Chapter 15

Bayesian Optimal Experimental Design

Chapter details

Chapter DOI:

https://doi.org/10.4322/978-65-86503-88-3.c15

Chapter suggested citation / reference style:

Carlon, Andre G., et al. (2022). “Bayesian Optimal Experimental Design”. *In* Jorge, Ariosto B., et al. (Eds.) *Uncertainty Modeling: Fundamental Concepts and Models*, Vol. III, UnB, Brasilia, DF, Brazil, pp. 479–501. Book series in Discrete Models, Inverse Methods, & Uncertainty Modeling in Structural Integrity.

P.S.: DOI may be included at the end of citation, for completeness.

Book details

Book: Uncertainty Modeling: Fundamental Concepts and Models

Edited by: Jorge, Ariosto B., Anflor, Carla T. M., Gomes, Guilherme F., & Carneiro, Sergio H. S.

Volume III of Book Series in:

Discrete Models, Inverse Methods, & Uncertainty Modeling in Structural Integrity

Published by: UnB City: Brasilia, DF, Brazil Year: 2022

DOI: https://doi.org/10.4322/978-65-86503-88-3

Bayesian Optimal Experimental Design

Andre G. Carlon[1], Rafael H. Lopez[2*], Leandro F. F. Miguel[2], André J. Torii[3]

[1]Computer, Electrical and Mathematical Sciences & Engineering Division, King Abdullah University of Science and Technology, Saudi Arabia. E-mail: agcarlon@gmail.com
[2]Center for Optimization and Reliability in Engineering (CORE), Department of Civil Engineering, Federal University of Santa Catarina, Federal University of Santa Catarina, Brazil. E-mail: rafael.holdorf@ufsc.br, leandro.miguel@ufsc.br
[3] Latin American Institute for Technology, Infrastructure and Territory (ILATIT), Federal University for Latin American Integration (UNILA), Avenida Tancredo Neves, 6731, Foz do Iguaçu/PR, Brazil, 85867-970. E-mail: andre.torii@unila.edu.br
*Corresponding author

Abstract

Experiments are of fundamental importance both in academic sciences and the industry. In many cases, performing experiments is an expensive task requiring resources and time. Therefore, it is essential to design experiments as best as possible to obtain the most information. However, in many cases, the optimal design of an experiment is not obvious, being even counter-intuitive in some situations. The goal of optimal experimental design is to formalize the statistical information obtained by experiments as an objective function to be maximized using optimization methods. Here, we focus on the Bayesian setting, using a discrepancy between the information before and after an experiment as its quality. Specifically, we use the Shannon's Expected Information Gain (SEIG), the mutual information between the experimental observations and the parameters of interest, to measure the quality of experiments.

Keywords: Optimal Experimental Design; Expected Information Gain; Uncertainty Quantification; Bayesian Inference; Monte Carlo Sampling

1 Introduction

Many fields of science heavily rely on information obtained by experiments. Statistics about quantities of interest are inferred from data resulting from experiment observations. Hence, it is important that experiments provide informative data. For example, in structural engineering, it is useful to have statistical information about the material properties, such as the Young modulus, Poisson modulus, yield stress, so that the engineer can take informative decisions. However, to obtain statistically relevant data, experiments must have large enough samples. For example, if a three-point flexural test is performed on concrete beams to evaluate the fracture toughness of concrete, a significant number of beams must be built beforehand and let to cure for 28 days. Even

if a hundred beams are built, cured, and tested, the standard error in the fracture toughness estimate is still the standard deviation of one of the samples divided by ten. To reduce the standard error of the mean by one digit requires increasing the sample size by a hundred times. A three-point bending experiment for fracture toughness determination and its design parameters are presented in Figure 1. Properly defining design parameters can reduce dispersion of the quantity of interest

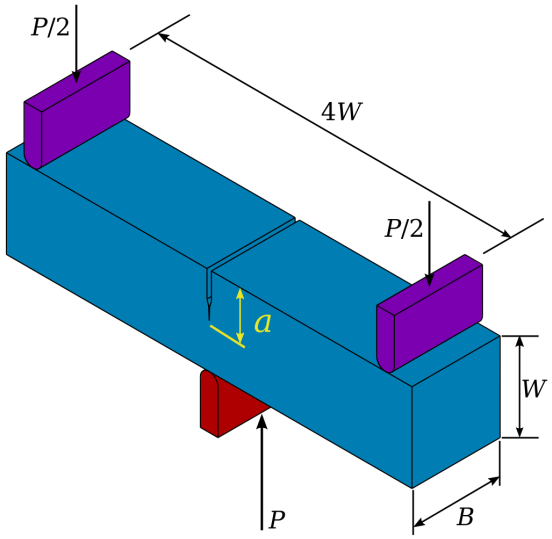


Figure 1: Example of fracture toughness testing. Source: Wikimedia commons Commons [2011]

estimation or reduce the cost required to achieve the same precision.

Another example of practical interest is the one of verifying the quality of composite laminates. Some composite materials can have improved mechanical properties in some directions by alternating plies of orthotropic materials with some specific angles between them. One way of testing these laminates is by electrical impedance tomography (EIT). However, the information obtained depends heavily on the currents imposed on the electrodes, as can be seen in the results obtained by Beck et al. [2018]. The set-up of an EIT experiment is presented in Figure 2, where the electrodes are illustrated in black, the first ply in blue and the second ply in red. We optimize the currents in an EIT experiment in Section **??**.

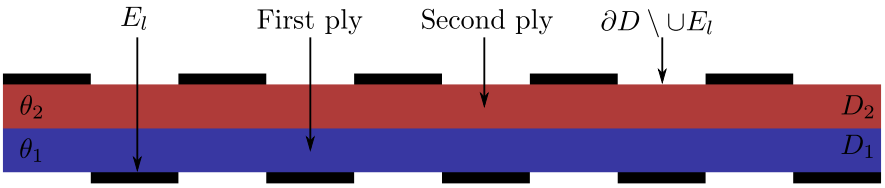


Figure 2: Electrical impedance tomography example. Source: adapted from Beck et al. [2018]

Given the difficulties and the costs involved in the experimental process, it might be interesting to optimally tune experiments parameters to maximize their efficiency, which is the main concern of optimal experimental design (OED). Thus, in OED, we seek the optimum design parameters that maximize some measure of efficiency of an experiment.

The classical optimality criteria for experiments are known as the alphabetic optimality criteria. However, they require the model to be linear with respect to the random parameters. The al-

phabetic optimality criteria use the Fisher information matrix, a measure of information inversely proportional to the covariance matrix, to qualify an experiment. The better known of these criteria, D-optimality, consists in minimizing the determinant of the Fisher information matrix, as presented in Chaloner and Verdinelli [1995].

In this chapter, we aim at presenting an OED framework that can optimize experiments with nonlinear black-box forward models. Thus, we opt to measure the performance of an experiment by its Shannon's expected information gain (EIG) [Shannon, 1948]. The EIG is related to the relative entropy of information and, in the OED context, is a measure on how much information an experiment provides [Huan and Marzouk, 2013]. Lindley [1956] was the first to use EIG as an utility function for OED. However, for a general experiment with nonlinear model, estimating EIG is cumbersome.

Ryan [2003] develop an EIG estimator based on Monte Carlo integration (MCI), however, his estimator requires the evaluation of two nested integrals. We refer to Ryan's estimator as the double-loop Monte Carlo (DLMC). The DLMC estimator has the disadvantages of being expensive to evaluate and numerically unstable. The evaluation of the two nested MCI requires a large number of experiment simulations.Indeed, if the two outer and inner MCI have sample size of, respectively, N and M, the DLMC estimator requires $N(M+1)$ forward model evaluations to approximate the EIG. Moreover, for some cases, M needs to be large to avoid numerical underflow [Beck et al., 2018]. Long et al. [2013] propose a method to estimate the EIG using a Laplace approximation, furnishing the Monte Carlo with Laplace approximation (MCLA) estimator. The MCLA estimator does not require the evaluation of one of the two nested integrals, thus, being less expensive. The disadvantage of MCLA is that the Laplace approximation introduces a bias that might not be acceptable depending on the situation. Beck et al. [2018] propose an importance sampling for DLMC that uses Laplace approximation to draw more informative samples, reducing the cost of DLMC without adding the bias of MCLA. The DLMC with the importance sampling is called double-loop Monte Carlo with importance sampling (DLMCIS).

The main goal of this chapter is to present the DLMC, MCLA and DLMCIS estimators for the SEIG in a didactic manner. To achieve this goal, we first present the basis of Bayesian inference and how to measure information using the SEIG. Then, the procedure for the numerical evaluation of the DLMC, MCLA and DLMCIS estimators is presented, together with their the pseudocodes.

2 Optimal Experimental Design (OED)

The goal of OED is to find the best experimental setup [Chaloner and Verdinelli, 1995]. From the perspective of optimization, it is important to define a criterion to measure the efficiency of experiments: the objective function to be maximized. Here, we use the Shannon's SEIG [Shannon, 1948] of an experiment to evaluate its performance. Thus, on the next section, we introduce SEIG and other concepts related to it.

2.1 Experiment model

We model the evaluation of N_e repetitive experiments as

$$\boldsymbol{y}_i(\boldsymbol{\xi}, \boldsymbol{\theta}_t, \boldsymbol{\epsilon}_i) = \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}_t) + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, N_e, \tag{1}$$

where $\boldsymbol{y}_i \in \mathbb{R}^q$ is the vector of experiment observations, $\boldsymbol{\theta}_t \in \mathbb{R}^d$ is the vector of parameters of interest to be recovered, $\boldsymbol{\xi}$ is the vector with experiment parameters to be optimized, $\boldsymbol{g}$ is the experiment model, and $\boldsymbol{\epsilon}$ is the additive noise from measurements. For N_e experiments performed with the same setup $\boldsymbol{\xi}$, the set of observed data is gathered in a set $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^{N_e}$. Since we cannot observe $\boldsymbol{\theta}_t$ directly, we use the random variable $\boldsymbol{\theta}$: $\Theta \to \mathbb{R}^m$ with prior distribution $\pi(\boldsymbol{\theta})$ instead

of $\boldsymbol{\theta}_t$. Thus, through observations $\boldsymbol{Y}$, we calculate statistics about $\boldsymbol{\theta}$. Here, our goal is to find the optimal experimental design $\boldsymbol{\xi} \in \Xi \subset \mathbb{R}^n$ that provides more information about $\boldsymbol{\theta}$, where Ξ is the space of experiment designs. Moreover, we consider $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ to be an additive noise, i.e. independent of $\boldsymbol{g}$, $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$, for some positive-definite and symmetric covariance matrix $\boldsymbol{\Sigma}_\epsilon$. This model for experiments is the same used in [Long et al., 2013, Beck et al., 2018].

For example, consider the case of a single three-point test where the stiffness modulus (E) of some material is estimated, as illustrated in Figure 3. In this case, the parameter of interest (θ) is

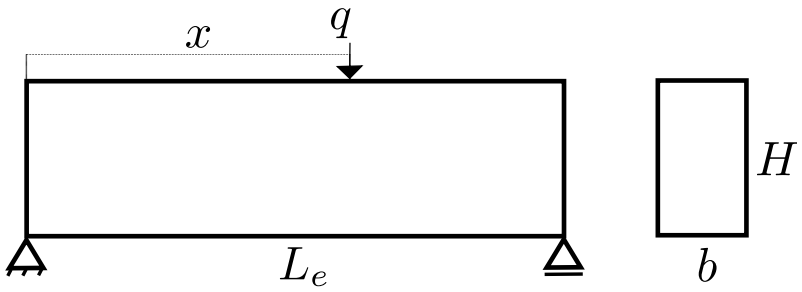


Figure 3: Example of the setup for a three-point flexural test.

E, the measurement y is the observed deflection in mm, and ϵ is the error in the observation of y. The design parameter of the experiment to be optimized is $\xi = x$, the position where the point load q is applied, and where the deflection is measured. The model relating E and x to the observation y is $g(x, E) = \frac{qx}{6L_e E I_e}(2L_e - x)$, where I_e is the moment of inertia of the cross-section of the beam. This simple problem is used to illustrate the concepts introduced in the present chapter; the proof that the optimum is at $x^* = L_e/2$ is trivial.

2.2 Bayesian Inference

To evaluate the quality of experiments, we use a Bayesian framework of analysis. Thus, in this section, we introduce essential concepts of Bayesian inference. The main idea behind Bayesian inference is to use new data to update some previously known probability distribution of a parameter. This update is done using Bayes' theorem:

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{\xi}) = \frac{p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\xi})\pi(\boldsymbol{\theta})}{p(\boldsymbol{Y}|\boldsymbol{\xi})}, \tag{2}$$

where $\pi(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{\xi})$ is the posterior distribution, the updated probability density function (pdf) of some random variable $\boldsymbol{\theta}$ given observations $\boldsymbol{Y}$; $\pi(\boldsymbol{\theta})$ is the prior distribution, the pdf of $\boldsymbol{\theta}$ before the experiment; $p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\xi})$ is the likelihood, the probability of observing $\boldsymbol{Y}$ given the previously known prior $\pi(\boldsymbol{\theta})$; and $p(\boldsymbol{Y}|\boldsymbol{\xi})$ is the evidence, the probability of $\boldsymbol{Y}$ being observed.

On the context of experiments, Bayesian inference is used to estimate a posterior pdf of a parameter of interest given a prior pdf and some observations $\boldsymbol{Y}$ provided by the experiment. For example, consider the previously mentioned three-point flexural test; in Figure 4 we present the pdf for the stiffness both before and after the flexural experiment, i.e., the prior and posterior pdfs of (2). It can be seen that the dispersion is reduced after the experiment, meaning that the experiment provided useful information about the parameters of interest.

In this example, we modeled the prior knowledge about the parameter of interest as being Gaussian-distributed; however, other distributions can be used as well. Using the Bayes' theorem in (2), the posterior can be obtained by multiplying the prior by the division between the likelihood and the evidence. For the experiment model with additive noise in (1), the likelihood of observing

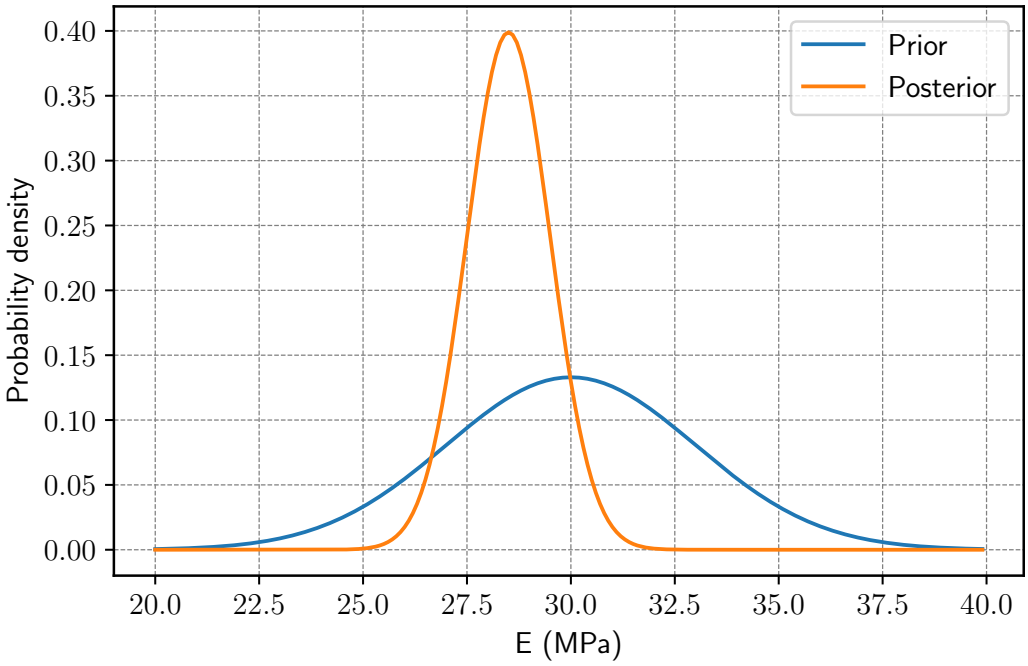


Figure 4: Prior and posterior pdfs for the three-point flexural experiment.

$\boldsymbol{Y}$ given $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ is a multivariate Gaussian with mean $\boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta})$ and covariance matrix $\boldsymbol{\Sigma_\epsilon}$,

$$p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\xi}) = \det\left(2\pi\boldsymbol{\Sigma_\epsilon}\right)^{-\frac{N_e}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_e} \|\boldsymbol{y}_i - \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta})\|^2_{\boldsymbol{\Sigma_\epsilon}^{-1}}\right), \tag{3}$$

where $\boldsymbol{\theta}$ is not necessarily the same used to calculate $\boldsymbol{Y}$. Using (3), the likelihood in (2) can be evaluated, yet, the evidence $p(\boldsymbol{Y}|\boldsymbol{\xi})$ is not known. To estimate the evidence, we follow the same procedure as Ryan [2003]. We introduce a new variable $\boldsymbol{\theta}^*$ that is independent and identically distributed with respect to $\boldsymbol{\theta}$ and marginalize its likelihood with respect to $\boldsymbol{\theta}^*$ as

$$p(\boldsymbol{Y}|\boldsymbol{\xi}) = \int_\Theta p(\boldsymbol{Y}|\boldsymbol{\theta}^*, \boldsymbol{\xi})\pi(\boldsymbol{\theta}^*)\mathrm{d}\boldsymbol{\theta^*}. \tag{4}$$

The likelihood in (4) can be calculated from (3) as

$$p(\boldsymbol{Y}|\boldsymbol{\theta}^*, \boldsymbol{\xi}) = \det\left(2\pi\boldsymbol{\Sigma_\epsilon}\right)^{-\frac{N_e}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_e} \|\boldsymbol{y}_i - \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}^*)\|^2_{\boldsymbol{\Sigma_\epsilon}^{-1}}\right). \tag{5}$$

Thus, substituting in the Bayes' equation in (2) the likelihood and the evidence respectively presented in (3) and (4), one can obtain the posterior distribution of the experiment, i.e., the probability distribution for the parameter of interest after the experiment.

To estimate how much information an experiment provides, we use the Kullback–Leibler divergence (D_{KL}) between the prior and posterior pdfs.

2.3 Kullback–Leibler Divergence

According to Cover and Thomas [2012], entropy is a measure of uncertainty of a random variable. Cover and Thomas define the differential entropy of a random variable $\boldsymbol{\theta}$ with pdf $f_{\boldsymbol{\theta}}$ as

$$-\int_\Theta f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log f_{\boldsymbol{\theta}}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{6}$$

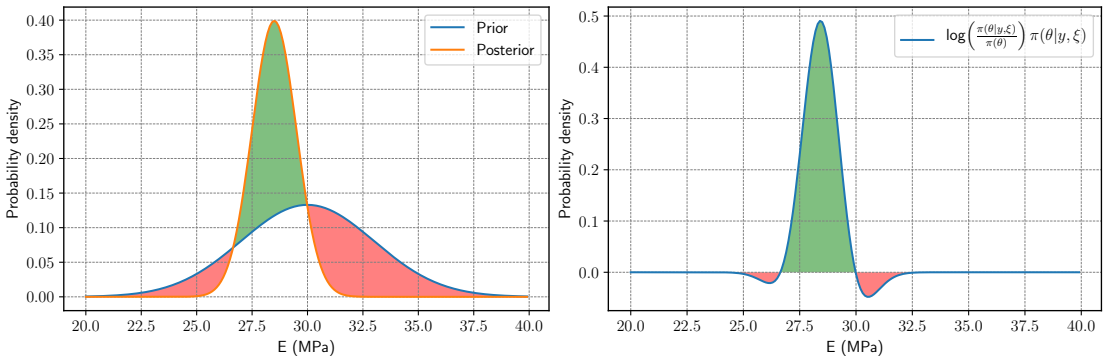
The larger the entropy of $\pi(\boldsymbol{\theta})$ is, the larger its uncertainty with respect to $\boldsymbol{\theta}$ is. The D_{KL} is the entropy of a probability measure with respect to another [Kullback and Leibler, 1951], and, for two probability measures $f_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\theta}}$ on $\boldsymbol{\theta}$ with the same support Θ, the D_{KL} is defined as

$$\begin{aligned} D_{\mathrm{KL}}(f_{\boldsymbol{\theta}}(\boldsymbol{\theta})\|g_{\boldsymbol{\theta}}(\boldsymbol{\theta})) &= -\int_{\Theta} f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log g_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} + \int_{\Theta} f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\ &= -\int_{\Theta} f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \left(\frac{g_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{f_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\Theta} \log \left(\frac{f_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{g_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{7}$$

The larger the D_{KL} of $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ with respect to $g_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is, the more different is $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ with respect to $g_{\boldsymbol{\theta}}(\boldsymbol{\theta})$; the D_{KL} between two distributions is a measure of the discrepancy between them. As a way of measuring the efficiency of an experiment, we use the D_{KL} of the posterior with respect of the prior:

$$D_{\mathrm{KL}}(\pi(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\xi})\|\pi(\boldsymbol{\theta})) = \int_{\Theta} \log \left(\frac{\pi(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\xi})}{\pi(\boldsymbol{\theta})} \right) \pi(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\theta}. \tag{8}$$

For the sake of simplicity, the Kullback–Leibler divergence between the prior and posterior pdfs is referred simply as Kullback–Leibler divergence and denoted as D_{KL}. In Figure 5, at the left, we present the prior and posterior pdf of the three-point flexural test. The regions where the probability density of the posterior distribution is greater than the one of the prior are shaded in green, whereas the regions where the posterior pdf is less than the prior pdf are shaded in red. In Figure 5, at the right, we present the integrand in (8) over the domain of θ for the three-point flexural test. The regions where the integrand is positive are shaded in green, and the regions where the integrand is negative are shaded in red. The more informative an experiment is, the

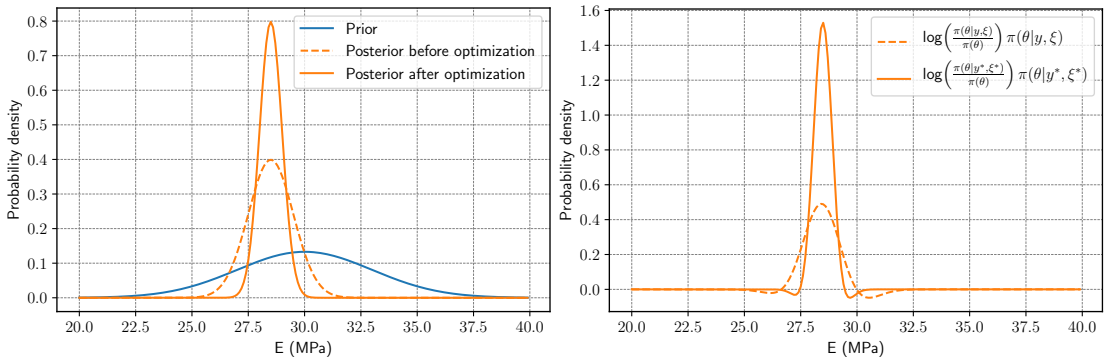


(a) **(b)**

Figure 5: For the three-point flexural test: prior and posterior distributions (a), and the integrand of the D_{KL} (b).

larger is the integral of the function in Figure 5 (b). Thus, we aim to find the $\boldsymbol{\xi}^*$ that provides the more informative observations $\boldsymbol{Y}^*$; the $\boldsymbol{Y}^*$ that maximize difference between the green area and the red area in Figure 5 (b).

In Figure 6, we present a comparison between the optimized and non-optimized cases for the three-point flexural test. The non-optimized case is evaluated with $x = L_e/4$, whereas, in the optimized case, the load and the measurement are in the middle of the beam, i.e., $x = x^* = L_e/2$. In the left plot of Figure 6, it can be observed that the posterior pdf for the optimized configuration is more concentrated than the posterior pdf before before optimization. In the right plot of Figure 6, the integrand of the D_{KL} is presented for both the optimized and non-optimized configurations. It



(a) **(b)**

Figure 6: The pdfs for prior, posterior, and optimized posterior (a) and the D_{KL} integrand for the non-optimized and for the optimized cases (b).

can be observed that the D_{KL} for the optimized case is greater than the D_{KL} for the non-optimized configuration, as can be inferred from the areas under the integrands.

When modeling a real experiment, one might need to consider the noise inherent to experiment observations. To estimate the information gain considering the noise in observations, we use the SEIG [Shannon, 1948].

2.4 Shannon's Expected Information Gain

The D_{KL} does not consider the noise in observations due to measurement uncertainties. To estimate the information gain on this context, we need to marginalize the D_{KL} with respect to $\boldsymbol{Y}$, thus, obtaining the Shannon's expected information gain [Shannon, 1948] as

$$I = \int_{\mathcal{Y}} \int_{\Theta} \log\left(\frac{\pi(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\xi})}{\pi(\boldsymbol{\theta})}\right) \pi(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\theta} p(\boldsymbol{Y}|\boldsymbol{\xi}) \mathrm{d}\boldsymbol{Y}. \tag{9}$$

Lindley [1956] is the first to use SEIG as an utility function for optimal experimental design. An example of the integrand in (9) with respect to both the parameter of interest and the observations for the three-point flexural test is presented in Figure 7. Estimating I requires the solution of a double integral over the sample space of the parameters of interest (Θ), and the observations ($\mathcal{Y}$).

Using the Bayes' equation presented in (2), we can rewrite (9) as

$$I = \int_{\Theta} \int_{\mathcal{Y}} \log\left(\frac{p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\xi})}{p(\boldsymbol{Y}|\boldsymbol{\xi})}\right) p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\xi}) \mathrm{d}\boldsymbol{Y} \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \tag{10}$$

Presenting SEIG as in (10) has the advantage that, for our experiment model, the likelihood can be calculated directly using (3). Moreover, using Eq.4 to estimate the evidence results in

$$I = \int_{\Theta} \int_{\mathcal{Y}} \log\left(\frac{p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\xi})}{\displaystyle\int_{\Theta} p(\boldsymbol{Y}|\boldsymbol{\theta}^*,\boldsymbol{\xi})\pi(\boldsymbol{\theta}^*)\mathrm{d}\boldsymbol{\theta}^*}\right) p(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\xi}) \mathrm{d}\boldsymbol{Y} \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \tag{11}$$

Here, we use (11) to calculate the SEIG.

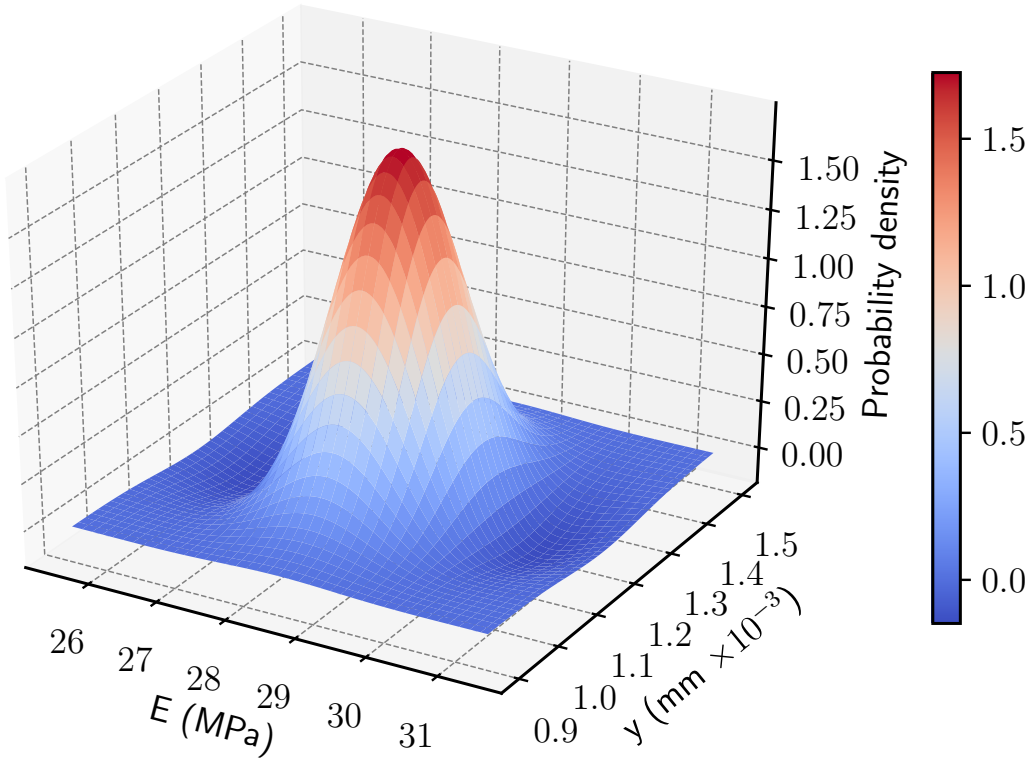


Figure 7: The SEIG integrand.

3 Shannon's Expected Information Gain Estimators

Evaluating I in (11) requires the solution of the double integral over $\boldsymbol{\theta}$ and $\boldsymbol{Y}$, moreover, it requires the solution of the integral used to marginalize the evidence. In most cases, these integrals do not have closed form, thus, numerical methods are needed to approximate them. When the number of random parameters is small, quadrature methods can be used to approximate the integrals in (11), however, as pointed out by Robert and Casella [2013], these methods suffer from the curse of dimensionality. Hamada et al. [2001] note that using deterministic integration methods to approximate (11) becomes unfeasible if the dimensionality of $\boldsymbol{\theta}$ exceeds three.

3.1 Error analysis

We focus on the case where the integrals in (11) do not have closed form solution and Monte Carlo integration (MCI) is needed to approximate them. In this section, we introduce Monte Carlo-based estimators for SEIG that can deal with high dimensional parameter spaces. We present the error analysis and complexity of the SEIG estimators with respect to the evaluation of $\boldsymbol{g}$ by FEM with a mesh discretization parameter h. As $h \to 0$, the discretization bias asymptotically converges as

$$\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{\xi},\boldsymbol{\theta}) - \boldsymbol{g}_h(\boldsymbol{\xi},\boldsymbol{\theta})\|_2\right] = \mathcal{O}\left(h^{\eta}\right), \tag{12}$$

where $\boldsymbol{g}_h$ is the FEM approximation of $\boldsymbol{g}$ using h, and $\eta > 0$ is the rate of convergence of the discretization error. For the computational effort analysis, we assume that the cost of evaluating $\boldsymbol{g}_h$ is $\mathcal{O}(h^{-\varrho})$, for the constant $\varrho > 0$. Both η and ϱ are constants that depend only on the numerical approach used in approximating $\boldsymbol{g}$ by $\boldsymbol{g}_h$.

In the present case, the error of estimators can be decomposed in two terms: their bias and variances. The variance of the estimator is a measure of the dispersion from different and independent

estimations of the same quantity, e.g., for an estimator $\mathcal{I}$, its variance is $\mathbb{V}[\mathcal{I}] \stackrel{\text{def}}{=} \mathbb{E}\left[(\mathcal{I} - \mathbb{E}[\mathcal{I}])^2\right]$. The bias of an estimator is the difference between its expected value and the true value that is being estimated, e.g., for an estimator $\mathcal{I}$ of a quantity I, the bias of $\mathcal{I}$ is $|I - \mathbb{E}[\mathcal{I}]|$. Figure 8 illustrates the bias and variance of an hypothetical estimator of I, where the curve in blue is the probability density that a value of $\mathcal{I}$ is estimated. Moreover, the distance between the expected value of $\mathcal{I}$ and I, i.e., the bias, is presented, as is the square root of the variance, σ, the standard error of the estimator. The bias of the estimator includes the bias from numerical approximation of the model.

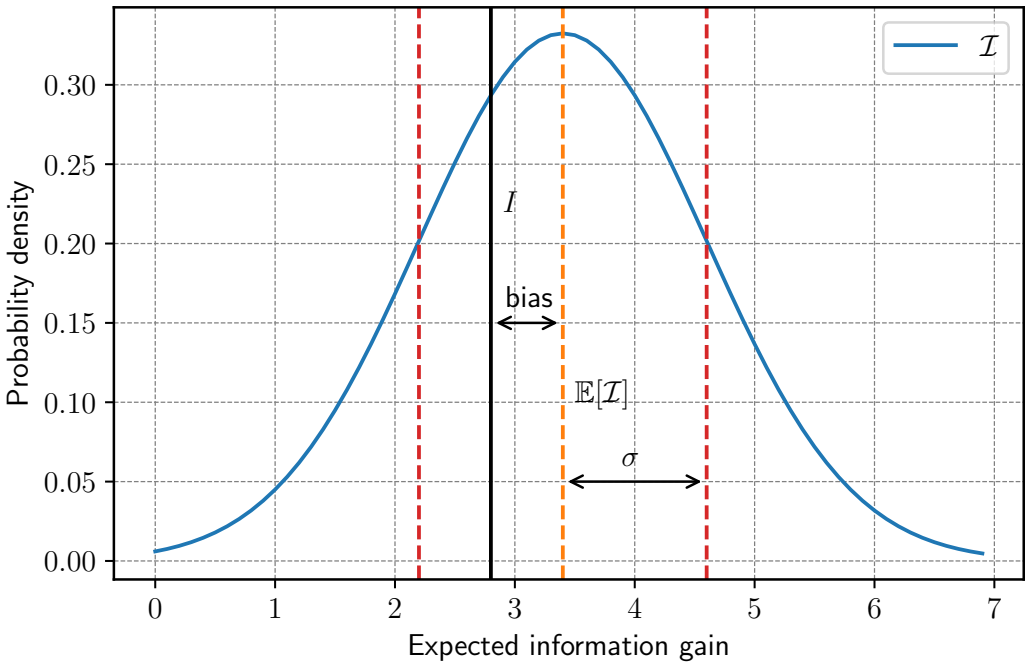


Figure 8: Bias and variance of an estimator

For further information about the bias and variance of each SEIG estimator, the reader is referred to Beck et al. [2018].

3.2 Monte Carlo integration

The MCI is a method for approximating integrals by sampling the integrand and averaging the samples. For example, let $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^n$, and $f : \boldsymbol{\theta} \to \mathbb{R}$ with $\boldsymbol{\theta}$ with probability $\pi(\boldsymbol{\theta})$. In this case, the expectation of $f(\boldsymbol{\theta})$ can be approximated as

$$\mathbb{E}[f(\boldsymbol{\theta})] = \int_{\Theta} f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \approx \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}_i), \tag{13}$$

where $\boldsymbol{\theta}$ is sampled from $\pi(\boldsymbol{\theta})$. From the strong law of large numbers, the Monte Carlo estimator converges to the real value of the integral as $N \to \infty$ [Rubinstein and Kroese, 2016]. Moreover, from the central limit theorem, as $N \to \infty$, the error in approximating an integral by MCI converges to zero with rate $1/\sqrt{N}$ [Rubinstein and Kroese, 2016].

3.3 Double-loop Monte Carlo estimator

The double integral in (10) can be approximated using MCI by sampling $\boldsymbol{\theta}$ from the prior pdf and $\boldsymbol{Y}$ from the likelihood (the measures that the integrands are being integrated over in (10)). Then,

for N samples, the Monte Carlo estimator for OED is defined as

$$\mathcal{I}_{MC}(\boldsymbol{\xi}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} \log\left(\frac{p(\boldsymbol{Y}_n|\boldsymbol{\theta}_n,\boldsymbol{\xi})}{p(\boldsymbol{Y}_n|\boldsymbol{\xi})}\right). \tag{14}$$

Also, we use (4) and approximate the evidence integral by a MCI as

$$\int_{\Theta} p(\boldsymbol{Y}|\boldsymbol{\theta}^*,\boldsymbol{\xi})\pi(\boldsymbol{\theta}^*)\mathrm{d}\boldsymbol{\theta}^* \approx \frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{Y}|\boldsymbol{\theta}^*_m,\boldsymbol{\xi}). \tag{15}$$

Thus, the DLMC estimator for SEIG is defined as

$$\mathcal{I}_{DLMC}(\boldsymbol{\xi}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} \log\left(\frac{p(\boldsymbol{Y}_n|\boldsymbol{\theta}_n,\boldsymbol{\xi})}{\frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{Y}_n|\boldsymbol{\theta}^*_m,\boldsymbol{\xi})}\right). \tag{16}$$

The first to use DLMC in OED is Ryan [2003].

Monte Carlo estimators are generally not biased, however, the DLMC estimator has a bias resulting from the logarithm of the inner loop. However, DLMC is a consistent estimator because the bias goes to zero as the number of inner loop samples M goes to infinity. The DLMC estimator has bias and variance respectively given by

$$|I - \mathbb{E}[\mathcal{I}_{\text{DLMC}}]| \leq C_{DL,1}h^{\eta} + \frac{C_{DL,2}}{M} + o(h^{\eta}) + \mathcal{O}\left(\frac{1}{M^2}\right), \quad \text{and} \tag{17}$$

$$\mathbb{V}[\mathcal{I}_{\text{DLMC}}] = \frac{C_{DL,3}}{N} + \frac{C_{DL,4}}{NM} + \mathcal{O}\left(\frac{1}{NM^2}\right) \tag{18}$$

for the constants $C_{DL,1}$, $C_{DL,2}$, $C_{DL,3}$, and $C_{DL,4}$ (cf. [Beck et al., 2018]).

To evaluate the DLMC estimator in (16), one needs the likelihood of observing $\boldsymbol{Y}_n$ given $\boldsymbol{\theta}_n$ and the likelihoods of observing $\boldsymbol{Y}_n$ given each $\boldsymbol{\theta}^*_m$. Using (3) to estimate the likelihood of observing $\boldsymbol{Y}_n$ given $\boldsymbol{\theta}_n$ furnishes

$$p(\boldsymbol{Y}_n|\boldsymbol{\theta}_n,\boldsymbol{\xi}) = \det\left(2\pi\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\right)^{-\frac{N_e}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_e}\left\|\boldsymbol{y}_i^{(n)}(\boldsymbol{\xi}) - \boldsymbol{g}(\boldsymbol{\xi},\boldsymbol{\theta}_n)\right\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}{}^{-1}}\right), \tag{19}$$

$$= \det\left(2\pi\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\right)^{-\frac{N_e}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_e}\left\|\boldsymbol{g}(\boldsymbol{\xi},\boldsymbol{\theta}_n) + \boldsymbol{\epsilon}_i - \boldsymbol{g}(\boldsymbol{\xi},\boldsymbol{\theta}_n)\right\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}{}^{-1}}\right), \tag{20}$$

$$= \det\left(2\pi\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\right)^{-\frac{N_e}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_e}\left\|\boldsymbol{\epsilon}_i\right\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}{}^{-1}}\right). \tag{21}$$

Since $\boldsymbol{Y}$ and $\boldsymbol{g}$ are evaluated using the same $\boldsymbol{\theta}$, the model evaluation $\boldsymbol{g}$ in (20) cancels out. Thus, evaluating the likelihood of observing $\boldsymbol{Y}_n$ given $\boldsymbol{\theta}_n$ does not require any model evaluation. However, for the evidence evaluation, we must evaluate the model g. For example, consider the likelihood in (3), where $\boldsymbol{Y}_n$ is evaluated from $\boldsymbol{\theta}_n$ and ϵ (sampled in the outer loop), and $\boldsymbol{\theta}^*_m$ is sampled in the inner loop (independently from $\boldsymbol{\theta}_n$),

$$p(\boldsymbol{Y}_n|\boldsymbol{\theta}^*_m,\boldsymbol{\xi}) = \det\left(2\pi\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\right)^{-\frac{N_e}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_e}\left\|\boldsymbol{y}_i^{(n)}(\boldsymbol{\xi}) - \boldsymbol{g}(\boldsymbol{\xi},\boldsymbol{\theta}^*_m)\right\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}{}^{-1}}\right). \tag{22}$$

Each observation of $p(\boldsymbol{Y}_n|\boldsymbol{\theta}^*_m,\boldsymbol{\xi})$ requires an evaluation of $\boldsymbol{g}(\boldsymbol{\xi},\boldsymbol{\theta}^*_m)$.

Algorithm 1 Pseudocode for the DLMC estimator for SEIG.

function DLMC($\boldsymbol{\xi}, N, M$)
  for $n = 1, 2, ..., N$ **do** ▷ Outer loop
    Sample $\boldsymbol{\theta}_n$ from $\pi(\boldsymbol{\theta})$
    Evaluate $\boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}_n)$
    for $i = 1, 2, .., .N_e$ **do**
      Sample $\boldsymbol{\epsilon}_i$ from $\mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$
      $\boldsymbol{y}_i \leftarrow \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}_n) + \boldsymbol{\epsilon}_i$
    end for
    $\boldsymbol{Y}_n \leftarrow \{\boldsymbol{y}_i\}_{i=1}^{N_e}$
    $p(\boldsymbol{Y}_n|\boldsymbol{\theta}_n, \boldsymbol{\xi}) \leftarrow \det(2\pi\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-\frac{N_e}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_e} \|\boldsymbol{\epsilon}_i\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}{}^{-1}}\right)$
    for $m = 1, 2, ..., M$ **do** ▷ Inner loop
      Sample $\boldsymbol{\theta}^*_m$ from $\pi(\boldsymbol{\theta})$
      Evaluate $\boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}^*_m)$
      $p(\boldsymbol{Y}_n|\boldsymbol{\theta}^*_m, \boldsymbol{\xi}) \leftarrow \det(2\pi\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-\frac{N_e}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_e} \left\|\boldsymbol{y}_i^{(n)}(\boldsymbol{\xi}) - \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}^*_m)\right\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}{}^{-1}}\right)$
    end for
    $p(\boldsymbol{Y}_n|\boldsymbol{\xi}) \leftarrow \frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{Y}_n|\boldsymbol{\theta}^*_m, \boldsymbol{\xi})$
  end for
  $\mathcal{I}_{DLMC}(\boldsymbol{\xi}) \leftarrow \frac{1}{N}\sum_{n=1}^{N} \log\left(\frac{p(\boldsymbol{Y}_n|\boldsymbol{\theta}_n, \boldsymbol{\xi})}{p(\boldsymbol{Y}_n|\boldsymbol{\xi})}\right)$
  return $\mathcal{I}_{DLMC}(\boldsymbol{\xi})$
end function

In Algorithm 1, we present the pseudocode for DLMC, where the inputs are the experiment setup, $\boldsymbol{\xi}$, the sample-size for the outer loop, N, and the sample-size for the inner loop, M. Problem parameters, e.g., $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, $\pi(\boldsymbol{\theta})$, $\boldsymbol{g}$, N_e, are considered to be known. The DLMC returns the estimation of I, $\mathcal{I}_{DLMC}$.

From Algorithm 1, it can be seen that the cost of evaluating the DLMC estimator is $N(M+1)$ forward model evaluations. Considering that each model evaluation requires the solution of a PDE using FEM with a mesh of size h, the computational cost of evaluating the DLMC estimator is of order $N(M+1)h^{-\varrho}$.

The DLMC estimator can suffer from numerical instabilities, namely, *numerical underflow* [Beck et al., 2018]. The marginalization of the likelihood of observing $\boldsymbol{Y}$ (evaluated using $\boldsymbol{\theta}^*$ sampled in the outer loop) given $\boldsymbol{\theta}$ sampled in the inner loop can result in numerical underflow. If the likelihood is zero for all the M inner samples, the evidence also becomes zero. If $\boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}^*_m)$ is too distant to each $\boldsymbol{y}^{(n)}$, or if $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ has small eigenvalues, the likelihood can get smaller than floating-point precision. Thus, to avoid numerical underflow, M needs to be large enough as to guarantee that at least one of the M likelihoods is not numerically evaluated to zero. If the evidence is evaluated to zero, then (14) cannot be evaluated. Figure 9 illustrates the numerical underflow for the flexural test for five inner loop samples; for each loop the model is evaluated and the likelihood of y being observed is drawn in red. It can be observed that, for all g evaluated at each inner loop, the likelihood of observing the $y^{(n)}$ evaluated at the outer loop is near to zero. In the next sections we present two estimators of SEIG that are able to overcome the high cost and numerical instabilities of DLMC.

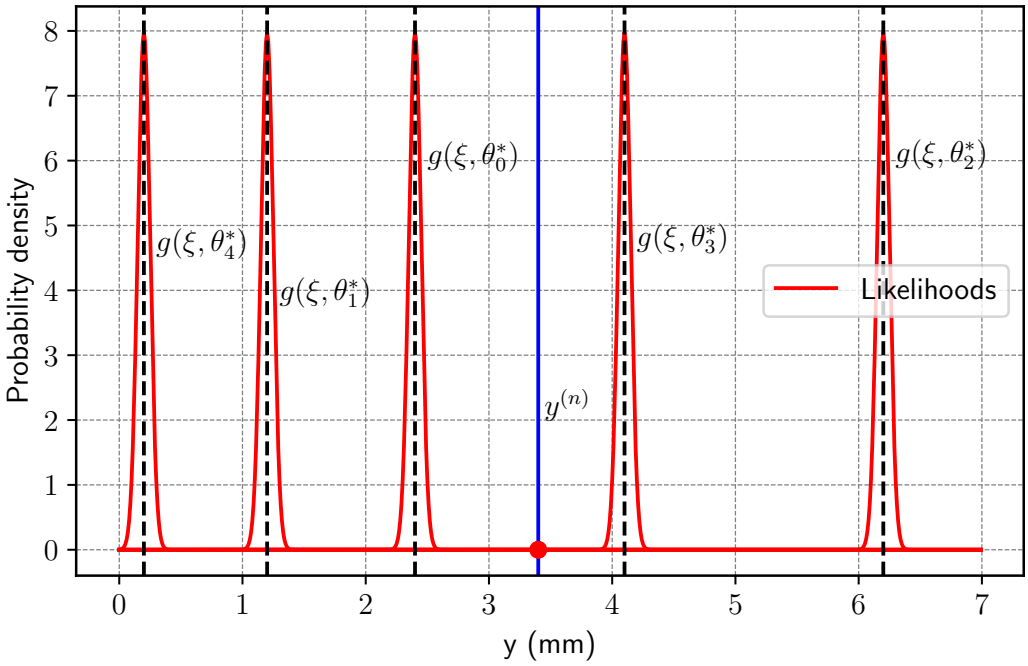


Figure 9: Numerical underflow illustrated for the three-point flexural test.

3.4 Monte Carlo estimator with Laplace approximation

The cost of solving the two-loop Monte Carlo required for DLMC can be large. Since MCI is a method for approximating integrals, one might think of alternative methods to approximate one of the integrals. Long et al. [2013] propose the use of the Laplace method to approximate the the logarithm of the posterior distribution by a second-order Taylor expansion, thus, avoiding the evaluation of the evidence. We follow the same approach as Long et al. [2013] to devise the Monte Carlo with Laplace method (MCLA) estimator. One advantage of the MCLA is that the approximated posterior pdf is a Gaussian function

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{\xi}) \approx \det(2\pi\boldsymbol{\Sigma}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}))^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\boldsymbol{\xi}, \boldsymbol{Y})\|^2_{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}})}\right), \tag{23}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum a posteriori (MAP) of $\boldsymbol{\theta}$, and $\boldsymbol{\Sigma}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}})$ is the covariance matrix of the posterior at the MAP. The MAP is the $\boldsymbol{\theta}$ that maximizes the posterior pdf, i.e., the $\boldsymbol{\theta}$ more likely to be $\boldsymbol{\theta}_t$ after the experiment data is considered. Thus, $\hat{\boldsymbol{\theta}}$ is the $\boldsymbol{\theta}$ that solves

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\xi}, \boldsymbol{Y}) \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta}\in\Theta}{\arg\min} \left[\frac{1}{2}\sum_{i=1}^{N_e} \|\boldsymbol{y}_i - \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta})\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}{}^{-1}} - \log(\pi(\boldsymbol{\theta}))\right]. \tag{24}$$

Long et al. [2013] show that

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

The covariance matrix of the posterior $\boldsymbol{\Sigma}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}})$ is the Hessian matrix of the negative logarithm of the posterior pdf evaluated at $\boldsymbol{\xi}$ and $\hat{\boldsymbol{\theta}}$,

$$\boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}) = N_e \nabla_{\boldsymbol{\theta}}(\boldsymbol{g}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}))^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}{}^{-1} \nabla_{\boldsymbol{\theta}} \boldsymbol{g}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \log(\pi(\hat{\boldsymbol{\theta}})) + \mathcal{O}_{\mathbb{P}}\left(\sqrt{N_e}\right). \tag{26}$$

A detailed deduction of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}})$ is presented in Beck et al. [2018].

According to Long et al. [2013], using the approximation $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}_t$ yields the Laplace-approximated SEIG as

$$I(\boldsymbol{\xi}) = \int_{\Theta} \left[-\frac{1}{2} \log(\det(2\pi \boldsymbol{\Sigma}(\boldsymbol{\xi}, \boldsymbol{\theta}_t))) - \frac{\dim(\boldsymbol{\theta})}{2} - \log(\pi(\boldsymbol{\theta}_t)) \right] \pi(\boldsymbol{\theta}_t) \mathrm{d}\boldsymbol{\theta}_t + \mathcal{O}\left(\frac{1}{N_e}\right). \quad (27)$$

Compared to (9), the approximated SEIG in (27) has only one integral, not requiring the integration over $\boldsymbol{Y}$ nor the integral to calculate the evidence. This SEIG with Laplace approximation is consistent with Chaloner and Verdinelli [1995], according to whom, maximizing SEIG is equivalent to minimizing the determinant of the posterior covariance matrix.

Estimating I in (27) by MCI results in the MCLA:

$$\mathcal{I}_{\text{MCLA}}(\boldsymbol{\xi}) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \left[-\frac{1}{2} \log(\det(2\pi \boldsymbol{\Sigma}(\boldsymbol{\xi}, \boldsymbol{\theta}_n))) - \frac{\dim(\boldsymbol{\theta})}{2} - \log(\pi(\boldsymbol{\theta}_n)) \right]. \quad (28)$$

According to Beck et al. [2018], the bias and variance of the MCLA estimator are, respectively,

$$|I - \mathbb{E}[\mathcal{I}_{\text{MCLA}}]| \leq C_{LA,1} h^{\eta} + \frac{C_{LA,2}}{N_e} + o(h^{\eta}), \quad \text{and} \quad (29)$$

$$\mathbb{V}[\mathcal{I}_{\text{MCLA}}] = \frac{C_{LA,3}}{N} \quad (30)$$

for the constants $C_{LA,1}$, $C_{LA,2}$, and $C_{LA,3}$. Although DLMC is a consistent estimator, MCLA is not, because the bias from the Laplace approximation does not vanish as the number of samples increases, being dependent on the number of repetitive experiments N_e.

Algorithm 2 presents the pseudocode for the MCLA estimator. Since MCLA does not have an inner loop, it only needs the outer loop sample-size, N. The evaluation of (26) requires a Jacobian of the model $\boldsymbol{g}$ with respect to the parameters $\boldsymbol{\theta}$; the main cost of MCLA evaluation.

Algorithm 2 Pseudocode for the MCLA estimator for SEIG.

function MCLA($\boldsymbol{\xi}, N$)
  for $n = 1, 2, ..., N$ **do** ▷ Outer loop
    Sample $\boldsymbol{\theta}_n$ from $\pi(\boldsymbol{\theta})$
    Evaluate $\nabla_{\boldsymbol{\theta}} \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}_n)$
    Use $\nabla_{\boldsymbol{\theta}} \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}_n)$ to evaluate $\boldsymbol{\Sigma}(\boldsymbol{\xi}, \boldsymbol{\theta}_n)$ using (26)
  end for
  $\mathcal{I}_{MCLA}(\boldsymbol{\xi}) \leftarrow \frac{1}{N} \sum_{n=1}^{N} \left[-\frac{1}{2} \log(\det(2\pi \boldsymbol{\Sigma}(\boldsymbol{\xi}, \boldsymbol{\theta}_n))) - \frac{\dim(\boldsymbol{\theta})}{2} - \log(\pi(\boldsymbol{\theta}_n)) \right]$
  return $\mathcal{I}_{MCLA}(\boldsymbol{\xi})$
end function

If forward-Euler is used for the estimation of the Jacobian in (26), the evaluation of the MCLA estimator cost has order $N(\dim(\boldsymbol{\theta}) + 1)h^{-\varrho}$. Thus, in comparison to DLMC, MCLA requires lower computational effort if $\dim(\boldsymbol{\theta})$ is less than M. For most cases, the M required to achieve a certain precision with DLMC is large, therefore, for these cases, MCLA is is computationally more efficient than DLMC.

3.5 Double-loop Monte Carlo estimator with Laplace-based importance sampling

Based on the method proposed by Beck et al. [2018], we use an importance sampling in the MCI of the evidence. Instead of sampling $\boldsymbol{\theta}^*$ from $\pi(\boldsymbol{\theta})$, we sample $\boldsymbol{\theta}^*$ from $\tilde{\pi}(\boldsymbol{\theta})$, where

$\tilde{\pi}(\boldsymbol{\theta}) \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}))$, and $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}$ are given in (24) and (26). Therefore, we are using a Gaussian approximation of the posterior at its MAP to draw more informative samples for the inner loop. The advantage of using the importance sampling is that it avoids the approximation error from the Laplace approximation and the numerical underflow problem from DLMC with the drawback of requiring, for each outer loop, to solve the sub-problem of finding the MAP $\hat{\boldsymbol{\theta}}$. The DLMCIS estimator is defined as

$$\mathcal{I}_{DLMCIS}(\boldsymbol{\xi}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} \log\left(\frac{p(\boldsymbol{Y}_n|\boldsymbol{\theta}_n, \boldsymbol{\xi})}{\frac{1}{M}\sum_{m=1}^{M}\mathcal{L}(\boldsymbol{Y}_n; \boldsymbol{\xi}; \boldsymbol{\theta}_m^*)}\right), \tag{31}$$

where

$$\mathcal{L}(\boldsymbol{Y}; \boldsymbol{\xi}; \boldsymbol{\theta}) = \frac{p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\xi})\pi(\boldsymbol{\theta})}{\tilde{\pi}(\boldsymbol{\theta})}. \tag{32}$$

The bias and variance of the DLMCIS estimator are deduced on the original paper by Beck et al. [2018] and are proven to be the same as of DLMC for a given tolerance, thus the DLMCIS estimator is also consistent. However, to achieve the desired tolerance, the inner loop sample-size M is significantly smaller for DLMCIS than for DLMC. Like in MCLA, the evaluation of (26) requires a Jacobian of the model $\boldsymbol{g}$ with respect to the parameters $\boldsymbol{\theta}$, moreover, finding $\hat{\boldsymbol{\theta}}$ requires solving the optimization problem in (24). We use a steepest descent search to find $\hat{\boldsymbol{\theta}}$. For that we use the gradient of the function to be minimized in (24),

$$\hat{\nabla}_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} -(\nabla_{\boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}))^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{Y} - \boldsymbol{g}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}})) - \nabla_{\boldsymbol{\theta}}\log\pi(\hat{\boldsymbol{\theta}}). \tag{33}$$

Algorithm 3 Pseudocode for finding MAP using steepest descent.

function FINDMAP($\boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{Y}, \alpha_{\boldsymbol{\theta}}$, TOL)
  $\hat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}$
  for $j = 1, 2, 3..$ **do**
    $\hat{\nabla}_{\boldsymbol{\theta}} \leftarrow -(\nabla_{\boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}))^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{Y} - \boldsymbol{g}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}})) - \nabla_{\boldsymbol{\theta}}\log\pi(\hat{\boldsymbol{\theta}})$
    $\hat{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}} - \alpha_{\boldsymbol{\theta}}\hat{\nabla}_{\boldsymbol{\theta}}$
    if $\|\hat{\nabla}_{\boldsymbol{\theta}}\|_2 <$ TOL **then**
      Break
    end if
  end for
  return $\hat{\boldsymbol{\theta}}$
end function

If forward-Euler is used for the estimation of the Jacobians in (26) and Algorithm 3, the evaluation of the DLMCIS estimator costs $N((C_{MAP}+1)(\dim(\boldsymbol{\theta})+1)+M)h^{-\varrho}$, where C_{MAP} is the number of iterations to estimate $\hat{\boldsymbol{\theta}}$ in Eq. in Algorithm 3.

Besides the advantage of reducing the number of samples of the inner loop, M, the DLMCIS estimator is more robust to numerical underflow than the DLMC estimator. The change of measure in the sampling of $\boldsymbol{\theta}^*$ guarantees that $\boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}^*)$ is close to $\boldsymbol{Y}$, evaluated using the $\boldsymbol{\theta}$ sampled in the outer loop. Thus, it is not likely that, for all inner loops, the likelihood of observing $\boldsymbol{Y}$ given $\boldsymbol{\theta}^*$ is numerically evaluated to zero. Figure 10 illustrates how the importance sampling mitigates the numerical underflow problem for the three-point flexural point; the likelihoods of observing $y^{(n)}$ for each g are larger than zero.

A pseudocode for the DLMCIS is presented in Algorithm 4, where the line where importance sampling happens is shaded in gray.

Algorithm 4 Pseudocode for the DLMCIS estimator for SEIG.

function DLMCIS($\boldsymbol{\xi}, N, M$)
  for $n = 1, 2, ..., N$ **do** ▷ Outer loop
    Sample $\boldsymbol{\theta}_n$ from $\pi(\boldsymbol{\theta})$
    for $i = 1, 2, .., .N_e$ **do**
      Sample $\boldsymbol{\epsilon}_i$ from $\mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$
      $\boldsymbol{y}_i \leftarrow \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}_n) + \boldsymbol{\epsilon}_i$
    end for
    $\boldsymbol{Y}_n \leftarrow \{\boldsymbol{y}_i\}_{i=1}^{N_e}$
    $p(\boldsymbol{Y}_n|\boldsymbol{\theta}_n, \boldsymbol{\xi}) \leftarrow \det(2\pi\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-\frac{N_e}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_e} \|\boldsymbol{\epsilon}_i\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}{}^{-1}}\right)$
    Find $\hat{\boldsymbol{\theta}}_n(\boldsymbol{\xi}, \boldsymbol{Y}_n)$ using Algorithm 3
    Evaluate $\nabla_{\boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}_n)$
    Use $\nabla_{\boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}_n)$ to evaluate $\boldsymbol{\Sigma}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}_n)$ using (26)
    for $m = 1, 2, ..., M$ **do** ▷ Inner loop
      Sample $\boldsymbol{\theta}^*_m$ from $\tilde{\pi}(\boldsymbol{\theta}) \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\Sigma}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}_n))$ ▷ Importance sampling
      Evaluate $\boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}^*_m)$
      $p(\boldsymbol{Y}_n|\boldsymbol{\theta}^*_m, \boldsymbol{\xi}) \leftarrow \det(2\pi\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-\frac{N_e}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_e} \left\|\boldsymbol{y}_i^{(n)}(\boldsymbol{\xi}) - \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}^*_m)\right\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}{}^{-1}}\right)$
      $\mathcal{L}(\boldsymbol{Y}_n; \boldsymbol{\xi}; \boldsymbol{\theta}^*_m) \leftarrow p(\boldsymbol{Y}_n|\boldsymbol{\theta}^*_m, \boldsymbol{\xi})\pi(\boldsymbol{\theta}^*_m)/\tilde{\pi}(\boldsymbol{\theta}^*_m)$
    end for
    $p(\boldsymbol{Y}_n|\boldsymbol{\xi}) \leftarrow \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}(\boldsymbol{Y}_n; \boldsymbol{\xi}; \boldsymbol{\theta}^*_m)$
  end for
  $\mathcal{I}_{DLMCIS}(\boldsymbol{\xi}) \leftarrow \frac{1}{N}\sum_{n=1}^{N} \log\left(\frac{p(\boldsymbol{Y}_n|\boldsymbol{\theta}_n, \boldsymbol{\xi})}{p(\boldsymbol{Y}_n|\boldsymbol{\xi})}\right)$
  return $\mathcal{I}_{DLMCIS}(\boldsymbol{\xi})$
end function

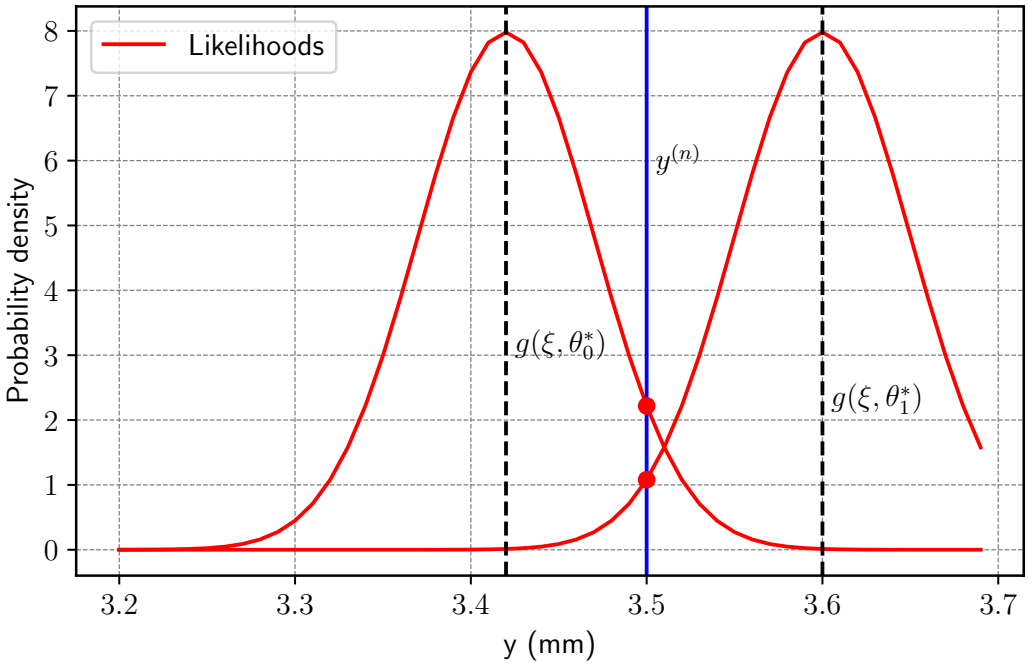


Figure 10: Avoiding numerical underflow using importance sampling illustrated for the three-point flexural test.

4 Numerical example

This numerical example addresses the problem of positioning a strain gauge on a beam modeled following Timoshenko theory for estimating the beam's mechanical properties. Its purpose is to show that the proposed OED framework can reproduce the engineer's intuition. In other words, the optimized design found is consistent with what an engineer would intuitively expect.

4.1 Strain gauge positioning on Timoshenko beam

The OED problem here consists in finding the optimal placement of a strain-gauge on a beam to estimate the material's Young and shear deformation moduli. It is opted to employ the MCLA estimator of the EIG, given its lower cost in comparison to DLMC and DLMCIS.

It is assumed that the strain-gauges provide (noisy) strains observations in the vertical and longitudinal axes for a given point of the domain of the beam. The beam mechanical properties to be estimated are the the Young modulus E and the shear modulus G, given measurements obtained from the strain gauge using the Timoshenko beam model. The beam's length, height and width are, respectively, $L_e = 10$ m, $H = 2$ m and $b = 0.1$ m. Although the beam's geometry is not consistent with engineering practice, it is devised to furnish an interesting OED problem: we want an optimization problem where the optimum is not in a vertex of the beam. A uniform load q_o of 1.00 kN/mm is imposed on the beam's vertical axis and distributed along its main axis. The geometry of the beam, the load, and the position of the strain gauge are illustrated in Figure 11.

We aim to locate a strain gauge on the beam that maximizes the information on E and G. W employ the Timoshenko's theory Timoshenko [1921] as the forward model of the problem, a mechanical model that captures the strains resulting from both normal and shear stresses. Such a

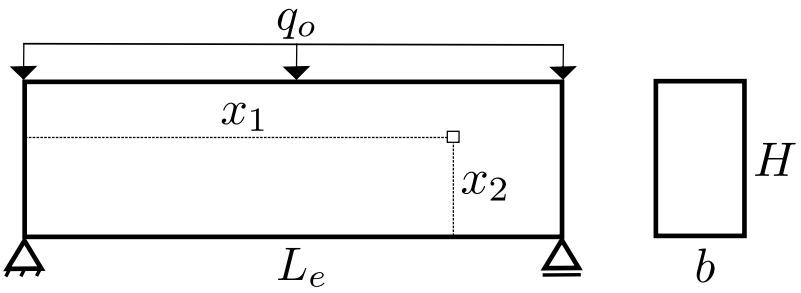


Figure 11: (Example 3) Geometry of the Timoshenko beam.

model is defined as

$$\begin{cases} K_s G A_r \varepsilon_{12} = \frac{q_o L_e}{2} - q_o x_1, \\ E I_n \varepsilon_{11} = \frac{q_o x_1 (L_e - x_1)}{2} x_2, \end{cases} \tag{34}$$

where ε_{11} is the normal strain, ε_{12} is the shear strain, x_1 and x_2 are the positions of the strain gauge on the horizontal and vertical axes respectively, q_o is the uniform load, L_e is the length of the beam, I_n is the inertia moment of the cross section, K_s is the Timoshenko constant ($K_s = 5/6$ in all test cases), and A_r is the cross-section area.

4.2 Bayesian formulation

The optimal position for the strain gauge that provides the maximum information about E and G is denoted by $\boldsymbol{\xi}^* = (x_1^*, x_2^*)$. The longitudinal strain on the main axis of the beam, denoted by ε_{11}, together with the transverse strain ε_{12}, compose the output of the forward model. Therefore, based on (34), we find that

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta}) &= (\varepsilon_{11}(\boldsymbol{\xi}, \boldsymbol{\theta}), \varepsilon_{12}(\boldsymbol{\xi}, \boldsymbol{\theta})) \\ &= \left(\frac{\xi_2 \left(q_o L_e \xi_1 - q_o \xi_1^2 \right)}{2 \theta_1 I_n}, \frac{\frac{L_e}{2} q_o - q_o \xi_1}{K_s \theta_2 A_r} \right), \end{aligned} \tag{35}$$

where (x_1, x_2) and (E, G) are replaced by (ξ_1, ξ_2) and (θ_1, θ_2), respectively. The additive error of the measurement is Gaussian $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon)$, where the noise covariance matrix is $\boldsymbol{\Sigma}_\epsilon = \text{diag}\left\{\sigma_{\epsilon_1}^2, \sigma_{\epsilon_2}^2\right\}$.

4.3 Test cases

We investigate the EIG of this problem in four test cases, in which we attempt to locate the optimal strain-gauge placement on a beam. We test all the different cases, changing the variance of the prior pdf of $\boldsymbol{\theta}$, the dispersion of the measurement noise, and the number of experiments. The prior pdf of $\boldsymbol{\theta}$ is Gaussian with the distribution $\pi(\boldsymbol{\theta}) \sim \mathcal{N}\left((\mu_{pr}^E, \mu_{pr}^G)^T, \text{diag}\left\{(\sigma_{pr}^G)^2, (\sigma_{pr}^E)^2\right\}\right)$, where $\mu_{pr}^E = 30.00$ GPa and $\mu_{pr}^G = 11.54$ GPa.

Table 1 presents the parameters used in each of the four cases.

We devised the parameters of the four cases with the intention of having four different OED problems. In the first two cases, both E and G have standard deviation of, respectively, 30% and 20% of their means. Also, case 1 has larger observation noise than case 2, and more experiments, 3. The third and fourth cases are exactly like case 2, except that case 3 has significantly less dispersion in G, a coefficient of variation of 4%, and case 4 has the same coefficient of variation for E.

Table 1: Parameters for the Timoshenko beam problem (Example 3).

Parameter	N_e	σ^E_{pr}(GPa)	σ^G_{pr}(GPa)	$\sigma_{\epsilon_1}(\times 10^{-4})$	$\sigma_{\epsilon_2}(\times 10^{-4})$
Case 1	3	9.00	3.46	6.25	1.30
Case 2	1	6.00	2.31	3.75	0.78
Case 3	1	6.00	0.46	3.75	0.78
Case 4	1	1.20	2.31	3.75	0.78

We evaluate the EIG of each case using a grid of uniformly distributed points, which provides contour plots of the expected information gain across the beam's domain. These plots are illustrated in Figure 12.

In cases 1 and 2, the optima are similarly located near the bottom of the beam, between the middle and the end. In case 3, the optimum is located in the bottom-middle of the beam; in case 4, the optimum is located on the supports. These placements are expected, as the Young modulus depends on the bending moment (that is maximum at the middle of the beam ($x_1 = L_e/2$)), and the shear modulus depends on the shear stress (that is maximum at the beam supports ($x_1 = 0$ and $x_1 = L_e$)). In case 3, the prior information about G is more accurate; consequently, the optimum is at the middle of the beam where more information about E can be collected. Conversely, in case 4, the optimum is at the supports, where data is more informative about G.

In Table 2, we present the non-optimized points randomly chosen on the design domain (Non-Opt.), the optimized setups (Opt.), the respective expected information gains in relation to the prior, and the standard deviations of the posterior pdfs of the parameters E and G for the four cases. The posteriors are evaluated at $\hat{\boldsymbol{\theta}} = (\mu^E_{pr}, \mu^G_{pr})$ for the four cases are presented in Figure 13.

We observe a reduced variance in the optimized experiment, compared to the original, reflecting the importance of an informative experiment. In cases 3 and 4, no information is acquired about G and E, respectively, since the variances in the axes are not reduced, compared to the prior.

Table 2: Results from the Timoshenko beam problem (Example 3).

		x^*_1(mm)	x^*_2(mm)	$\mathcal{I}_{\text{MCLA}}$	σ^E_{post} (GPa)	σ^G_{post} (GPa)
Case 1	Non-Opt.	5500.00	-100	0.14	8.00	2.40
	Opt.	8022.59	-1000.00	2.43	2.48	0.54
Case 2	Non-Opt.	5500.00	-100	0.23	2.38	1.38
	Opt.	7962.77	-1000.00	3.35	1.60	0.74
Case 3	Non-Opt.	5500.00	-100	0.06	5.70	0.46
	Opt.	5004.47	-1000.00	1.28	1.72	0.46
Case 4	Non-Opt.	5500.00	-100	0.22	1.20	1.93
	Opt.	10000.00	-1000.00	1.94	1.20	0.33

Because we use the biased and inconsistent MCLA estimator of the gradient, as a sanity check, we evaluate the gradient at the optima we found (the first two cases) using the full gradient of the DLMCIS estimator with $N = 10^3$ and $M = 10^2$. In both cases, the gradient norm is below 10^{-3}, meaning that the bias of the Laplace approximation is considerably small at the optima. We conclude that the biased solutions found are not significantly distant to the real optima.

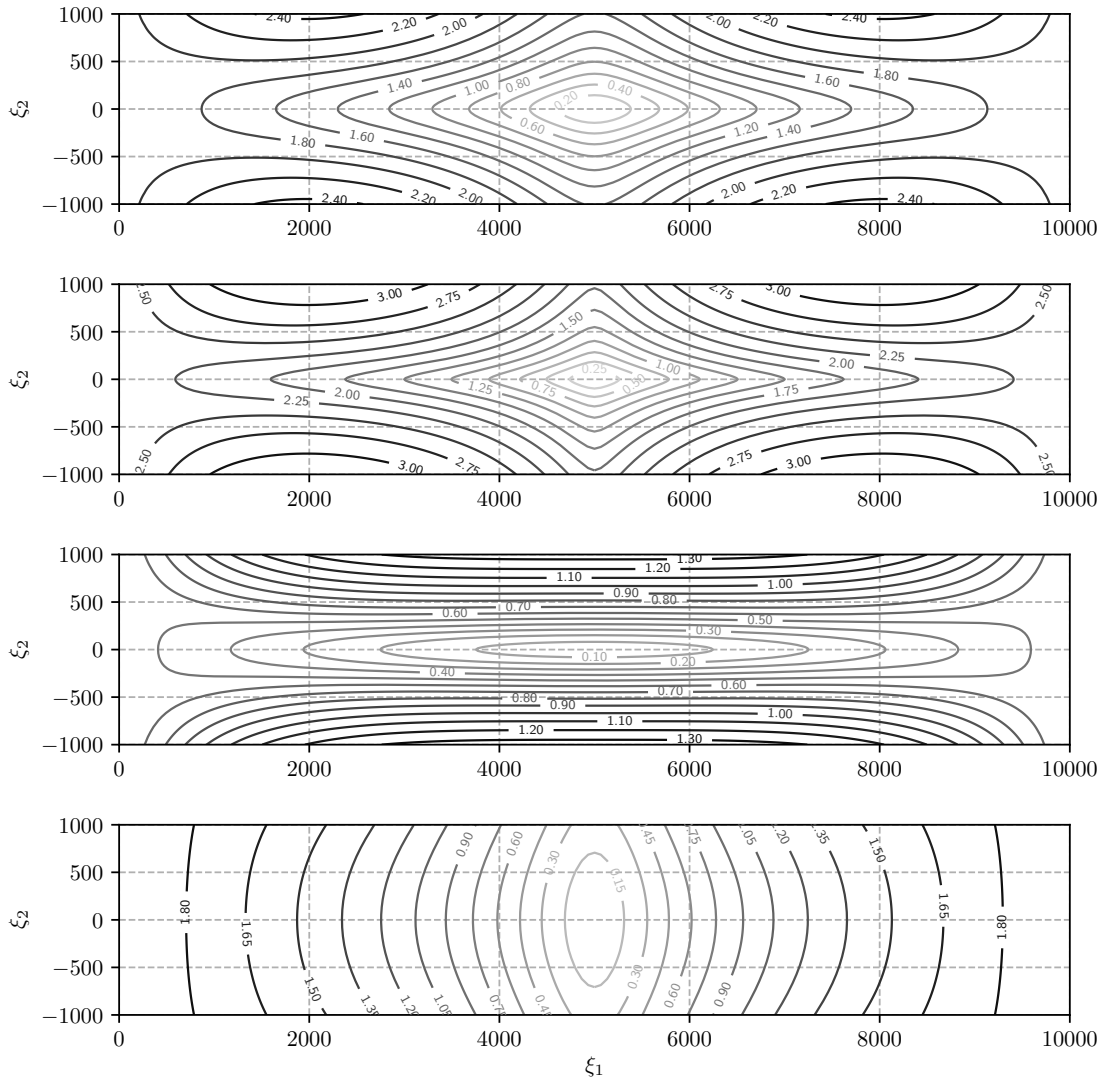


Figure 12: From top to bottom, cases 1 to 4 (summarized in Table 2). Expected information gain contours computed with MCLA.

The main goal of this example is to visualize how the optimum results are consistent with the input data given.

5 Chapter summary

In this chapter, we introduced key concepts related to Bayesian inference and OED. In Section 2.1, we presented the model of experiments with additive noise. Also, in Section 2.1, we defined a measure of efficiency of experiments: D_{KL} between the prior and posterior pdfs. Moreover, we presented, from an experiment with additive noise, the SEIG, the quantity we want to maximize.

In Section 3, we presented SEIG estimators based in MCI. Applying MCI in SEIG furnishes the DLMC estimator, introduced by Ryan [2003]. The DLMC estimator is consistent and can estimate SEIG for any general case, however, depending on the problem it can be expensive and unstable [Beck et al., 2018]. As an alternative to DLMC, we presented the MCLA estimator pro-

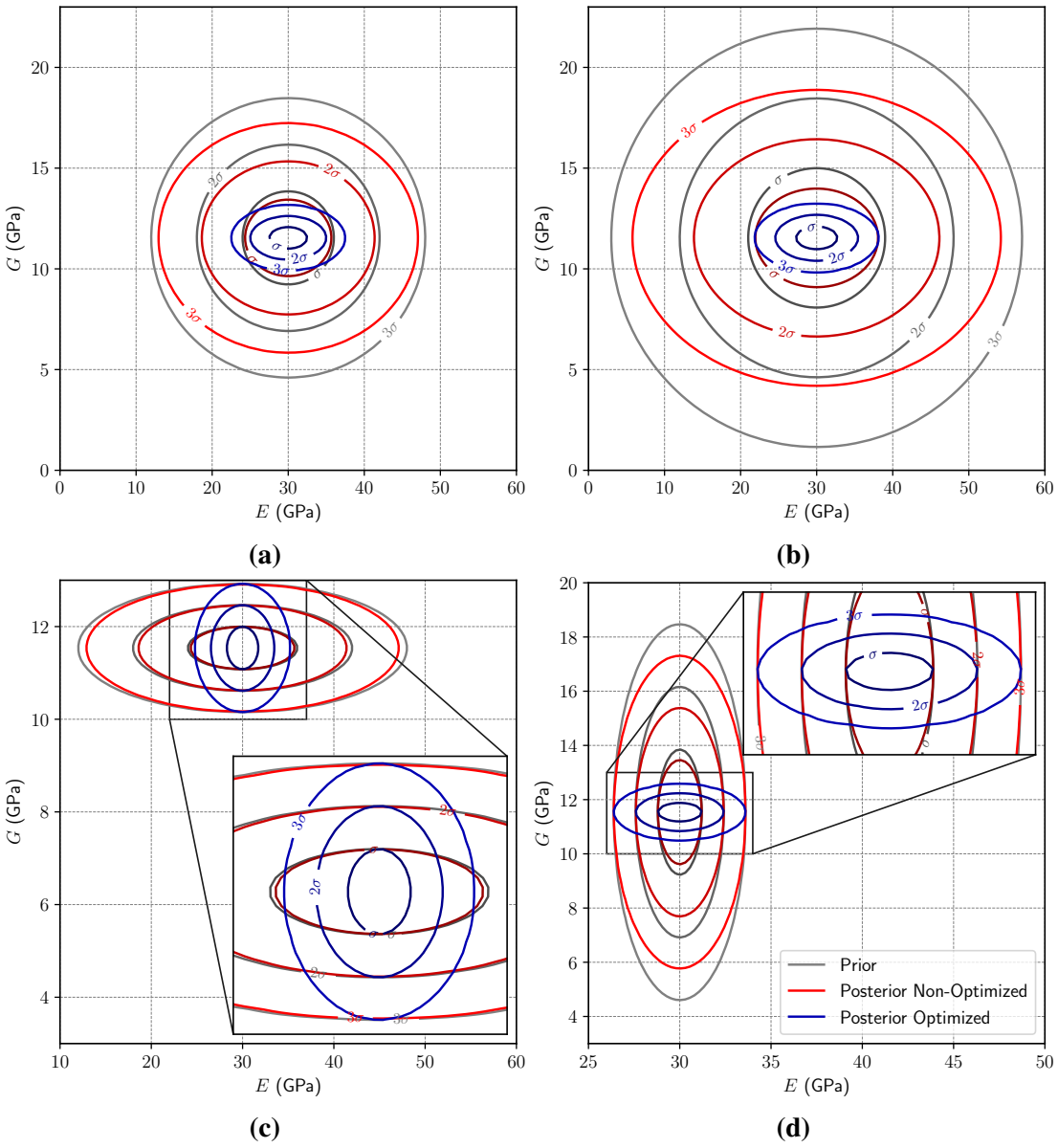


Figure 13: Prior, posterior, and optimized posterior pdfs for the Young modulus E and the shear modulus G for cases 1 (a), 2 (b), 3 (c), and 4 (d).

posed by Long et al. [2013] in Section 3. This estimator uses the Laplace approximation of the posterior, thus, avoiding the inner-loop in DLMC. The resulting algorithm, MCLA, is an inexpensive estimator for SEIG, with the drawback of introducing bias due to the Laplace approximation of the posterior distribution. This bias resulting from the Laplace approximation does not vanish as sample-sizes go to infinity, thus, MCLA is an inconsistent estimator. This fact is then investigated in the numerical example section. For the cases where the bias of MCLA is not acceptable, we introduced an importance sampling based on the Laplace approximation, proposed by Beck et al. [2018], resulting in the DLMCIS estimator.

Acknowledgements

Acknowledgements are optional, at the discretion of the chapter co-authors.

References

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

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

W. Commons. Single-edge notch-bending specimen (also called three-point bending specimen) for fracture toughness testing., 2011. URL `https://commons.wikimedia.org/wiki/File:SingleEdgeNotchBending.svg`.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

M. Hamada, H. Martz, C. Reese, and A. Wilson. Finding near-optimal bayesian experimental designs via genetic algorithms. *The American Statistician*, 55(3):175–181, 2001.

X. Huan and Y. M. Marzouk. Simulation-based optimal bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 232(1):288–317, 2013.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

Q. Long, M. Scavino, R. Tempone, and S. Wang. Fast estimation of expected information gains for bayesian experimental designs based on laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, 259:24–39, 2013.

C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons, 2016.

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): 585–603, 2003.

C. E. Shannon. A mathematical theory of communication, part i, part ii. *Bell Syst. Tech. J.*, 27: 623–656, 1948.

S. P. Timoshenko. Lxvi. on the correction for shear of the differential equation for transverse vibrations of prismatic bars. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 41(245):744–746, 1921.