Global optimization of expensive-to-evaluate functions: an empirical comparison of two sampling criteria

Julien Villemonteix, Emmanuel Vazquez, Maryan Sidorkiewicz, Eric Walter

Noname manuscript No.
(will be inserted by the editor)

Global optimization of expensive-to-evaluate functions: an empirical comparison of two sampling criteria

Julien Villemonteix · Emmanuel Vazquez · Maryan Sidorkiewicz · Eric Walter

the date of receipt and acceptance should be inserted later

Abstract In many global optimization problems motivated by engineering applications, the number of function evaluations is severely limited by time or cost. To ensure that each of these evaluations usefully contributes to the localization of good candidates for the role of global minimizer, a stochastic model of the function can be built to conduct a sequential choice of evaluation points. Based on Gaussian processes and Kriging, the authors have recently introduced the informational approach to global optimization (IAGO) which provides a one-step optimal choice of evaluation points in terms of reduction of uncertainty on the location of the minimizers. To do so, the probability density of the minimizers is approximated using conditional simulations of the Gaussian process model behind Kriging. In this paper, an empirical comparison between the underlying sampling criterion called conditional minimizer entropy (CME) and the standard expected improvement sampling criterion (EI) is presented. Classical test functions are used as well as sample paths of the Gaussian model and an industrial application. They show the interest of the CME sampling criterion in terms of evaluation savings.

Keywords expected improvement, global optimization, Kriging

1 Introduction

To minimize an expensive-to-evaluate function f, a common approach is to use a cheap approximation of this function, which can lead to significant savings over traditional methods. In this context, global optimization techniques based on Gaussian processes and Kriging (see, e.g., [4]) are often preferred, for they provide an appealing probabilistic framework to account for the uncertainty on the function approximation. Expensive-to-evaluate functions are often encountered in industrial optimization problems, where the function value may be the output of complex computer simulations, or the result of costly measurements on prototypes.

Most Kriging-based strategies proposed in the past few years (see, e.g., [9] and the references therein) *implicitly* seek a likely value for a global minimizer, and then assume it to be a suitable location for the next evaluation of f. Yet, making full use of Kriging, it is possible to *explicitly* account for the uncertainty on the global minimizers, and the most likely location of a global optimizer is not necessarily a good evaluation point to improve the accumulated knowledge on the global minimizers.

Based on these considerations, the Informational Approach to Global Optimization (IAGO) strategy recently proposed in [20] evaluates f where the potential for reduction of the uncertainty on the location of

Julien Villemonteix · Maryan Sidorkiewicz
Renault S.A., Energy Systems Department, 78298 Guyancourt, France

Emmanuel Vazquez
Supelec, Département Signaux et Systèmes Électroniques, 91192 Gif-sur-Yvette, France
E-mail: emmanuel.vazquez@supelec.fr

Eric Walter
Laboratoire des Signaux et Systèmes, CNRS-Supelec-Univ Paris-Sud, Supelec 91192 Gif-sur-Yvette, France

the minimizers is deemed to be highest. The entropy of the conditional distribution of the global minimizers is taken as the uncertainty measure, and is approximated using conditional simulations of the Gaussian process modeling f. This approach has two main advantages over classical Kriging-based global optimization methods, such as the Efficient Global Optimization (EGO) algorithm (see [10]). First, it should lead to significant savings on the number of evaluations of f. Second, results under the form of probability distributions are particularly attractive. The purpose of this paper is to evidence the evaluation savings that can be obtained via the use of IAGO.

EGO and IAGO differ only by the sampling criterion used for choosing the next evaluation point. These two criteria, namely *expected improvement* (EI) for EGO and *conditional minimizer entropy* (CME) for IAGO, undergo a series of numerical experiments. The first experiments are conducted on four classical test functions. Later on, empirical convergence rates are estimated using sample paths of a Gaussian process. A final comparison is performed on a real-case application to the design of intake ports in the automotive industry, for which a single evaluation of the function to be optimized requires about ten hours of computer time.

The Kriging framework is briefly recalled in Section 2, as well as the definitions of the EI and CME criteria. A brief description of computational aspects of the IAGO approach is also presented. Section 3 reports the empirical comparison of these two criteria. Finally, Section 4 presents conclusions and offers perspectives for future work.

2 Kriging-based global optimization

Let $\mathbb{X}$, the factor space, be a compact subset of $\mathbb{R}^d$ and $f : \mathbb{X} \to \mathbb{R}$ be the function to be minimized. The objective is to find $\mathbf{x}^*$ a global minimizer of f over $\mathbb{X}$ when the evaluation of f is expensive. To do so, a cheap model of f (also known as surrogate approximation) based on previous evaluations will be used. Even if deterministic models have been discussed (as in the response surface methodology, see, e.g., [15]), it is stochastic models that will retain our attention, and more precisely the Bayesian approach to global optimization (see, e.g., [14]). In this framework, f is viewed as a realisation (or sample path) of a stochastic process F (F can also be viewed as a Bayesian prior on f). The distribution of F conditionally to past evaluation results for f is used to design a *sampling criterion* to be optimized to choose an additional evaluation point for f.

When F is Gaussian (we make this assumption in the rest of the paper), the conditional distribution of F at an untried point is also Gaussian with mean and variance that can be obtained analytically using Kriging (prediction based on Gaussian processes has been known for more than 50 years as Kriging in geostatistics and we shall keep to this terminology). Gaussian models and Kriging have been introduced in the field of Bayesian optimization in [10], through the Efficient Global Optimization (EGO) algorithm. Since then (see [14] for an overview of previous work in the field), Gaussian processes and Kriging have been the object of most publications in the field of Bayesian global optimization, with improvements of the EGO algorithm (see, e.g., [22] or [8]) and comparative studies (see, e.g. [9] or [17]). Our contribution to the field is also based on Kriging.

2.1 Linear prediction

In this section, we recall some well-known facts about Kriging on which the rest of the paper is based (for more details, see [4, 20] and the references therein).

Let $k(.,.)$ be the covariance function of F, and $\mathbf{x}$ be a point in $\mathbb{X}$ where F is to be predicted. The mean of $F(\mathbf{x})$ is assumed to be a finite linear combination of known functions p_i of $\mathbf{x}$, $m(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{p}(\mathbf{x})$, where $\boldsymbol{\beta}$ is a vector of fixed coefficients to be computed, and $\mathbf{p}(\mathbf{x}) = [p_1(\mathbf{x}), ..., p_l(\mathbf{x})]^\top$. Usually the functions p_i are monomials of low degree in the components of $\mathbf{x}$ (in practice, their degrees do not exceed two).

Given the vector $\mathbf{f}_n = [f(\mathbf{x}_1), ..., f(\mathbf{x}_n)]^\top$ of past evaluations at points in $\mathbb{S}_n = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \in \mathbb{X}^n$ (a sample value of $\mathbf{F}_n = [F(\mathbf{x}_1), ..., F(\mathbf{x}_n)]^\top$), the Kriging predictor $\hat{F}(\mathbf{x})$ of $F(\mathbf{x})$ is the minimum-variance unbiased linear predictor in the vector space span$\{F(\mathbf{x}_1), ..., F(\mathbf{x}_n)\}$. It can be written as

$$\hat{F}(\mathbf{x}) = \boldsymbol{\lambda}(\mathbf{x})^\top \mathbf{F}_n \,, \tag{1}$$

with $\mathbf{F}_n = [F(\mathbf{x}_1), ..., F(\mathbf{x}_n)]^\mathsf{T}$, and $\boldsymbol{\lambda}(\mathbf{x})$ the vector of Kriging coefficients for the prediction at $\mathbf{x}$.

The vector of coefficients $\boldsymbol{\lambda}(\mathbf{x})$ is solution of the linear system of equations

$$\begin{pmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^\mathsf{T} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}(\mathbf{x}) \\ \boldsymbol{\mu}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{k}(\mathbf{x}) \\ \mathbf{p}(\mathbf{x}) \end{pmatrix}, \tag{2}$$

with $\mathbf{0}$ a matrix of zeros, $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{1 \leq i,j \leq n}$ the covariance matrix of F at all evaluation points in $\mathbb{S}_n$, $\mathbf{k}(\mathbf{x}) = [k(\mathbf{x}_1, \mathbf{x}), ..., k(\mathbf{x}_n, \mathbf{x})]^\mathsf{T}$, the vector of covariances between $F(\mathbf{x})$ and $\mathbf{F}_n$, and

$$\mathbf{P} = \begin{pmatrix} \mathbf{p}(\mathbf{x}_1)^\mathsf{T} \\ \vdots \\ \mathbf{p}(\mathbf{x}_n)^\mathsf{T} \end{pmatrix}.$$

The Kriging coefficients at $\mathbf{x}$ can thus be computed without evaluating $f(\mathbf{x})$, along with the variance of the prediction error

$$\hat{\sigma}^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T} \mathbf{k}(\mathbf{x}) - \mathbf{p}(\mathbf{x})^\mathsf{T} \boldsymbol{\mu}(\mathbf{x}), \tag{3}$$

as these quantities only depend on the covariance of F. Once f has been evaluated at all $\mathbf{x}_i$ in $\mathbb{S}_n$, the prediction of $f(\mathbf{x})$ is the conditional mean of F, given by

$$\hat{f}(\mathbf{x}) = \mathbb{E}[\hat{F}(\mathbf{x}) | \mathcal{F}_n] = \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T} \mathbf{f}_n,$$

with $\mathcal{F}_n = \{\mathbf{F}_n = \mathbf{f}_n\}$ the evaluation results. When f is evaluated exactly, Kriging is an interpolation ($\forall \mathbf{x}_i \in \mathbb{S}_n$, $\hat{F}(\mathbf{x}_i) = F(\mathbf{x}_i)$). Although noise on the evaluation results could easily be taken into account in the prediction, in what follows, the evaluations are assumed to be noise-free (see [20] for the noisy case).

As advocated in [18], the covariance of F is chosen within the Matérn class of covariance functions (cf. [20] and the reference therein for more details on the choice of a covariance), and the covariance parameters are either set a priori or estimated from the data using the maximum-likelihood method.

After the evaluations in $\mathbb{S}_n$, $f(\mathbf{x})$ is viewed as a sample path of F that interpolates the data $\mathbf{f}_n$. Such sample paths, known as *conditional sample paths*, are realizations of F conditionally to $\mathcal{F}_n$ and are essential to the IAGO approach. They represent all the behaviors that are deemed possible for f given the results of evaluations in $\mathbb{S}_n$. Figure 1(a) illustrates the relationships between f, $\hat{f}$, $\hat{\sigma}$ and the conditional sample paths.

2.2 Kriging-based sampling criteria

Among the many sampling criterion available in the literature, we feel that expected improvement (EI), which has been the object of most publications in the field for the last ten years, is the most suited for a comparison with the one we proposed in [20].

2.2.1 Expected improvement

This sampling criterion corresponds to a one-step optimal strategy given the Gaussian prior F on the unknown function f. Let $f^* = \min_{\mathbf{x} \in \mathbb{X}} f(\mathbf{x})$ be the global minimum of f, $\mathbb{S}_n$ be a set of n evaluation points in $\mathbb{X}$, and consider $M_n = \min_{\mathbf{x}_i \in \mathbb{S}_n} F(\mathbf{x}_i)$ an estimator for f^*. For the loss function

$$L(\mathbb{S}_n, F) = M_n - f^*,$$

the risk, or expected loss for a candidate point $\mathbf{c}$ for the evaluation of f, given the evaluation results $\mathbf{f}_n$, is given by

$$\mathbb{E}(L(\mathbb{S}_n \cup \{\mathbf{c}\}, F) | \mathcal{F}_n) = \mathbb{E}(\min\{M_n, F(\mathbf{c})\} | \mathcal{F}_n) - f^*. \tag{4}$$

One can show that minimizing (4) is equivalent to maximizing the EI criterion as presented for example in [9], i.e.,

$$\mathrm{EI}(\mathbf{c}) = \mathbb{E}\left[I(\mathbf{c}) | \mathcal{F}_n\right], \tag{5}$$

with

$$I(\mathbf{c}) = \begin{cases} 0 & \text{if} \quad F(\mathbf{c}) \geq M_n \\ M_n - F(\mathbf{c}) & \text{otherwise} \end{cases} .$$

One can easily rewrite (5) as

$$\mathrm{EI}(\mathbf{c}) = \hat{\sigma}(\mathbf{c}) \left[u\Phi(u) + \Phi'(u) \right], \tag{6}$$

with

$$u = \frac{m_n - \hat{f}(\mathbf{c})}{\hat{\sigma}(\mathbf{c})},$$

$m_n = \mathbb{E}[M_n|\mathcal{F}_n] = \min_{\mathbf{x}_i \in \mathbb{S}_n} f(\mathbf{x}_i)$ the current estimation of the minimum, and Φ the normal cumulative distribution. The new evaluation point is then chosen as a global maximizer of $\mathrm{EI}(\mathbf{c})$.

2.2.2 Conditional minimizer entropy

The IAGO approach is based on two complementary principles, that set it apart from previous work in Bayesian global optimization. First, a one-step optimal sampling criterion for the reduction of the uncertainty on the minimizers. Second, the use of Kriging to evaluate this sampling criterion by approximating the distribution of the minimizers conditionally to past evaluations. We now briefly present our sampling criterion, and refer to [20] for computational details.

In [20], conditional entropy has been introduced to measure the information gained on the minimizers by an additional evaluation of f. This *Stepwise Uncertainty Reduction* (SUR) strategy [6], chooses the point that potentially brings the largest reduction in entropy (seen as a measure of uncertainty).

More formally, given our Gaussian prior F on the function f to be minimized, the uncertainty on the minimizer $\mathbf{x}^*$ can be measured by the entropy of the global minimizers

$$H(\mathbf{X}^*) = -\sum_{\mathbf{x} \in \mathbb{G}} p_{\mathbf{X}^*}(\mathbf{x}) \log(p_{\mathbf{X}^*}(\mathbf{x})),$$

with $\mathbf{X}^*$ a random vector uniformly distributed in the set of the global minimizers of F over a discrete approximation $\mathbb{G}$ of $\mathbb{X}$, and $p_{\mathbf{X}^*}$ the point mass density of $\mathbf{X}^*$.

Now, given a vector f_n of evaluation results, the uncertainty left on $\mathbf{x}^*$ is the entropy of $p_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ the point mass density of $\mathbf{X}^*$ conditionally to the evaluation results $\mathcal{F}_n$ (or in short *conditional minimizer density*),

$$H(\mathbf{X}^*|\mathcal{F}_n) = -\sum_{\mathbf{x} \in \mathbb{G}} p_{\mathbf{X}^*}(\mathbf{x}|\mathcal{F}_n) \log(p_{\mathbf{X}^*}(\mathbf{x}|\mathcal{F}_n)).$$

The idea of the IAGO strategy is iteratively to ensure a one-step optimal reduction of the entropy of this distribution.

The risk associated with a candidate evaluation at $\mathbf{c} \in \mathbb{X}$ is then chosen as the differential entropy of the global minimizers conditionally to the potential result of an evaluation at $\mathbf{c}$ (in short CME for *conditional minimizer entropy*)

$$H_n(\mathbf{c}) = H(\mathbf{X}^*|\mathcal{F}_n, F(\mathbf{c})),$$

and the evaluation is performed at

$$\mathbf{x}_{n+1} = \arg\min_{\mathbf{c} \in \mathbb{X}} H_n(\mathbf{c}).$$

From the definition of conditional entropy [5], we can write

$$H_n(\mathbf{c}) = \int_{y \in \mathbb{R}} p_{F(\mathbf{c})}(y|\mathcal{F}_n) \left(-\sum_{x \in \mathbb{G}} p_{\mathbf{X}^*}(x|\mathcal{F}_n, F(\mathbf{c}) = y) \log(p_{\mathbf{X}^*}(x|\mathcal{F}_n, F(\mathbf{c}) = y)) \right) dy, \tag{7}$$

with $p_{F(\mathbf{c})}(\cdot|\mathcal{F}_n)$ the distribution of $F(\mathbf{c})$ and $p_{\mathbf{X}^*}(\cdot|\mathcal{F}_n, F(\mathbf{c}) = y)$ the distribution of $\mathbf{X}^*$ conditionally to $\mathcal{F}_n$ and $\{F(\mathbf{c}) = y\}$. The CME $H_n(\mathbf{c})$, as written in (7), can be viewed as an expected loss, the loss function being the entropy of $p_{\mathbf{X}^*}(\cdot|\mathcal{F}_n, F(\mathbf{c}) = y)$ the conditional minimizer density after $n+1$ evaluations.

2.2.3 Practical aspects

The distribution $p_{F(\mathbf{c})}(\cdot|\mathcal{F}_n)$ is Gaussian, with mean and variance simply obtained by Kriging. There is, however, no result in the literature that we can use to describe analytically any useful property of the conditional minimizer density. To compute (7), we resort to an approximation that is conducted via Monte-Carlo simulations of F conditionally to available evaluation results $\mathcal{F}_n$ and to a potential evaluation result y at $\mathbf{c}$ (this approximation as well as recommendations for the choice of $\mathbb{G}$ are described in details in [20]). This approximation leads to a complexity in $O(N)$ for the computation of $H_n(\mathbf{c})$, with N the size of the discrete approximation of $\mathbb{X}$. Note that in IAGO the conditional minimizer density is thus available at each step and provides (at least for low-dimensional problems) a clear view of the progress achieved in the optimization process (cf. Figure 1(b)).

In the Bayesian optimization framework, the expensive-to-evaluate function is replaced by a cheap criterion, updated after each evaluation, which has to be optimized for a new evaluation point to be chosen. Up to now, we have focused on the choice of criterion, but no attention has been paid to the entire procedure for global optimization, including for example an update process for the Kriging prediction. To keep this paper focused on a comparison between sampling criteria, we shall only mention the classical framework of the Efficient Global Optimization (EGO) (Algorithm 1).

EGO (see, e.g. [10]) starts with a small initial design used to get a first estimate of the parameters of the covariance and to compute a first Kriging model. Based on this model, an additional point is selected in the design space to be the location of the next evaluation of f in order to maximize the EI criterion. The parameters of the covariance are then re-estimated, the Kriging model is re-computed, and the process of choosing new points continues until the improvement expected from sampling additional points has become sufficiently small. The CME criterion can easily be inserted in a similar algorithm in place of EI to transform EGO into IAGO.

Algorithm 1: Efficient global optimization framework
Input: Initial design of evaluation points and corresponding values of f
Output: Additional evaluations
1. **while** the evaluation budget is not exhausted or some other convergence condition is not satisfied
2. **do** Estimate the parameters of the covariance
3. Compute the Kriging model
4. Optimize the sampling criterion (EI or CME here)
5. Evaluate f

3 Empirical comparison between EI and CME

As presented in the previous section, EI and CME are both Kriging-based sampling criteria and both one-step optimal in some sense. CME should lead to faster convergence rates, and this for three major reasons.

First, EI aims at estimating the *minimum*, while CME concentrates on the *minimizers*. The search is therefore likely to be more global when based on the latter. Second, EI aims at improving the estimation of the minimum by sampling where its appearance is most probable. It seems more reasonable to try diminishing the uncertainty associated with its position. For example, it might be excessively costly to refine the estimation in a small neighborhood of a *potential* minimum, which may only be local, while evaluations using H_n could show that a large part of the search space has a very low probability of containing the global minimum (this idea will be confirmed in Section 3.3). Third, the computation of CME involves the statistical properties of the sample paths of F, while, by contrast the computation of EI involves only the conditional mean and variance of F at $\mathbf{c}$. A more thorough use of the available information on the function is indeed appealing in this context of expensive, and therefore sparse, evaluations.

To substantiate these intuitions, a comparison of EI and CME is in order.

3.1 Experimental conditions

To make this comparison fair, we propose to study the behaviors of EI and CME independently Algorithm 1) and from the optimization method to be used to optimize the sampling criteria (Step 4 in Algorithm 1).

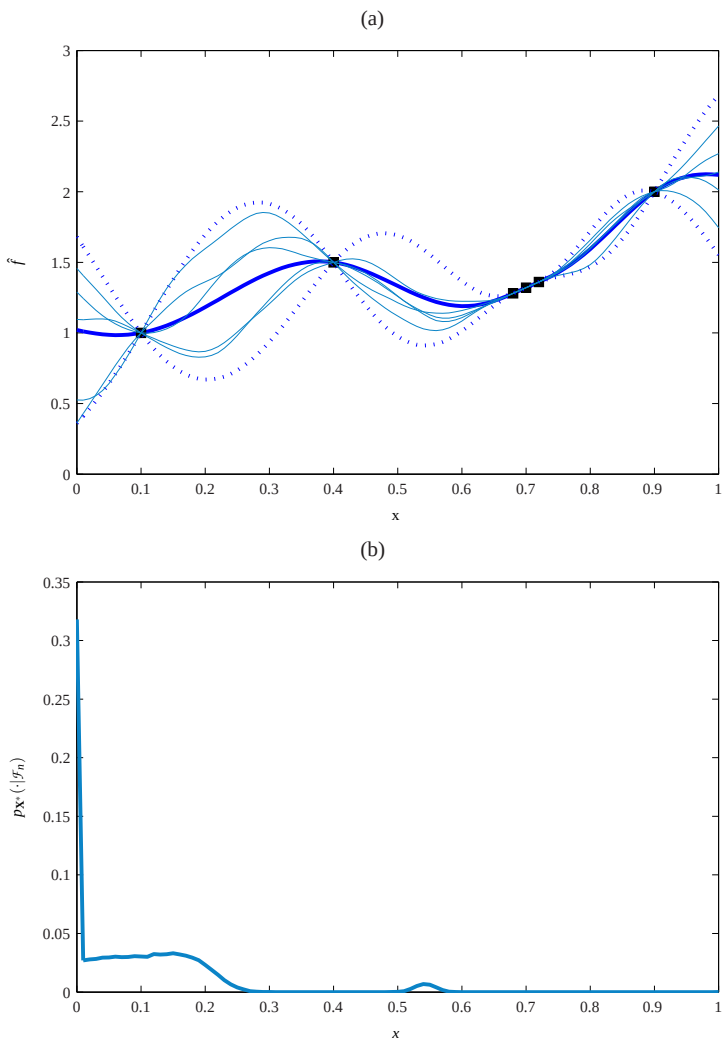

Fig. 1 (*a*): Conditional sample paths of F, and corresponding Kriging prediction. The squares represent available values of f, the bold line is the conditional mean $\hat{f}$ as computed by the Kriging predictor, the dotted lines provide 95% confidence intervals for the prediction ($\hat{f} \pm 1.96\hat{\sigma}$) and the thin lines are conditional sample paths. (*b*): Estimated conditional minimizer density ($p_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$) associated with the Kriging prediction.

These aspects are quite complex, and ad-hoc strategies have been proposed in the literature (see [9] for an example of optimization method for the sampling criteria). However, our first objective here is to motivate the choice of a sampling criterion.

Therefore, we conducted our experiments using the same Matérn covariance with the same values, fixed a priori, for the covariance parameters. The set $\mathbb{G}$ of potential evaluation points was identical for both criteria, and the choice of the next evaluation point was carried out via an exhaustive computation of the relevant sampling criterion over this set. The question to be addressed in what follows then boils down to the following interrogation: Given the same prior information on the function, which sampling criterion chooses the best point (in a sense to be discussed later) amongst a finite set of possible evaluations points?

3.2 Tests on classical benchmarks

The four test functions used in this section are taken from [8], where a comparison was conducted between EI and classical global optimization schemes such as DIRECT (see, e.g., [16]). The problem dimensions range from two to five, and all functions present several local minimizers (see Table 2 in Appendix 2). The covariance parameters are estimated beforehand on the results of 200 evaluations randomly chosen in

Table 1 Comparison of expected improvement and conditional minimizer entropy on four test functions taken from [8]*.

	G_i when points are chosen using EI			G_i when points are chosen using CME		
	$i = 20$	$i = 50$	$i = 100$	$i = 20$	$i = 50$	$i = 100$
Six-Hump Camel Back	0.65	1	1	**0.76**	1	1
Tilted Branin	0.83	0.92	0.98	**0.89**	**0.95**	0.97
Hartman 3	0.64	0.98	1	**0.82**	0.99	1
Ackley 5	0.36	**0.75**	0.73	0.34	0.59	0.72

*For each criterion, the convergence measure G_i is averaged over 50 runs (the estimated standard error for the estimation of these figures is always smaller than 0.01).

search space (using a latin hyper cube sampler), and the two criteria are optimized over a latin hyper cube design containing 1000 points randomly re-sampled after every evaluation.

A single point $\mathbf{x}_1$is randomly chosen in search space as a common starting point for both criteria, and 50 runs are conducted for each function to reduce the dependency on the starting point. After the i-th evaluation of f, the efficiency of each criteria is measured by

$$G_i = \frac{f(\mathbf{x}_1) - m_i}{f(\mathbf{x}_1) - f^*},$$

with $m_i = \min_{\mathbf{x} \in \{\mathbf{x}_1, \ldots \mathbf{x}_i\}} f(\mathbf{x})$ the current estimate of the global minimum. G_i (a modified version of the quality measure used in [2]) thus describes the reduction, after i iterations of the optimization process, of the initial estimation error for the global minimum $f(\mathbf{x}_1) - f^*$. Table 1 presents, for each criterion, the averaged efficiency after 20, 50 and 100 evaluations. EI beats CME for the Ackley function when $i = 40$, but for the other three test functions CME converges faster towards the optimum than EI, and significantly so for the Hartman 3 function.

3.3 Tests on Gaussian processes simulations

Even if a comparison on classical test functions gives some perspectives on the qualities of each of the criteria, the variability of the results from one test problem to the next may be significant, so one can hardly use them to decide beforehand which sampling criterion to use on a specific problem. It therefore would be best to derive some analytical convergence rates for both criteria under reasonable hypotheses on the function to be optimized. In our context of expensive-to-evaluate functions, these convergence rates would have to be non-asymptotic, and we do not know of any such results in the literature. However, the probabilistic framework considered here makes it possible to estimate empirical convergence rates. Since the function to be optimized is assumed to be a sample path of a Gaussian process, we can estimate the convergence rates with both criteria when optimizing sample paths of a Gaussian process whose covariance is the same as that chosen for the optimization algorithm.

For the sake of brevity, we shall limit our presentation to two Gaussian processes, one with very smooth sample paths, and the other with irregular sample paths.

Two sets of 1000 sample paths were generated over a regular grid of 1500 points in $[0,1]^2$. 15 evaluations are then performed on each sample paths using both criteria. After each new evaluation, and for each criterion, estimation errors are computed for the global minimum and the minimizer, as well as the entropy of the conditional minimizer density. Two estimators of the global minimum are considered here, namely $m_n = \min_{\mathbf{x} \in \{\mathbf{x}_1, \ldots \mathbf{x}_n\}} f(\mathbf{x})$, the best evaluation result obtained so far, and

$$\tilde{m}_n = \hat{f}(\arg\max_{\mathbf{x} \in \mathbb{G}} p_{\mathbf{X}^*}(\mathbf{x}|\mathcal{F}_n)),$$

the predicted value associated with the point where the conditional minimizer density is the highest. The average convergence rates for irregular sample paths are presented on Figure 2(a) in terms of the entropy of the conditional minimizer density, on Figure 2(b) in terms of the estimation error for m_n, and on Figure 2(c) in terms of the estimation error for $\tilde{m}_n$.

As expected, since the entropy of the conditional minimizer is the loss function behind CME, CME performs significantly better than EI in terms of the entropy of the conditional minimizer density and, in average, the uncertainty on the positions of the global minimizers diminishes faster if points are chosen

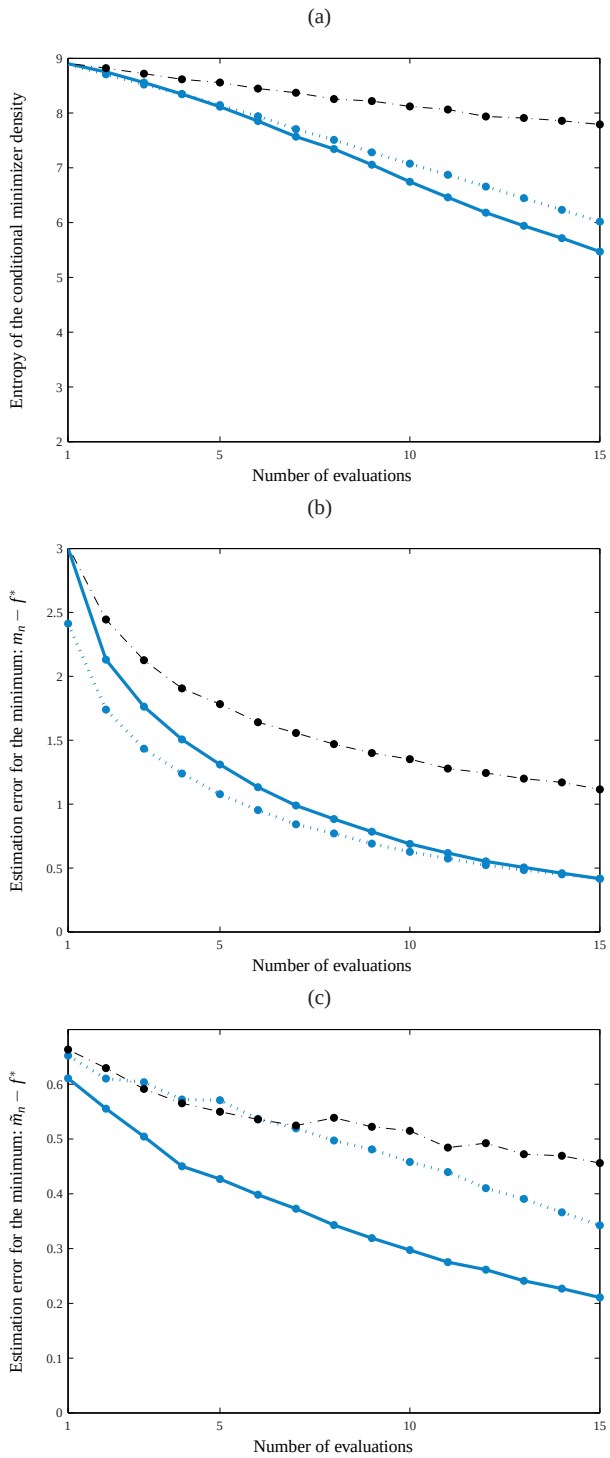

Fig. 2 Convergence rates using EI (dotted line) and IAGO (bold line), when convergence is measured by $H(\mathbf{X}^*|\mathcal{F}_n)$ (*a*), by the estimation error for the global minimum with the best value obtained so far (m_n) as an estimator (*b*), by the estimation error for the global minimum with $\tilde{m}_n$ as an estimator (*c*). The convergence measures are averaged over 1000 sample paths of a Gaussian process with a Matérn covariance with parameters $\nu = 1$, $\rho = 0.3$ and $\sigma = 1$ (see Appendix 5.1). The dashed line represents, as a reference, the convergence rate for a random choice of evaluation points.

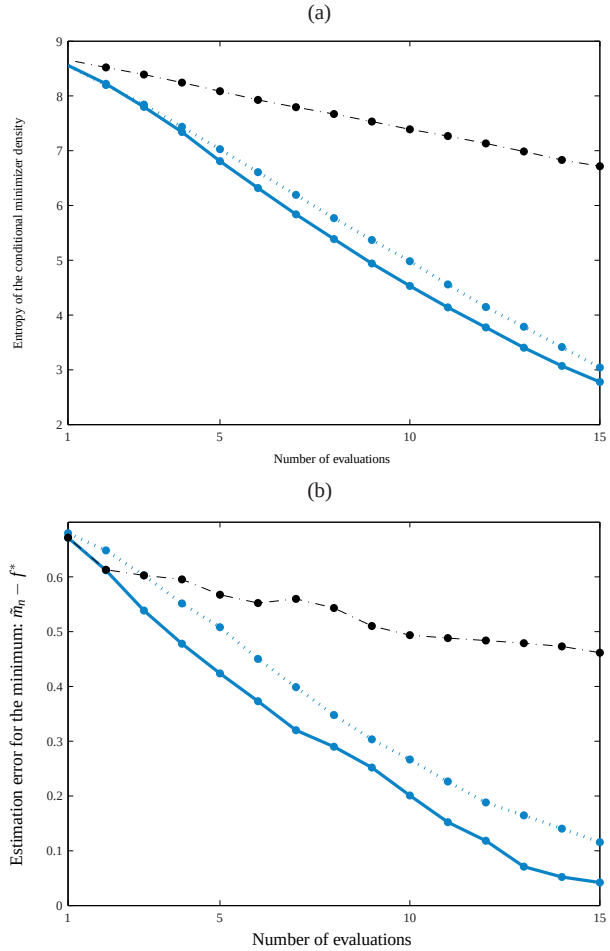

Fig. 3 Convergence rates using EI (dotted line) and IAGO (bold line), when convergence is measured by $H(\mathbf{X}^*|\mathcal{F}_n)$ (*a*), and when convergence is measured by the estimation error for the minimum with $\tilde{m}_n$ as estimator (*b*). The dashed line represents, as reference, the convergence rate for a random choice of evaluation points. The sample paths used here are smoother than those used for Figure 2 (the parameter for the Matérn covariance are $\nu = 5$, $\rho = 0.3$ and $\sigma = 1$).

using CME (cf. Figure 2(a)). This fact was guaranteed for the first evaluation since CME is one-step optimal for this loss function, but it had to be checked for several evaluations.

Similarly, if the convergence is measured by the estimation error $m_n - f^*$, EI is bound to perform better if we consider only the first evaluation, since the convergence measure is the loss function behind EI. However, it appears that after 15 evaluations, the performance of EI and CME are similar (cf. Figure 2(b)), suggesting that even if EI is one-step optimal, in the long run, CME will bring the largest reduction for $m_n - f^*$ (this is confirmed by computations, not presented here, with a larger number of evaluations).

EI would thus seem to be a better criterion in a context where very few evaluations are allowed. However, m_n is estimator actually a rather poor estimator of the global minimum, and it appears that when a faster-to-converge estimator is used instead of m_n, CME performs significantly better than EI (Figure 2(c)), and this right form the start. This estimator is $\tilde{m}_n$, whose interest is apparent for the three search strategy considered here, for which $m_n - f^*$ is significantly bigger than $\tilde{m}_n - f^*$, and more than three times so after the first evaluation (Figure 2(b) and Figure 2(c)). CME should therefore be preferred to EI when one is confronted with irregular sample paths, since it allows a better estimation of f^*.

If we look at what happens on a typical sample path (see Figure 4), the drawbacks of EI are clearly evidenced. As intuitively stated at the beginning of the section, EI stalls on a local optimum because with (4) as a loss function, it is better to ensure a small improvement near a minimum already found than to check that it is effectively a global minimum. In the case of irregular sample paths this might be highly dangerous, and IAGO performs better simply because it first addresses the question of whether a minimizer

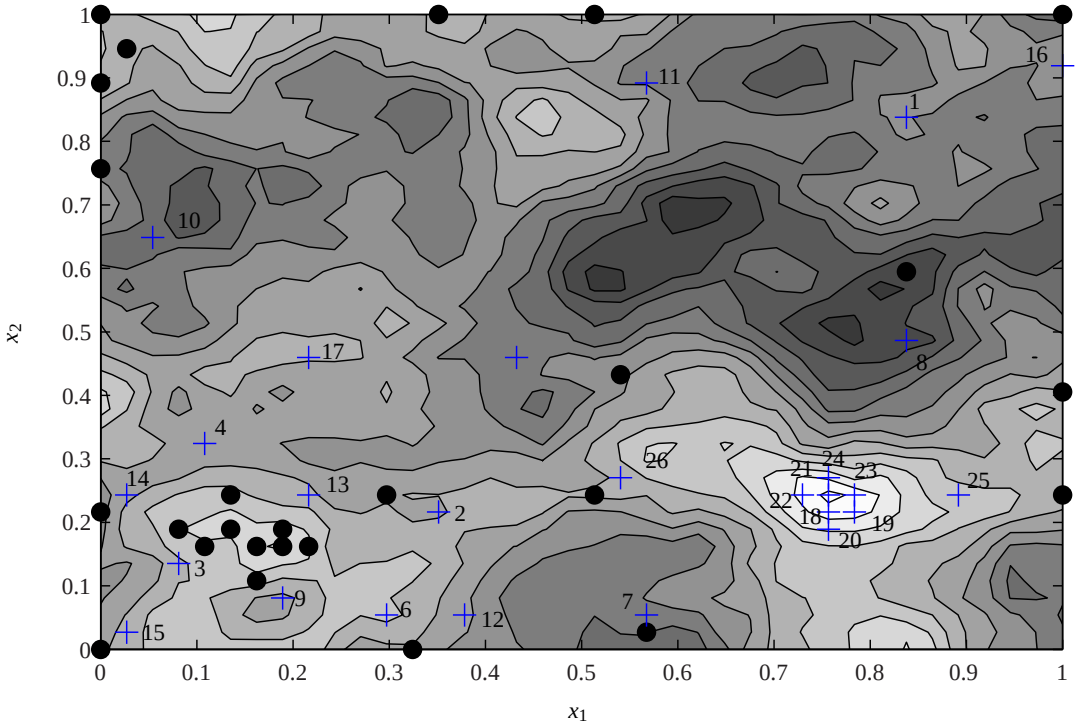

Fig. 4 Minimization of a sample path from the Gaussian process used to evaluate the convergence rates of Figure 2. The dots indicate the evaluation points chosen by EI. The crosses indicate the evaluation points chosen by CME. The order in which the evaluations are carried out is also indicated for CME.

is global before improving the precision on its exact position. When the sample paths are more regular, this advantage diminishes (cf. Figure 3), as the local optima are scarcer.

A significant problem is left aside here, namely what happens in practice if the parameters of the covariance are poorly estimated? Does the optimization strategy still perform well? *Robustness* to a poor choice of covariance parameters is of course a major issue but it is not considered here.The EI and CME criteria should have similar robustness properties and may both be deceived by a poor choice of covariance as demonstrated in [9]. We feel that this problem should be tackled from a Bayesian point of view, with some prior on the covariance parameters. This will be done in future work, where we shall compare an extended version of IAGO to the methods in [9] designed to be robust to a poor choice of covariance.

3.4 Test on an industrial application: intake port design

This section presents an industrial optimization problem in the automotive field, also used for the comparison of CME and EI.

3.4.1 Problem description

Intake ports (Figure 5) are engine components that convey a mixture of air and fuel to the combustion chambers. The importance of this type of component lies in the properties of the flow it induces in the combustion chamber, which has a direct impact on both the performance and the emissions of pollutant by the engine. To comply with new emission standards (Euro V and Euro VI), while satisfying the ever increasing need for engine performance, the shape of intake ports has to be carefully optimized. Two often-conflicting objectives have to be maximized simultaneously, namely the flow rate and a scalar characteristic of the turbulent flow known as *tumble* [13]. Physics tells us that the higher the flow rate, the larger the amount of fuel that can be burnt, and consequently the larger the power delivered by the engine. Similarly,

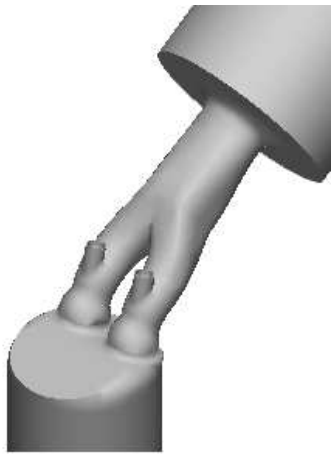

Fig. 5 Intake port. The component itself is in the middle. Below is the combustion chamber. The upper cylinder is a tranquilizing volume necessary for the convergence of finite-element simulations.

pollutants as nitrogen oxides (NO_x) and carbon monoxide (CO) are, to a large extent, created when the air/fuel mix is not homogeneous. Therefore, the larger the turbulence (and tumble accounts for the relevant properties of it), the smaller the pollution.

The specifications for these two objectives are liable to change during conception. Therefore, it is important to determine not only an optimal geometry for a given set of preferences but rather the full Pareto front. However, building prototypes for tests is exceedingly expensive, and each flow simulation by finite-element methods takes about ten hours on powerful servers. The approach advocated in this paper is therefore particularly attractive given the general will for reduction of duration and cost associated with development.

3.4.2 Computational issues

To extend our sampling criteria to a multi-objective problem, we use a standard procedure and consider several linear combinations of the objective functions (a.k.a. aggregations), each accounting for a different zone of the Pareto front. In [11], this approach has been used to extend the EI criterion to a multi-objective problem by randomly selecting a new aggregation after each evaluation of f. In this paper, we follow the same route, but use the IAGO framework to compute the entropy of the conditional density of the minimizers for the mono-objective optimization problem corresponding to each aggregation in a given set. The search can thus be directed towards the most uncertain regions of the Pareto front.

The resulting multi objective extensions of CME and EI have been applied to the optimization of six shape parameters of an intake port (these parameters are not detailed here for confidentiality reasons). To improve the number of simulations achievable in a given time, the geometry and mesh are automatically generated for each finite-element simulation. The optimization algorithm is then directly interfaced with the solver, limiting human intervention to the initialization of the procedure.

The initial value for the parameters of the Matérn covariance are estimated on simulations that have been collected during the design of previous intake ports. It thus becomes possible to initialize the algorithms with very few randomly chosen points (here with five points).

One thousand candidate points are used ($N = 1000$) and the parameters of the Matèrn covariance are fixed.

3.4.3 Results

For comparison purposes, simulations were conduced on twenty intake ports whose parameters were chosen using EI, CME, or a Latin Hyper Cube (LHC) as a reference. The Pareto-optimal points within each of the three sets of evaluation results are presented in Figure 6. Comparison between sets estimates of Pareto fronts is a tricky process, which may involve various quality measures [12]. Here however, the comparison is clearly in favor of CME, as

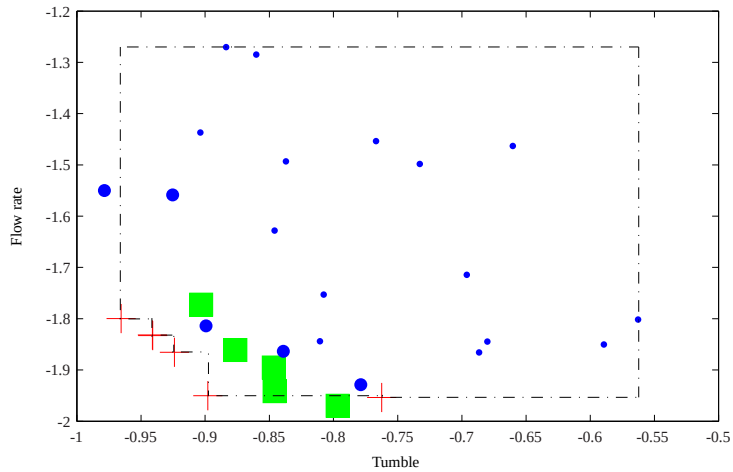

Fig. 6 Results of 20 evaluations randomly chosen using an LHC (dots), or optimally chosen by CME (crosses) or by EI (squares). For CME and EI, only the Pareto-optimal points are presented. For LHC, a larger dot size indicates points that are, within the considered set of points, Pareto-optimal. Tumble and flow rate have been rescaled for confidentiality reasons. The dashed lines delimit the set of points that are dominated by the Pareto-optimal points obtained using CME and which dominate the "worst point", i.e. the point with worst values attained for both objectives

.

- among all evaluation results, the point closest to an "ideal" solution (i.e. with best value yet obtained for both objectives, here $[-0.97, -1.97]^{\mathsf{T}}$) has been chosen by CME;
- all but one point chosen by EI are dominated by points chosen by CME. In other words, almost any good solution found by EI is bettered by a solution found by CME;
- the volume of the set of points dominated by the Pareto-optimal points (cf. [12] for details on this quality measure) is 0.31 for CME (this volume is represented on Figure 6), while it is only 0.26 for EI (the reference for the computation is the point with worst values attained for both objectives as coordinates).

This test case confirms that Bayesian global optimization is easily applicable to an industrial problem, even with a very small evaluation budget. The interest of CME is apparent after only a few evaluations, as predicted by the convergence rates of Section 3.3.

3.5 Computational burden

The comparison made so far dealt only with convergence rates. The superiority of CME over EI was demonstrated for sample paths of Gaussian random processes, at least for two very different regularities of the sample paths (cf. Section 3.3), while the convergence rates are generally in favor of CME when applied to some classical test functions (cf. Section 3.2) or to an actual industrial problem. However, EI is easier to compute since it only requires the mean and variance of the prediction at the candidate point, while the complexity of the computation of IAGO is in $O(N)$, with N the size of a discrete approximation of $\mathbb{X}$ used for the estimation of the conditional density.

In practice, with our implementation of IAGO (cf. [20]) around 40s are required on an AMD opteron 285 server to choose an additional evaluation point for the sample paths of Section 3.3 (by extensive computation of H_n over 1500 candidate points, which is enough in practice since this set is randomly re-sampled after each evaluation). By comparison, choosing a point with EI takes less than half a second under the same conditions. To broaden the range of potential applications, we tried to limit the computational expense by testing other approximations for the conditional minimizer density (since the manipulation of sample paths, necessary for the approximation proposed here, is responsible for most of the computational burden). We proposed for example, to estimate the derivatives of f by Kriging (as in, e.g., [19]) and to compute the probability for a given point to be a local optimum and under a certain threshold. It was then easy to build a relatively accurate approximation of the conditional minimizer density, but the approximation had a

detrimental effect on the convergence rate, so that EI then became more efficient. In fact, the quality of the approximation of the conditional minimizer density is important for CME to perform better than EI.

IAGO therefore remains destined to the optimization of functions that require a large amount of computer time (or more generally a significant expense) to be evaluated, which is after all what it was designed for and is the case in many applications in the industrial world including the one we presented here.

4 Conclusions and perspectives

In this paper, we have evidenced a clear superiority of CME over EI, especially when the function to be optimized is irregular. The comparison has been conducted using classical test functions and an actual industrial application, but above all using sample paths of the model behind Kriging. The use of sample paths indeed allows the computation of empirical convergence rates that can also be useful to tune other components of any Kriging-based algorithms (e.g., the optimization of the sampling criterion). Now that the interest of the CME criterion has been demonstrated, attention should turn to other crucial aspects of any Kriging-based optimization algorithm. One of these aspects is the improvement of robustness against a bad choice of covariance for the Gaussian process model. We feel that the Bayesian framework that we have used until now should also be useful in this respect.

5 Appendices

5.1 Appendix 1: Matèrn covariance

In this paper, we follow Stein (1999) and use of the isotropic Matérn covariance:

$$k(\mathbf{x},\mathbf{y}) = k(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}h}{\rho}\right)^{\nu} \mathcal{K}_\nu \left(\frac{2\nu^{1/2}h}{\rho}\right) \ \forall\,(\mathbf{x},\mathbf{y}) \in \mathbb{X}^2\,, \tag{8}$$

with h the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$, and $\mathcal{K}_\nu$ the modified Bessel function of the second kind [23]. The parameters of this covariance are easy to interpret, as ν controls regularity, σ^2 is the variance ($k(0) = \sigma^2$), and ρ represents the *range* of the covariance, *i.e.*, the characteristic correlation distance. They can either be fixed using prior knowledge on the system, or be estimated from experimental data. In geostatistics, estimation is carried out using the adequacy between the empirical and model covariance (see, e.g., [4]). In other areas, cross validation (cf. [21]) and maximum likelihood (cf. [18]) are mostly employed. For simplicity and generality reasons (cf. [18]), the maximum-likelihood method is preferred here.

5.2 Appendix 2: Test functions

References

1. Ackley, D.: A Connectionist Machine for Genetic Hill-climbing. Kluwer Academic Publishers (1987)
2. Barton, R.: Minimization algorithms for functions with random noise. Am. J. Math. Manag. Sci. **4**, 109–138 (1984)
3. Branin, F.: Widely convergent methods for finding multiple solutions of simultaneous nonlinear equations. IBM J. Res. Dev. (1972)
4. Chilès, J., Delfiner, P.: Geostatistics, Modeling Spatial Uncertainty. John Wiley & Sons, Inc, New York (1999)
5. Cover, T.M., Thomas, A.J.: Elements of Information Theory. John Wiley & Sons, Inc, New York (1991)
6. Geman, D., Jedynak, B.: An active testing model for tracking roads in satellite images. Tech. Rep. 2757, Institut National de Recherche en Informatique et en Automatique (INRIA) (1995)
7. Hartman, J.: Some experiments in global optimization. Nav. Res. Logist. Q. **20**, 569–576 (1973)
8. Huang, D., Allen, T., Notz, W., Zeng, N.: Global optimization of stochastic black-box systems via sequential kriging meta-models. J. Global Optim. **34**, 441–466 (2006)

Table 2 Test functions to be minimized [8]*.

Name	Description
Six-Hump Camel Back [3]	$d=2$ $f(\mathbf{x}) = 4x_1^2 - 2.1x_1^4 + 1/3x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4$ $-1.6 \leq x_1 \leq 2.4, -0.8 \leq x_2 \leq 1.2$ $N_{\text{local}} = 6, N_{\text{global}} = 2$ $\mathbf{x}^* = [0.089, 0.713]^\mathsf{T}$ and $[0.089, 0.713]^\mathsf{T}$, $f^* = -1.03$
Tilted Branin [8]	$d=2$ $f(\mathbf{x}) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos x_1 + 10 + 0.5x_1$ $-5 \leq x_1 \leq 10, 0 \leq x_2 \leq 15$ $N_{\text{local}} = 3, N_{\text{global}} = 1$ $\mathbf{x}^* = [-3.2, 12.3]^\mathsf{T}$ $f^* = -1.17$
Hartman 3 [7]	$d=3$ $f(\mathbf{x}) = -\sum_{i=1}^4 d_i \exp\left[-\sum_{j=1}^3 \alpha_{ij}(x_j - p_{ij})\right]$, where $\alpha = \begin{pmatrix} 3 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3 & 10 & 30 \\ 0.1 & 10 & 35 \end{pmatrix}$ $\mathbf{d} = \begin{pmatrix} 1 \\ 1.2 \\ 3 \\ 3.2 \end{pmatrix}$ $\mathbf{p} = \begin{pmatrix} 0.3689 & 0.1170 & 0.2673 \\ 0.4699 & 0.4387 & 0.7470 \\ 0.1091 & 0.8732 & 0.5547 \\ 0.03815 & 0.5743 & 0.8828 \end{pmatrix}$ $0 \leq x_i \leq 1$, $i = 1,2,3$ $N_{\text{local}} > 1, N_{\text{global}} = 1$ $\mathbf{x}^* = (0.114, 0.556, 0.852)^\mathsf{T}$, $f^* = -3.86$
Ackley 5 [1]	$d=5$ $f(\mathbf{x}) = -20\exp[-0.2\sqrt{\frac{1}{d}\sum_{i=1}^d x_i^2}] - \exp[\frac{1}{d}\sum_{i=1}^d \cos(2\pi x_i)] + 20 + e$ $\forall i \in [\![1,3]\!]$ $-32.8 \leq x_i \leq 32.8$ $N_{\text{local}} > 1, N_{\text{global}} = 1$ $\mathbf{x}^* = \mathbf{0}$, $f^* = 0$

* N_{local} stands for the number of local minima, while N_{global} stands for the number of global minima.

9. Jones, D.: A taxonomy of global optimization methods based on response surfaces. J. Global Optim. **21**, 345–383 (2001)
10. Jones, D., Schonlau, M., William, J.: Efficient global optimization of expensive black-box functions. J. Global Optim. **13**, 455–492 (1998)
11. Knowles, J.: Parego: A hybrid algorithm with on-line landscape approxiamtion for expensive multiobjective optimization problems. IEEE T. Evolut. Comput. **7**(2), 100–116 (2003)
12. Knowles, J., Thiele, L., Zitzler, E.: A tutorial on the performance assessment of stochastive multiobjective optimizers. Tech. Rep. 214, Computer Engineering and Networks Laboratory, ETH Zurich (2006)
13. Lumley, J.: Engines: An Introduction. Cambridge University Press (1999)
14. Mockus, J.: Bayesian Approach to Global Optimization: Theory and Applications. Kluwer Academic Publishers (1989)
15. Myers, R., Montgomery, D.: Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley-Interscience (2002)
16. Perttunen, C.: A computational geometric approach to feasible region division in constrained global optimization. In: Proceedings of the 1991 IEEE Conference on Systems, Man. and Cybernetics (1991)
17. Sasena, M., Papalambros, P., Goovaerts, P.: Exploration of metamodeling sampling criteria for constrained global optimization. Eng. Opt. **34**, 263–278 (2002)
18. Stein, M.: Interpolation of Spatial Data: Some Theory for Kriging. Springer, New-York (1999)
19. Vazquez, E., Walter, É.: Estimating derivatives and integrals with kriging. In: Proceedings of the joint 44th IEEE Conference on Decision and European control conference, pp. 8156–8161. Séville (2005)
20. Villemonteix, J., Vazquez, E., Walter, É.: An informational approach to the global optimization of expensive-to-evaluate functions. Submitted to J. Global Optim. (2006)
21. Wahba, G.: Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning, vol. 6, pp. 69–87. MIT Press, Boston (1998)
22. Williams, B., Santner, T., Notz, W.: Sequential design of computer experiments to minimize integrated response functions. Stat. Sinica **10**, 1133–1152 (2000)

23. Yaglom, A.: Correlation Theory of Stationary and Related Random Functions I: Basic Results. Springer Series in Statistics. Springer-Verlag, New-York (1986)