Engineering with Computers (2013) 29:127–138
DOI 10.1007/s00366-011-0249-3

ORIGINAL ARTICLE

Identification of quasi-optimal regions in the design space using surrogate modeling

Ivo Couckuyt · Jef Aernouts · Dirk Deschrijver · Filip De Turck · Tom Dhaene

Received: 17 June 2011 / Accepted: 20 December 2011 / Published online: 5 January 2012

Abstract The use of Surrogate Based Optimization (SBO) is widely spread in engineering design to find optimal performance characteristics of expensive simulations (forward analysis: from input to optimal output). However, often the practitioner knows a priori the desired performance and is interested in finding the associated input parameters (reverse analysis: from desired output to input). A popular method to solve such reverse (inverse) problems is to minimize the error between the simulated performance and the desired goal. However, there might be multiple quasi-optimal solutions to the problem. In this paper, the authors propose a novel method to efficiently solve inverse problems and to sample Quasi-Optimal Regions (QORs) in the input (design) space more densely. The development of this technique, based on the probability of improvement criterion and kriging models, is driven by a real-life problem from bio-mechanics, i.e., determining the elasticity of the (rabbit) tympanic membrane, a membrane that converts acoustic sound wave into vibrations of the middle ear ossicular bones.

I. Couckuyt (✉) · D. Deschrijver · F. De Turck · T. Dhaene
Department of Information Technology (INTEC),
Ghent University-IBBT, Gaston Crommenlaan 8,
Ghent 9050, Belgium
e-mail: ivo.couckuyt@ugent.be

D. Deschrijver
e-mail: dirk.deschrijver@ugent.be

T. Dhaene
e-mail: tom.dhaene@ugent.be

J. Aernouts
Laboratory of Biomedical Physics, University of Antwerp,
Groenenborgerlaan 171, Antwerp 2020, Belgium
e-mail: jef.aernouts@ua.ac.be

Keywords Inverse problem · Surrogate modeling · Kriging · Bio-mechanics

1 Introduction

This paper is concerned with efficiently solving complex, computational expensive design problems using surrogate modeling techniques [28]. Surrogate models, also known as metamodels, are cheap approximation models for computational expensive (black-box) simulations. Surrogate modeling techniques are well-suited to handle, for example, expensive finite element (FE) simulations, computational fluid dynamic (CFD) simulations and, of course, physical experiments. In particular, the research in this paper is concerned with deterministic computer codes, in contrast to non-deterministic (stochastic) problems.

Depending on the construction and usage of surrogate models several modeling flavors can be distinguished. Surrogate models can be built upfront to approximate the simulation code accurately over the entire input (design) space and, hence, can afterwards be used to replace the expensive code for design, analysis and optimization purposes. On the other hand, the construction of surrogate models can also be integrated in the optimization process. Usually, the latter case, known as Surrogate Based Optimization (SBO), generates surrogate models on the fly that are only accurate in certain regions of the input space, e.g., around optimal regions.

The construction of surrogate models as efficiently as possible is an entire research domain in itself. In order to come to an acceptable model, numerous problems and design choices need to be overcome (what data collection strategy to use, which variables are relevant, how to integrate domain knowledge, etc.). Other aspects of surrogate modeling include choosing the right type of approximation model for the problem at hand, a tuning strategy for the surrogate model parameters (=hyperparameters), and a performance measure to asses the accuracy of the surrogate model [14].

The general work-flow of surrogate modeling is illustrated in Fig. 1. First, an experimental design, e.g., from Design of Experiments (DOE), is specified and evaluated. Subsequently, surrogate models are built to fit this data as well as possible, according to a set of measures (e.g., cross-validation). The hyperparameters are estimated using an optimization algorithm. The accuracy of the set of surrogate models is improved until no further improvement can be obtained (or when another stopping criterion, such as a time limit, is met). If the stopping criteria are satisfied the process is halted and the final, best surrogate model is returned. On the other hand, when no stopping criterion is met, a sequential design strategy, also known as active learning or adaptive sampling, will select new data points to be evaluated and the surrogate models are updated with this new data.

Most often, surrogate models are used to solve so-called "*forward problems*". The practitioner is interested in the output or performance characteristics of the simulation system given the input (design) parameters. The surrogate models define the mapping between the input space (design space) and the output space (performance space). Examples of forward problems are found in validation and verification, sensitivity analysis, and optimization.

In contrast, the focus of the "*reverse (inverse) problem*" is on exploring the input space. Ideally, a surrogate model could be created that maps the output parameters to the input parameters (as opposite to forward modeling) of the complex system over the entire output space. However, many inverse problems are ill-posed. Considering Hadamard's definition of ill-posedness [15], the two outstanding problems hampering the creation of a full inverse surrogate model are non-uniqueness and instability. A good overview of the associated intricacies is presented by Barton in [3]. For all the above reasons, the inverse problem is often reduced to the task of finding one (or more) input parameter combination for a certain output characteristic. Still, it is possible that,

1. No such input parameter combination exists.
2. More than one input parameter combination satisfies the given output characteristic(s).

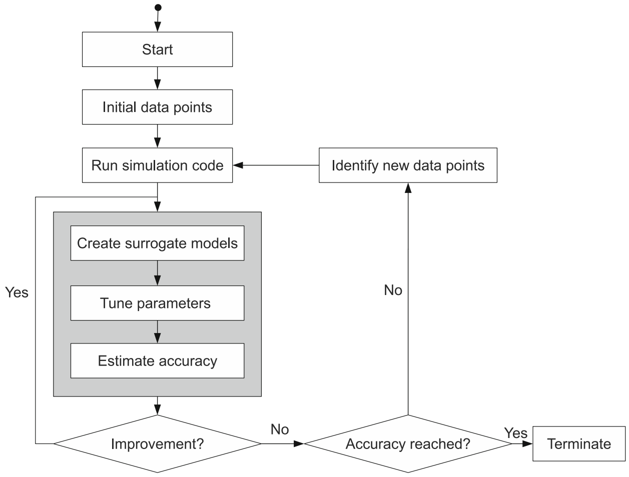


Fig. 1 Flow chart of the surrogate modeling process [14]

A typical inverse problem is the estimation of some (physical) material or design parameter, e.g., the permittivity of a substrate [5] or the elasticity of rubber [1], given the desired output or system behavior. A popular solution is to convert the reverse problem to a (forward) optimization problem. Namely, a simulation model is constructed, parametrized by the properties or design parameters of interest. By minimizing the error between the parametrized simulation model and the measured data the input parameters (material properties) of the simulation model are obtained that correspond with the measurements or desired output, see Fig. 2.

The focus of this paper is to efficiently solve inverse problems where an infinite number of input parameter combinations is possible, i.e., whole regions in the input space that satisfy the cost function sufficiently well. From now on these regions in the input space will be denoted Quasi-Optimal Regions (QORs). Hence, traditional (surrogate-based) optimization of the cost function is insufficient.

Recently, Picheny et al. [20] presented a scheme to sample the input regions that correspond to an output region of interest. While using a similar approach, our work focusses on finding QORs as efficiently as possible. Moreover, the approach of Picheny et al. requires expensive numerical integration and the kriging model must be updated for every new sample point (though alternative, faster approaches are discussed). We present in this work a simple and cheap criterion in combination with a space-filling criterion to ensure proper coverage of the input domain.

The main contribution of this paper is a novel sequential design strategy, denoted as the "*QOR sampling algorithm*", that is able to efficiently sample QORs densely, in a quasi-uniform way. The surrogate model of choice is the Gaussian Process (GP) based kriging. Kriging is a popular surrogate model for the approximation of deterministic

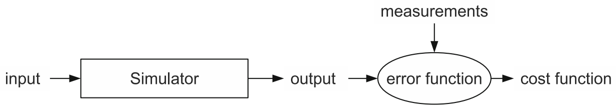


Fig. 2 The inverse problem is often solved by minimizing the error function between the simulation output and the measured data

computer code [22]. GPs enable the use of statistical infill criteria in a sequential design strategy. The presented method consists of the extension of such a statistical infill criterion, namely, the Probability of Improvement (PoI) [17], and a new search strategy to exploit the infill criterion.

The QOR sampling algorithm has been implemented in a flexible research platform for surrogate modeling, the **SU**rrogate **MO**deling (SUMO) Toolbox [14] (The SUMO Toolbox can be downloaded from: http://sumo.intec.ugent.b. An AGPL open source license is available for research purposes) and has been applied to a real-life problem from bio-mechanics, i.e., determining the elasticity of the (rabbit) tympanic membrane, a membrane that converts acoustic sound wave into vibrations of the middle ear ossicular bones.

Section 2 gives a brief overview of kriging, including the extension of kriging to handle noisy cost functions, which are typical for inverse problems. Section 3 describes the use of infill criteria and, in particular, describes an extension of PoI needed to solve the engineering problem at hand. A new search strategy to exploit infill criteria is described in Sect. 4. The QOR sampling algorithm is applied to the Branin and Hartman functions in, respectively, Sects. 5 and 6. The engineering problem from bio-mechanics is presented in Sect. 7. Details of the SUMO Toolbox configuration are found in Sect. 7.2. Results of the engineering problem and conclusions form the last two sections of this paper, i.e., Sects. 7.3 and 8.

2 Kriging

Kriging is a popular surrogate model to approximate deterministic noise-free data. First conceived by Danie Krige in geostatistics, these Gaussian Process [13] based surrogate models are compact and cheap to evaluate, and have proven to be very useful for tasks such as optimization [18], design space exploration, visualization, prototyping, and sensitivity analysis [28].

A thorough mathematically treatment of kriging is given by [11, 22, 23]. Basically, kriging is a two-step process: first a regression function $f(\mathbf{x})$ is constructed based on the data, and, subsequently, a Gaussian process $Z(\mathbf{x})$ is constructed through the residuals.

$$Y(\mathbf{x}) = f(\mathbf{x}) + Z(\mathbf{x}), \tag{1}$$

where $f(\mathbf{x})$ is a regression function and Z is a Gaussian process with mean 0, variance σ^2 and a correlation matrix Ψ.

Assume a set of n samples, $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ in d dimensions (see Eq. 2) and associated function values, $\mathbf{y} = (y_1, \ldots, y_n)'$, where $(\cdot)'$ is the transpose of a vector or matrix.

$$X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)' = \begin{pmatrix} x_{1,1} & \ldots & x_{1,d} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \ldots & x_{n,d} \end{pmatrix} \tag{2}$$

Essentially, the regression part is encoded in the $n \times p$ model matrix F using basis functions $b_i(\mathbf{x})$ for $i = 1 \ldots p$ (e.g., a power base for polynomials),

$$F = \begin{pmatrix} b_1(\mathbf{x_1}) & b_2(\mathbf{x_1}) & \cdots & b_p(\mathbf{x_1}) \\ \vdots & \vdots & \vdots & \vdots \\ b_1(\mathbf{x_n}) & b_2(\mathbf{x_n}) & \cdots & b_p(\mathbf{x_n}) \end{pmatrix},$$

while the stochastic process is mostly defined by the $n \times n$ correlation matrix Ψ,

$$\Psi = \begin{pmatrix} \psi(\mathbf{x}_1, \mathbf{x}_1) & \ldots & \psi(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \psi(\mathbf{x}_n, \mathbf{x}_1) & \ldots & \psi(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix},$$

where $\psi(\cdot,\cdot)$ is the correlation function. $\psi(\cdot,\cdot)$ is parametrized by a set of hyperparameters $\boldsymbol{\theta}$, which are identified by Maximum Likelihood Estimation (MLE) (though other approaches are possible). Subsequently, the Best Linear Unbiased Predictor (BLUP) of kriging is derived as,

$$\hat{y}(\mathbf{x}) = M\alpha + r(\mathbf{x}) \cdot \Psi^{-1} \cdot (\mathbf{y} - F\boldsymbol{\alpha}), \tag{3}$$

where $M = \big(b_1(\mathbf{x})\, b_2(\mathbf{x}) \ldots b_p(\mathbf{x})\big)$ is the model matrix of the predicting point $\mathbf{x}$, α is a $p \times 1$ vector denoting the coefficients of the regression function, determined by Generalized Least Squares (GLS), and $r(\mathbf{x})$ is an $1 \times n$ vector of correlations between the point $\mathbf{x}$ and the samples X.

Note that kriging is an interpolation technique. This is easily seen by substituting the ith sample point $\mathbf{x_i}$ in the BLUP (Equation 3) and considering that $r(\mathbf{x_i})$ is the ith column of Ψ, hence, $r(\mathbf{x_i}) \cdot \Psi^{-1}$ is an unit vector e_i with a 1 at the ith position,

$$\hat{y}(\mathbf{x_i}) = M\alpha + e_i \cdot (\mathbf{y} - F\alpha) = M\boldsymbol{\alpha} + y_i - M\alpha = y_i. \tag{4}$$

While this is a nice property for many simulation problems, it might produce undesired results when dealing with stochastic simulations and/or in the presence of noise. Therefore, the formal work of Staum et al. [27] is adapted in this paper to extend kriging for approximation instead of interpolation, also known as regression kriging or stochastic kriging. To that end, the noise is modeled as a separate Gaussian process $\xi(\mathbf{x})$,

$$Y(\mathbf{x}) = f(\mathbf{x}) + Z(\mathbf{x}) + \xi(\mathbf{x}), \tag{5}$$

where ξ is a Gaussian process with mean 0, variance τ^2 and correlation matrix $\sum$. The BLUP then becomes,

$$\hat{y}(\mathbf{x}) = M\alpha + r(\mathbf{x}) \cdot \left(\Psi + \frac{1}{\sigma^2}\sum\right)^{-1} \cdot (\mathbf{y} - F\alpha), \tag{6}$$

where $\frac{1}{\sigma^2}\sum$ is a matrix resembling signal-to-noise ratios. Depending on the type and distribution of noise the matrix $\sum$ has different forms. For the problem in this paper it can be assumed that the noise is homogeneous distributed across the input space. This results in a scalar value (10^{λ}) on the diagonal of the correlation matrix, i.e., $\sum = 10^{\lambda} I_n$.

While in stochastic simulation the matrix $\sum$, and thus λ, can be created based on repeated simulations this is not true for a deterministic simulation problem, as presented in this paper. Therefore the variable λ is estimated as part of the likelihood optimization of kriging. After the kriging surrogate model has been constructed an estimate of the noise variance τ^2 is calculated as $\hat{\tau}^2 = 10^{\lambda}\sigma^2$.

3 Infill criteria

In engineering, infill criteria are (sampling) functions, also known as figures of merit or metrics, that measure how *interesting* a data point is in the input space. Starting from an initial approximation of the simulation system, new sample points (infill or update points) are selected based on an infill criterion. The scope of infill criteria ranges from increasing the accuracy of the prediction (e.g., for creating globally accurate surrogate models) to the prediction itself to facilitate optimization. In global SBO, it is crucial that the infill criterion is a balance between exploration (enhancing the overall accuracy of the surrogate model) and exploitation (enhancing the accuracy of the surrogate model solely in the region of the (current) optimum).

A well-known infill criterion that is able to effectively solve this trade-off is Expected Improvement (EI), which has been popularized by Jones et al. [11, 18, 25] as the Efficient Global Optimization (EGO) algorithm. Jones wrote an excellent discussion regarding the infill criteria approach in [17]. Subsequently, Sasena compared different infill criteria for optimization and investigated extensions of those infill criteria for constrained optimization problems in [24].

3.1 Probability of Improvement (PoI)

Among several statistical infill criteria investigated by Jones the Probability of Improvement (PoI) is used and generalized in this work. The PoI Eq. (7), defined below, can be interpreted graphically (see Fig. 3). At $\mathbf{x} = 0.5$, a Gaussian probability density function (PDF) is drawn and expresses the uncertainty about the predicted function value of a sampled and unknown function $y = f(\mathbf{x})$. Thus, the uncertainty at any point $\mathbf{x}$ is treated as the realization of a random variable $Y(\mathbf{x})$ with mean $\hat{y} = \hat{f}(\mathbf{x})$ (=prediction) and variance $\hat{s}^2 = \hat{\sigma}^2(\mathbf{x})$ (=prediction variance). The prediction variance can be corrected to account for the estimation of the hyperparameters [19]. Assuming the random variable $Y(\mathbf{x})$ is normally distributed, then the shaded area under the Gaussian PDF is the *PoI* of any newly calculated function value $f(\mathbf{x})$ over the intermediate minimum function value $\hat{f}_{\min}$ (the dotted line). PoI is denoted as $P(Y(\mathbf{x}) \leq \hat{f}_{\min})$ i.e.,

$$\begin{aligned} PoI = P(Y(\mathbf{x}) \leq \hat{f}_{\min}) &= \int_{-\infty}^{\hat{f}_{\min}} Y(\mathbf{x})dY \\ &= \Phi\left(\frac{\hat{f}_{\min} - \hat{y}}{\hat{s}}\right), \end{aligned} \tag{7}$$

where $\Phi(t)$ is the standard normal cumulative distribution function $\Phi(t) = \frac{1}{2}[1 + erf(\frac{t}{\sqrt{2}})]$ and $erf(\cdot)$ is the error function.

PoI or any other statistical infill criteria (e.g., EI) are optimized over $\mathbf{x}$ to find the subsequent data point to evaluate. Note, however, that besides the prediction $\hat{y} = \hat{f}(\mathbf{x})$ of the surrogate model, a point-wise variance estimation $\hat{s}^2 = \hat{\sigma}^2(\mathbf{x})$ of the surrogate is also required. Both predictions ($\hat{y}$ and $\hat{s}^2$) are provided by the kriging surrogate model.

3.2 Generalized Probability of Improvement (gPoI)

While PoI is a very useful infill criterion for optimization, it only focuses on the global optimum, not on a range of output values. The authors extend the idea of the PoI criterion to allow identification of an arbitrarily band in the output space. Let $[T_1,T_2]$ be the range of interest in the output space. The generalized Probability of Improvement (gPoI) is defined as the probability that the function value $f(\mathbf{x})$ at a point $\mathbf{x}$ lies with the output range $[T_1,T_2]$,

$$\begin{aligned} gPoI(\mathbf{x}) = P(T_1 \leq Y(\mathbf{x}) \leq T_2) &= \int_{T_1}^{T_2} Y(\mathbf{x})dY \\ &= P(T_2 \leq Y(\mathbf{x})) - P(T_1 \leq Y(\mathbf{x})) \\ &= \int_{-\infty}^{T_2} Y(\mathbf{x})dY - \int_{-\infty}^{T_1} Y(\mathbf{x})dY \\ &= \Phi\left(\frac{T_2 - \hat{y}}{\hat{s}}\right) - \Phi\left(\frac{T_1 - \hat{y}}{\hat{s}}\right), \end{aligned} \tag{8}$$

where $\Phi(t)$ is the standard normal cumulative distribution function $\Phi(t) = \frac{1}{2}[1 + erf(\frac{t}{\sqrt{2}})]$ and $erf(\cdot)$ is the error function. Note that the standard abbreviation "PoI" is not well-suited anymore as the focus is now on sampling an interval instead of improving the optimum.

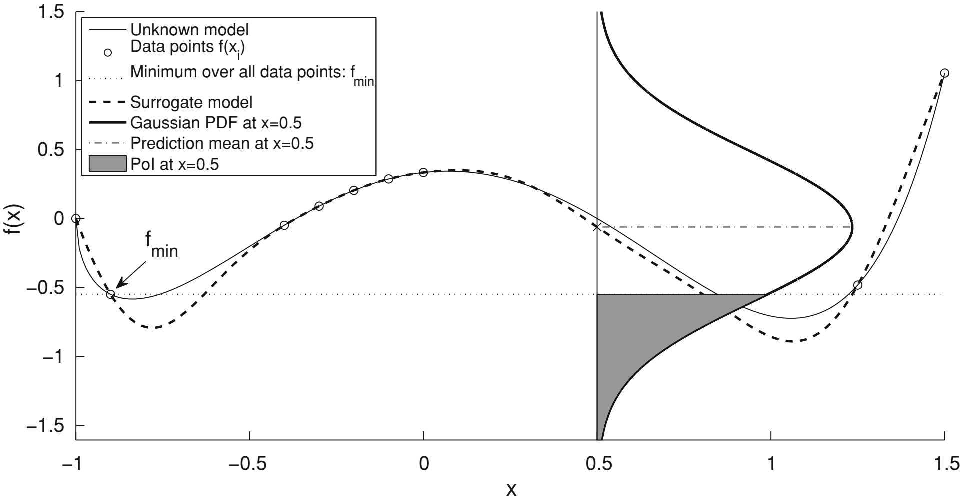


Fig. 3 Graphical illustration of a Gaussian Process (GP) and Probability of Improvement (PoI). A surrogate model (*dashed line*) is constructed based on some data points (*circles*). For each point the surrogate model predicts a Gaussian probability density function (PDF). E.g., at $x = 0.5$ an example of such a PDF is drawn. The volume of the shaded area is the PoI over the minimum function value $f_{\min}$

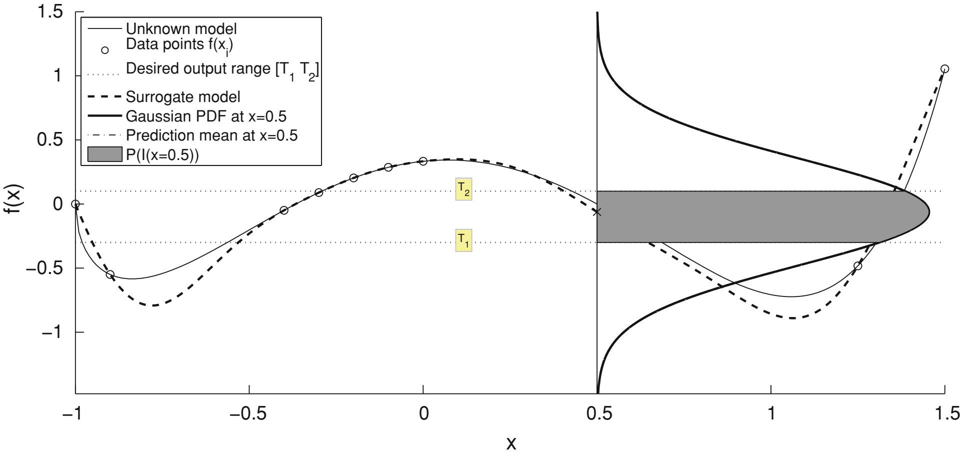


Fig. 4 Graphical illustration of a Gaussian Process (GP) and the generalized Probability of Improvement (gPoI). A surrogate model (*dashed line*) is constructed based on some data points (*circles*). For each point the surrogate model predicts a Gaussian probability density function (PDF). E.g., at $x = 0.5$ an example of such a PDF is drawn. The volume of the *shaded area* is the gPoI based on the desired output range $[T_1,T_2]$

The gPoI criterion has recently been successfully applied to quasi-uniformly sample the region(s) in the input space that correspond to a desired interval $[T_1,T_2]$ in the output space [6] (T_1 and T_2 are defined upfront). In essence, input (design) parameters are sought that correspond with a certain set of performances (=inverse problem).

In this paper, one wants to find the QORs which include all near-optimal solutions. So, the lower bound can be defined as $T_1 = -\infty$ and T_2 is defined on the fly. By varying and tightening the upper bound T_2 of the integral (see Eq. 8) dynamically during the optimization process the QOR can be accurately identified.

The authors suggests to use the intermediate minimal function value $\hat{f}_{\min}$ plus a percentage p of $|\hat{f}_{\min}|$ as upper bound, namely, $T_2 = \hat{f}_{\min} + p \cdot |\hat{f}_{\min}|$. Thus, all input parameter combinations are sought that lie within a desired percentage p of $\hat{f}_{\min}$.

Furthermore noise can be taking into account by adding an extra offset to the upper bound. This might be required as many inverse problems involve noisy measurements, as will be shown in the application of Sect. 7. To that end, regression kriging [27] is used, as explained in Sect. 2, where a parameter 10^{λ} gives an indication of the amount of noise. Furthermore, as λ is determined during the MLE of kriging an estimate of the noise variance τ^2 can be calculated as $\hat{\tau}^2 = 10^{\lambda}\sigma^2$.

Assuming the noise being Gaussian distributed, the 68% confidence interval on the exact (intermediate) minimum function value $f_{\min}$ is given by $[f_{\min} - \alpha\hat{\tau}, f_{\min} + \alpha\hat{\tau}]$, where $\alpha = 1$ (95% confidence intervals can be obtained by using $\alpha = 2$). Assuming the intermediate lowest (noisy) function value $\hat{f}_{\min}$ is the lower bound, namely, $\hat{f}_{\min} = f_{\min} - \alpha\hat{\tau}$, then it is easy to see that the upper bound can be expressed in terms of $\hat{f}_{\min}$, namely, $f_{\min} + \alpha\hat{\tau} = \hat{f}_{\min} + 2\alpha\hat{\tau}$. Thus, in sum, assuming that the measurements errors are homogeneous distributed in the input space and the estimated λ is correct, the upper bound is defined by $T_2 = (\hat{f}_{\min} + 2\alpha\hat{\tau}) + p \cdot |\hat{f}_{\min} + 2\alpha\hat{\tau}|$. Note that this is by no means an unerring formula. While no extra parameters are introduced the estimate of λ may not be accurate and the type of error is often unknown.

Unlike PoI, the gPoI cannot be simply optimized over **x** to identify new data samples. Available samples that are already lying in the desired output range [T_1, T_2] have a high probability (see Fig. 6a), and, hence, straightforward optimization of the criterion might result in duplicate samples. Other (space-filling) strategies have to be devised that makes fully use of the information provided by gPoI. This problem is further explored in the next section.

4 Search strategies

The techniques explained in the previous sections (e.g., EI, PoI, gPoI, etc.) are all utility functions. Such utility functions are used to identify interesting new points in a sequential design strategy. Various search strategies exist to exploit these utility functions.

For instance, the original EGO algorithm [18] simply optimizes the EI. In particular, the deterministic branch and bound methodology was used to find the global optimum. To that end, a convex upper bound had to be calculated for the EI. However, if one wants to use other utility functions (or other types of surrogate models) this upper bound must be redefined. In later work more black-box optimization methods were used with similar results, e.g., the DIviding REctangles (DIRECT) [16] or an extensive pattern search. Moreover, multi-modal optimization methods are suggested in literature (e.g., in [21, 26]) to select multiple samples in one iteration, taking full advantage of parallel computing.

Algorithm 1 Pseudocode of the QOR sampling algorithm.

```
$samples_0 =$ generateInitialDesign() // e.g., a Latin Hypercube Design
$values_0 =$ simulate($samples_0$) // call the simulation code

$T_1 =$ lowerbound // fraction of the maximal function value, e.g., 0.5
$T_2 =$ upperbound // e.g., infinity

$i = 0$
while $|samples_i| < maxSamples$
  $krige_i =$ fitKriging($samples_i, values_i$) // Build surrogate model

  $p_{test} =$ generateTestPoints($n = 100 \times |samples_i|$) // Generate n test points
  $score_1 =$ gPoI($krige_i, p_{test}, T_1, T_2$) // evaluate gPoI, see Equation (8)
  $score_2 =$ MD($p_{test}$) // evaluate the minimum distance, see Equation (9)
  $score = w_1 \times score_1 + w_2 \times score_2$ // weighted global score, e.g., $w_1 = w_2 = 0.5$
  $p_{new} =$ selectBestPoints($p_{test}, score, m = 10$) // select $m$ best points from $p_{test}$
  $y_{new} =$ simulate($p_{new}$) // call the simulation code

  $samples_{i+1} = samples_i \cup p_{new}$
  $values_{i+1} = values_i \cup y_{new}$
  $i = i + 1$
end
```

Global optimization methods are not suited to directly exploit the new gPoI criterion because there might exist multiple (or even an infinite number of) solutions. Therefore, the authors adapt a generic sampling framework for sequential design [7]. The algorithmic flow is depicted in Fig. 5. The search strategy is configured as followed: First, n candidate samples are drawn from the uniform distribution. Subsequently, these candidates are ranked according to two ($k = 2$) criteria: the gPoI criterion (see Fig. 6a) and a Minimum Distance (MD) criterion that calculates the Euclidean distance to the closest sample. The MD criterion is defined by,

$$\mathrm{MD}(\mathbf{x}) = \frac{(l+1)^{(1/l-1)}}{2} \min_{\mathbf{p} \in \text{samples}} \sqrt{\sum_{j=1}^{d} (x_j - p_j)^2}, \tag{9}$$

where d is the number of input parameters and the factor $\frac{(l+1)^{(1/l-1)}}{2}$ (upper bound estimate) scales the MD criterion into the same range as the gPoI criterion, namely [0, 1]. The estimate on the upper bound is calculated as follows, if the current number of samples is l, the optimal maximin configuration of these samples is a $\sqrt{l} \times \sqrt{l}$ uniform grid. Hence, the maximin distance of this layout is a maximum and can be used as an upper bound. Maximizing the MD criterion takes care of the space-filling properties in the input space (see Fig. 6b). Furthermore, kriging implies that new input points close to existing samples result in highly correlated output values. Hence, dense clusters of points do not provide much new valuable information and are avoided by the search strategy. The combined sequential sampling criterion is now defined as the weighted average of the two criteria, see Fig. 6c. Pseudocode of the QOR sampling algorithm is found in Algorithm 1.

This approach is a nice balance between exploration (space-filling) and exploitation (uniform sampling in the input range that satisfies the QORs). If the gPoI criterion is low across the whole input space the MD criterion will dominate and, hence, the input space will be further

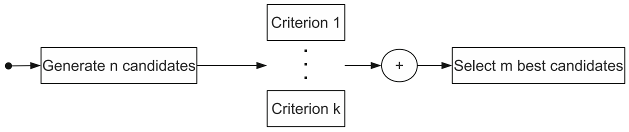


Fig. 5 General flow of a sequential design strategy

explored, enhancing the accuracy of the surrogate model. On the other hand, as the number of samples increases, the influence of the MD score will decrease, enabling the exploitation of the gPoI criterion.

5 Example 1: determining the QORs of the Branin function

5.1 Problem setting

The Branin function is a well-known benchmark function for optimization, it has two input variables (x_1, x_2) and its equation is given by,

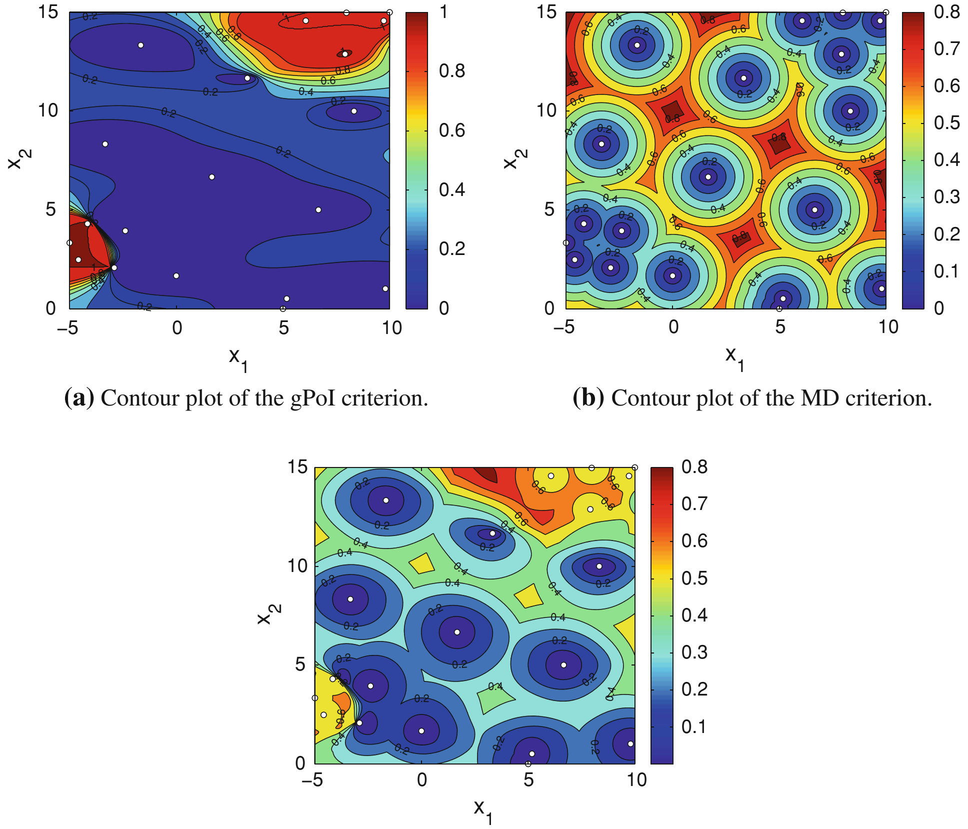
$$f(x_1,x_2) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10. \qquad (10)$$

In this research, the goal is not to identify the unique optimum of the Branin function over the design space, but the goal is to sample the regions corresponding to the 50% highest function values densely in a space-filling way. So, in this section the QORs correspond with the top 50% of the Branin function.

5.2 Experimental setup

Version 7.0.2 of the SUMO toolbox is used to determine the QORs of (10). An initial set of 10 samples is generated by an optimal maximin Latin Hypercube Design (LHD; [8]). Subsequently, 90 infill points are selected based on the gPoI and MD figures of merit as discussed in Sect. 4. To find the QORs, we adapt the bounds $[T_1, T_2]$ of the gPoI criterion in consecutive steps, namely, $[T_1, T_2] = [\hat{f}_{\max} - 0.5 \cdot |\hat{f}_{\max}|, \infty]$, where $\hat{f}_{\max}$ is the intermediate maximal function value. Samples are selected in batches of $m = 10$ and the sequential sampling is halted when the number of samples reaches 100. Thus, after the initial set of 10 samples, we have a total of $20, 30, \ldots, 90, 100$ samples. The kriging surrogate model is configured using the standard Gaussian correlation function and a constant regression function. The hyperparameters, including the λ parameter, are efficiently estimated using SQPLab [4] (http://www-rocq.inria.fr/~gilbert/modulopt/optimization-routines/sqplab/sqplab.html), utilizing likelihood derivative information.

Fig. 6 Example 1: Snapshot of the two sequential sampling criteria (gPoI and MD) and the combined weighted average at 20 samples (*white dots*) for the 2D Branin function



(a) Contour plot of the gPoI criterion.

(b) Contour plot of the MD criterion.

(c) Contour plot of the combined criteria.

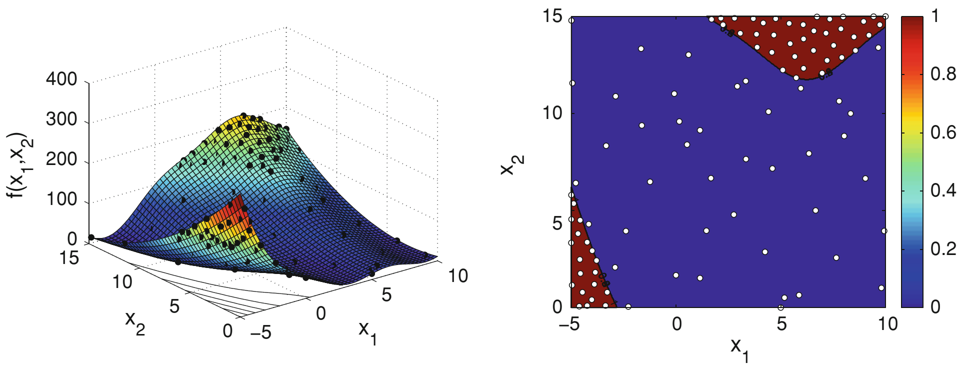


Fig. 7 Example 1: Final results of the Branin function (100 samples). It is seen that the QORs of interest are densely samples in an uniform way (samples are denoted by the *dots*)

(a) Landscape plot of the kriging model.

(b) Contour plot of the gPoI criterion (based on the kriging model).

5.3 Results

An intermediate snapshot of the different sampling criteria (at 20 samples) is shown in Fig. 6. The gPoI emphasizes the highest regions of the Branin function (i.e., exploiting the function behavior) while the MD criterion takes care of the exploration aspect. The combination of the two criteria is the actual metric used to select new samples in a sequential way. A landscape plot of the final kriging model and corresponding gPoI contour plot is shown in Fig. 7. Note that, while the QORs are sampled densely, other regions of the input space have not been neglected completely.

6 Example 2: determining the QORs of the Hartman function

6.1 Problem setting

The six-dimensional Hartman function is another well-known benchmark function for optimization, the Hartman equations are given by,

$$A = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix},$$

$$c = (1 \;\; 1.2 \;\; 3 \;\; 3.2)',$$

$$p = \begin{pmatrix} 0.1312 & 0.1696 & 0.5569 & 0.0124 & 0.8283 & 0.5886 \\ 0.2329 & 0.4135 & 0.8307 & 0.3736 & 0.1004 & 0.9991 \\ 0.2348 & 0.1451 & 0.3522 & 0.2883 & 0.3047 & 0.6650 \\ 0.4047 & 0.8828 & 0.8732 & 0.5743 & 0.1091 & 0.0381 \end{pmatrix},$$

$$f(x_1,\ldots,x_6) = -\sum_{i=1}^{4} c_i \cdot \exp\left(-\sum_{j=1}^{6} A_{i,j}(x_j - p_{i,j})^2\right). \qquad (11)$$

The goal is to sample the QORs corresponding to the 50% lowest function values densely in a space-filling way.

6.2 Experimental setup

Version 7.0.2 of the SUMO toolbox is used to determine the QORs of (11). An initial set of 51 samples is generated by an optimal maximin Latin Hypercube Design (LHD; [8]). Subsequently, 950 infill points are selected based on the gPoI and MD figures of merit as discussed in Sect. 4. To find the QORs, we adapt the bounds [T_1, T_2] of the gPoI criterion in consecutive steps, namely, $[T_1, T_2] = [-\infty, \hat{f}_{\min} + 0.5 \cdot |\hat{f}_{\min}|]$, where $\hat{f}_{\min}$ is the intermediate minimal function value. Samples are selected in batches of $m = 10$ and the sequential sampling is halted when the number of samples exceeds 1000. Hence, after the initial set of 51 samples, we have a total of $61, 71, \ldots, 511, 521, \ldots, 991$, 1001 samples. The kriging surrogate model is configured using the standard Gaussian correlation function and a constant regression function. The hyperparameters, including the λ parameter, are estimated using SQPLab.

6.3 Results

The number of samples (in percent) that are inside the QORs versus the number of evaluated samples is given in Fig. 8. Only 1.9% (=1 sample) of the initial design of 51 samples satisfies the QORs, using the exact global minimal function value $f_{\min}$ in the bound calculation (in contrast to the estimated $\hat{f}_{\min}$). As the search progresses the number of samples that satisfies the output range increases rapidly. At 301 samples 53% of the output values (=160 samples) lie within the desired range. The slight decline from 600 samples onwards is due to the saturation of the QORs with samples, hence, the QOR sampling algorithm starts focussing more on exploring the input domain. A similar trend is observed when using the estimated $\hat{f}_{\min}$ to calculate the bounds of the QORs.

Of particular interest is the observation that identifying QORs is less prohibited by the curse of dimensionality than

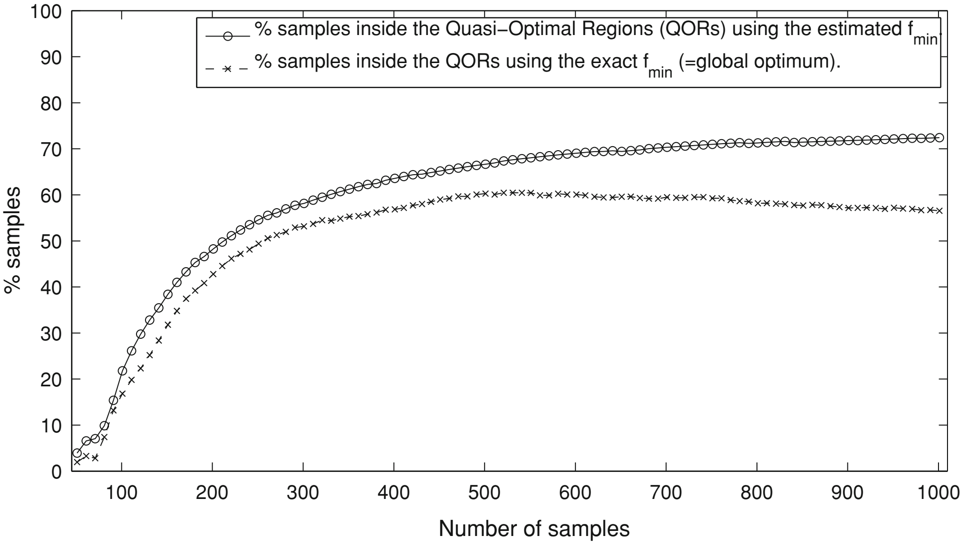


Fig. 8 Evolution of the number of samples (in percent) inside the QORs for the Hartman function

creating a global accurate surrogate model, but more expensive than SBO. While SBO methods only need to evaluate a series of points towards finding the optimum, a global accurate surrogate model needs exponentially more data points to cover the whole input domain, whereas the QOR sampling algorithm only needs enough data points to identify and uniformly cover the QORs.

7 Example 3: elasticity of the middle ear tympanic membrane

7.1 Problem setting

In hearing science, finite element modeling is commonly used to study the mechanical behavior of the middle ear, e.g. [12]. In such models, tympanic membrane elasticity parameters have a significant influence on the output [10]. However, good data for the mechanical properties of the tympanic membrane are still lacking [9].

In order to fill this gap, a setup was developed to determine tympanic membrane elasticity in situ by Aernouts et al. [1]. The characterization method consists of four steps: (1) doing a point indentation perpendicular on the membrane surface; (2) measuring the indentation depth, the resulting force and the three-dimensional shape data; (3) simulating the experiment with a finite element model, and (4) adapting the model to fit the measurements using optimization procedures. A detailed description of the application of this method on a rabbit tympanic membrane sample is given in [2].

The tympanic membrane sample (in this case obtained from a rabbit) was placed on a translation and rotation stage, a schematic drawing is shown in Figure 9a. Indentations in and out in a direction perpendicular to the surface membrane were carried out using a stepper motor with

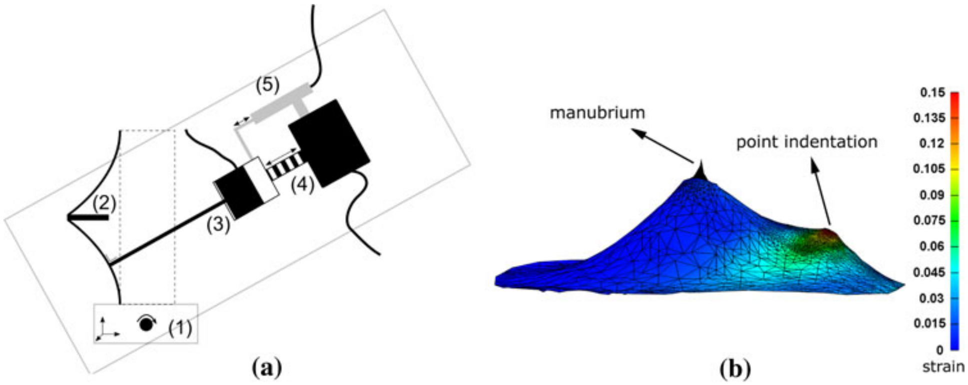
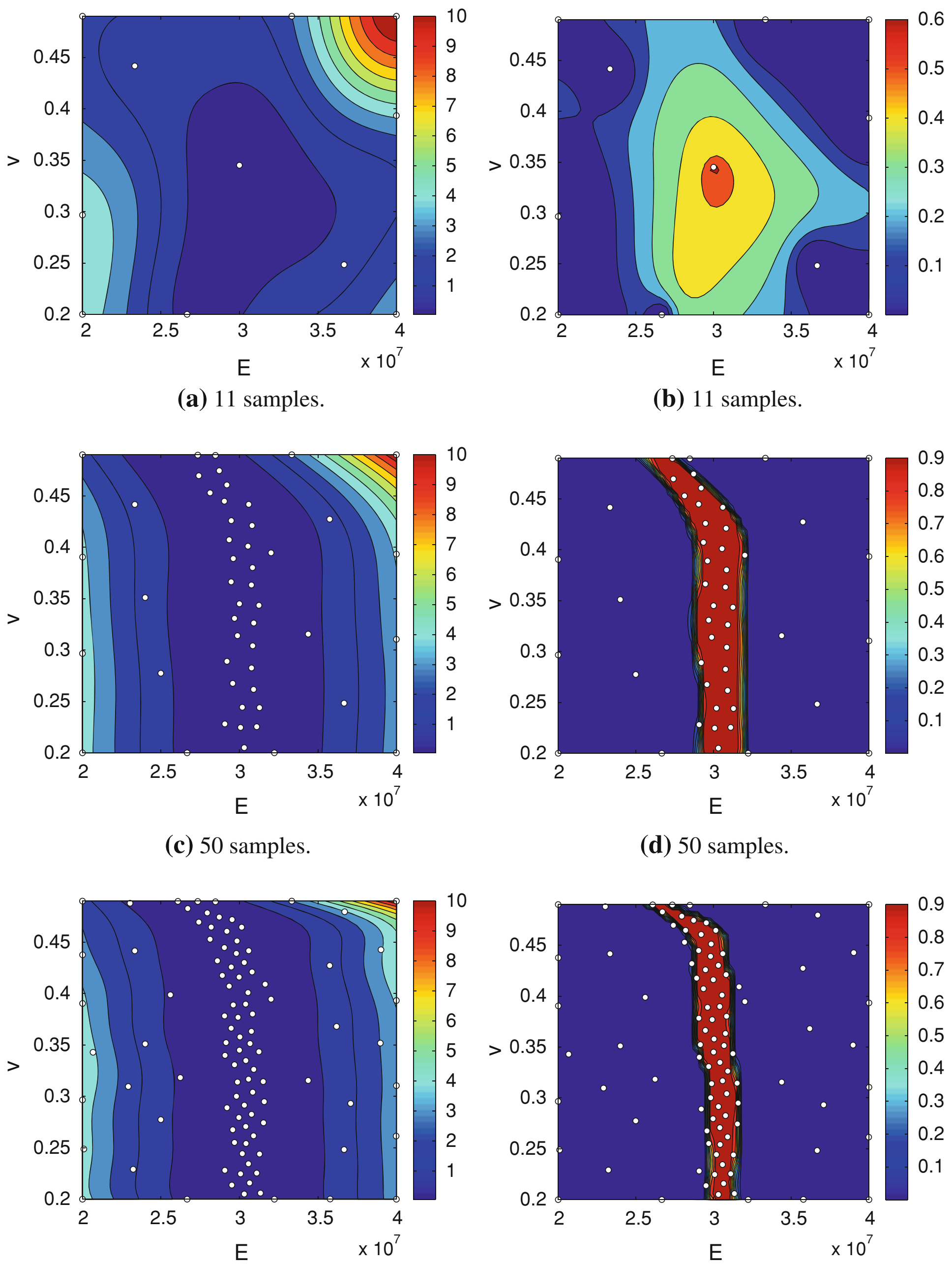


Fig. 9 Example 2: Bio-mechanical characterization example. **a** Schematic drawing of the point indentation setup: (*1*) translation and rotation stage, (*2*) cross-section of tympanic membrane sample, (*3*) needle connected to a load cell, (*4*) stepper motor and (*5*) Linear Variable Differential Transformer. **b** Finite element model of the tympanic membrane with indentation. The number of membrane shell elements equal to 5,988. The effective strain in the point indentation area after indentation rises up to approximately 15%



(a) 11 samples.

(b) 11 samples.

(c) 50 samples.

(d) 50 samples.

(e) 100 samples.

(f) 100 samples.

Fig. 10 Example 2: *Left column*: Contour plots of the intermediate kriging model of $\text{error}_{\text{force}}$ at various stages in the sampling process (samples are denoted by *dots*). The QOR is densely and quasi-uniformly sampled. *Right column*: Corresponding contour plots of the new gPoI criterion. The QOR is identified properly by the criterion

indentation depths up to 400 µm. The resulting force was measured with a load cell and the exact indentation depth was assessed with a Linear Variable Differential Transformer (LVDT).

In order to construct a finite element model, an LCD-Moiré profilometer was used to obtain a three-dimensional shape of the membrane before and during indentation. On the basis of these Moiré shape images a highly detailed non-uniform finite element mesh was created. In the needle indentation area and in the manubrium neighborhood, mesh density was increased. This is illustrated in Fig. 9. The tympanic membranes was modeled as a linear isotropic homogeneous elastic material which is described with two independent elasticity parameters: Young's modulus E and Poisson's ratio ν. The numerical simulations were performed with the finite element code FEBio (http://mrl.sci.utah.edu/software/febio), which is specifically designed for bio-mechanical applications.

Determining the value of the linear elasticity parameters is done by minimizing the discrepancy between the model and the experimental measurements. Namely, by calculating,

$$\arg\min_{E,m}(\text{error}_{\text{force}}), \tag{12}$$

where,

$$\text{error}_{\text{force}} = \frac{1}{N}\sum_{j=1}^{N}\left(F_{\exp}(q_j) - F_{\text{mod}}(q_j)\right)^2, \tag{13}$$

with N the number of measured points, q_j the indentation depth, $F_{\exp}(q_j)$ the experimental force and $F_{\text{mod}}(q_j)$ the simulated force.

7.2 Experimental setup

Version 7.0.2 of the SUMO toolbox is used to determine the QOR of (13). An initial set of samples is generated by an optimal maximin Latin Hypercube Design (LHD; [8]) of 7 points together with 4 corner points, adding up to a total of 11 initial points. Subsequently, infill points are selected based on the gPoI and MD figures of merit. For this problem the upper bound is defined as $T_2 = (\hat{f}_{\min} + 2\hat{\tau}) + 0.5 \cdot |\hat{f}_{\min} + 2\hat{\tau})|$, namely we are interested in all quasi-optimal solutions that deviate maximally 50% of the best solution found, taking noise into account. Samples are selected and evaluated one by one ($m = 1$) to ensure optimal space-fillingness.

The kriging surrogate model is configured using the standard Gaussian correlation function and a constant regression function. The hyperparameters, including the λ parameter, are efficiently estimated using SQPLab. The process is halted when the number of samples exceeds 100. The average computation time of FEBio for one simulation is about 5–10 min.

7.3 Results

Contour plots of the intermediate kriging model of $\text{error}_{\text{force}}$, and the associated gPoI, at various stages in the sampling process are shown in Fig. 10. The initial kriging model based on 11 samples has not yet discovered the QOR completely. After a stage of mostly exploration-based sampling, the full QOR is outlined after approximately 50 samples. The focus is now shifted to sampling the identified QOR densely (exploitation) until the stopping criterion of 100 samples is reached.

A contour plot of the final kriging surrogate model is shown in Fig. 10e. Obviously, the QOR is quite densely sampled in comparison with other parts of the input domain. Moreover, in Fig. 10f the contour plot of the gPoI of the final kriging model shows a clearly defined band in the input domain (=optimal curve) with probability one. The gray zone of uncertainty is reduced to a very small region at the edge of the optimal curve.

8 Conclusion

This paper introduced a simple but powerful method to solve (inverse) problems consisting of multiple quasi-optimal solutions. The Quasi-Optimal Regions (QORs) are identified with a limited number of expensive function evaluations. The QORs offers the user a trade-off between several solutions, similar to the Pareto front in multi-objective optimization. The QOR sampling method, based on the generalized Probability of Improvement (gPoI) criterion, is implemented in the SUMO Matlab toolbox [14], and successfully applied on the Branin and Hartman functions, and used to determine the elasticity of the middle ear tympanic membrane.

Within the QOR sampling algorithm framework several variations are possible. For instance, the gPoI has been successfully applied to identify input regions in the design space that correspond to a certain band in the output space, providing a tool to solve inverse problems directly. Furthermore, the method is relatively dimension-free, i.e., it does not pose any extra restrictions than those already inherent in the kriging surrogate model.

Acknowlegments Ivo Couckuyt and Jef Aernouts are funded by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen)

References

1. Aernouts J, Couckuyt I, Crombecq K, Dirckx J (2010) Elastic characterization of membranes with a complex shape using point indentation measurements and inverse modelling. Int J Eng Sci 48:599–611
2. Aernouts J, Soons J, Dirckx J (2010) Quantification of tympanic membrane elasticity parameters from in situ point indentation measurements: validation and preliminary study. Hear Res 263: 177–182
3. Barton R (2005) Issues in development of simultaneous forward-inverse metamodels. In: Proceedings of the Winter simulation conference, pp 209–217. doi:10.1109/WSC.2005.1574253
4. Bonnans J, Gilbert J, Lemaréchal C, Sagastizábal C (2006) Numerical optimization: theoretical and practical aspects. Springer, Berlin
5. Couckuyt I, Declercq F, Dhaene T, Rogier H (2010) Surrogate-based infill optimization applied to electromagnetic problems. Advan Des Optimiz Microwave/rf Circuits Systems (special issue) 20(5):492
6. Couckuyt I, Gorissen D, DeTurck F, Dhaene T (2010) Inverse surrogate modeling: output performance space sampling. In: 13th AIAA/ISSMO Multidisciplinary Analysis Optimization Conference
7. Crombecq K, Dhaene T (2010) Generating sequential space-filling designs using genetic algorithms and monte carlo methods. In: Simulated evolution and learning (SEAL-2010), pp 80–84
8. Dam E, van Husslage B, den Hertog D, Melissen J (2007) Maximin Latin hypercube designs in two dimensions. Operat Res 55(1):158–169

9. Decraemer W, Funnell W (2008) Anatomical and mechanical properties of the tympanic membrane, chronic otitis media. Pathogenesis-oriented therapeutic management. Kugler Publications, The Hague
10. Elkhouri N, Liu H, Funnell W (2006) Low-frequency finite-element modelling of the gerbil middle ear. J Assoc Res Otolaryngol 7:399–411
11. Forrester A, Sobester A, Keane A (2008) Engineering design via surrogate modelling: a practical guide. Wiley, Chichester
12. Gan R, Sun Q, Feng B, Wood M (2006) Acoustic-structural coupled finite element analysis for sound transmission in human ear—pressure distributions. Med Eng Phys 28:395–404
13. Gibbs M, Mackay DJC (1997) Efficient implementation of gaussian processes. Tech Rep, Department of Physics, Cavendish Laboratory, Cambridge University
14. Gorissen D, Crombecq K, Couckuyt I, Demeester P, Dhaene T (2010) A surrogate modeling and adaptive sampling toolbox for computer based design. J Mach Learn Res 11:2051–2055, http://sumo.intec.ugent.be/
15. Hadamard J (1902) Sur les problèmes aux dérivées partielles et leur signification physique. Tech Rep 49–52, Princeton University Bulletin
16. Jones D, Perttunen C, Stuckman B (1993) Lipschitzian optimization without the lipschitz constant. Optimiz Theory Appl 79(1): 157–181
17. Jones DR (2001) A taxonomy of global optimization methods based on response surfaces. Global Optimiz 21:345–383
18. Jones DR, Schonlau M, Welch WJ (1998) Efficient global optimization of expensive black-box functions. J Global Optimiz 13(4):455–492. doi:10.1023/A:1008306431147
19. Kleijnen J, van Beers W, van Nieuwenhuyse I (2011) Expected improvement in efficient global optimization through bootstrapped kriging. Journal of Global Optimization 51:1–15. doi: 10.1007/s10898-011-9741-y
20. Picheny V, Ginsbourger D, Roustant O, Haftka R (2010) Adaptive designs of experiments for accurate approximation of a target region. Mech Des 132(7):9
21. Ponweiser W, Wagner T, Vincze M (2008) Clustered multiple generalized expected improvement: A novel infill sampling criterion for surrogate models. In: Congress on Evolutionary Computation
22. Sacks J, Welch WJ, Mitchell T, Wynn HP (1989) Design and analysis of computer experiments. Stat Sci 4(4):409–435
23. Santner T, Williams B, Notz W (2003) The design and analysis of computer experiments. Springer series in statistics. Springer-Verlag, New York
24. Sasena M (2002) Flexibility and efficiency enhancements for constrainted global design optimization with kriging approximations. Ph.D. thesis, University of Michigan
25. Schonlau M (1997) Computer experiments and global optimization. Ph.D. thesis, University of Waterloo
26. Sóbester A, Leary SJ, Keane AJ (2004) A parallel updating scheme for approximating and optimizing high fidelity computer simulations. Struct Multidisciplinary Optimiz 27:371–383(13)
27. Staum J (2009) Better simulation metamodeling: The why, what, and how of stochastic kriging. In: Proceedings of the Winter Simulation Conference
28. Wang G, Shan S (2007) Review of metamodeling techniques in support of engineering design optimization. J Mech Des 129(4): 370–380. doi:10.1115/1.2429697