An informational approach to the global optimization of expensive-to-evaluate functions

Julien Villemonteix, Emmanuel Vazquez, Eric Walter

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

An informational approach to the global optimization of expensive-to-evaluate functions

Julien Villemonteix · Emmanuel Vazquez · 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 evaluation contributes to the localization of good candidates for the role of global minimizer, a sequential choice of evaluation points is usually carried out. In particular, when Kriging is used to interpolate past evaluations, the uncertainty associated with the lack of information on the function can be expressed and used to compute a number of criteria accounting for the interest of an additional evaluation at any given point. This paper introduces minimizers entropy as a new Kriging-based criterion for the sequential choice of points at which the function should be evaluated. Based on *stepwise uncertainty reduction*, it accounts for the informational gain on the minimizer expected from a new evaluation. The criterion is approximated using conditional simulations of the Gaussian process model behind Kriging, and then inserted into an algorithm similar in spirit to the *Efficient Global Optimization* (EGO) algorithm. An empirical comparison is carried out between our criterion and *expected improvement*, one of the reference criteria in the literature. Experimental results indicate major evaluation savings over EGO. Finally, the method, which we call IAGO (for Informational Approach to Global Optimization), is extended to robust optimization problems, where both the factors to be tuned and the function evaluations are corrupted by noise.

Keywords Gaussian process, global optimization, Kriging, robust optimization, stepwise uncertainty reduction

1 Introduction

This paper is devoted to global optimization in a context of expensive function evaluation. The objective is to find global minimizers in $\mathbb{X}$ (the factor space, a bounded subset of $\mathbb{R}^d$) of an unknown function $f : \mathbb{X} \to \mathbb{R}$, using a very limited number of function evaluations. Note that the global minimizer may not be unique (any global minimizer will be denoted as $\boldsymbol{x}^*$). Such a problem is frequently encountered in the industrial world. For instance, in the automotive industry, optimal crash-related parameters are obtained using costly real tests and time-consuming computer simulations (a single simulation of crash-related deformations may take up to 24 hours on dedicated servers). It then becomes essential to favor optimization methods that use the dramatically scarce information as efficiently as possible.

To make up for the lack of knowledge on the function, surrogate (also called meta or approximate) models are used to obtain cheap approximations [13]. They turn out to be convenient tools for visualizing

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

Emmanuel Vazquez
SUPELEC 91192 Gif-sur-Yvette, France
E-mail: emmanuel.vazquez@supelec.fr

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

the function behavior or suggesting the location of an additional point at which f should be evaluated in the search for $\boldsymbol{x}^*$. Surrogate models based on Gaussian processes have received particular attention. Known in geostatistics under the name of *Kriging* since the early 1960s [15], Gaussian process models provide a probabilistic framework to account for the uncertainty stemming from the lack of information on the system. When dealing with an optimization problem, this framework allows the set of function evaluations to be chosen efficiently [12–14].

In this context, several strategies have been proposed, with significant advantages over traditional optimization methods when confronted to expensive-to-evaluate functions. Most of them *implicitly* seek a likely value for $\boldsymbol{x}^*$, and then assume it to be a suitable location for a new evaluation of f. Yet, given existing evaluation results, the most likely location of a global minimizer is not necessarily a good evaluation point to improve our knowledge on $\boldsymbol{x}^*$. As we shall show, by making full use of Kriging, it is instead possible to *explicitly* estimate the probability distribution of the optimum location, which allows an information-based search strategy.

Based on these observations, the present paper introduces minimizers entropy as a criterion for the choice of new evaluation points. This criterion, directly inspired from *stepwise uncertainty reduction* [9], is then inserted in an algorithm similar to the *Efficient Global Optimization* (EGO) algorithm [14]. We call the resulting algorithm IAGO, for *Informational Approach to Global Optimization*.

Section 2 recalls the principle of Kriging-based optimization, along with some general ideas on Gaussian process modeling that are used in Section 3 to build an estimate of the distribution of the global minimizers. Section 4 details the stepwise uncertainty reduction approach applied to global optimization, while Section 5 describes the corresponding algorithm and its extensions to noisy problems. Section 6 illustrates the behavior of the new algorithm on some simple benchmark problems, along with its performances compared with those of the classical EGO algorithm, chosen for its good compromise between local and global search [17]. Finally, after a conclusion section and to make this paper self-contained, Section 8 recalls, as an appendix, some more results on Gaussian process modeling and Kriging.

2 Kriging-based global optimization

When dealing with expensive-to-evaluate functions, optimization methods based on probabilistic surrogate models (and Kriging in particular) have significant advantages over traditional optimization techniques, as they require fewer function evaluations to provide an acceptable solution. Kriging provides not only a cheap approximation of the function but also an estimate of the potential error in this approximation. Numerous illustrations of this superiority can be found in the literature (see, for instance, [6]) and many variations have been explored (for extensive surveys, see [13] and [17]). As explained in this section, these methods deal with the cost of evaluation using an adaptive sampling strategy, replacing the optimization of the expensive-to-evaluate function f by a series of optimizations of a cheap criterion.

2.1 Gaussian process modeling and Kriging

This section briefly recalls the principle of Gaussian process (GP) modeling, and lays down the necessary notation. A more detailed presentation is available in the appendix (Section 8).

When modeling with Gaussian processes, the function f is assumed to be a sample path of a Gaussian random process F, with mean function $m(\boldsymbol{x})$ and covariance function $k(\cdot,\cdot)$ defined over $\mathbb{X}^2$. If we denote $(\Omega, \mathcal{A}, \mathcal{P})$ the underlying probability space, this amounts to assuming that $\exists\, \omega \in \Omega$, such that $F(\omega,\cdot) = f(\cdot)$. Whenever possible, we shall omit the dependence of F in ω to simplify notation.

In particular, given a set of n evaluation points $\mathbb{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ (the *design*), $\forall \boldsymbol{x}_i \in \mathbb{S}$ the evaluation result $f(\boldsymbol{x}_i)$ is viewed as a sample value of the random variable $F(\boldsymbol{x}_i)$. Kriging computes an unbiased linear predictor of $F(\boldsymbol{x})$ in the vector space $\mathbb{H}_\mathbb{S} = \operatorname{span}\{F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)\}$, which can be written as

$$\hat{F}(\boldsymbol{x}) = \boldsymbol{\lambda}(\boldsymbol{x})^\mathsf{T} \boldsymbol{F}_\mathbb{S}\,, \tag{1}$$

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

Given the covariance function of F, the Kriging coefficients can be computed along with the variance of the prediction error

$$\hat{\sigma}^2(\boldsymbol{x}) = \mathrm{var}(\hat{F}(\boldsymbol{x}) - F(\boldsymbol{x})). \tag{2}$$

The covariance function of F is chosen within a parametrized class (for instance, the Matèrn class), and its parameters are either estimated from the data or chosen a priori (see Section 8.3.2 for details on the choice of a covariance function).

Once f has been evaluated at all evaluation points in $\mathbb{S}$, the predicted value of f at $\boldsymbol{x}$ is given by

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{f}_{\mathbb{S}}\,, \tag{3}$$

with $\boldsymbol{f}_{\mathbb{S}} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^{\mathsf{T}}$ ($\boldsymbol{f}_{\mathbb{S}}$ is viewed as a sample value of $\boldsymbol{F}_{\mathbb{S}}$). The same results could be derived in a Bayesian framework, where $F(\boldsymbol{x})$ is Gaussian conditionally to the evaluations carried out ($\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}$), with mean $\hat{f}(\boldsymbol{x})$ and variance $\hat{\sigma}^2(\boldsymbol{x})$.

Note that the random processes $F(\boldsymbol{x})$ and $\hat{F}(\boldsymbol{x})$ satisfy

$$\forall\, \boldsymbol{x}_i \in \mathbb{S},\ \ \hat{F}(\boldsymbol{x}_i) = F(\boldsymbol{x}_i), \tag{4}$$

and that the prediction at $\boldsymbol{x}_i \in \mathbb{S}$ is $f(\boldsymbol{x}_i)$. When f is assumed to be evaluated exactly, Kriging is thus an interpolation, with the considerable advantage over other interpolation methods that it also provides an explicit characterization of the prediction error (zero-mean Gaussian with variance $\hat{\sigma}^2(\boldsymbol{x})$).

2.2 Adaptive sampling strategies

The general principle of optimization using Kriging is iteratively to evaluate f at a point that optimizes a criterion based on the model obtained using previous evaluation results. The simplest approach would be to choose a minimizer of the prediction $\hat{f}$ as a new evaluation point. However, by doing so, too much confidence would be put in the current prediction and search is likely to stall on a local optimum (as illustrated by Figure 1). To compromise between local and global search, more emphasis has to be put on the prediction error, which can indicate locations where additional evaluations are needed to improve confidence in the model. This approach has led to a number of criteria to select additional evaluation points based on both prediction and prediction error.

A standard example of such a criterion is *expected improvement* (EI) [18]. As the name suggests, it involves computing how much improvement in the optimum is expected, if f is evaluated at a given additional point. Let $f_{\min}$ be the best function value obtained so far. The improvement expected from an additional evaluation of f at $\boldsymbol{x}$ given $\boldsymbol{f}_{\mathbb{S}}$, the results of past evaluations, can then be expressed as

$$\mathrm{EI}(\boldsymbol{x}) = \mathbb{E}\left[\max\left(f_{\min} - F\left(\boldsymbol{x}\right), 0\right) | \boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}\right].$$

Since $F(\boldsymbol{x})$ is conditionally Gaussian with mean $\hat{f}(\boldsymbol{x})$ and variance $\hat{\sigma}^2(\boldsymbol{x})$,

$$\mathrm{EI}(\boldsymbol{x}) = \hat{\sigma}(\boldsymbol{x}) \left[u\Phi(u) + \frac{d\Phi}{du}(u)\right], \tag{5}$$

with

$$u = \frac{f_{\min} - \hat{f}(\boldsymbol{x})}{\hat{\sigma}(\boldsymbol{x})}$$

and Φ the normal cumulative distribution function. The new evaluation point is then chosen as a global maximizer of $\mathrm{EI}(\boldsymbol{x})$. An example is given on Figure 2, where the problem that deceived the naive method of Figure 1 is directly solved with the EI criterion. This method has been used for computer experiments in [17], while modified criteria have been used in [11] and [25] to deal with noisy functions.

In [13] and [24], a fair number of alternative criteria are presented and compared. Although quite different in their formulation, they generally aim to answer the same question: What is the most likely position of $\boldsymbol{x}^*$? Another, and probably more relevant, question is: Where should the evaluation be carried out optimally to improve knowledge on the global minimizers?

In what follows, a criterion that addresses this question will be presented, along with its performances. The reference for comparison will be EI, which is a reasonable compromise between local and global search [17], and has been successfully used in many applications.

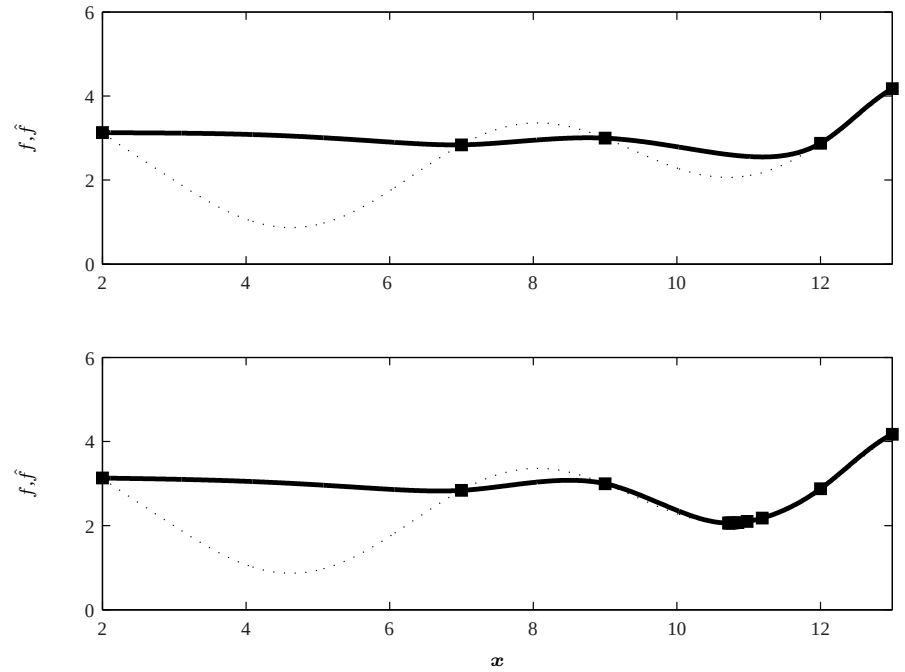


Fig. 1 Naive approach to optimization using Kriging: (*top*) prediction $\hat{f}$ (bold line) of the true function f (dotted line, supposedly unknown) obtained from an initial design materialized by squares; (*bottom*) prediction after seven iterations minimizing $\hat{f}$.

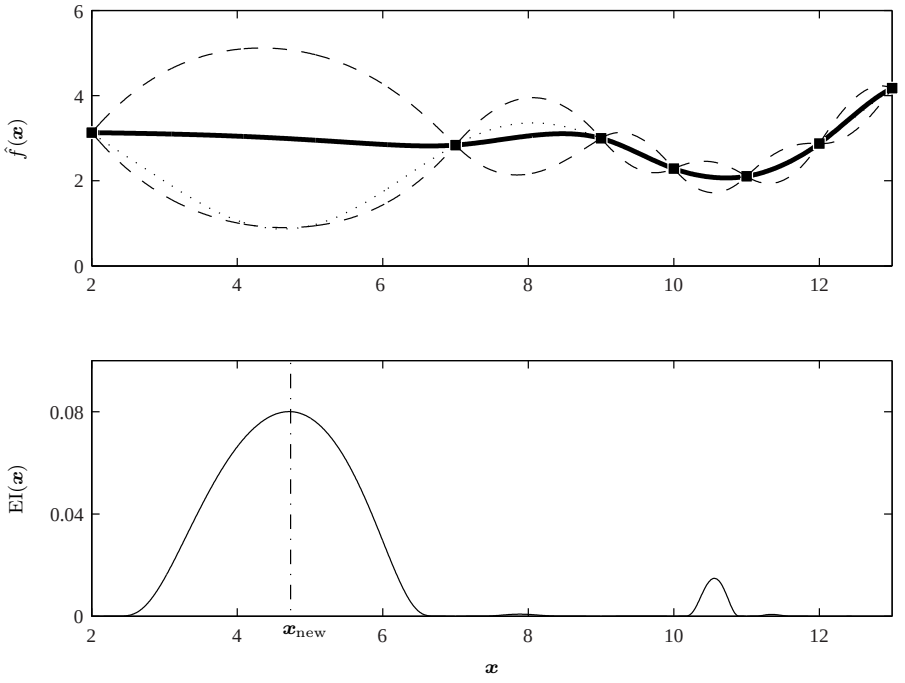


Fig. 2 EI approach to optimization using Kriging: (*top*) prediction $\hat{f}$ (bold line), 95% confidence intervals computed using $\hat{\sigma}$ (dashed line) and true function f (dotted line); (*bottom*) expected improvement.

3 Estimating the density of x^*

Once a Kriging surrogate model $\hat{f}$ has been obtained, any global minimizer of $\hat{f}$ is a natural approximation of $\boldsymbol{x}^*$. However, it might be excessively daring to trust this approximation as it does not take in account the uncertainty of the prediction. A more cautious approach to estimating $\boldsymbol{x}^*$ is to use the probabilistic framework associated with F. Of course, $\boldsymbol{x}^*$ is not necessarily unique, and we shall focus on describing the set of all global minimizers of f as efficiently as possible.

3.1 Probabilistic modeling of the global minimizers of f

According to the GP model, a global minimizer $\boldsymbol{x}^*$ of f corresponds to a global minimizer of this particular sample path of F. It seems therefore natural to use the GP model of f to obtain a probabilistic model for $\boldsymbol{x}^*$. Consider the *random* set $\mathcal{M}^*_{\mathbb{X}}$ of the global minimizers of F over $\mathbb{X}$, *i.e.* the set of all global minimizers for each sample path, which for any $\omega \in \Omega$ can be written as

$$\mathcal{M}^*_{\mathbb{X}}(\omega) = \{\boldsymbol{x}^* \in \mathbb{X} | F(\omega, \boldsymbol{x}^*) = \min_{\boldsymbol{u} \in \mathbb{X}} F(\omega, \boldsymbol{u})\}.$$

To ensure that $\mathcal{M}^*_{\mathbb{X}}(\omega)$ is not empty for all ω, we assume that F has continuous sample paths with probability one. This continuity can be ensured through a proper choice of covariance function (see, e.g., [1]).

Let $\boldsymbol{X}^*$ be a random vector uniformly distributed on $\mathcal{M}^*_{\mathbb{X}}$ (from now on, we omit the dependency of $\mathcal{M}^*_{\mathbb{X}}$ in ω). The probability density function of this random vector conditional to past evaluation results, that we shall thereafter call conditional density of the global minimizers and denote $p_{\boldsymbol{X}^*|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x})$, is of great interest, as it allows one not only to estimate the global minimizers of f (for example, through the maximization of their conditional density), but also to characterize the uncertainty associated with this estimation. In fact, $p_{\boldsymbol{X}^*|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x})$ contains all of what has been assumed and learned about the system. However, no tractable analytical expression for $p_{\boldsymbol{X}^*|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x})$ is available [2, 19]. To overcome this difficulty, the approach taken here is to consider a discrete version of the conditional distribution, and to approximate it using Monte Carlo simulations.

Let $\mathbb{G} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be a finite subset of $\mathbb{X}$, $\mathcal{M}^*_{\mathbb{G}}$ be the random set of global minimizers of F over $\mathbb{G}$, and $\boldsymbol{X}^*_{\mathbb{G}}$ be a random vector uniformly distributed on $\mathcal{M}^*_{\mathbb{G}}$. The conditional probability mass function of $\boldsymbol{X}^*_{\mathbb{G}}$ given $\boldsymbol{f}_{\mathbb{S}}$ (or simply minimizers distribution) is then $\forall \boldsymbol{x} \in \mathbb{G}$

$$P_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x}) \;=\; \mathrm{P}(\boldsymbol{X}^*_{\mathbb{G}} = \boldsymbol{x} \,|\, \boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}) \;.$$

It can be approximated using conditional simulations, *i.e.*, simulations of F that satisfy $\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}$. Assuming that non-conditional simulations are available, several methods exist to make them conditional [4]. Conditioning by Kriging seems the most promising of them in the present context and will be presented in the next section.

To keep the presentation simple, we assume in what follows that $\mathbb{S} \subset \mathbb{G}$.

3.2 Conditioning by Kriging

This method, due to G. Matheron, uses the unbiasedness of the Kriging prediction to transform non-conditional simulations into simulations interpolating the results $\boldsymbol{f}_{\mathbb{S}}$ of the evaluations. The idea is to sample from the conditional distribution of the prediction error $F - \hat{F}$ rather than from the conditional distribution of F, which is made easier by the fact that the statistical properties of the prediction error do not depend on the result of the evaluations, nor on the mean $m(\boldsymbol{x})$ of $F(\boldsymbol{x})$.

To present this more formally, let Z be a zero-mean Gaussian process with covariance function k (the same as that of F) and $\hat{Z}$ be its Kriging predictor based on the random variables $Z(\boldsymbol{x}_i)$, $\boldsymbol{x}_i \in \mathbb{S}$, and consider the random process

$$T(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}) + \left[Z(\boldsymbol{x}) - \hat{Z}(\boldsymbol{x})\right], \tag{6}$$

where $\hat{f}$ is the mean of the Kriging predictor based on the design points in $\mathbb{S}$. Since this Kriging predictor is an interpolator, at evaluation points in $\mathbb{S}$, we have $\hat{f}(\boldsymbol{x}_i) = f(\boldsymbol{x}_i)$. Equation (4) implies that $Z(\boldsymbol{x}_i) = \hat{Z}(\boldsymbol{x}_i)$,

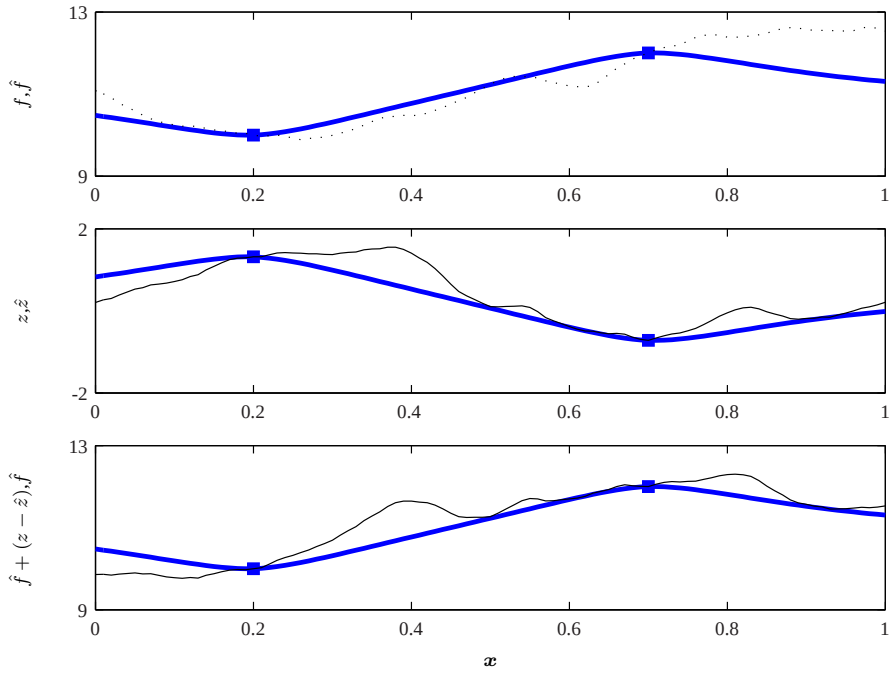


Fig. 3 Conditioning a simulation: (*top*) unknown real curve f (doted line), sample points (squares) and associated Kriging prediction $\hat{f}$ (bold line); (*middle*) non-conditional simulation z, sample points and associated Kriging prediction $\hat{z}$ (bold line); (*bottom*) the simulation of the Kriging error $z - \hat{z}$ is picked up from the non-conditional simulation and added to the Kriging prediction to get the conditional simulation (thin line).

which leads to $T(\boldsymbol{x}_i) = f(\boldsymbol{x}_i),\ \forall \boldsymbol{x}_i \in \mathbb{S}$. In other words, T is such that all its sample paths interpolate the known values of f. It is then easy to check that T has the same finite-dimension distributions as F conditionally to past evaluation results [7], simply because the prediction error $Z - \hat{Z}$, for Z, has the same distribution as the prediction error for F , $F - \hat{F}$. Note that the same vector $\boldsymbol{\lambda}(\boldsymbol{x})$ of Kriging coefficients is used to interpolate the data and the simulations at design points. Using (3), one can rewrite (6) as

$$T(\boldsymbol{x}) = Z(\boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \left[\boldsymbol{f}_{\mathbb{S}} - \boldsymbol{Z}_{\mathbb{S}}\right], \tag{7}$$

with $\boldsymbol{Z}_{\mathbb{S}} = [Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)]^{\mathsf{T}}$.

In summary, to simulate F over $\mathbb{G}$ conditionally to past evaluation results $\boldsymbol{f}_{\mathbb{S}}$, we can simulate a zero-mean Gaussian process Z over $\mathbb{G}$, compute the prediction error for each simulation and shift the prediction error around the desired mean $\hat{f}$. This is achieved by the following procedure (illustrated on Figure 3):

- compute, for every point in $\mathbb{G}$, the vector of Kriging coefficients based on the design points in $\mathbb{S}$,
- compute the Kriging prediction $\hat{f}(\boldsymbol{x})$ based on past evaluation results $\boldsymbol{f}_{\mathbb{S}}$ for every $\boldsymbol{x}$ in $\mathbb{G}$,
- collect non-conditional sample paths of Z over $\mathbb{G}$ (provided that a Gaussian sampler is available, setting the proper covariance for the simulated vector can be achieved using, for example, the Cholesky decomposition),
- apply (7) for each non conditional simulation and at every point in $\mathbb{G}$. That is, to generate $t(\boldsymbol{x})$, a conditional simulation of $T(\boldsymbol{x})$ from a non-conditional simulation $z(\boldsymbol{x})$ of $Z(\boldsymbol{x})$, apply

$$t(\boldsymbol{x}) = z(\boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}}[\boldsymbol{f}_{\mathbb{S}} - \boldsymbol{z}_{\mathbb{S}}], \tag{8}$$

  where $\boldsymbol{z}_{\mathbb{S}}$ is the sampled valued of Z over $\mathbb{S}$, which is available since $\mathbb{S} \subset \mathbb{G}$.

With this sampling method, it becomes straightforward to estimate $P_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}$. Let $\boldsymbol{x}^*_i$ be a global minimizer of the i-th conditional simulation ($i = 1, \ldots, r$) over $\mathbb{G}$ (if it is not unique, choose one randomly). Then, for any $\boldsymbol{x}$ in $\mathbb{G}$, a classical estimator is

$$\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x}) = \frac{1}{r}\sum_{i=1}^{r} \delta_{\boldsymbol{x}^*_i}(\boldsymbol{x}), \tag{9}$$

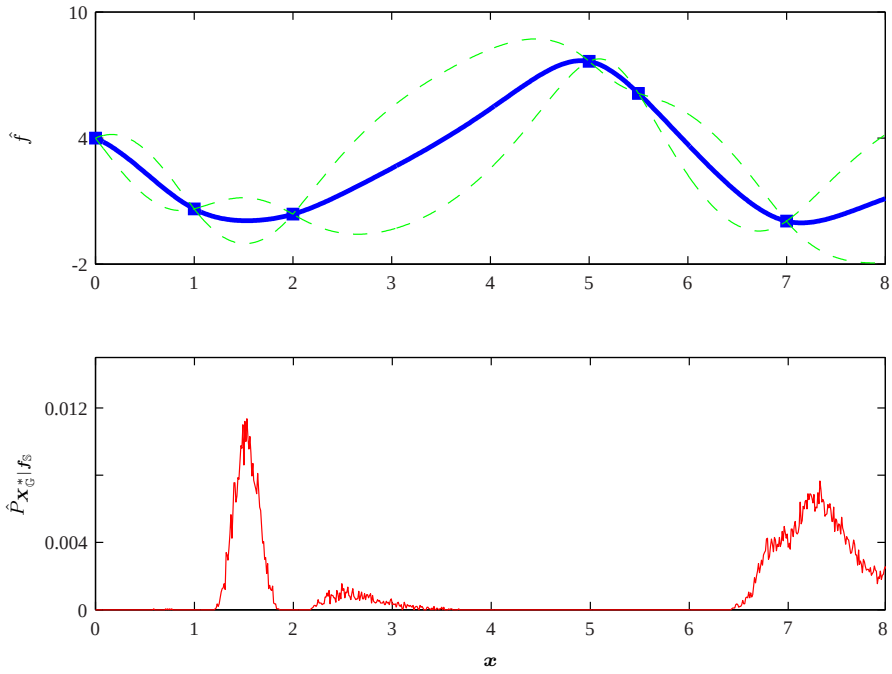


Fig. 4 Estimation of the distribution of $\boldsymbol{X}^*_{\mathbb{G}}$: (*top*) Kriging interpolation, 95% confidence intervals and sample points; (*bottom*) estimated distribution of $\boldsymbol{X}^*_{\mathbb{G}}$ using 10000 conditional simulations of F and a regular grid for $\mathbb{G}$.

with δ the Kronecker symbol. Figure 4 presents the approximation $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}$ for an example where locating a global minimizer is not easy. Knowing the conditional distribution of $\boldsymbol{X}^*_{\mathbb{G}}$ gives valuable information on the areas of $\mathbb{X}$ where a global minimizer might be located, and that ought to be investigated. This idea will be detailed in the next section.

4 The stepwise uncertainty reduction strategy

The knowledge about the global minimizers of f is summarized by $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}$. In order to evaluate the interest of a new evaluation of f at a given point, a measure of the expected information gain is required. An efficient measure is *conditional entropy*, as used in sequential testing [9] in the *Stepwise Uncertainty Reduction* (SUR) strategy. This section extends the SUR strategy to global optimization.

4.1 Conditional entropy

The entropy of a discrete random variable U (expressed in bits) is defined as:

$$H(U) \;=\; -\sum_{u} \mathrm{P}(U=u)\log_2\,\mathrm{P}(U=u).$$

$H(U)$ measures the spread of the distribution of U. It decreases as this distribution gets more peaked. In particular :

- $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x}) = 1/N \;\; \forall \boldsymbol{x} \in \mathbb{G} \;\Rightarrow\; H(\boldsymbol{X}^*_{\mathbb{G}}) = \log_2(N)$,
- $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x}) = \begin{cases} 0 \;\text{ if }\; \boldsymbol{x} \neq \boldsymbol{x}_0 \\ 1 \;\text{ if }\; \boldsymbol{x} = \boldsymbol{x}_0 \end{cases} \;\Rightarrow\; H(\boldsymbol{X}^*_{\mathbb{G}}) = 0$

Similarly, for any event $\mathcal{B}$, the entropy of U relative to the probability measure $\mathrm{P}(.|\mathcal{B})$ is

$$H(U|\mathcal{B}) \;=\; -\sum_{u} \mathrm{P}(U=u|\mathcal{B})\log_2\,\mathrm{P}(U=u|\mathcal{B}).$$

The conditional entropy of U given another discrete random variable V is

$$H(U|V) \;=\; \sum_{v} \mathrm{P}(V=v) H(U|V=v),$$

and the conditional entropy of U given $\mathcal{B}$ and V is

$$H(U|\mathcal{B},V) \;=\; \sum_{v} \mathrm{P}(V=v|\mathcal{B}) H(U|\mathcal{B},V=v). \tag{10}$$

Note that $H(U|V)$ and $H(U|\mathcal{B},V)$ are, despite the similarity of notation with conditional expectation, deterministic quantities. More details on conditional entropy can be found in [5].

4.2 Conditional minimizers entropy

Let $F_Q(\boldsymbol{x})$ be a discrete version of $F(\boldsymbol{x})$, defined as $F_Q(\boldsymbol{x}) = Q(F(\boldsymbol{x}))$ with Q a quantization operator. Q is characterized by a finite set of M real numbers $\{y_1, \ldots, y_M\}$, and defined $\forall u \in \mathbb{R}$ as

$$Q(u) = y_k \text{ with } k = \min_i |y_i - u|. \tag{11}$$

For optimization problems, the SUR strategy for the selection of the next value of $\boldsymbol{x} \in \mathbb{X}$ at which f will be evaluated will be based on $H(\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}, F_Q(\boldsymbol{x}))$, the conditional entropy of $\boldsymbol{X}^*_{\mathbb{G}}$ given the evaluation results $\{\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}\}$ and $F_Q(\boldsymbol{x})$ (we shall refer to it later on as conditional entropy of the minimizers, or simply minimizers entropy).

Using (10) we can write

$$H(\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}, F_Q(\boldsymbol{x})) \;=\; \sum_{i=1}^{M} \mathrm{P}(F_Q(\boldsymbol{x}) = y_i|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}) H(\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}, F_Q(\boldsymbol{x}) = y_i) \tag{12}$$

with

$$H(\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}, F_Q(\boldsymbol{x}) = y_i) \;=\; -\sum_{\mathbf{u}\in\mathbb{G}} P_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}},y_i}(\mathbf{u}) \log_2 P_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}},y_i}(\mathbf{u}) \;,$$

and

$$P_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}},y_i}(\boldsymbol{u}) = \mathrm{P}(\boldsymbol{X}^* = \boldsymbol{u}|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}, F_Q(\boldsymbol{x}) = y_i).$$

$H(\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}, F_Q(\boldsymbol{x}))$ is a measure of the anticipated uncertainty remaining in $\boldsymbol{X}^*_{\mathbb{G}}$ given the candidate evaluation point $\boldsymbol{x}$ and the result $\boldsymbol{f}_{\mathbb{S}}$ of the previous evaluations. Anticipation is introduced in (12) by considering the entropy of $\boldsymbol{X}^*_{\mathbb{G}}$ resulting from every possible sample value of $F_Q(\boldsymbol{x})$. At each stage of the iterative optimization, the SUR strategy retains for the next evaluation a point that minimizes the expected entropy of the minimizers distribution after the evaluation, i.e., a point that maximizes the expected gain in information about $\boldsymbol{X}^*_{\mathbb{G}}$.

The conditional entropy of the minimizers thus takes in account the conditional statistical properties of F and particularly the covariance function of the model. There lies the interest of the SUR strategy applied to global optimization. It makes use of what has been previously assumed and learned about f to pick up the most informative evaluation point. By contrast, the EI criterion (as most standard criteria) depends only on the conditional mean and variance of F at the design point being considered.

5 Implementing the SUR strategy

5.1 IAGO algorithm

Our algorithm is similar in spirit to the strategy for Kriging-based optimization known as *Efficient Global Optimization* (EGO) [14]. EGO starts with a small initial design, estimates the parameters of the covariance function of F and computes the 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 using the EI criterion. The parameters of the

Algorithm
Input: Set $\mathbb{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of evaluation points and corresponding values $\boldsymbol{f}_\mathbb{S}$ of the function f
Output: Additional evaluation point $\boldsymbol{x}_{\text{new}}$
1. Choose $\mathbb{G}$, a discrete representation of $\mathbb{X}$
2. Set covariance parameters either a priori or by maximum-likelihood estimation based on $\boldsymbol{f}_\mathbb{S}$
3. Compute r non-conditional simulations over $\mathbb{G}$
4. Compute $\hat{f}(\boldsymbol{x})$ and $\hat{\sigma}(\boldsymbol{x})$ over $\mathbb{G}$ by Kriging from $\boldsymbol{f}_\mathbb{S}$
5. **while** the set of candidate points has not been entirely explored
6. **do** Take an untried point $\boldsymbol{x}_\text{c}$ in the set of candidate points
7. Compute the parameters $\{y_1, \ldots, y_M\}$ of the quantization operator Q
8. Compute the Kriging coefficients at every point in $\mathbb{G}$ based on evaluation points in $\mathbb{S}$ and $\boldsymbol{x}_\text{c}$
9. **for** $i \leftarrow 1$ **to** M
10. **do** Construct conditional simulations using (7) and assuming that $f(\boldsymbol{x}_\text{c}) = y_i$
11. Find a global minimizer $\boldsymbol{x}_k^*$ of the k-th conditional simulation over $\mathbb{G}$ ($k = 1, \ldots, r$)
12. Estimate $P_{\boldsymbol{X}_\mathbb{G}^*|\boldsymbol{f}_\mathbb{S}, y_i}$ over $\mathbb{G}$ using (9)
13. Compute $H(\boldsymbol{X}_\mathbb{G}^*|\boldsymbol{F}_\mathbb{S} = \boldsymbol{f}_\mathbb{S}, F_Q(\boldsymbol{x}_\text{c}) = y_i)$
14. Compute the minimizers entropy given an evaluation at $\boldsymbol{x}_\text{c}$ using (12)
15. Output $\boldsymbol{x}_\text{new}$ that minimizes the conditional entropy over the set of candidate points

Table 1 Selection of a new evaluation point for f.

covariance function are then re-estimated, the model re-computed, and the process of choosing new points continues until the improvement expected from sampling additional points has become sufficiently small. The IAGO algorithm uses the same idea of iterative incorporation of the obtained information to the prior on the function, but with a different criterion.

To compute the minimizers entropy using (12), a different quantization operator Q_x is used for each value of $\boldsymbol{x}$ to improve the precision with which the empirical mean of entropy reduction over possible evaluation results is computed. We use the fact that $F(\boldsymbol{x})$ is conditionally Gaussian with mean $\hat{f}(\boldsymbol{x})$ and variance $\hat{\sigma}^2(\boldsymbol{x})$ obtained by Kriging, to select a set of values $\{y_1(\boldsymbol{x}), \ldots, y_M(\boldsymbol{x})\}$, such that

$$P(F_{Q_{\boldsymbol{x}}}(\boldsymbol{x}) = y_i | \boldsymbol{F}_\mathbb{S} = \boldsymbol{f}_\mathbb{S}) = \frac{1}{M} \; \forall \, i \in [\![1 : M]\!] \, . \tag{13}$$

Here we used a set of ten possible values ($M = 10$).

For each of these possible values (or hypotheses $F(\boldsymbol{x}) = y_i$), $\hat{P}_{\boldsymbol{X}_\mathbb{G}^*|\boldsymbol{f}_\mathbb{S}, y_i}$ is computed using conditional simulations. The minimizers entropy is then obtained using (12). These operations are carried out on a discrete set of candidate evaluation points (see Section 5.2 for some details on the choice of this set), and a new evaluation of f is finally performed at a point that minimizes minimizers entropy. Next, as in the EGO algorithm, the covariance parameters are re-estimated and the model re-computed. The procedure for the choice of an additional evaluation point is described in Table 1.

When the number of additional function evaluations is not specified beforehand, we propose to use as a stopping criterion the conditional probability that the global minimum of the GP model be no further apart of $f_\text{min} = \min_{\boldsymbol{x}_i \in \mathbb{S}} f(\boldsymbol{x}_i)$ (the best function value yet obtained) than a given tolerance threshold δ. The algorithm then stops when

$$\mathrm{P}(F^* < f_\text{min} + \delta | \boldsymbol{F}_\mathbb{S} = \boldsymbol{f}_\mathbb{S}) < P_\text{Stop} \, ,$$

with $F^* = \min_{\boldsymbol{x} \, \in \, \mathbb{G}} F(\boldsymbol{x})$, and $P_\text{Stop} \in [0, 1]$ a critical value to be chosen by the user. Proposed in [18], this stopping criterion is well suited here, since evaluating the repartition function of $f(\boldsymbol{x}^*)$ does not require any additional computation. We can indeed use the conditional simulations that have been performed to approximate the conditional distribution of $\boldsymbol{X}_\mathbb{G}^*$ for this purpose, provided that we keep track, for each of them, not only of a global minimizer, but also of the minimum. The histogram thus obtained can then easily be transformed into a simple approximation of the conditional repartition function of the minimum.

5.2 Computational complexity

With the previous notation, n the number of evaluation points, r the number of conditional simulations, N the number of points in $\mathbb{G}$ and M the number of discretized potential evaluation results for an evaluation, the computational complexity for the approximation of the minimizers entropy (Steps 7 to 14 in Table 1) is as follows:

- computing Kriging coefficients at every point in $\mathbb{G}$ (Step 8): $\mathrm{O}(n^2 N)$, as (20) (to be found in appendix) has to be solved N times while changing the $n+1$-st evaluation point each time. A large part of the factorization of the covariance matrix can be reused, and Kriging at an untried point is then simply in $\mathrm{O}(n^2)$,
- constructing conditional simulations (Step 10): $\mathrm{O}(nrN)$ (M is not involved since the main part of the conditioning procedure described by (8) can be carried out outside the loop on the discretized potential evaluation results),
- locating the global minimizers for each simulation by exhaustive search (Step 11): $\mathrm{O}(rNM)$,

Since all other operations are in $\mathrm{O}(N)$ at most, evaluating minimizers entropy at any given point requires $\mathrm{O}(N)$ operations.

To complete the description of an implementable algorithm, we must specify a choice for $\mathbb{G}$ and a policy for the minimization of minimizers entropy. What follows is just an example of a possible strategy, and many variants could be considered.

The simplest choice for $\mathbb{G}$ is a uniform grid on $\mathbb{X}$. However, as the number of evaluations of f increases, the spread of $P_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}$ diminishes along with the precision for the computation of the entropy. To keep a satisfactory precision over time, $\mathbb{G}$ can be a random sample of points in $\mathbb{X}$, re-sampled after every evaluation of f with the distribution $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}$. Re-sampling makes it possible to use a set $\mathbb{G}$ with a smaller cardinal and to escape, at least partly, the curse of dimensionality (to resample using $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}$, any non-parametric density estimator could be used along with a sampling method such as Metropolis-Hastings, see, e.g., [3]).

Ideally, to choose an additional evaluation point for f using IAGO, minimizers entropy should be minimized over $\mathbb{X}$. However, this of course is in itself a global optimization problem, with many local optima. It would be possible to design an ad-hoc optimization method (as in [13]), but this perspective is not explored here. Instead, we evaluate the criterion extensively over a chosen set of candidate points. Note that only the surrogate model is involved at this stage, which makes the approach practical. The idea is, exactly as for the choice of $\mathbb{G}$, to use a space-filling sample covering $\mathbb{X}$ and resampled after each new evaluation. The current implementation of IAGO simply uses a Latin Hyper Cube (LHC) sample, however, it would be easy to adapt this sample iteratively using the conditional distribution of the minimizers $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}$ as a prior. For instance, areas of the design space where the distribution is sufficiently small could be ignored. After a few evaluations, a large portion of the design space usually satisfies this property, and the computations saved could be used to improve knowledge on the criterion by sampling where $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}$ is high (using the same approach as for the choice of $\mathbb{G}$).

As dimension increases, trying to cover the factor space while keeping the same accuracy leads to an exponential increase in complexity. However, in a context of expensive function evaluation, the objective is less to specify exactly all global minimizers (which would be too demanding in function evaluations anyway), than to use available information efficiently to reduce the likely areas for the location of these minimizers. This is exactly the driving concept behind IAGO. In practice, within a set of one thousand candidate points, picking an additional evaluation point requires about five minutes with a standard personal computer (and this figure is relatively independent of the dimension of factor space). Moreover, the result obtained can be trusted to be a consistent choice within this set of candidate points, in regard of what has been assumed and learned about f.

5.3 Taking noise in account

Practical optimization problems often involve noise. This section discusses possible adaptations of the optimization algorithm that make it possible to deal with noisy situations, namely noise on the evaluation of f and noise on the factors.

5.3.1 Noise on the evaluation of f

When the results of the evaluations of f are corrupted by noise, the algorithm must take this fact into account. A useful tool to deal with such situations is *non-interpolative Kriging* (see Section 8.2).

If the evaluation at $\boldsymbol{x}_i \in \mathbb{S}$ is assumed to be corrupted by an additive Gaussian noise ε_i with known mean and variance, the Kriging prediction should no longer be interpolative. The optimization algorithm remains

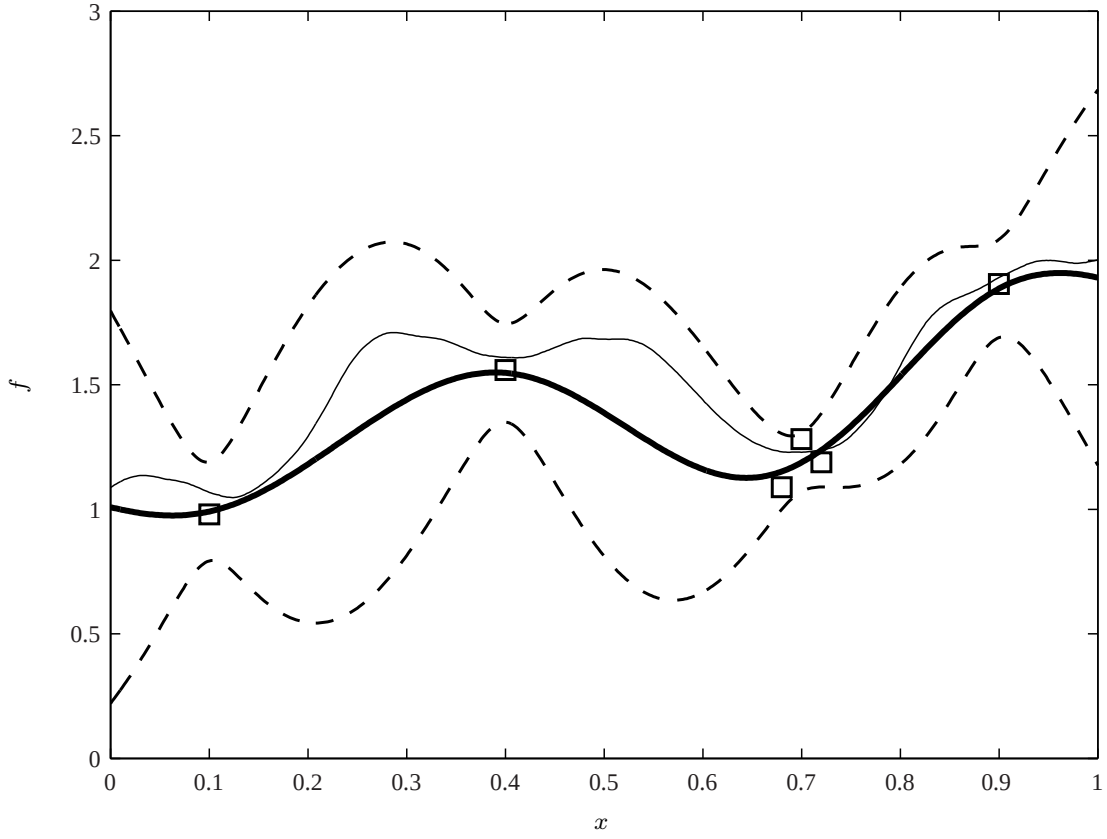


Fig. 5 Example of prediction by Kriging (bold line) of noisy measurements represented by squares. Dashed lines represent 95% confidence regions for the prediction and the thin solid line is an example of conditional simulation obtained using the method presented in Section 5.3.1.

nearly unchanged, except for the conditional simulations. Sample paths of F, should be built conditionally to evaluation results, *i.e.* realizations of the random variables $f(\boldsymbol{x}_i) + \varepsilon_i$ for $\boldsymbol{x}_i \in \mathbb{S}$. Since the variance of the prediction error is no longer zero at evaluation points (in other words, there is some uncertainty left on the values of f at evaluation points), we first have to sample, at each evaluation point, from the distribution of F conditionally to noisy evaluation results. An interpolative simulation, based on these samples, is then built using conditioning by Kriging. An example of such a simulation is presented on Figure 5 for a noise variance of 0.01.

5.3.2 Noise on the factors

In many industrial design problems, the variability of the values of the factors in mass production has a significant impact on performance. One might then want to design a system that optimizes some performance measure while ensuring that performance uncertainty (stemming from noise on the factors) remains under control. These so-called *robust optimization* problems can generally be written as

$$\underset{\boldsymbol{x} \in \mathbb{D}}{\arg\min}\, J(\boldsymbol{x})\,, \tag{14}$$

with $J(\boldsymbol{x})$ a cost function reflecting some statistical property of the corrupted performance measure $f(\boldsymbol{x} + \boldsymbol{\varepsilon})$, where $\boldsymbol{\varepsilon}$ is a random vector accounting for noise on the factors. Classical cost functions are:

- mean: $J(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\varepsilon}}[f(\boldsymbol{x} + \boldsymbol{\varepsilon})]$,
- standard deviation: $J(\boldsymbol{x}) = \sqrt{\mathrm{var}_{\boldsymbol{\varepsilon}}(f(\boldsymbol{x} + \boldsymbol{\varepsilon}))}$,
- linear combination of mean and standard deviation: $J(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\varepsilon}}[f(\boldsymbol{x} + \boldsymbol{\varepsilon})] + \sqrt{\mathrm{var}_{\boldsymbol{\varepsilon}}(f(\boldsymbol{x} + \boldsymbol{\varepsilon}))}$,

– α-quantile: $J(\boldsymbol{x}) = Q^\alpha(\boldsymbol{x})$ with $Q^\alpha(\boldsymbol{x})$ such that $\mathrm{P}(f(\boldsymbol{x}+\boldsymbol{\varepsilon}) < Q^\alpha(\boldsymbol{x})) = \alpha$.

Using, for example, the α-quantile as a cost function, it is possible to adapt our optimization algorithm to solve (14). Given a set of evaluation results $\boldsymbol{f}_\mathbb{S}$ at noise-free evaluation points, and if it is possible to sample from the distribution $p_{\boldsymbol{\varepsilon}}$ of $\boldsymbol{\varepsilon}$, a Monte Carlo approximation $\hat{Q}^\alpha(\boldsymbol{x})$ of $Q^\alpha(\boldsymbol{x})$ is easily obtained by computing $\hat{f}(\boldsymbol{x}+\boldsymbol{\varepsilon})$ over a set sampled from $p_{\boldsymbol{\varepsilon}}$. The global optimization algorithm can then be applied to $Q^\alpha(\boldsymbol{x})$ instead of f, using pseudo-evaluations $\hat{\boldsymbol{Q}}^\alpha_\mathbb{S} = [\hat{Q}^\alpha(\boldsymbol{x}_1), \ldots, \hat{Q}^\alpha(\boldsymbol{x}_n)]$ (recomputed after each evaluation of f) instead of $\boldsymbol{f}_\mathbb{S}$. This naive approach can certainly be improved, but is sufficient to show the feasibility of a robust approach and to illustrate on a simple example (to be presented in the next section) the impact of $\boldsymbol{\varepsilon}$ on the evaluation points to be chosen by IAGO.

It is of course possible to combine these ideas and to deal simultaneously with noise both on the factors and the function evaluations.

6 Illustrations

This section presents some simple examples of global optimization using IAGO, with a regular grid as a set of candidate evaluation points. An empirical comparison with global optimization using expected improvement is also presented. The Matérn covariance class will be used for Kriging prediction, as it facilitates the tuning of the variance, regularity and range of correlation of the underlying random process, but note that any kind of admissible covariance function could have been used. The parameters of the covariance may be estimated from the data using a maximum-likelihood approach (see Section 8.3).

6.1 A one-dimensional example

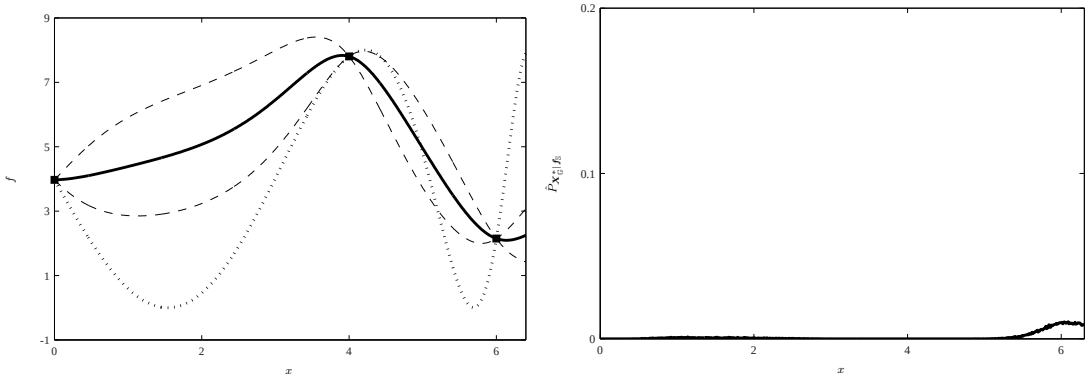
Consider the function with two global minimizers illustrated by Figure 6 and defined by $f : x \longmapsto 4[1 - \sin(x + 8\exp(x-7))]$. Given an initial design consisting of three points, the IAGO algorithm is used to compute six additional points iteratively. The final Kriging model is depicted in the left part of Figure 6, along with the resulting conditional distribution for the minimizers on the right part. After adding some noise on the function evaluations, the variant of IAGO presented in Section 5.3.1 is also applied to the function with the same initial design. In both cases, six additional evaluations have significantly reduced the uncertainty associated with the position of the global minimizers. The remaining likely locations reduce to small areas centered on the two actual global minimizers. In the noisy case, larger zones are identified, a direct consequence of the uncertainty associated with the evaluations.

Figure 7 illustrates robust optimization using the same function and initial design, but considering an additive zero-mean Gaussian noise on the factors with a standard deviation of 0.2. The cost function used is the 90%-quantile $Q^{90\%}$, which is computed on the surrogate model but also, and only for the sake of comparison, on the true function using Monte Carlo uncertainty propagation (the quantile is approximated using 5000 simulations). After six iterations of the robust optimization algorithm, the distribution of the robust minimizers is sufficiently peaked to give a good approximation of the true global robust minimizer.

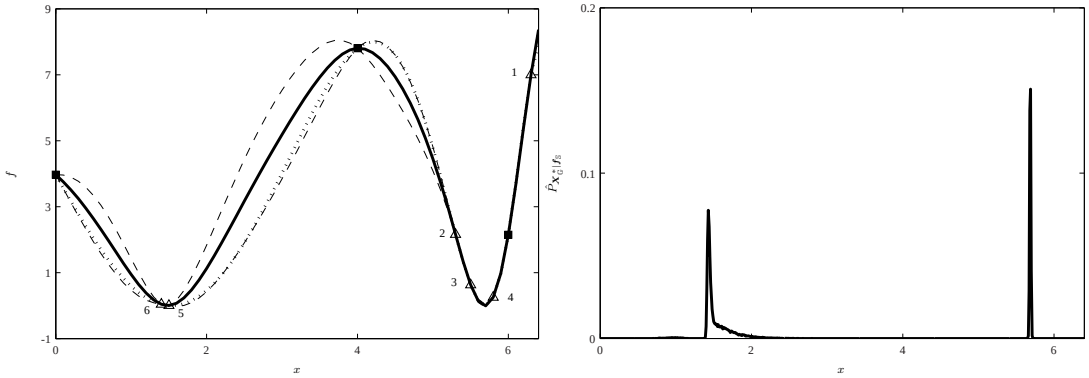
These result are encouraging as they show that the requirement of fast uncertainty reduction is met. The next section provides some more examples, along with a comparison with EGO, the EI-based global optimization algorithm.

6.2 Empirical comparison with expected improvement

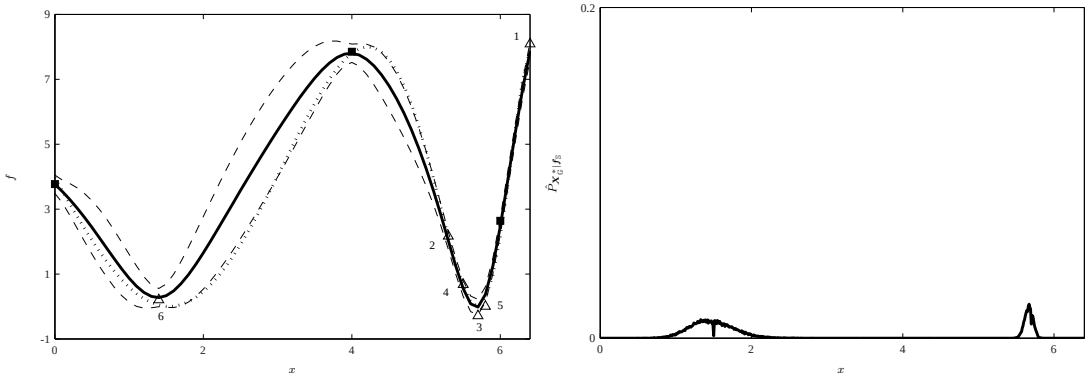
Consider first the function described by Figure 8. Given an initial design of three points, both EI and minimizers entropy are computed. Their optimization provides two candidate evaluation points for f, which are also presented on Figure 8, along with the post-evaluation prediction and conditional distribution for $\boldsymbol{X}^*_\mathbb{G}$. For this example, the regularity parameter of the Matérn covariance is set a priori to a high value (2.5). By taking in account the covariance function of F through conditional simulations, the minimizers entropy uses regularity to conclude faster. The resulting conditional distribution of the minimizers is then generally more peaked using the IAGO algorithm than using the EGO algorithm (as illustrated by Figure 8(c) and Figure 8(b)).



(a) Kriging prediction and conditional distribution of the global minimizers based on the initial design



(b) Standard IAGO algorithm (noise free case)



(c) IAGO algorithm for noisy evaluations (the additive noise is zero-mean Gaussian with standard deviation 0.2)

Fig. 6 Example of global optimization using IAGO on a function of one variable (dotted line), with an initial design consisting of three points (represented by squares). Six additional evaluations are carried out (triangles) using two versions of the IAGO algorithm. The graphs on the *left part* of the figure account for the predictions, while the *right part* presents the corresponding conditional distributions of the global minimizers.

Consider now the Branin function (see, for instance, [8]), defined as

$$f : [-5, 10] \times [0, 15] \longrightarrow \mathbb{R}$$
$$(x_1, x_2) \longmapsto \left(x_2 - \tfrac{5.1}{4\pi^2}x_1^2 + \tfrac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \tfrac{1}{8\pi}\right)\cos(x_1) + 10.$$

It has three global minimizers $\boldsymbol{x}_1^* \approx (-3.14, 12.27)^\mathsf{T}$, $\boldsymbol{x}_2^* \approx (3.14, 2.27)^\mathsf{T}$ and $\boldsymbol{x}_3^* \approx (9.42, 2.47)^\mathsf{T}$, and the global minimum is approximately equal to 0.4. Given an initial uniform design of sixteen points, fifteen additional points are iteratively selected and evaluated using the IAGO and EGO algorithms. The parameters of the Matèrn covariance are estimated on the initial design, and kept unchanged during both procedures.

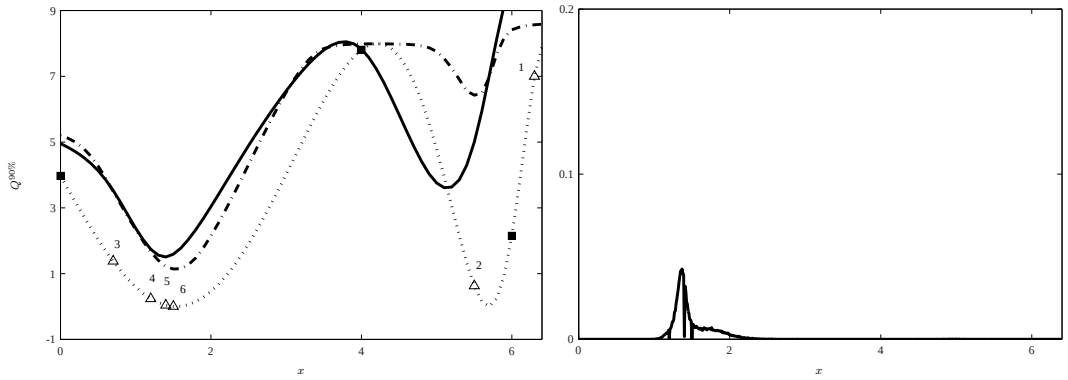


Fig. 7 Example of robust optimization using IAGO and the cost function $Q^{90\%}$. The function f (dotted line), corrupted by an additive Gaussian noise on the factor (zero mean with a standard deviation of 0.2), is studied starting from the initial design of three points already used in Figure 6. Six additional evaluations are carried out (triangles), which are used to estimate the cost function based on the Kriging model (bold line), along with the conditional distribution of the robust minimizers (*right*). The cost function $Q^{90\%}$ estimated, only for the sake of comparison, from the true function using Monte Carlo uncertainty propagation is also provided (mixed line).

Table 2 Estimation results for the Branin function using the evaluations of Figure 9

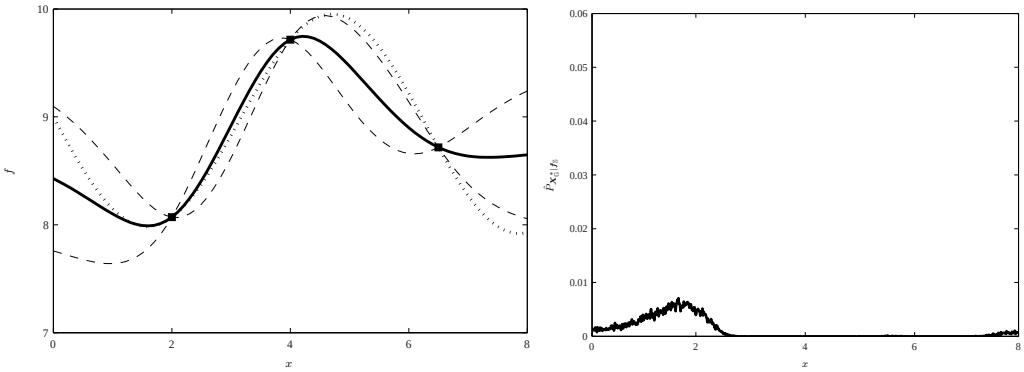
	EGO		IAGO	
	15 iterations	35 iterations	15 iterations	35 iterations
Euclidean distance between $\boldsymbol{x}_1^*$ and its final estimate	3.22	3.22	**2.18**	**0.23**
Value of the true function at estimated minimizer	17.95	17.95	**2.59**	**0.40**
Euclidean distance between $\boldsymbol{x}_2^*$ and its final estimate	2.40	2.40	**0.44**	**0.18**
Value of the true function at estimated minimizer	13.00	13.00	**0.85**	**0.42**
Euclidean distance between $\boldsymbol{x}_3^*$ and its final estimate	**0.04**	**0.04**	0.82	0.23
Value of the true function at estimated minimizer	**0.40**	**0.40**	1.94	0.44

The positions of the evaluation points are presented on Figure 9 (left), along with the three global minimizers. Table 2 summarizes the results obtained with EGO and IAGO, based on the final Kriging models obtained with both approaches. Note that the EI criterion in EGO is maximized with a high precision, while minimizers entropy in IAGO is computed over a thousand candidate evaluation points located on a regular grid. It appears nevertheless that the algorithm using EI stalls on a single global minimizer, while the minimizers entropy allows a relatively fast estimation of all three of them. Besides IAGO yields a better global approximation of the supposedly unknown function. If twenty additional evaluations are carried out (as presented in the right part of Figure 9), the final Kriging prediction using minimizers entropy estimates the minimum with an error of less than 0.05 for all three minimizers (cf. Table 2), while the use of EI does not improve the information on any minimizer any further. The difference between the two strategies is clearly evidenced. The EI criterion, overestimating the confidence in the initial prediction, has led to performing evaluations extremely close to one another, for a very small information gain. In a context of expensive function evaluation, this is highly detrimental. The entropy criterion, using the same covariance parameters, does not stack points almost at the same location before having identified the most likely zones for the minimizers. The use of what has been assumed and learned about the function is clearly more efficient in this case, and this property should be highly attractive when dealing with problems of higher dimension.

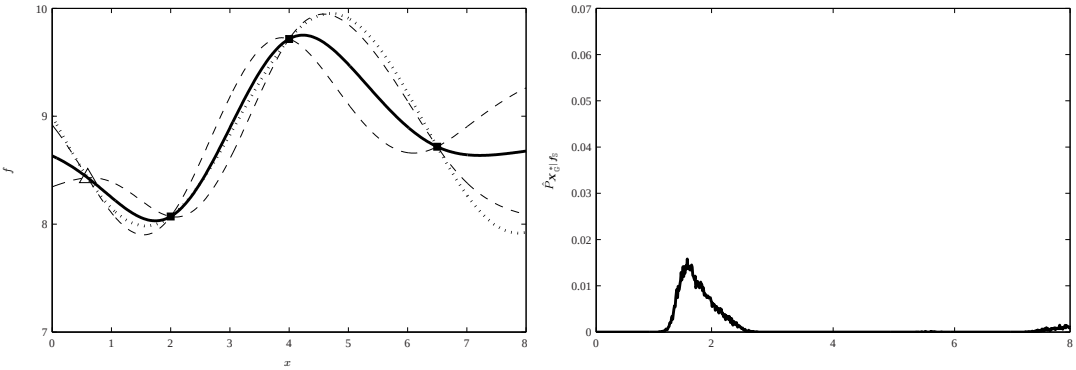
7 Discussion

7.1 Robustness to uncertainty on the covariance parameters

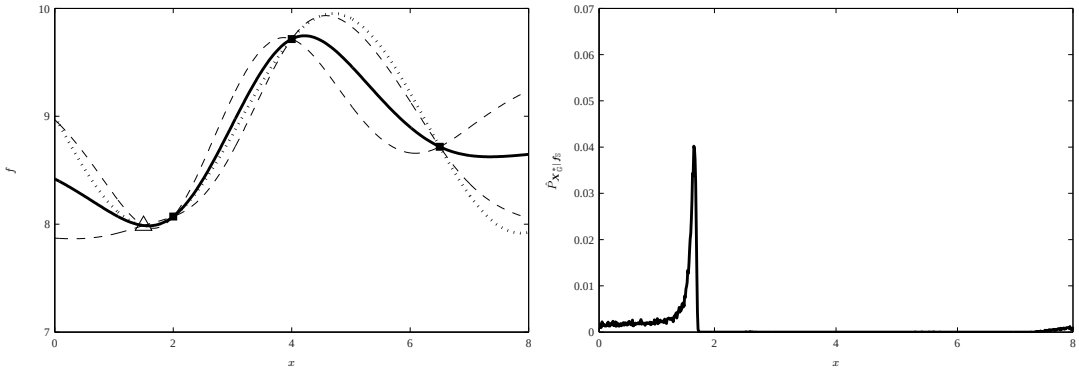
Jones studied in [13] the potential of Kriging-based global optimization methods such as EGO. One of his most important conclusion, is that these methods "*can perform poorly if the initial sample is highly*



(a) Initial prediction and minimizers distribution



(b) Prediction and minimizers distribution after an additional evaluation of f chosen with EI



(c) Prediction and minimizers distribution after an additional evaluation of f chosen with minimizers entropy

Fig. 8 Comparison between minimizers entropy and EI: the *left* side contains the Kriging predictions before and after an additional evaluation chosen with either EI or minimizers entropy, while the *right* side presents the corresponding conditional distribution of the global minimizers.

deceptive". An eloquent example is provided on page 373 [13], where a sine function is sampled using its own period, leading to a flat prediction over the domain, associated with a small prediction error.

This potential for deception is present throughout the IAGO procedure, and should not be ignored. To overcome this difficulty, several methods have been proposed (see, e.g., Enhanced Method 4 in [13] or [10]), which achieve some sort of robustness to an underestimation of the prediction error and more generally to a bad choice of covariance function. They seem to perform better than classical algorithms, including EGO.

Comparing the IAGO approach to such methods is an interesting topic for future research. The issue considered here was to demonstrate the interest of the minimizers entropy criterion, and we felt that this had to be done independently from the rest of the procedure.

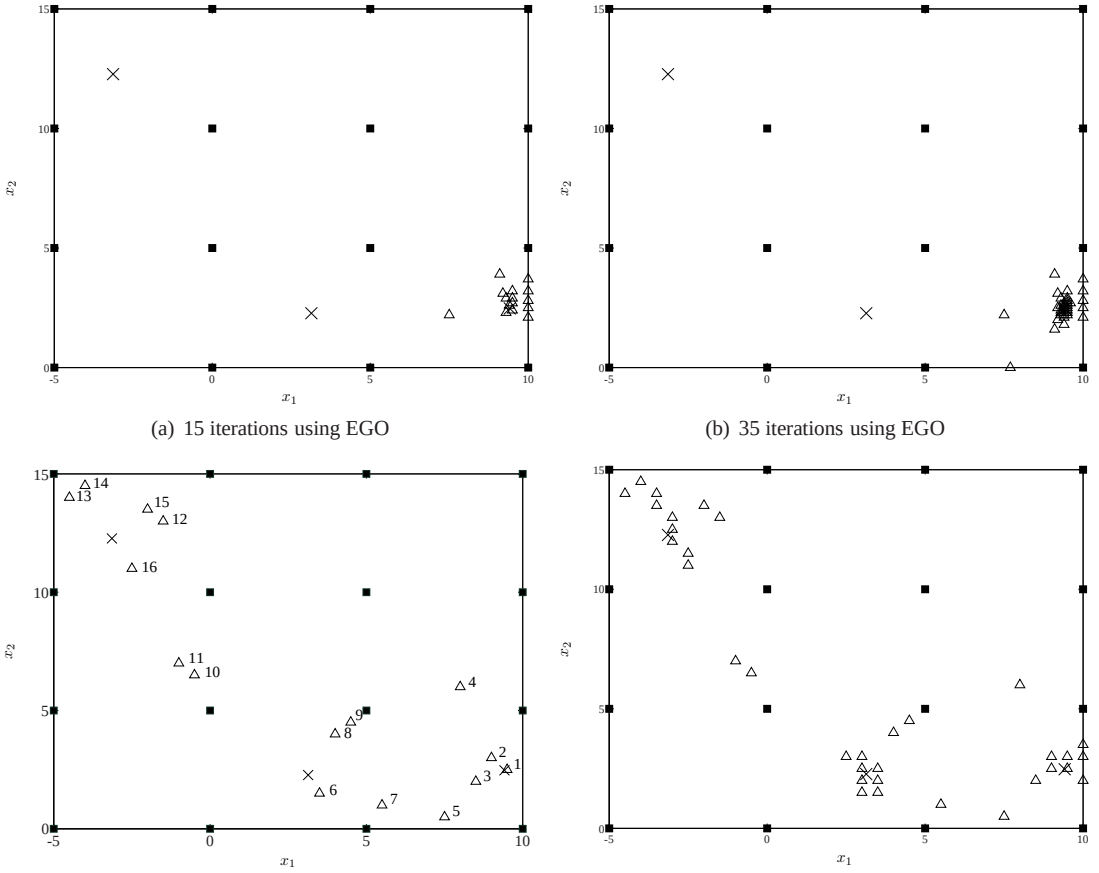


Fig. 9 Fifteen iterations of two optimization algorithms, that differ by their criteria for selecting evaluation points for f, on the Branin function: (*top*) the EI criterion is used, (*bottom*) the minimizers entropy criterion is used with a thousand candidate evaluation points for f set on a regular grid (squares account for initial data, triangles for new evaluations, and crosses give the actual locations of the three global minimizers).

It is of course essential to make IAGO robust to errors in the estimation of the covariance parameters. In many industrial problems, this can be easily done by using prior knowledge on the unknown function to restrict the possible values for these parameters. For example, experts of the field often have information regarding the range of values attainable by the unknown function. This information can be directly used to restrict the search space for the variance of the modeling process F, or even to choose it beforehand.

More generally, given the probabilistic framework used here, it should be relatively easy to develop a Bayesian or minimax extension of IAGO to guide the estimation of the parameters of the covariance function. A comparison with robust methods such as those detailed in [13] will then be essential.

7.2 Conclusions and perspectives

In this paper, a stepwise uncertainty reduction strategy has been used for the sequential global optimization of expensive-to-evaluate functions. This strategy iteratively selects a minimizer of the conditional minimizers entropy as the new evaluation point. To compute this entropy, a Gaussian random model of the function evaluations is used and the minimizers entropy is estimated through Kriging and conditional simulations. At each iteration, the result of the new evaluation is incorporated in the data base used to re-build the Kriging model (with a possible re-estimation of the parameters of its covariance function).

We have shown on some simple examples that, compared to the classical EI-based algorithm EGO, the method proposed significantly reduces the evaluation effort in the search for global optimizers. The

stepwise uncertainty reduction strategy allows the optimization method to adapt the type of search to the information available on the function. In particular, the minimizers entropy criterion makes full use of the assumed regularity of the unknown function to balance global and local searches.

Choosing an adequate set of candidate points is crucial, as it must allow a good estimation of a global minimizer of the criterion, while keeping computation feasible. Promising results have already been obtained with space-filling designs, and adaptive sampling based on the conditional density of the global minimizers should be useful as dimension increases.

Extension to constrained optimization is an obviously important topic for future investigations. When it is easy to discard the candidate points in $\mathbb{X}$ that do not satisfy the constraints, the extension is trivial. For expensive-to-evaluate constraints, the extension is a major challenge.

Finally, the stepwise uncertainty reduction strategy associated with conditioning by Kriging is a promising solution for the robust optimization of expensive-to-evaluate functions, a problem that is central to many industrial situations, for which an efficient product design must be found in the presence of significant uncertainty on the values actually taken by some factors in mass production. In addition, robustness to the uncertainty associated with the estimation of the parameters of the covariance function should also be sought.

8 Appendix: modeling with Gaussian processes

This section recalls the main concepts used in this paper, namely Gaussian process modeling and Kriging. The major results will be presented along with the general framework for the estimation of the model parameters.

8.1 Kriging when f is evaluated exactly

Kriging [4, 15] is a prediction method based on random processes that can be used to approximate or interpolate data. It can also be understood as a kernel regression method, such as *splines* [23] or *Support Vector Regression* [20]. It originates from geostatistics and is widely used in this domain since the 60s. Kriging is also known as the *Best Linear Unbiased Prediction* (BLUP) in statistics, and has been more recently designated as Gaussian Processes (GP) in the 90s in the machine learning community.

As mentioned in Section 2.1, it is assumed that the function f is a sample path of a Gaussian random process F. Denote by $m(\boldsymbol{x}) = E[F(\boldsymbol{x})]$ the mean function of $F(\boldsymbol{x})$ and by $k(\boldsymbol{x}, \boldsymbol{y})$ its covariance function, written as

$$k(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{cov}(F(\boldsymbol{x}), F(\boldsymbol{y})).$$

Kriging then computes the BLUP of $F(\boldsymbol{x})$, denoted by $\hat{F}(\boldsymbol{x})$, in the vector space generated by the evaluations $\mathbb{H}_{\mathbb{S}} = \mathrm{span}\{F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)\}$. As an element of $\mathbb{H}_{\mathbb{S}}$, $\hat{F}(\boldsymbol{x})$ can be written as

$$\hat{F}(\boldsymbol{x}) = \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{F}_{\mathbb{S}}\,. \tag{15}$$

As the BLUP, $\hat{F}(\boldsymbol{x})$ must have the smallest variance for the prediction error

$$\hat{\sigma}^2(\boldsymbol{x}) = \mathbb{E}[(\hat{F}(\boldsymbol{x}) - F(\boldsymbol{x}))^2], \tag{16}$$

among all unbiased predictors. The variance of the prediction error satisfies

$$\hat{\sigma}^2(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{K} \boldsymbol{\lambda}(\boldsymbol{x}) - 2\boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{k}(\boldsymbol{x}), \tag{17}$$

with

$$\boldsymbol{K} = (k(\boldsymbol{x}_i, \boldsymbol{x}_j))\,,\ (i,j) \in [\![1, n]\!]^2$$

the $n \times n$ covariance matrix of F at evaluation points in $\mathbb{S}$, and

$$\boldsymbol{k}(\boldsymbol{x}) = [k(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, k(\boldsymbol{x}_n, \boldsymbol{x})]^{\mathsf{T}}$$

the vector of covariances between $F(\boldsymbol{x})$ and $\boldsymbol{F}_{\mathbb{S}}$

The prediction method [16] assumes that the mean of $F(\boldsymbol{x})$ can be written as a finite linear combination

$$m(\boldsymbol{x}) = \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{p}(\boldsymbol{x}),$$

where $\boldsymbol{\beta}$ is a vector of fixed but unknown coefficients, and

$$\boldsymbol{p}(\boldsymbol{x}) = [p_1(\boldsymbol{x}), \ldots, p_l(\boldsymbol{x})]^{\mathsf{T}}$$

is a vector of known functions of the factor vector $\boldsymbol{x}$. Usually these functions are monomials of low degree in the components of $\boldsymbol{x}$ (in practice, their degree does not exceed two). These functions may be used to reflect some prior knowledge on the unknown function. As we have none for the examples considered here, we simply use an unknown constant.

The Kriging predictor at $\boldsymbol{x}$ is then the best linear predictor subject to the unbiasedness constraint $\mathbb{E}(\hat{F}(\boldsymbol{x})) = m(\boldsymbol{x})$, whatever the unknown $\boldsymbol{\beta}$. The unbiasedness constraint translates into

$$\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{P}^{\mathsf{T}} \boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{p}(\boldsymbol{x}), \tag{18}$$

with

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

For (18) to be satisfied for all $\boldsymbol{\beta}$, the Kriging coefficients must satisfy the linear constraints

$$\boldsymbol{P}^{\mathsf{T}} \boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{p}(\boldsymbol{x}), \tag{19}$$

called *universality constraints* by Matheron. At this point, Kriging can be reformulated as follows: find the vector of Kriging coefficients that minimizes the variance of the prediction error (17) subject to the constraints (19). This problem can be solved via a Lagrangian formulation, with $\boldsymbol{\mu}(\boldsymbol{x})$ a vector of l Lagrange multipliers for the constraints in (19). The coefficients $\boldsymbol{\lambda}(\boldsymbol{x})$ are then solutions of the linear system of equations

$$\begin{pmatrix} \boldsymbol{K} & \boldsymbol{P} \\ \boldsymbol{P}^{\mathsf{T}} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}(\boldsymbol{x}) \\ \boldsymbol{\mu}(\boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{k}(\boldsymbol{x}) \\ \boldsymbol{p}(\boldsymbol{x}) \end{pmatrix}, \tag{20}$$

with $\boldsymbol{0}$ a matrix of zeros. A convenient expression for the variance of the prediction error is obtained by substituting $\boldsymbol{k}(\boldsymbol{x}) - \boldsymbol{P}\boldsymbol{\mu}(\boldsymbol{x})$ for $\boldsymbol{K}\boldsymbol{\lambda}(\boldsymbol{x})$ in (17) as justified by (20), to get

$$\hat{\sigma}^2(\boldsymbol{x}) = \mathbb{E}\left[F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x})\right]^2 = k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{k}(\boldsymbol{x}) - \boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{\mu}(\boldsymbol{x}). \tag{21}$$

The variance of the prediction error at $\boldsymbol{x}$ can thus be computed without any evaluation of f, using (20) and (21). It provides a measure of the quality associated with the Kriging prediction. Evaluations of f remain needed to estimate the parameters of the covariance function of F (if any), as will be seen in Section 8.3.2.

Once f has been evaluated at all evaluation points, the prediction of the value taken by f at $\boldsymbol{x}$ becomes

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{f}_{\mathbb{S}}, \tag{22}$$

with $\boldsymbol{f}_{\mathbb{S}} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^{\mathsf{T}}$ ($\boldsymbol{f}_{\mathbb{S}}$ is viewed as a sample value of $\boldsymbol{F}_{\mathbb{S}}$).

It is easy to check that (20) implies that

$$\forall\, \boldsymbol{x}_i \in \mathbb{S},\ \ \hat{F}(\boldsymbol{x}_i) = F(\boldsymbol{x}_i).$$

The prediction of f at $\boldsymbol{x}_i \in \mathbb{S}$ is then $f(\boldsymbol{x}_i)$, so Kriging is an interpolation with the considerable advantage that it also accounts for model uncertainty through an explicit characterization of the prediction error.

Remark: The Bayesian framework (see, for instance, [26]) is an alternative approach to derive the BLUP, in which F is viewed as a Bayesian prior on the output. In the case of a zero-mean model, the conditional distribution of the function is then Gaussian with mean

$$\mathbb{E}\left[F(\boldsymbol{x}) | \, \boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}\right] = \boldsymbol{k}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{K}^{-1} \boldsymbol{f}_{\mathbb{S}}, \tag{23}$$

and variance

$$\mathrm{Var}\left[F(\boldsymbol{x}) | \, \boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}\right] = k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{K}^{-1} \boldsymbol{k}(\boldsymbol{x}),$$

which are exactly the mean (22) and variance (21) of the Kriging predictor for a model F with zero mean. The Kriging predictor can also be viewed as the conditional mean of $F(\boldsymbol{x})$ in the case of an unknown mean, if the universality constraints are viewed as a non-informative prior on $\boldsymbol{\beta}$.

8.2 Kriging when f is evaluated approximately

The Kriging predictor was previously defined as the element of the space $\mathbb{H}_{\mathbb{S}}$ generated by the random variables $F(\boldsymbol{x}_i)$ that minimizes the prediction error. A natural step is to extend this formulation to the case of a function whose evaluations are corrupted by additive independent and identically distributed Gaussian noise variables ε_i with zero mean and variance σ_ε^2. The model of the observations then becomes $F^{\text{obs}}_{\boldsymbol{x}_i} = F(\boldsymbol{x}_i) + \varepsilon_i,\ i = 1, \ldots, n$, and the Kriging predictor for $F(\boldsymbol{x})$ takes the form $\hat{F}(\boldsymbol{x}) = \boldsymbol{\lambda}(\boldsymbol{x})^\mathsf{T} \boldsymbol{F}^{\text{obs}}_{\mathbb{S}}$ with $\boldsymbol{F}^{\text{obs}}_{\mathbb{S}} = \left[F^{\text{obs}}_{\boldsymbol{x}1}, \ldots, F^{\text{obs}}_{\boldsymbol{x}_n}\right]^\mathsf{T}$. The unbiasedness constraint (19) remain unchanged, while the mean-square error (2) becomes

$$\mathbb{E}[\hat{F}(\boldsymbol{x}) - F(\boldsymbol{x})]^2 = k(\boldsymbol{x}, \boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{K} + \sigma_\varepsilon^2 \boldsymbol{I}_n)\boldsymbol{\lambda}(\boldsymbol{x}) - 2\boldsymbol{\lambda}(\boldsymbol{x})^\mathsf{T}\boldsymbol{k}(\boldsymbol{x}),$$

with $\boldsymbol{I}_n$ the identity matrix. Finally, using Lagrange multipliers as before, it is easy to show that the coefficients $\boldsymbol{\lambda}(\boldsymbol{x})$ of the prediction must satisfy

$$\begin{pmatrix} \boldsymbol{K} + \sigma_\varepsilon^2 \boldsymbol{I}_n & \boldsymbol{P} \\ \boldsymbol{P}^\mathsf{T} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}(\boldsymbol{x}) \\ \boldsymbol{\mu}(\boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{k}(\boldsymbol{x}) \\ \boldsymbol{p}(\boldsymbol{x}) \end{pmatrix}. \tag{24}$$

The resulting prediction is no longer interpolative, but can still be viewed as the mean of the conditional distribution of F. The variance of the prediction error is again obtained using (21).

8.3 Covariance choice

Choosing a suitable covariance function $k(\cdot, \cdot)$ for a given f is a recurrent and fundamental question. It involves the choice of a parametrized class (or model) of covariance, and the estimation of its parameters.

8.3.1 Covariance classes

The asymptotic theory of Kriging [21] stresses the importance of the behaviour of the covariance near the origin. This behaviour is indeed linked with the quadratic-mean regularity of the random process. For instance, if the covariance function is continuous at the origin, then the process will be continuous in quadratic mean. In practice, one often uses covariances that are *invariant by translation* (or equivalently *stationary*), *isotropic*, and such that regularity can be adjusted. Non-stationary covariances are seldom used in practice, as they make parameter estimation particularly difficult [4]. Isotropy, however, is not required and can even be inappropriate when the factors are of different natures. An example of an anisotropic, stationary covariance class is $k(\boldsymbol{x}, \boldsymbol{y}) = k(h)$, with $h = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\mathsf{T} \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{y})}$ where $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{X}^2$ and $\boldsymbol{A}$ is a symmetric positive definite matrix.

A number of covariance classes are classically used (for instance, exponential $h \mapsto \sigma^2 \exp(-\theta|h|^\alpha)$, product of exponentials, or polynomial). The *Matérn covariance* class offers the possibility to adjust regularity with a single parameter [21]. Stein (1999) advocates the use of the following parametrization of the Matérn class:

$$k(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}h}{\rho}\right)^\nu \mathcal{K}_\nu \left(\frac{2\nu^{1/2}h}{\rho}\right), \tag{25}$$

where $\mathcal{K}_\nu$ is the modified Bessel function of the second kind [27]. This parameterization is 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. To stress the significance and relevance of the regularity parameter, Figure 10 shows the influence of ν on the covariance function, and Figure 11 demonstrates its impact on the sample paths. Since Kriging assumes that f is a sample path of F, a careful choice of the parameters of the covariance is essential.

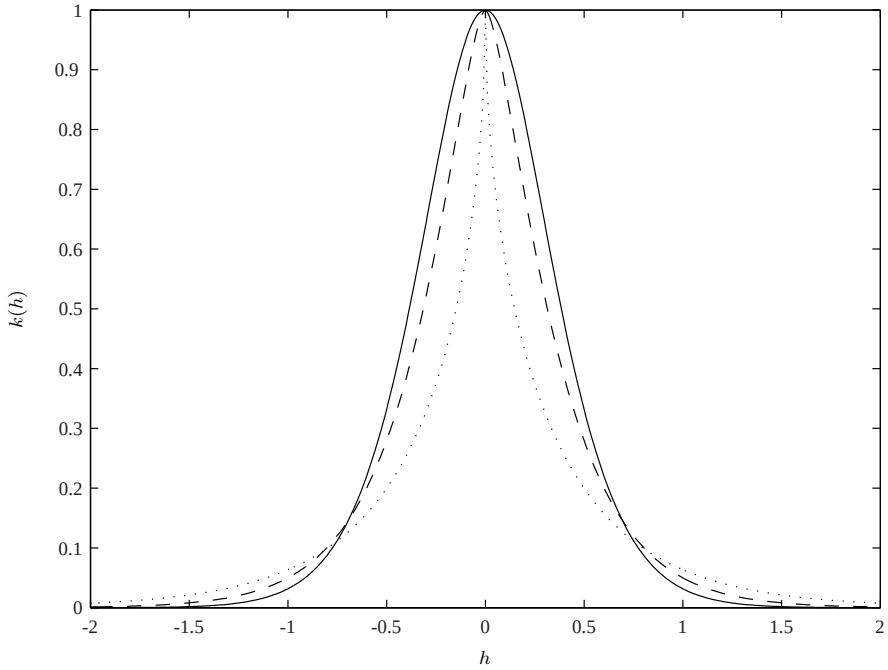


Fig. 10 Matérn covariances with $\rho = 0.5$, $\sigma^2 = 1$. Solid line corresponds to $\nu = 4$, dashed line to $\nu = 1$ and dotted line to $\nu = 0.25$.

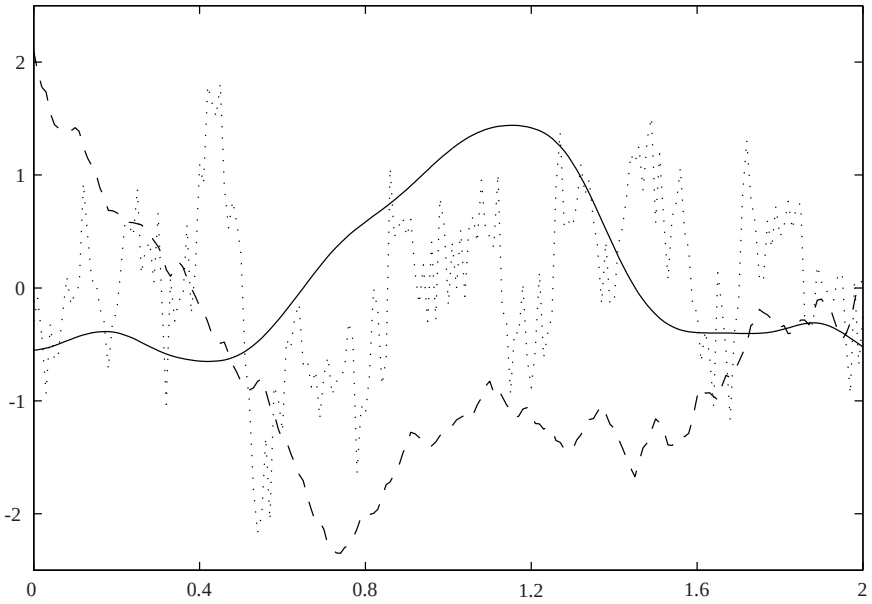


Fig. 11 Three sample paths of a zero-mean Gaussian process with a Matérn covariance. Conventions are as in Figure 10: $\nu = 4$ for the solid line, $\nu = 1$ for the dashed line and $\nu = 0.25$ for the dotted line.

8.3.2 Covariance parameters

The parameters for a given covariance class 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 covariances [4]. In other areas, cross validation [23] and maximum likelihood [21] are mostly employed. For simplicity and generality reasons [21], the maximum-likelihood method is preferred here. Using the joint probability density of the observed Gaussian vector, and assuming that the mean of

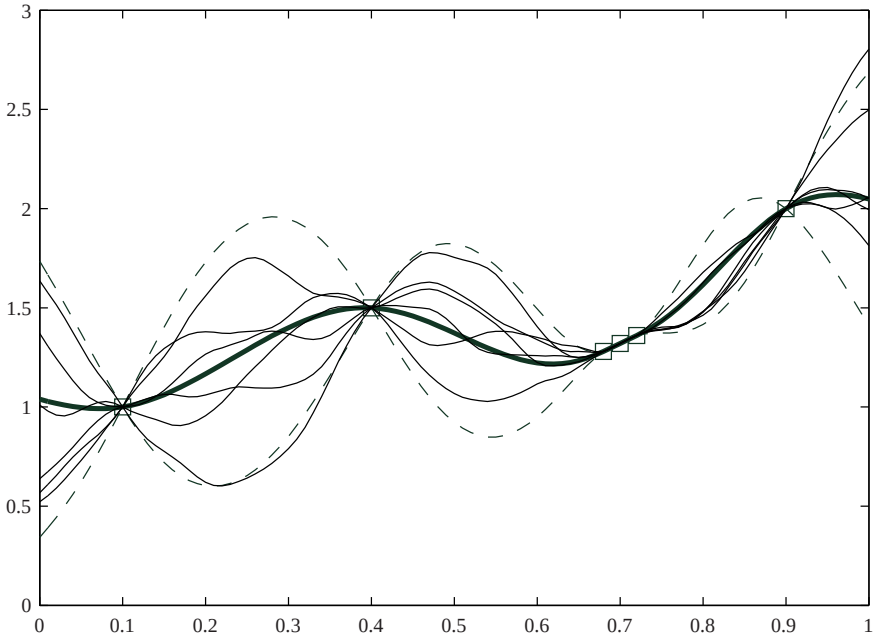


Fig. 12 Example of Kriging interpolation (bold line) for a function of one variable. The data are represented by squares, and the covariance parameters were estimated by REML. Dashed lines delimit 95% confidence region for the prediction. The thin solid lines are examples of conditional simulations.

$F(\boldsymbol{x})$ is zero for the sake of simplicity, one obtains the maximum-likelihood estimate of the vector $\boldsymbol{\theta}$ of the covariance parameters (see, for instance, [22]) by minimizing the negative log-likelihood

$$l(\boldsymbol{\theta}) = \frac{n}{2}\log 2\pi + \frac{1}{2}\log\det \boldsymbol{K}(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{f}_{\mathbb{S}}^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{\theta})^{-1}\boldsymbol{f}_{\mathbb{S}}\,. \tag{26}$$

When the mean for $F(\boldsymbol{x})$ is unknown, the parameters can be estimated, using for example the *REstricted Maximum Likelihood* (REML, see [21]). This is the approach used for the examples in this paper.

Figure 12 illustrates prediction by Kriging with a Matérn covariance, the parameters of which have been estimated by REML. The prediction interpolates the data, and confidence intervals are deduced from the square root of the variance of the prediction error to assess the quality of the prediction between data. Figure 12 also contains a series of conditional simulations (obtained with the method explained in Section 3.2), namely sample paths of F that interpolate the data. As implied by (23), the Kriging prediction is the mean of these conditional simulations.

Acknowledgements

The authors wish to thank Donald R. Jones for his comments that greatly contributed to improving the accuracy and clarity of this paper.

References

1. Abrahamsen, P.: A review of Gaussian random fields and correlation functions. Tech. rep., Norwegian Computing Center (1997). URL www.math.ntnu.no/ omre/TMA4250/V2007/abrahamsen2.ps
2. Adler, R.: On excursion sets, tubes formulas and maxima of random fields. Ann. Appl. Probab. **10**(1), 1–74 (2000)
3. Chib, S., Greenberg, E.: Understanding the Metropolis-Hastings algorithm. Amer. Statist. **49**(4), 327–335 (1995)

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. Cox, D., John, S.: Sdo: a statistical method for global optimization. In: Alexandrov, N., Hussaini, M.Y. (eds.) Multidisciplinary Design Optimization: State of the Art, pp. 315–329. SIAM, Philadelphia (1997). URL citeseer.ifi.unizh.ch/cox97sdo.html
7. Delfiner, P.: Shift invariance under linear models. Ph.D. thesis, Princetown University, New Jersey (1977)
8. Dixon, L., Szegö, G.: The global optimisation problem: an introduction. In: Dixon, L., Szegö, G. (eds.) Towards Global Optimization 2. North-Holland Publishing Company (1978)
9. 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)
10. Gutmann, H.: A radial basis function method for global optimization. J. Global Optim. **19**(3), 201–227 (2001)
11. Huang, D.: Experimental planning and sequential Kriging optimization using variable fidelity data. Ph.D. thesis, Ohio State University (2005)
12. 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)
13. Jones, D.: A taxonomy of global optimization methods based on response surfaces. J. Global Optim. **21**, 345–383 (2001)
14. Jones, D., Schonlau, M., William, J.: Efficient global optimization of expensive black-box functions. J. Global Optim. **13**, 455–492 (1998)
15. Matheron, G.: Principles of geostatistics. Economic Geology **58**, 1246–1266 (1963)
16. Matheron, G.: Le krigeage universel. In: Cahiers du Centre de Morphologie Mathématique de Fontainebleau. Ecole des Mines de Paris (1969). Fasc. 1
17. Sasena, M., Papalambros, P., Goovaerts, P.: Exploration of metamodeling sampling criteria for constrained global optimization. Eng. Opt. **34**, 263–278 (2002)
18. Schonlau, M.: Computer experiments and global optimization. Ph.D. thesis, University of Waterloo (1997)
19. Sjö, E.: Crossings and maxima in Gaussian fields and seas. Ph.D. thesis, Lund Institute of Technology (2000)
20. Smola, A.: Learning with kernels. Ph.D. thesis, Technische Universität Berlin (1998)
21. Stein, M.: Interpolation of Spatial Data: Some Theory for Kriging. Springer, New-York (1999)
22. Vechia, A.: Estimation and model identification for continuous spatial processes. J. Royal Statist. Soc. **B**(50), 297–312 (1998)
23. 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)
24. Watson, A., Barnes, R.: Infill sampling criteria to locate extremes. Math. Geol. **27(5)**, 589–698 (1995)
25. Williams, B., Santner, T., Notz, W.: Sequential design of computer experiments to minimize integrated response functions. Stat. Sinica **10**, 1133–1152 (2000)
26. Williams, C., Rasmussen, C.: Gaussian processes for regression. In: Touretzky, D., Mayer, M., Hasselmo, M. (eds.) Advances in neural information processing systems, vol. 8. MIT Press (1996)
27. Yaglom, A.: Correlation Theory of Stationary and Related Random Functions I: Basic Results. Springer Series in Statistics. Springer-Verlag, New-York (1986)