STOCHASTIC OPTIMIZATION FOR DESIGN OF EXPERIMENTS

André G. Carlon

Rafael H. Lopez

a.g.carlon@posgrad.usfc.br

rafael.holdorf@ufsc.br

Department of Civil Engineering – UFSC

Rua Joao Pio Duarte, s/n, 88034-000, Florianopolis, SC, Brazil

Luis Espath

espath@gmail.com

CEMSE, KAUST

23955-6900, Thuwal, Saudi Arabia

Abstract. *We present a stochastic optimization method to converge to local maxima of the Shannon's expected information gain of experiments using a Bayesian framework. We avoid the high cost of evaluating several double loop Monte Carlo simulation (DLMC) each iteration by employing the stochastic gradient. Our method has proven to converge to local maxima with a fraction of the cost of the classical approach, making possible to optimize experiments with more expensive models. We confirm the convergence of our method optimizing experimental design problems with analytical ODE models and with finite elements approximations of PDEs.*

Keywords: *Optimal experimental design, Bayesian design, Stochastic optimization*

INTRODUCTION

Many fields of science rely heavily on information obtained through experiments. Usually, we will expect probabilistic information from experiments on the form of a probability density function (PDF) of some quantity of interest (QoI) that we want to have information about, e.g. a Gaussian PDF defined by mean and standard deviation. In some cases, this experiments might be expensive and/or require a large number of executions to achieve the desired confidence about the QoI. On these cases, it is interesting to use the parameters of the experiment that minimize the uncertainties, main concern of the field of science called Optimal Experimental Design (OED). For example, one might want to measure the Young modulus and yielding point of some linear elastic material through a tensile test. On this case, the QoI are the Young modulus and the yielding point of the material, and the parameters are the set-up of the experiment, e.g. geometry of the sample, direction and speed of the loading, number of tests.

The decision on the parameters will directly influence the variance of the information about the QoI resulting from the observation. On the present paper, we will denote the QoI by $\boldsymbol{\theta}$, the parameters of the experiment by $\boldsymbol{\xi}$ and the observations by $\mathbf{y}$. We want to define the parameters $\boldsymbol{\xi}$ that will provide the observations $\mathbf{y}$ that give us more information about the QoI $\boldsymbol{\theta}$.

The goal of OED is to find an optimal set of parameters that is the maximizer of some measure of the information about the QoI. On the context of Bayesian inference, the Shannon expected information gain – a measure of the distance between prior and posterior PDFs – is usually employed. The advantage of this approach is that it allows the analysis of priors and posteriors with any shape. The flexibility of this approach has the drawback of requiring the solution of a double integral over the sample space of the QoI. The straightforward approach to estimate the Shannon expected information gain is to approximate the solution of the double integrals with a Double Loop Monte Carlo (DLMC), a process that might be very expensive depending on the modeling of the forward model. By forward model we mean the simulation of the experiment, by analytical or numerical methods, a process that can be computationally expensive. On the case of the aforementioned tensile test, the evaluation of the forward model is equivalent to numerically simulate the test with a sample, using a specific set-up $\boldsymbol{\xi}$ and properties $\boldsymbol{\theta}$ to evaluate observations $\mathbf{y}$. Observing the expected information gain requires the simulation of the experiment with a sample of different mechanical properties and the same set-up, resulting in different observations. After a certain number of simulations, the observations $\mathbf{y}$ can be used to estimate a PDF of $\boldsymbol{\theta}$, using the DLMC model. Moreover, estimating the gradient of the expected information gain in relation to parameters $\boldsymbol{\xi}$ with a forward Euler finite difference requires $\dim(\boldsymbol{\xi}) + 1$ DLMC evaluations. Supposing that a numerical simulation of the tensile test takes 1 second to be performed and that 1,000 simulations are required for each loop, the estimation of the gradient of the expected information gain with respect to the 4 different parameters would take 5×10^6 seconds or 58 days! Considering that the optimization is an iterative process requiring an estimation of the gradient for each iteration, a whole life would be needed to optimize a simple experiment like the tensile test. Our goal is to use the stochastic gradient(SG) to reduce the cost in OED.

OPTIMAL EXPERIMENTAL DESIGN

Since we intend to optimize experiments, we must define a formulation for the realization of experiments. For that, we will define an observation of the experiment as $\mathbf{y}$, a set of quantities

of interest as $\boldsymbol{\theta}$, a set of parameters of the experiment as $\boldsymbol{\xi}$ and the noise on observations as $\boldsymbol{\epsilon}$. Using a forward model $\mathbf{g}$ we define our observations as

$$\mathbf{y} = \mathbf{g}(\boldsymbol{\xi}, \boldsymbol{\theta}) + \boldsymbol{\epsilon}.$$

Throughout the present paper the noise $\boldsymbol{\epsilon}$ will be considered to be additive – independent of $\mathbf{g}, \boldsymbol{\xi}$ and $\boldsymbol{\theta}$ –, and Gaussian, with distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Sigma_\epsilon)$.

Bayesian inference

To estimate the efficiency of different set-ups of experiments we will use concepts from Bayesian inference. We will consider that we have some prior information about $\boldsymbol{\theta}$ and that the experiment will provide more information about it. To estimate this information we will use Bayes formula:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{y})} \tag{1}$$

where:

- $\pi(\boldsymbol{\theta}|\mathbf{y})$—posterior distribution: the updated PDF of some random variable $\boldsymbol{\theta}$ given an observation $\mathbf{y}$;
- $\pi(\boldsymbol{\theta})$—prior distribution: the probability of $\boldsymbol{\theta}$ before the observation $\mathbf{y}$;
- $p(\mathbf{y}|\boldsymbol{\theta})$—likelihood: the probability of observing $\mathbf{y}$ given the previously known prior $\pi(\boldsymbol{\theta})$;
- $p(\mathbf{y})$—evidence: the probability of $\mathbf{y}$ being observed.

To estimate the distance between the prior and posterior distribution we will use the Kullback-Leibler Divergence.

Kullback Leibler divergence

The Kullback Leibler Divergence (D_{KL}) is a measure of the distance between two PDFs p and q:

$$D_{\mathrm{KL}}(P|Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

We want to use the D_{KL} to estimate the distance between the prior and the posterior as a way of measuring the efficiency of an experiment, so we can define the D_{KL} as:

$$D_{\mathrm{KL}}(\pi(\boldsymbol{\theta}|\mathbf{y})|\pi(\boldsymbol{\theta})) = \int_{\Theta} \pi(\boldsymbol{\theta}|\mathbf{y}) \log\left(\frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{\pi(\boldsymbol{\theta})}\right) d\boldsymbol{\theta}.$$

Shannon expected information gain

The D_{KL} does not consider the noise in observations due to measure uncertainties. To estimate the information gain on this context, we need to integrate the D_{KL} over the probability of observing $\mathbf{y}$—its evidence:

$$I = \int_{\mathcal{Y}} \int_{\Theta} \pi(\boldsymbol{\theta}|\mathbf{y}) \log \left(\frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{\pi(\boldsymbol{\theta})} \right) d\boldsymbol{\theta} p(\mathbf{y}) d\mathbf{y}. \tag{2}$$

We can rewrite the eq. (2) using the Bayes equation presented in eq. (1) and substitute the relation between the PDFs of the posterior and prior by the relation between the likelihood and the evidence:

$$I = \int_{\Theta} \int_{\mathcal{Y}} p(\mathbf{y}|\boldsymbol{\theta}) \log \left(\frac{p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})} \right) d\mathbf{y} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{3}$$

We will use eq. (3) instead of eq. (2) to derive the DLMC equation used to estimate the information gain on this paper.

Double loop Monte Carlo

To estimate the Shannon expected information gain, the integrals on eq. (3) can be approximated using MC method:

$$\mathcal{I}_{dlmc} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \log \left(\frac{p(\mathbf{y}_n|\boldsymbol{\theta}_n)}{p(\mathbf{y}_n)} \right)$$

where $p(\mathbf{y}_n)$ needs to be approximated by another MC:

$$p(\mathbf{y}_n) = \frac{1}{M} \sum_{m=1}^{M} p(\mathbf{y}_n|\boldsymbol{\theta}_m^*).$$

It is important to notice that the random variable $\boldsymbol{\theta}^*$ used to estimate the evidence is i.i.d. in relation to $\boldsymbol{\theta}$, i.e. the samples of $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$ are different. The DLMC estimator for the Shannon expected information gain is defined as:

$$\mathcal{I}_{dlmc} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \log \left(\frac{p(\mathbf{y}_n|\boldsymbol{\theta}_n)}{\frac{1}{M} \sum_{m=1}^{M} p(\mathbf{y}_n|\boldsymbol{\theta}_m^*)} \right). \tag{4}$$

According to Ryan (2003), the the variance of the DLMC estimator presented in eq. (4) is $\mathcal{O}(N^{-1})$ and its bias is $\mathcal{O}(M^{-1})$.

STOCHASTIC GRADIENT

The SG is a method to estimate the gradient of a stochastic function on iterative processes. The method is based on the Stochastic Approximation proposed by Robbins and Monro (1951) where the root of the average of several functions is to be found. Considering that on local search methods the goal is to find regions where the gradient is a null vector, the two methods – SG and SA – are analogous. On the steepest descent method, the gradient is summed up throughout iterations, so deviations are averaged out. The idea of the SG is that the precision on the estimation of the gradients can be relaxed because errors will average out as iterations go. Considering a function $f(\boldsymbol{\xi}, \boldsymbol{\theta})$ where $\boldsymbol{\xi}$ is a vector containing the deterministic variables—which we will optimize in relation to—and $\boldsymbol{\theta}$ is a vector with random variables, the gradient of f in respect to $\boldsymbol{\xi}$ would require an expectation over $\boldsymbol{\theta}$, i.e. $\nabla_{\boldsymbol{\xi}} f(\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\xi}} f(\boldsymbol{\xi}, \boldsymbol{\theta})]$. We define the estimator $\tilde{\nabla} f$ as the stochastic approximation of the gradient ∇f. The estimator can be used as long as it is not biased, i.e.

$$\mathbb{E}_{\boldsymbol{\theta}}[\tilde{\nabla}_{\boldsymbol{\xi}} f(\boldsymbol{\xi}, \boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\xi}} f(\boldsymbol{\xi}, \boldsymbol{\theta})]. \tag{5}$$

For a step-size α, we define the update rule for $\boldsymbol{\xi}$ on the k-th iteration as:

$$\boldsymbol{\xi}_{k+1} = \boldsymbol{\xi}_k - \alpha_k \tilde{\nabla}_{\boldsymbol{\xi}}\ f(\boldsymbol{\xi}_k, \boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i$ is the $\boldsymbol{\theta}$ sampled on the i-th iteration. Consequently, the k-th $\boldsymbol{\xi}$ will be:

$$\boldsymbol{\xi}_k = \boldsymbol{\xi}_0 - \sum_{i=1}^{k-1} \alpha_i \tilde{\nabla}_{\boldsymbol{\xi}} f(\boldsymbol{\xi}_i, \boldsymbol{\theta}_i). \tag{6}$$

Observing eq. (6) we can conclude that, if eq. (5) is satisfied and $\mathbb{E}_{\boldsymbol{\theta}} f$ is convex on $\boldsymbol{\xi}$, as $k \to \infty$, the k-th $\boldsymbol{\xi}$ will converge to the optimum: $\boldsymbol{\xi}_k \to \boldsymbol{\xi}^*$, i.e. $||\boldsymbol{\xi}_k - \boldsymbol{\xi}^*|| \to 0$. On the present paper the use of the steepest descent using the SG is called Stochastic Gradient Descent(SGD).

The advantage of using the SG to perform optimization is that less samples of the random variables are needed to estimate the gradient, because we can admit a larger error. On the case of the Shannon expected information gain estimator of eq. (4), we can reduce N to a smaller integer $N^* < N$. Robbins and Monro propose the use of $N^* = 1$:

$$\tilde{\nabla}_{\boldsymbol{\xi}} \mathcal{I}_{dlmc}(\boldsymbol{\xi}) = \nabla_{\boldsymbol{\xi}} \left(\log \left(\frac{p(\mathbf{y}_n | \boldsymbol{\theta}_n, \boldsymbol{\xi})}{\frac{1}{M} \sum_{m=1}^{M} p(\mathbf{y}_n | \boldsymbol{\theta}_m^*, \boldsymbol{\xi})} \right) \right) \tag{7}$$

EXAMPLES

On this section three examples will be studied: two based on benchmark functions with no physical meaning and a third based on an engineering experiment to measure material properties.

Benchmark problem 1

The SG and its ramifications are built upon the assumption of convexity of the function to be minimized/maximized on its domain. To test its efficiency, we will evaluate their convergence on the problem of maximizing the expected information gain of a convex model. Even though not all convex forward models will result in a convex expected information gain, this is true for the present case. Noting that, for this example, θ is a scalar and $\boldsymbol{\xi}$ has size 2, the function is

$$g(\boldsymbol{\xi}, \theta) = \boldsymbol{\xi}^T \mathbf{A} \boldsymbol{\xi}\, \theta - \frac{1}{50} \boldsymbol{\xi}^T \mathbf{A} \theta^2 - 10\theta - 1$$

with

$$\mathbf{A} = \begin{bmatrix} 1 & -0.2 \\ -0.2 & 0.5 \end{bmatrix}.$$

The prior is Gaussian and defined as $\pi(\theta) \sim \mathcal{N}(0, 10)$ and the additive error is also Gaussian: $\epsilon \sim \mathcal{N}(0, 1)$. To solve this problem the SGD will be used to maximize the estimation of the Shannon expected information gain. On Fig. 1 is presented a contour of the estimation of the Shannon expected information with the path of the optimization, the norm of the gradient and the distance to the optimum.

The SGD algorithm successfully converged to the maximum even with an error on the estimation of the gradient and an additive noise.

Benchmark problem 2

On this problem, the information gain is measured for the combination of 2 different experiments with the same forward model. These experiments are to be performed with two different set-ups ξ_1 and ξ_2:

$$\begin{bmatrix} y_1(\theta, \xi_1) \\ y_2(\theta, \xi_2) \end{bmatrix} = \begin{bmatrix} \theta^3 {\xi_1}^2 + \theta \text{exp}(-|0.2 - \xi_1|) \\ \theta^3 {\xi_2}^2 + \theta \text{exp}(-|0.2 - \xi_2|) \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

considering the parameters bounded as $0 \leq \xi_{1,2} \leq 1$, the random variable $\theta \sim \mathcal{U}(0, 1)$ and the additive error $\epsilon \sim \mathcal{N}(0.0, 10^{-4})$. This problem has been studied by some authors like Huan and Marzouk (2013) and Ben Issaid (2015). Its minima are $(0.2, 1)$ and $(1, 0.2)$, however there are local minima in $(0.2, 0.2)$ and $(1, 1)$. On Fig. 2 the contour of the estimation of the Shannon expected information gain is presented along with 4 paths of optimization using the SG. The convergence of the norms and distances to optima for the 4 cases are presented on the same figure. On Fig. 3 the prior and posteriors are shown, however, the difference between the posteriors is not visible. That is why Fig 4 is included, showing a detail of the mode of the posteriors in log scale.

The SGD converged to local optima on the 4 cases, showing its efficiency on local search of stochastic problems.

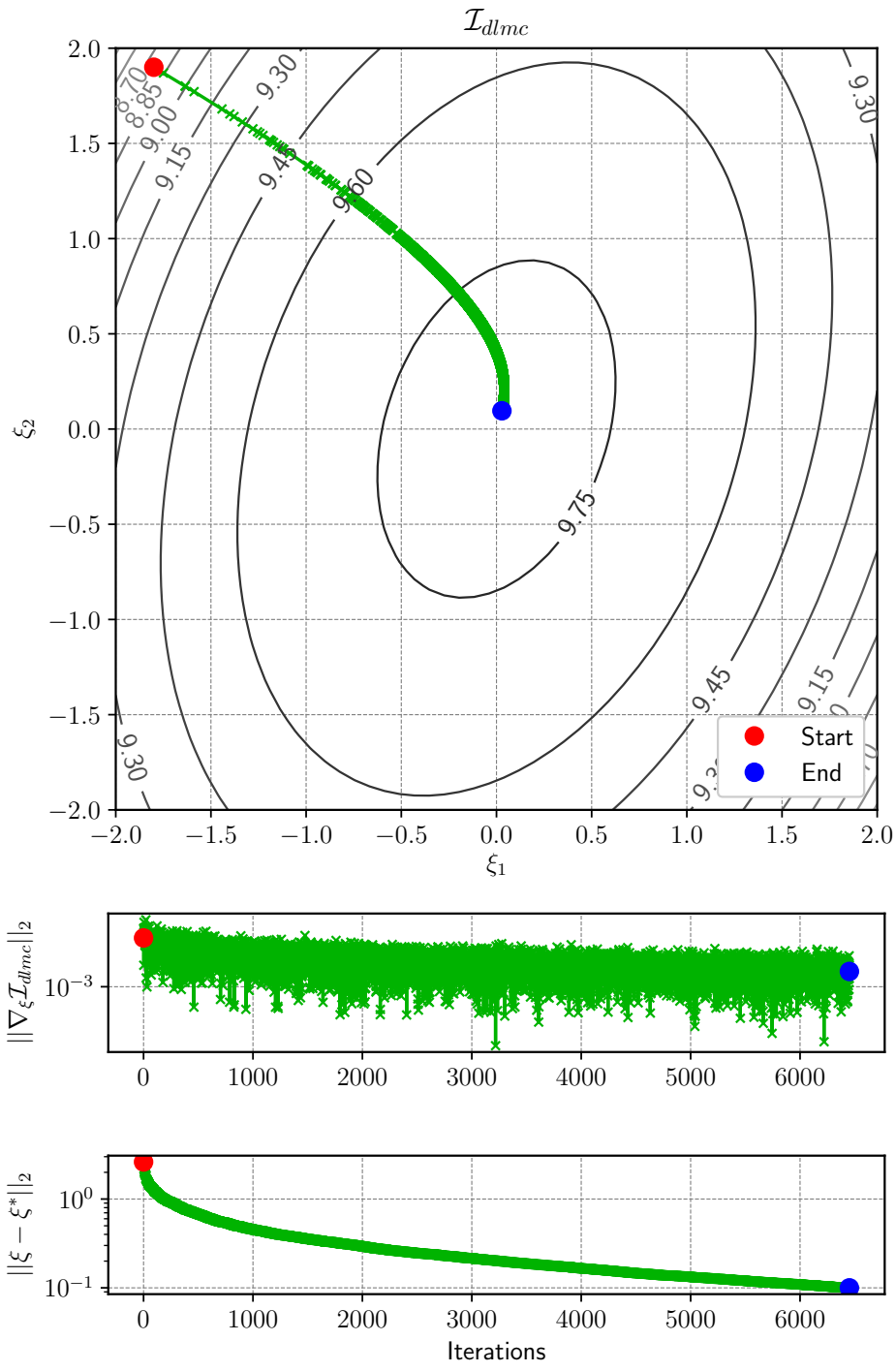

Figure 1: Benchmark problem 1 contour and convergence.

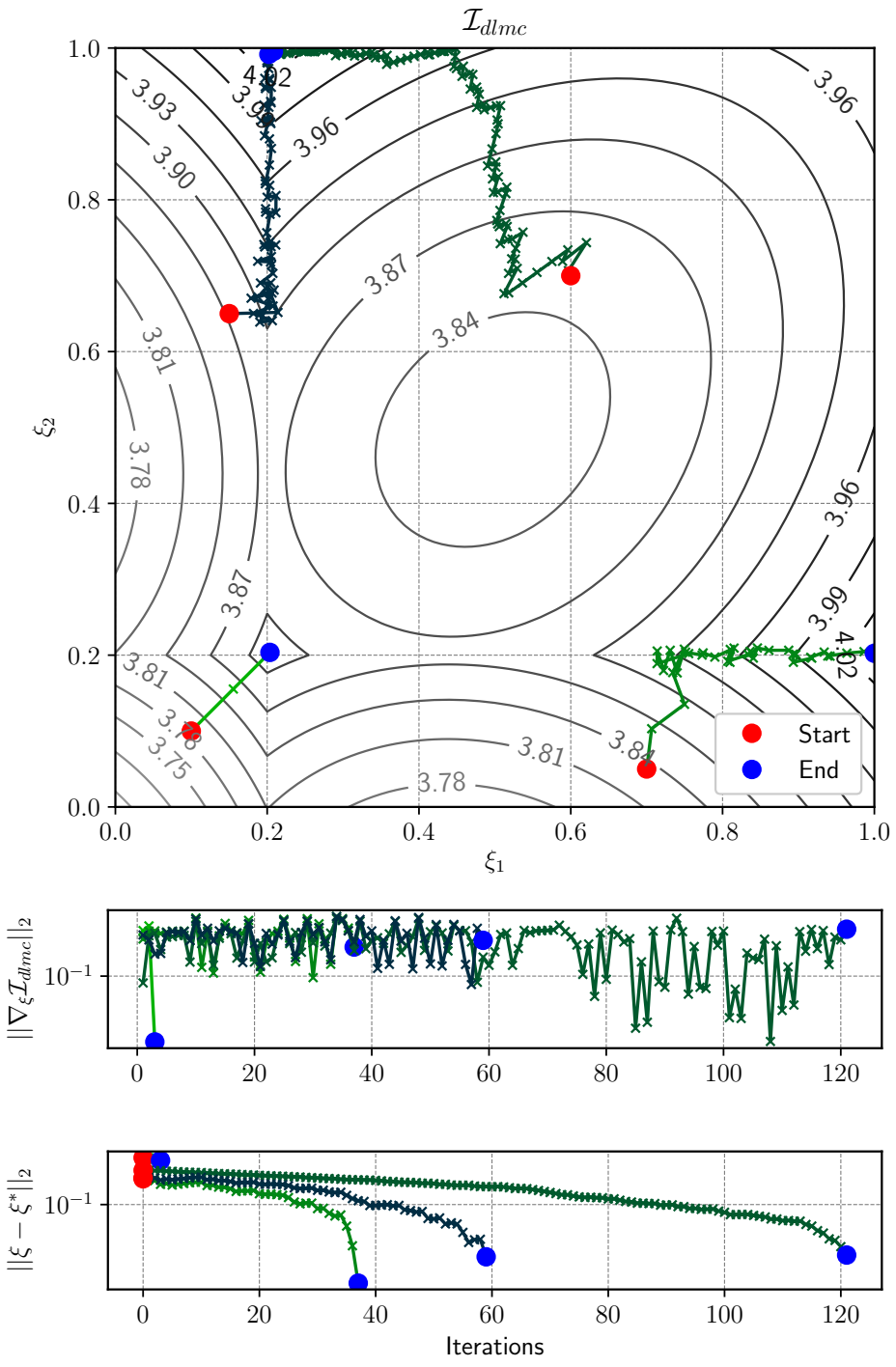

Figure 2: Benchmark problem 2 with 4 starting points.

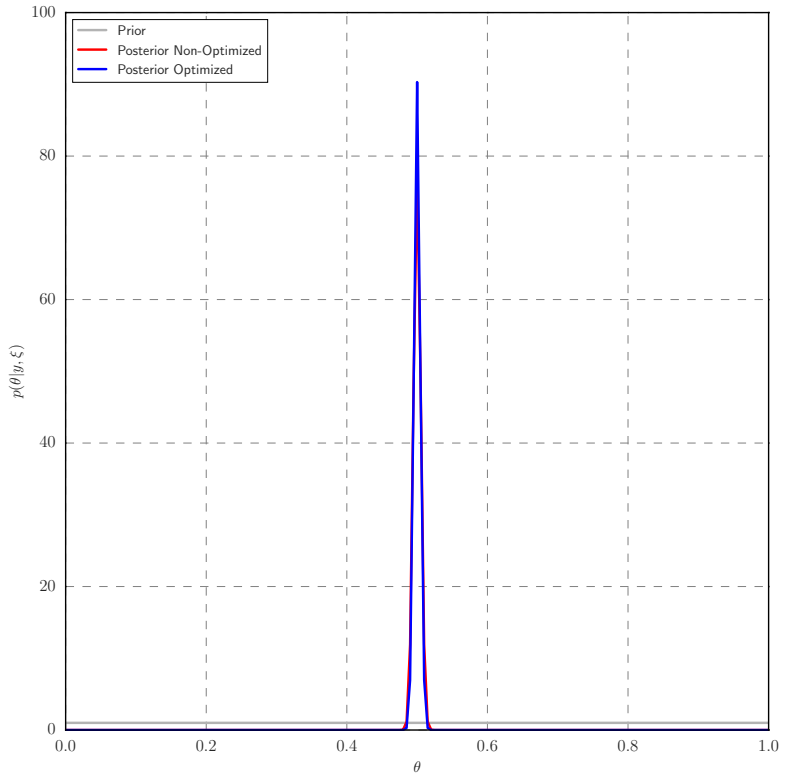

Figure 3: Benchmark problem 2 prior and posteriors before and after optimization.

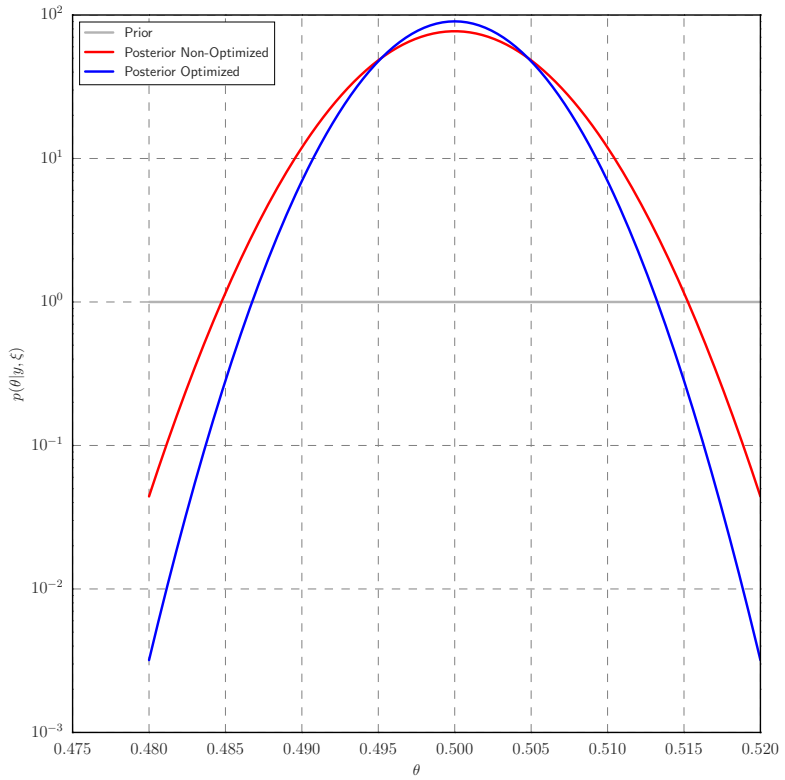

Figure 4: Benchmark problem 2 detail of posteriors before and after optimization.

Timoshenko beam

The Timoshenko beam is a beam where the shear strain is considered to be linear on its cross section. We propose the optimization of the positioning of a strain gauge that measures the horizontal strain – ε_{11} – and the $45°$strain – ε_{12} – to optimize the information about the Young and shear modulus – E and G respectively. Its governing equations are:

$$\begin{aligned} -GAK_s\frac{d}{dx_1}\left(\frac{dw_2}{dx_1}-\boldsymbol{\theta}\right) &= q_o \\ -EI\frac{d^2\boldsymbol{\theta}}{dx_1^2} - GAK_s\left(\frac{dw_2}{dx_1}-\boldsymbol{\theta}\right) &= 0 \end{aligned}$$

where

- L: length of the beam
- x_1: horizontal position of strain gauge
- x_2: vertical position of strain gauge
- K_s: Timoshenko's shear factor
- I: inertia moment
- A: cross section area.

For a simple supported beam of length L with uniform load q the forward model can be written as:

$$g(\boldsymbol{\xi},\boldsymbol{\theta}) = (\varepsilon_{11}(\boldsymbol{\xi},\boldsymbol{\theta}),\varepsilon_{12}(\boldsymbol{\xi},\boldsymbol{\theta})) = \left(\frac{x_2(\boldsymbol{\xi})\left(qLx_1(\boldsymbol{\xi})-qx_1^2(\boldsymbol{\xi})\right)}{2E(\boldsymbol{\theta})I}, \frac{\frac{L}{2}q-qx_1(\boldsymbol{\xi})}{K_sG(\boldsymbol{\theta})A}\right)$$

Parameters for the four cases studied are presented in Table 1 and results are presented in Table 2. Four cases will be tested. A case with greater variance and 3 experiments, a second case with less uncertainties on the prior information but with only one experiment, and cases 3 and 4 with, respectively, smaller variances on E and G. Figures 5-12 show the results of the 4 cases with the contours of the estimations of the expected information gain, norms of the gradients, distances to optima. The contours of the PDFs of the prior and posterior distributions before and after optimization are also shown for all cases.

Observing the results of the numerical tests it is clear that the optimization of the parameters improved the efficiency of the experiments, reducing the uncertainties about the material properties. It is interesting to note the influence of the prior knowledge on the optimal experimental set-up. On case 3 the prior standard deviation of E is reduced to $1/5$ of the on on case 2 and that shifted the optima to the supports of the beam, where more information is obtained about G. The opposite happens on case 4, where the standard deviation of G is reduced, instead of E. This results in a change on the expected information gain measured, shifting the optima to the mean of the beam. Moreover, the SGD performed well on finding the solution of the proposed problem.

Table 1: Parameters for the Timoshenko beam problem.

Parameter	μ_{pr}^{E}(GPa)	σ_{pr}^{E}(GPa)	μ_{pr}^{G}(GPa)	σ_{pr}^{G}(GPa)
Case 1	30.00	9.00	11.54	3.46
Case 2	30.00	6.00	11.54	2.31
Case 3	30.00	6.00	11.54	0.46
Case 4	30.00	1.20	11.54	2.31
Parameter	L (mm)	H (mm)	b (mm)	q (N/mm)
Case 1	10,000	2,000	100	-1,000
Case 2	10,000	2,000	100	-1,000
Case 3	10,000	2,000	100	-1,000
Case 4	10,000	2,000	100	-1,000
Parameter	Num. of Experiments	$\epsilon_{\varepsilon_{11}}$	$\epsilon_{\varepsilon_{12}}$	K_s
Case 1	3	0.625	0.130	5/6
Case 2	1	0.375	0.078	5/6
Case 3	1	0.375	0.078	5/6
Case 4	1	0.375	0.078	5/6

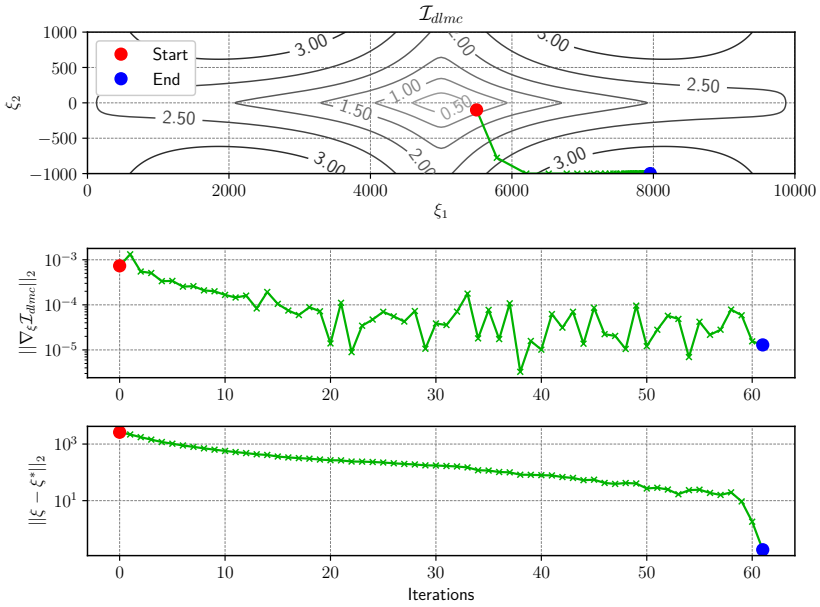

Figure 5: Information gain optimization of a Timoshenko beam for case 1.

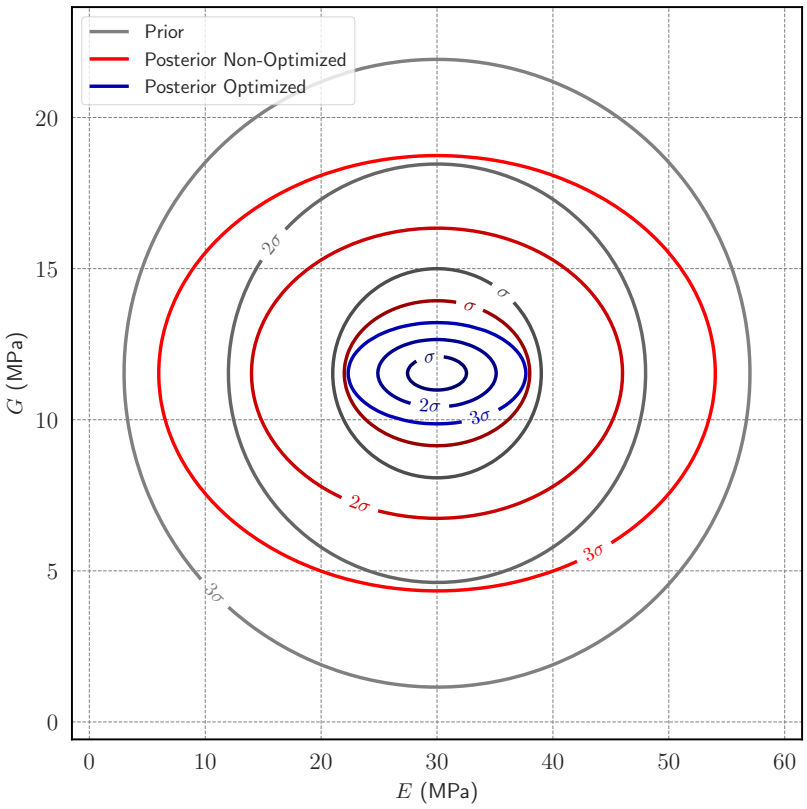

Figure 6: Information gain PDFs of a Timoshenko beam for case 1.

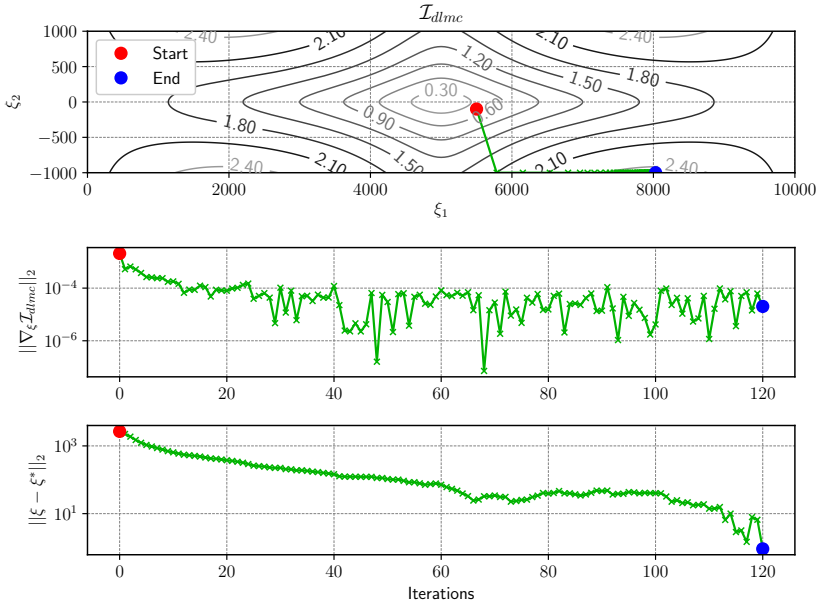

Figure 7: Information gain optimization of a Timoshenko beam for case 2.

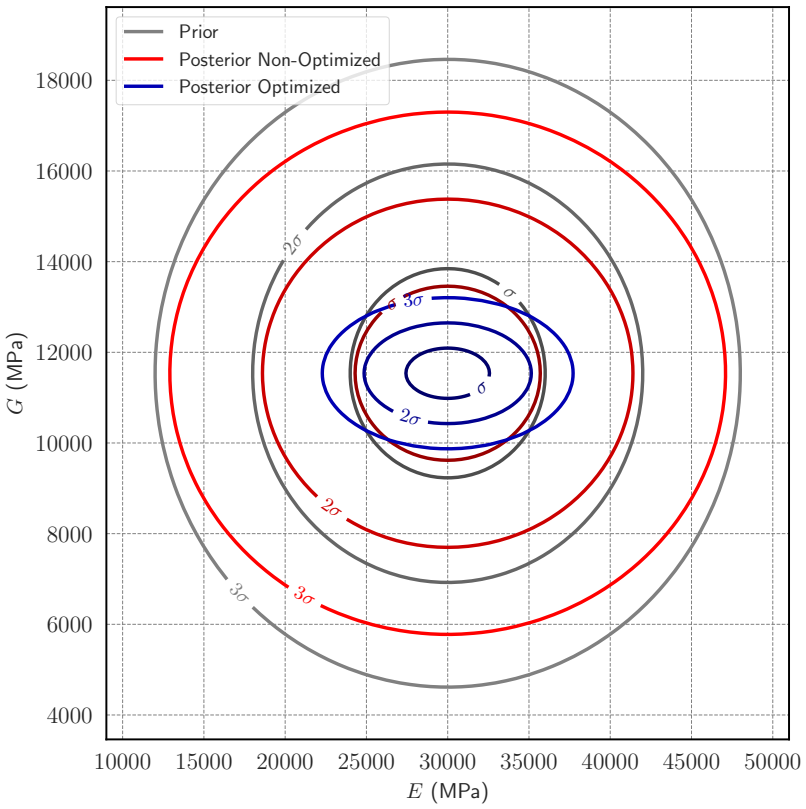

Figure 8: Information gain PDFs of a Timoshenko beam for case 2.

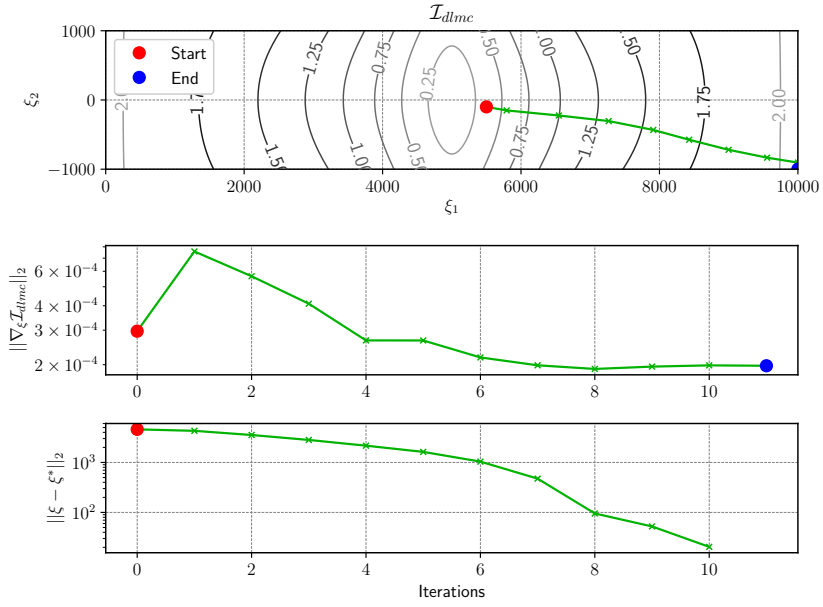

Figure 9: Information gain optimization of a Timoshenko beam for case 3.

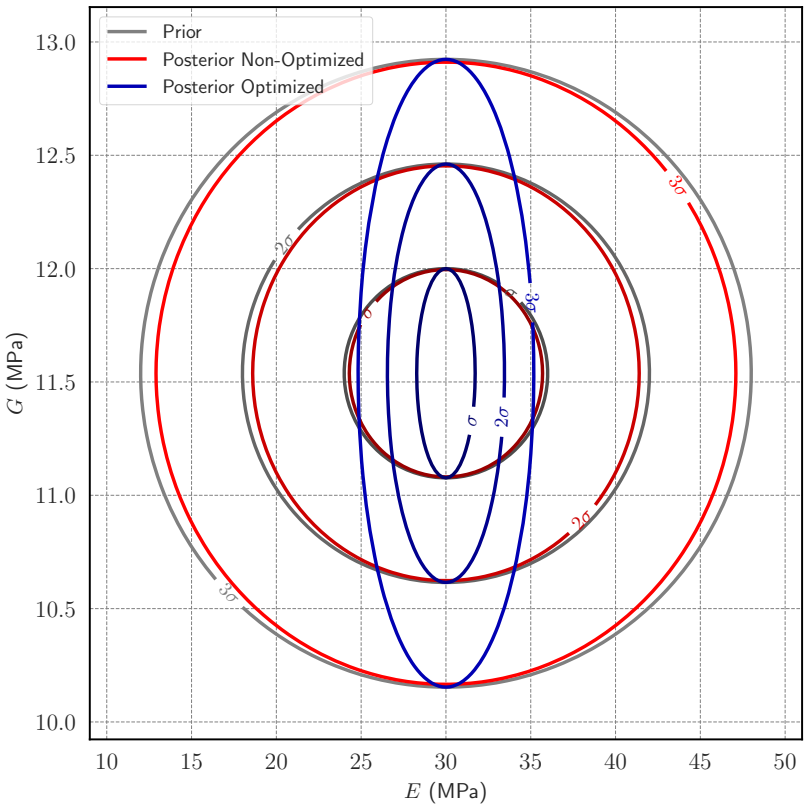

Figure 10: Information gain PDFs of a Timoshenko beam for case 3.

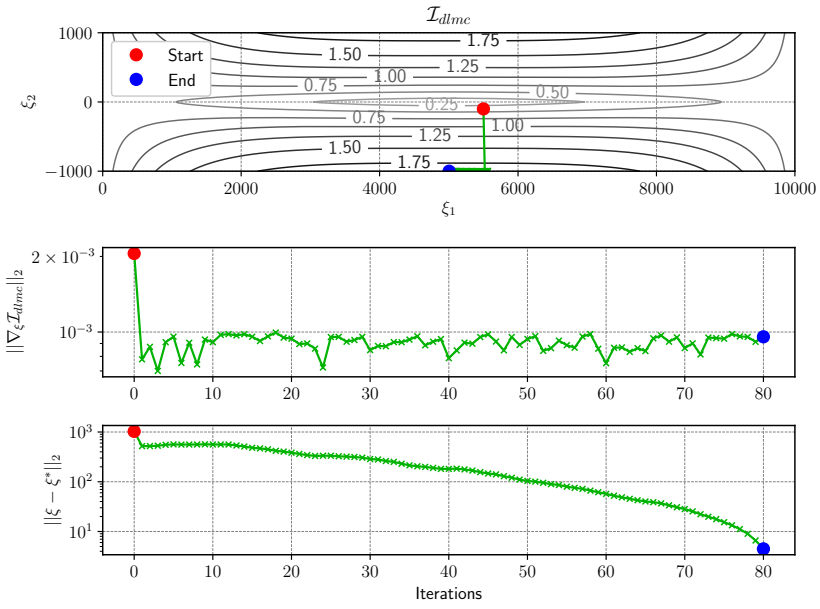

Figure 11: Information gain optimization of a Timoshenko beam for case 4.

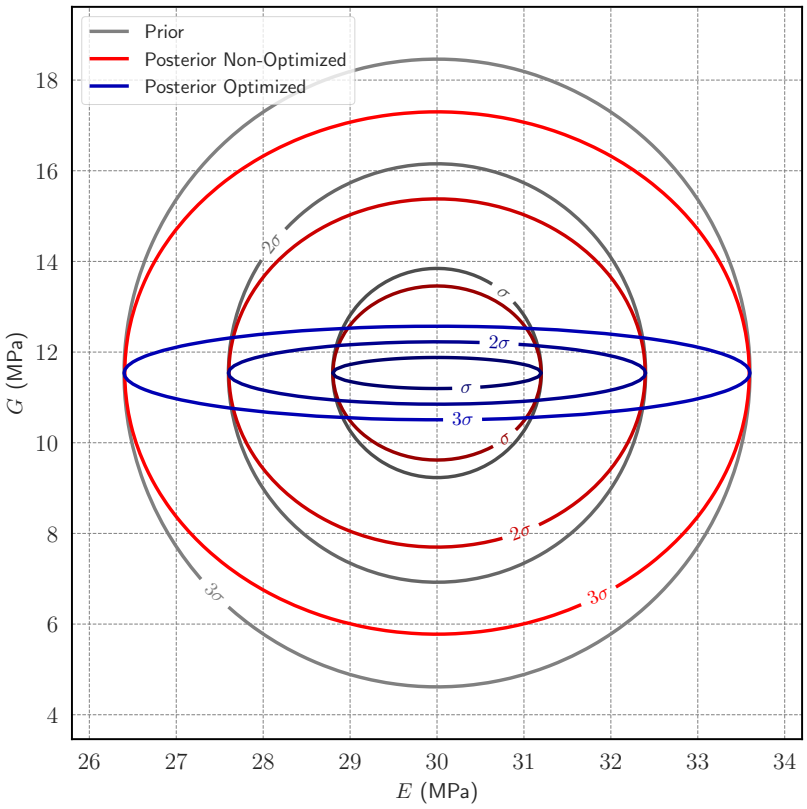

Figure 12: Information gain PDFs of a Timoshenko beam for case 4.

Table 2: Results for the Timoshenko beam problem.

		x_1^*(mm)	x_2^*(mm)	$I(\boldsymbol{\xi})$	σ_{post}^E (GPa)	σ_{post}^G (GPa)
Case 1	Non Opt.	5500.00	-100	0.14	8.00	2.40
	Opt.	7955.47	-1000.00	2.73	2.40	0.55
Case 2	Non Opt.	5500.00	-100	0.23	2.38	1.38
	Opt.	8028.77	-1000.00	2.35	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

CONCLUSION

On this paper we presented a method to reduce the cost of the optimization of Bayesian design by using the stochastic gradient. The SG reduces the cost of the optimization process in several orders of magnitude, even avoiding one of the two nested Monte Carlos needed to be evaluated each iteration. We tested the efficiency of the SGD on the solution of three different problems: two based on benchmark forward models and a third based on a physical model. On three cases the SGD succeeded in converging to minima, reducing uncertainty about the quantities of interest at the end of optimization even with the variance on the estimation of the gradient. On the third example, different prior distributions were tested and the consequences on the expected gain are clear. The expected information gain – and its maxima – depends heavily on the prior information, so badly formulated or wrong prior information can lead to bad estimates of the solution. Observing the results, we conclude that the SGD has shown to be a promising alternative to solve Bayesian design optimization.

REFERENCES

Ben Issaid, C. 2015. *Bayesian Optimal Experimental Design Using Multilevel Monte Carlo*. PhD thesis.

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

Robbins, H. and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Ryan, K. J. 2003. 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.