}{NM^2} \mathbb{E} \left[\mathbb{V} \left[\frac{\sum_{m=1}^M p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_m)}{p(\mathbf{Y})} \middle| \mathbf{Y} \right] \right] \\
&= \frac{1}{NM} \mathbb{E} \left[\mathbb{V} \left[\frac{p(\mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{Y})} \middle| \mathbf{Y} \right] \right]. \tag{A.6}
\end{aligned}$$

The proof is completed by combining the two terms into the variance. $\square$

Appendix B. A leading-order expansion of the likelihood with respect to N_e

Proof. The likelihood is defined as

$$p(\mathbf{Y}_n|\boldsymbol{\theta}) = \prod_{i=1}^{N_e} p(\mathbf{y}_{n,i}|\boldsymbol{\theta}), \quad \forall n \leq N, \text{ where } \mathbf{Y}_n = \{\mathbf{y}_{n,i}\}_{i=1}^{N_e},$$

and the errors are distributed as $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon)$, $(\boldsymbol{\Sigma}_\epsilon)_{jj} = \sigma_{\epsilon_j}^2$, and the unknown parameter value $\boldsymbol{\theta}$ follows the prior pdf, $\pi(\boldsymbol{\theta})$. Thus, the likelihood for the inner loop can be written as

$$p(\mathbf{Y}_n|\tilde{\boldsymbol{\theta}}_{n,m}) = (2\pi|\boldsymbol{\Sigma}_\epsilon|)^{-\frac{N_e}{2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{N_e} \left\| \mathbf{y}_{n,i} - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2\right), \quad (\text{B.1})$$

where

$$\begin{aligned} \left\| \mathbf{y}_{n,i} - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2 &= \left(\mathbf{g}(\boldsymbol{\theta}_n) + \boldsymbol{\epsilon}_i - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right)^T \boldsymbol{\Sigma}_\epsilon^{-1} \left(\mathbf{g}(\boldsymbol{\theta}_n) + \boldsymbol{\epsilon}_i - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right) \\ &= \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right)^T \boldsymbol{\Sigma}_\epsilon^{-1} \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right) \\ &\quad + \boldsymbol{\epsilon}_i^T \boldsymbol{\Sigma}_\epsilon^{-1} \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right) + \boldsymbol{\epsilon}_i^T \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{\epsilon}_i \\ &= \left\| \mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2 + \boldsymbol{\epsilon}_i^T \boldsymbol{\Sigma}_\epsilon^{-1} \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right) + \|\boldsymbol{\epsilon}_i\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2. \end{aligned}$$

We know that

$$\sum_{i=1}^{N_e} \left\| \mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2 = N_e \left\| \mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2,$$

which we expand to

$$\sum_{i=1}^{N_e} \boldsymbol{\epsilon}_i^T \boldsymbol{\Sigma}_\epsilon^{-1} \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right) \quad \text{and} \quad \sum_{i=1}^{N_e} \|\boldsymbol{\epsilon}_i\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2.$$

We will use Kolmogorov's inequality for a random variable X :

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq k\sqrt{\mathbb{V}[X]}\right) < \frac{1}{k^2}, \quad (\text{B.2})$$

where k is a real number greater than 1.

First,

$$\begin{aligned} \sum_{i=1}^{N_e} \boldsymbol{\epsilon}_i^T \boldsymbol{\Sigma}_\epsilon^{-1} \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right) &= \sum_{i=1}^{N_e} \sum_{j=1}^q \frac{\epsilon_{i,j}}{\sigma_{\epsilon_j}^2} \left(\mathbf{g}_j(\boldsymbol{\theta}_n) - \mathbf{g}_j(\tilde{\boldsymbol{\theta}}_{n,m}) \right) \\ &= \sum_{j=1}^q \left(\sum_{i=1}^{N_e} \epsilon_{i,j} \right) \frac{1}{\sigma_{\epsilon_j}^2} \left(\mathbf{g}_j(\boldsymbol{\theta}_n) - \mathbf{g}_j(\tilde{\boldsymbol{\theta}}_{n,m}) \right). \end{aligned}$$

We apply (B.2) to attain $\mathbb{P}\left(\left|\sum_{i=1}^{N_e} \epsilon_{i,j}\right| \geq k\sigma_{\epsilon_j}\sqrt{N_e}\right) < \frac{1}{k^2}$, which leads to $\sum_{i=1}^{N_e} \epsilon_{i,j} = \sigma_{\epsilon_j}\mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right)$ and, thus,

$$\begin{aligned} \sum_{i=1}^{N_e} \boldsymbol{\epsilon}_i^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right) &= \sum_{j=1}^q \left(\sigma_{\epsilon_j} \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right) \right) \frac{1}{\sigma_{\epsilon_j}^2} \left(g_j(\boldsymbol{\theta}_n) - g_j(\tilde{\boldsymbol{\theta}}_{n,m}) \right) \\ &= \sum_{j=1}^q \frac{1}{\sigma_{\epsilon_j}} \left(g_j(\boldsymbol{\theta}_n) - g_j(\tilde{\boldsymbol{\theta}}_{n,m}) \right) \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right) \\ &= \mathbf{v}_{\boldsymbol{\epsilon}}^T \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right) \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right), \end{aligned}$$

where $\mathbf{v}_{\boldsymbol{\epsilon}}$ is the vector of the diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}$. For the other term, we have

$$\begin{aligned} \sum_{i=1}^{N_e} \|\boldsymbol{\epsilon}_i\|_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}}^2 &= \sum_{i=1}^{N_e} \boldsymbol{\epsilon}_i^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{\epsilon}_i = \sum_{i=1}^{N_e} \sum_{j=1}^q \frac{\epsilon_{i,j}^2}{\sigma_{\epsilon_j}^2} \\ &= \sum_{j=1}^q \frac{1}{\sigma_{\epsilon_j}^2} \left(\sum_{i=1}^{N_e} \epsilon_{i,j}^2 \right). \end{aligned}$$

The random variable $\sum_{i=1}^{N_e} \epsilon_{i,j}^2$ follows the chi-squared distribution with mean $\sigma_{\epsilon_j} N_e$ and variance $2\sigma_{\epsilon_j}^2 N_e$. Similarly, we apply (B.2) to obtain $\sum_{i=1}^{N_e} \epsilon_{i,j}^2 = -\sigma_{\epsilon_j} N_e + \sum_{i=1}^{N_e} (\epsilon_{i,j}^2 + \sigma_{\epsilon_j}^2) = -\sigma_{\epsilon_j} N_e + \sigma_{\epsilon_j} \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right)$; as a result,

$$\sum_{i=1}^{N_e} \|\boldsymbol{\epsilon}_i\|_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}}^2 = \sum_{j=1}^q \frac{1}{\sigma_{\epsilon_j}^2} \left(\sum_{i=1}^{N_e} \epsilon_{i,j}^2 \right) = \sum_{j=1}^q \sigma_{\epsilon_j}^{-1} \left(N_e + \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right) \right).$$

Finally, we end up with

$$\begin{aligned} p(\mathbf{Y}_n | \tilde{\boldsymbol{\theta}}_{n,m}) &= (2\pi|\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}|)^{-\frac{N_e}{2}} \exp \left(-\frac{1}{2} N_e \left\| \mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right\|_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}}^2 \right. \\ &\quad \left. - \frac{1}{2} \mathbf{v}_{\boldsymbol{\epsilon}}^T \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right) \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right) - \frac{1}{2} \sum_{j=1}^q \sigma_{\epsilon_j}^{-1} \left(N_e + \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right) \right) \right). \end{aligned}$$

□

Appendix C. Laplace approximation of posterior distributions

We derive the Laplace posterior pdf $\pi_{la}(\boldsymbol{\theta} | \mathbf{Y}) \sim \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)$, which is an approximation of the posterior pdf, $\pi(\boldsymbol{\theta} | \mathbf{Y})$.

Proof. We let F be the negative logarithmic of the posterior distribution:

$$F(\boldsymbol{\theta}) = -\log(\pi(\boldsymbol{\theta}|\mathbf{Y})). \quad (\text{C.1})$$

The mode $\hat{\boldsymbol{\theta}}$ of the posterior pdf is given by

$$\hat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \arg \min_{\boldsymbol{\theta} \in \Theta} F(\boldsymbol{\theta}), \quad (\text{C.2})$$

which implies that $\nabla_{\boldsymbol{\theta}} F(\hat{\boldsymbol{\theta}}) = 0$. By matching the second-order Taylor expansion of F around $\hat{\boldsymbol{\theta}}$ and $-\log(\pi_{la}(\boldsymbol{\theta}|\mathbf{Y}))$ we obtain $\hat{\Sigma}$ as the inverse Hessian matrix of the negative logarithm of the posterior pdf evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, i.e.,

$$\hat{\Sigma} = \left(\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} F(\hat{\boldsymbol{\theta}}) \right)^{-1}. \quad (\text{C.3})$$

We use $\nabla_{\boldsymbol{\theta}} F(\hat{\boldsymbol{\theta}}) = 0$ in the evaluation at $\boldsymbol{\theta} = \boldsymbol{\theta}_t$ of the Taylor expansion, of $\nabla_{\boldsymbol{\theta}} F$, around $\hat{\boldsymbol{\theta}}$ to write

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_t - \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}_t) (\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}_t))^{-1} + \mathcal{O} \left(\left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t \right\|^2 \right). \quad (\text{C.4})$$

In order to express these moments, (C.3) and (C.4), in terms of $\mathbf{g}$ rather than F , we further expand by introducing

$$\mathbf{E}_{\epsilon}(\boldsymbol{\theta}) = \sum_{i=1}^{N_e} \mathbf{r}_i^T(\boldsymbol{\theta}), \quad \mathbf{J}(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \mathbf{g}(\boldsymbol{\theta}), \quad \mathbf{H}(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \mathbf{g}(\boldsymbol{\theta}), \quad \text{and} \quad h(\boldsymbol{\theta}) = \log(\pi(\boldsymbol{\theta})),$$

where $\mathbf{J}(\boldsymbol{\theta})$ and $\mathbf{H}(\boldsymbol{\theta})$ are the Jacobian and the Hessian of $-\mathbf{g}(\boldsymbol{\theta})$, respectively. For $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, we have $\mathbf{E}_{\epsilon}(\hat{\boldsymbol{\theta}}) \sim \mathcal{N}(N_e(\mathbf{g}(\boldsymbol{\theta}_t) - \mathbf{g}(\hat{\boldsymbol{\theta}})), N_e \Sigma_{\epsilon})$.

This allows us to recast the moments as

$$\begin{aligned} \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_t - & \left(\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_t) + \mathbf{E}_{\epsilon} \Sigma_{\epsilon}^{-1} \mathbf{J}(\boldsymbol{\theta}_t) \right) \left(N_e \mathbf{J}(\boldsymbol{\theta}_t)^T \Sigma_{\epsilon}^{-1} \mathbf{J}(\boldsymbol{\theta}_t) \right. \\ & \left. - \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}_t) + \mathbf{H}(\boldsymbol{\theta}_t)^T \Sigma_{\epsilon}^{-1} \mathbf{E}_{\epsilon} \right)^{-1} + \mathcal{O} \left(\left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t \right\| \right), \end{aligned} \quad (\text{C.5})$$

and

$$\hat{\Sigma} = \left(N_e \mathbf{J}(\hat{\boldsymbol{\theta}})^T \Sigma_{\epsilon}^{-1} \mathbf{J}(\hat{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}) + \mathbf{H}(\hat{\boldsymbol{\theta}})^T \Sigma_{\epsilon}^{-1} \mathbf{E}_{\epsilon} \right)^{-1}. \quad (\text{C.6})$$

Furthermore, Long et al. [10] show that

$$\begin{aligned} \mathbf{J}(\hat{\boldsymbol{\theta}})^T \Sigma_{\epsilon}^{-1} \mathbf{E}_{\epsilon} &= \mathcal{O}_{\mathbb{P}}(\sqrt{N_e}), \quad \mathbf{H}(\hat{\boldsymbol{\theta}})^T \Sigma_{\epsilon}^{-1} \mathbf{E}_{\epsilon} = \mathcal{O}_{\mathbb{P}}(\sqrt{N_e}) \\ \text{and } N_e \mathbf{J}(\hat{\boldsymbol{\theta}})^T \Sigma_{\epsilon}^{-1} \mathbf{J}(\hat{\boldsymbol{\theta}}) &= \mathcal{O}_{\mathbb{P}}(N_e). \end{aligned}$$

Taking into account these probabilistic rates with respect to N_e , we state the approximations of the moments to complete the proof as

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_t - (N_e \mathbf{J}^T(\boldsymbol{\theta}_t) \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{J}(\boldsymbol{\theta}_t) + \mathbf{H}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{E}_\epsilon - \nabla \nabla h(\boldsymbol{\theta}_t))^{-1} \mathbf{J}^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{E}_\epsilon + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{N_e}\right) \quad (\text{C.7})$$

and

$$\hat{\boldsymbol{\Sigma}}^{-1} = N_e \mathbf{J}(\hat{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{J}(\hat{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}) + \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right). \quad (\text{C.8})$$

□

Appendix D. Expected information gain with Laplace method

The derivation of the expected information gain with the Laplace method.

Proof. We consider

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y}) = (2\pi)^{-\frac{d}{2}} |\hat{\boldsymbol{\Sigma}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_{\hat{\boldsymbol{\Sigma}}^{-1}}^2\right), \quad (\text{D.1})$$

which is the Gaussian approximation of the posterior pdf, for rewriting the Kullback-Leibler divergence,

$$D_{kl} = \int_{\Theta} \pi(\boldsymbol{\theta}|\mathbf{Y}) \log\left(\frac{\pi(\boldsymbol{\theta}|\mathbf{Y})}{\pi(\boldsymbol{\theta})}\right) d\boldsymbol{\theta}, \quad (\text{D.2})$$

into

$$D_{kl} = \int_{\Theta} \log\left(\frac{\pi(\boldsymbol{\theta}|\mathbf{Y})}{\pi(\boldsymbol{\theta})}\right) \tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} + \underbrace{\int_{\Theta} \log\left(\frac{\pi(\boldsymbol{\theta}|\mathbf{Y})}{\pi(\boldsymbol{\theta})}\right) (\pi(\boldsymbol{\theta}|\mathbf{Y}) - \tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})) d\boldsymbol{\theta}}_{\varepsilon_{int}} \quad (\text{D.3})$$

Then, by using the decomposition

$$\log\left(\frac{\pi(\boldsymbol{\theta}|\mathbf{Y})}{\pi(\boldsymbol{\theta})}\right) = \log\left(\frac{\pi(\boldsymbol{\theta}|\mathbf{Y})}{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})}\right) + \log\left(\frac{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})}{\pi(\boldsymbol{\theta})}\right), \quad (\text{D.4})$$

we can write

$$D_{kl} = \int_{\Theta} \left[\underbrace{\log\left(\frac{\pi(\boldsymbol{\theta}|\mathbf{Y})}{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})}\right)}_{\varepsilon_{la}} + \log\left(\frac{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})}{\pi(\boldsymbol{\theta})}\right) \right] \tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} + \varepsilon_{int}. \quad (\text{D.5})$$

The logarithmic error introduced by replacing the posterior density $\pi(\boldsymbol{\theta}|\mathbf{Y})$ by its Gaussian approximation $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})$ is here denoted by ε_{la} . Furthermore, we know that

$$\begin{aligned}\log\left(\frac{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})}{\pi(\boldsymbol{\theta})}\right) &= \log(\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y})) - \underbrace{\log(\pi(\boldsymbol{\theta}))}_{h(\boldsymbol{\theta})} \\ &= -\frac{1}{2}\log((2\pi)^d|\hat{\Sigma}|) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \hat{\Sigma}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) - h(\boldsymbol{\theta}),\end{aligned}\quad (\text{D.6})$$

which allows us to express D_{kl} as

$$D_{kl} = \int_{\Theta} \left[-\frac{1}{2}\log((2\pi)^d|\hat{\Sigma}|) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \hat{\Sigma}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) - h(\boldsymbol{\theta}) + \varepsilon_{la} \right] \tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} + \varepsilon_{int}.$$

In [10], it was shown that

$$\int_{\Theta} \varepsilon_{la} \tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{N_e^2}\right), \quad \varepsilon_{int} = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{N_e^2}\right),$$

and

$$\int_{\Theta} h(\boldsymbol{\theta}) \tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} = -h(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\text{tr}\left(\hat{\Sigma} : \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}})\right),$$

which leads us to

$$\begin{aligned}D_{kl} &= \int_{\Theta} \left[-\frac{1}{2}\log((2\pi)^d|\hat{\Sigma}|) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \hat{\Sigma}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) - h(\boldsymbol{\theta}) \right] \tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{N_e^2}\right), \\ &= -\frac{1}{2}\log((2\pi)^d|\hat{\Sigma}|) - \frac{d}{2} - h(\hat{\boldsymbol{\theta}}) - \frac{\text{tr}\left(\hat{\Sigma} : \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}})\right)}{2} + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{N_e^2}\right).\end{aligned}$$

Now, the expected information gain, $I = \mathbb{E}[D_{kl}]$, can be written as:

$$\begin{aligned}I &= \int_{\mathbf{Y}} D_{kl} p(\mathbf{Y}) d\mathbf{Y} \\ &= \int_{\Theta} \int_{\mathbf{Y}} D_{kl} p(\mathbf{Y}|\boldsymbol{\theta}_t) d\mathbf{Y} \pi(\boldsymbol{\theta}_t) d\boldsymbol{\theta}_t \\ &= \int_{\Theta} \int_{\mathbf{Y}} \left[-\frac{1}{2}\log((2\pi)^d|\hat{\Sigma}|) - \frac{d}{2} - h(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\text{tr}\left(\hat{\Sigma} : \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}})\right) \right] p(\mathbf{Y}|\boldsymbol{\theta}_t) d\mathbf{Y} \pi(\boldsymbol{\theta}_t) d\boldsymbol{\theta}_t + \mathcal{O}\left(\frac{1}{N_e^2}\right).\end{aligned}$$

Finally, to conclude the proof we use approximation (C.7) to obtain

$$I = \int_{\Theta} \left[-\frac{1}{2}\log((2\pi)^d|\hat{\Sigma}|) - \frac{d}{2} - h(\boldsymbol{\theta}_t) \right] p(\boldsymbol{\theta}_t) d\boldsymbol{\theta}_t + \mathcal{O}\left(\frac{1}{N_e}\right). \quad (\text{D.7})$$

□

Appendix E. Effect of the change of measure on the arithmetic underflow

We show the effect of the proposed importance sampling approach on the inner loop computation of DLMCIS with respect to the number of repetitive experiments, N_e . The derivation below follows closely the derivation in Appendix B.

Proof. By using the importance sampling based on the Laplace approximation, we obtain

$$L(\mathbf{Y}_n; \tilde{\boldsymbol{\theta}}_{n,m}) = \frac{p(\mathbf{Y}_n | \tilde{\boldsymbol{\theta}}_{n,m}) \pi(\tilde{\boldsymbol{\theta}}_{n,m})}{\tilde{\pi}_n(\tilde{\boldsymbol{\theta}}_{n,m})}, \quad (\text{E.1})$$

where

$$\begin{aligned} p(\mathbf{Y}_n | \tilde{\boldsymbol{\theta}}_{n,m}) &= (2\pi |\boldsymbol{\Sigma}_\epsilon|)^{-\frac{N_e}{2}} \exp \left(-\frac{1}{2} N_e \left\| \mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2 \right. \\ &\quad \left. - \frac{1}{2} \mathbf{v}_\epsilon^T \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m}) \right) \mathcal{O}_{\mathbb{P}} \left(\sqrt{N_e} \right) - \frac{1}{2} \sum_{j=1}^q \sigma_{\epsilon_j}^{-1} \left(N_e + \mathcal{O}_{\mathbb{P}} \left(\sqrt{N_e} \right) \right) \right). \end{aligned}$$

The distribution $\tilde{\pi}$ is the Gaussian distribution $\tilde{\pi}_n \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}_n))$:

$$\tilde{\pi}_n(\tilde{\boldsymbol{\theta}}_{n,m}) = \left(2\pi |\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}_n)| \right)^{-\frac{1}{2}} \exp \left(-\frac{1}{2} \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right)^T \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\theta}}_n) \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right) \right),$$

where the inverse of the covariance matrix is given by

$$\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\theta}}_n) = N_e \mathbf{J}(\hat{\boldsymbol{\theta}}_n)^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{J}(\hat{\boldsymbol{\theta}}_n) - \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}_n) + \mathcal{O}_{\mathbb{P}} \left(\sqrt{N_e} \right).$$

This allows us to write

$$\begin{aligned} \tilde{\pi}_n(\tilde{\boldsymbol{\theta}}_{n,m}) &= \left(2\pi |\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}_n)| \right)^{-\frac{1}{2}} \exp \left(-\frac{N_e}{2} \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right)^T \mathbf{J}(\hat{\boldsymbol{\theta}}_n)^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{J}(\hat{\boldsymbol{\theta}}_n) \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right) \right. \\ &\quad \left. + \frac{1}{2} \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right)^T \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}_n) \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right) + \frac{1}{2} \mathcal{O}_{\mathbb{P}} \left(\sqrt{N_e} \right) \right) \\ &= \left(2\pi |\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}_n)| \right)^{-\frac{1}{2}} \exp \left(-\frac{N_e}{2} \left(\mathbf{J}(\hat{\boldsymbol{\theta}}_n) \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right) \right)^T \boldsymbol{\Sigma}_\epsilon^{-1} \left(\mathbf{J}(\hat{\boldsymbol{\theta}}_n) \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right) \right) \right. \\ &\quad \left. + \frac{1}{2} \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right)^T \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}_n) \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right) + \mathcal{O}_{\mathbb{P}} \left(\sqrt{N_e} \right) \right) \\ &= \left(2\pi |\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}_n)| \right)^{-\frac{1}{2}} \exp \left(-\frac{N_e}{2} \left\| \mathbf{J}(\hat{\boldsymbol{\theta}}_n) \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right) \right\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2 \right. \\ &\quad \left. + \frac{1}{2} \left\| \tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n \right\|_{\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}_n)}^2 + \mathcal{O}_{\mathbb{P}} \left(\sqrt{N_e} \right) \right). \end{aligned}$$

Dividing $p(\mathbf{Y}_n|\tilde{\boldsymbol{\theta}}_{n,m})\pi(\tilde{\boldsymbol{\theta}}_{n,m})$ by $\tilde{\pi}_n(\tilde{\boldsymbol{\theta}}_{n,m})$ yields

$$\begin{aligned} L(\mathbf{Y}_n; \tilde{\boldsymbol{\theta}}_{n,m}) &= \frac{(2\pi|\boldsymbol{\Sigma}_\epsilon|)^{-\frac{N_e}{2}}}{\left(2\pi|\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}_n)|\right)^{-\frac{1}{2}}} \pi\left(\tilde{\boldsymbol{\theta}}_{n,m}\right) \exp\left(-\frac{N_e}{2} \left\|\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m})\right\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2\right. \\ &\quad \left. + \frac{N_e}{2} \left\|\mathbf{J}(\hat{\boldsymbol{\theta}}_n) \left(\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n\right)\right\|_{\boldsymbol{\Sigma}_\epsilon^{-1}}^2 - \frac{1}{2} \left\|\tilde{\boldsymbol{\theta}}_{n,m} - \hat{\boldsymbol{\theta}}_n\right\|_{\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}_n)}^2 + \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right)\right. \\ &\quad \left. - \frac{1}{2} \mathbf{v}_\epsilon^T \left(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\tilde{\boldsymbol{\theta}}_{n,m})\right) \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right) - \frac{1}{2} \sum_{j=1}^q \sigma_{\epsilon_j}^{-1} \left(N_e + \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right)\right)\right). \end{aligned}$$

Then we apply the logarithm to obtain the log-likelihood expression. □