Identification of expensive-to-simulate parametric models using Kriging and Stepwise Uncertainty Reduction

Julien Villemonteix, Emmanuel Vazquez, Eric Walter

Identification of expensive-to-simulate parametric models using Kriging and Stepwise Uncertainty Reduction

Julien Villemonteix, Emmanuel Vazquez and Eric Walter[1]

***Abstract*— This paper deals with parameter identification for expensive-to-simulate models, and presents a new strategy to address the resulting optimization problem in a context where the budget for simulations is severely limited. Based on Kriging, this approach computes an approximation of the probability distribution of the optimal parameter vector, and selects the next simulation to be conducted so as to optimally reduce the entropy of this distribution. A continuous-time state-space model is used to illustrate the method.**

I. Introduction

The vector $\mathbf{x}$ of the parameters of a parametric model is usually estimated by optimizing some cost function $f(\mathbf{x})$ that quantifies the difference between a vector $\mathbf{y}$ of experimental data and the results $\mathbf{y}_\mathrm{m}(\mathbf{x})$ of model simulation ([14]). Except in some important but very specific cases where the optimal parameter vector $\mathbf{x}^*$ can be computed explicitly, this optimization requires a large number of model simulations. This paper is concerned with the case where the number of model simulations effectively achievable is severely limited by either time or cost.

In this context, it becomes essential to favor optimization methods that use the scarce information as efficiently as possible. Such methods often use an approximation based on available evaluations, as a cheap proxy for the function to be optimized. We shall refer to this proxy as a *surrogate approximation* to avoid confusion with the parametric model $\mathbf{y}_\mathrm{m}(\mathbf{x})$. Surrogate approximations based on Gaussian processes and *Kriging* (initially introduced in geostatistics [7]) have received particular attention [5], mainly for the underlying probabilistic framework, witch allows the set of function evaluations to be chosen efficiently.

In this context, the authors have introduced [12] the *Informational Approach to Global Optimization* (IAGO, [12]), which provides an *explicit* estimated probability distribution for the minimizers of f, allowing an information-based search strategy. In comparison, most alternative strategies *implicitly* seek a likely value for $\mathbf{x}^*$ and then assume it to be a suitable location for evaluating f ([4], [5], [6]).

This paper aims at drawing the attention of the control community on the pertinence and performances of the IAGO to be presented in Section III. Section II will recall the principles of Kriging, on which IAGO is based, and Section IV will illustrate the potential evaluations savings of the methodology on a simple but not uniquely identifiable continuous-time state-space model.

[1]Julien Villemonteix is with the Department of Energetic Systems, Renault SA. 78298 Guyancourt, France.
Emmanuel Vazquez is with the Department of signal and electronic systems, Supélec. 91192 Gif-sur-Yvette, France.
Eric walter is with the "Laboratoire des Signaux et Systèmes", CNRS, Supélec, Univ. Paris-Sud. 91192 Gif-sur-Yvette, France. Emails: {julien.villemonteix,emmanuel.vazquez}@supelec.fr, eric.walter@lss.supelec.fr

II. Estimating probability density for $\mathbf{x}^*$

A. Kriging and linear prediction

Kriging ([1], [10]) is a prediction method based on random processes, which can be used to approximate or interpolate data. It can also be understood as a kernel regression method, such as *splines* [13] or *Support Vector Regression* (SVR, [8]). It originates from geostatistics and has been 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.

When modeling with Gaussian processes, the function of interest $f : \mathbb{X} \to \mathbb{R}$ is assumed to be a sample path of a second-order Gaussian random process F with covariance $k(.,.)$. The mean of $F(\mathbf{x})$ is assumed to be a finite linear combination of known functions p_i of $\mathbf{x}$, $m(\mathbf{x}) = \boldsymbol{\beta}^\mathsf{T}\mathbf{p}(\mathbf{x})$, where $\boldsymbol{\beta}$ is a vector of fixed but unknown coefficients, and $\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), ..., p_l(\mathbf{x}))^\mathsf{T}$. Usually the functions p_i are monomials of low degree in the components of $\mathbf{x}$ (in practice, their degree does not exceed two).

Kriging consists in computing an unbiased linear prediction of $F(\mathbf{x})$ in the vector space $\mathbb{H}_\mathbb{S} = \mathrm{span}\{F(\mathbf{x}_1), ..., F(\mathbf{x}_n)\}$, which can be written as

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

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

To compute an unbiased prediction with minimal variance, a Lagrangian formulation is adopted, with $\boldsymbol{\mu}(\mathbf{x})$ a vector of l Lagrange multipliers. The coefficients $\boldsymbol{\lambda}(\mathbf{x})$ are then solutions of the linear system of equations

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

with $\mathbf{0}$ a matrix of zeros, $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))$, $(i,j) \in \{1, ..., n\}^2$ the $n \times n$ covariance matrix of F at all evaluation points in $\mathbb{S}$, $\mathbf{k}(\mathbf{x}) = [k(\mathbf{x}_1, \mathbf{x}), ..., k(\mathbf{x}_n, \mathbf{x})]^\mathsf{T}$, the vector of covariances between $F(\mathbf{x})$ and $\mathbf{F}_\mathbb{S}$, 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

$$\begin{aligned}\hat{\sigma}^2(\mathbf{x}) &= \mathrm{var}(\hat{F}(\mathbf{x}) - F(\mathbf{x})) \\ &= k(\mathbf{x},\mathbf{x}) - \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T}\mathbf{k}(\mathbf{x}) - \mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\mu}(\mathbf{x})\,,\end{aligned} \tag{3}$$

as these quantities only depend on the covariance of F. Once f has been evaluated at all $\mathbf{x}_i$ in $\mathbb{S}$, the prediction of $f(\mathbf{x})$ becomes $\hat{f}(\mathbf{x}) = \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T}\mathbf{f}_\mathbb{S}$, with $\mathbf{f}_\mathbb{S} = [f(\mathbf{x}_1),...,f(\mathbf{x}_n)]^\mathsf{T}$.

Note that, in the case of exact evaluations of f, Kriging is an interpolation ($\forall \mathbf{x}_i \in \mathbb{S}\, \hat{F}(\mathbf{x}_i) = F(\mathbf{x}_i)$).

B. Density of the global minimizers

According to the GP model, a global minimizer $\mathbf{x}^*$ of f corresponds to a global minimizer of a sample path of F. Hence the intuitive idea to consider a random quantity accounting for the knowledge on the global minimizers of F conditionally to past evaluations.

More formally, consider the random set $\mathcal{M}^*_\mathbb{G}$ of the global minimizers of F over $\mathbb{G}$ (a finite subset of $\mathbb{X}$), i.e,

$$\mathcal{M}^*_\mathbb{G} = \left\{\mathbf{x}^* \in \mathbb{G} : F(\mathbf{x}^*) = \min_{\mathbf{x}\in\mathbb{G}} F(\mathbf{x})\right\}.$$

Let then $\mathbf{X}^*_\mathbb{G}$ be a random vector uniformly distributed on $\mathcal{M}^*_\mathbb{G}$.

The probability density function $p_{\mathbf{X}^*_\mathbb{G}|\mathbf{f}_\mathbb{S}}$ of $\mathbf{X}^*$ conditionally to $\mathbf{f}_\mathbb{S}$, designated as the conditional density of the global minimizers in [12] (or in short minimizers density), can be viewed as the current solution of the global optimization problem as it contains all of what has been learnt about the function and its minimizers. In what follows, we propose a simulation-based approximation for the density of the minimizers.

C. Conditionning by Kriging

Initially, f is only assumed to be a sample path of F. As evaluations become available, f is assumed to be a sample path of F that interpolates the data, namely a *conditional sample path*, which can be viewed as a possible version of f (the Kriging prediction is in fact the mean of these sample paths). The simulation of these sample paths (known as *conditional simulation*) is of remarkable interest when one wishes to estimate quantities non-linear in the studied function, such as the minimizer [1]. Examples of such simulations are presented on Figure 1, along with the corresponding Kriging prediction. In this paper, we propose to use such simulations to compute an approximation $\hat{p}_{\mathbf{X}^*_\mathbb{G}|\mathbf{f}_\mathbb{S}}$ of the minimizers density.

Among the many available methods for generating conditional simulations [1], we use, mainly for simplicity and computational reasons, the unbiasedness of the Kriging prediction to transform non-conditional simulations into simulations interpolating the evaluations $\mathbf{f}_\mathbb{S}$.

Let Z be a zero-mean Gaussian process with covariance k (the same as that of F), $\hat{Z}$ be its Kriging predictor based on the random variables $Z(\mathbf{x}_i)$, $\mathbf{x}_i \in \mathbb{S}$, and consider the random process

$$T(\mathbf{x}) = \hat{f}(\mathbf{x}) + \left[Z(\mathbf{x}) - \hat{Z}(\mathbf{x})\right], \tag{4}$$

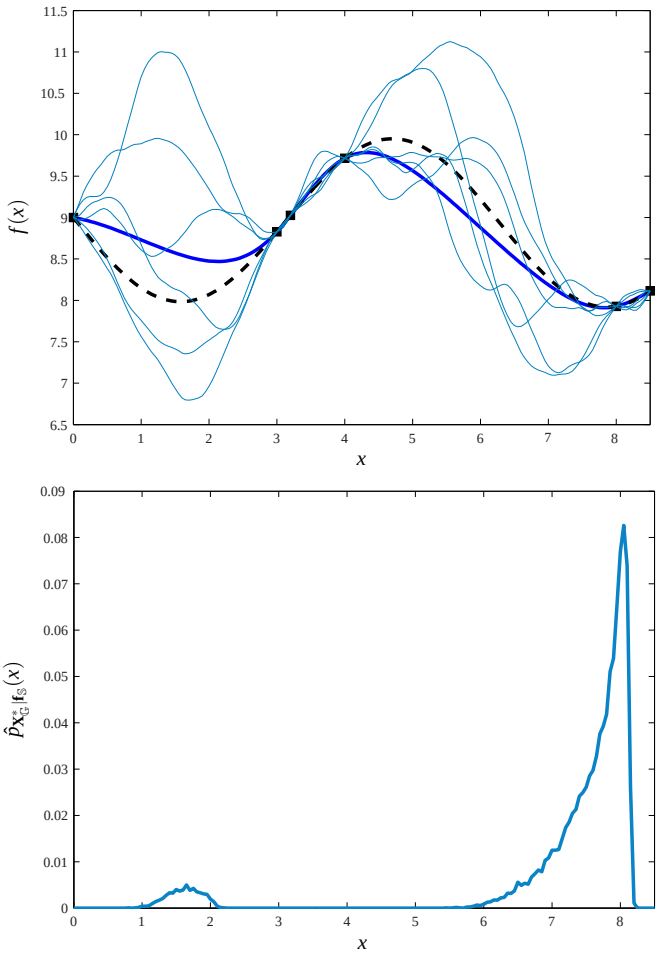

Fig. 1. Top: Kriging prediction (bold line) based on scarce evaluations (squares), along with conditional simulations (thin lines). Bottom: Conditional density of the minimizers approximated using conditional simulations.

where $\hat{f}$ is the mean of the Kriging predictor for F based on the design points in $\mathbb{S}$. It can then be easily verified that, as a result of the unbiasedness of $\hat{Z}$, the sample paths of T are also conditional simulations of F.

Using equation (1), one can rewrite (4) as

$$T(\mathbf{x}) = Z(\mathbf{x}) + \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T}\left[\mathbf{f}_\mathbb{S} - \mathbf{Z}_\mathbb{S}\right], \tag{5}$$

with $\mathbf{Z}_\mathbb{S} = [Z(\mathbf{x}_1),...,Z(\mathbf{x}_n)]^\mathsf{T}$. So the same vector $\boldsymbol{\lambda}(\mathbf{x})$ of Kriging coefficients is used for the interpolation of the data and for the simulations of Z.

In summary, to simulate F over $\mathbb{G}$ conditionally to past evaluations $\mathbf{f}_\mathbb{S}$, one can simply simulate a zero-mean Gaussian process Z over $\mathbb{G}$, compute, for every point in $\mathbb{G}$, the vector of Kriging coefficients based on the design points in $\mathbb{S}$, and apply (5). Obtaining an approximation for $p_{\mathbf{X}^*_\mathbb{G}|\mathbf{f}_\mathbb{S}}$ is then simply a matter of computing the global minimizers for a sufficient number of conditional simulations. An example of the resulting distribution is presented on Figure 1 along with the corresponding Kriging prediction (top).

III. Kriging-based global optimization

We have seen that the Kriging framework is well suited for an estimation of the minimizers density. Before describing the new IAGO search strategy, let us recall the optimization approaches that are standard when dealing with expensive-to-evaluate functions using a Kriging surrogate approximation.

A. Standard approaches

Most Kriging-based optimization algorithms are built on the same principle, and sequentially evaluate f at a point that optimizes a criterion based on the surrogate approximation obtained using the previous evaluations. In a sense, the expensive-to-evaluate cost function f is replaced by a cheaper cost function based on the surrogate approximation, which we refer to as the criterion to avoid confusion with f. A simple example of such a criterion is the prediction $\hat{f}$. However, too much confidence is then put in the current prediction, and search may stall on a local minimizer if the initial prediction is too distant from a the global minimizer.

To improve this basic criterion, a compromise between local and global search has to be struck. This compromise is generally achieved by putting more emphasis on the prediction error that indicates locations where additional evaluations are needed to improve confidence in the model. This approach has led to a number of criteria [5], and chiefly the *expected improvement* criterion (EI, cf. [6]) that we shall briefly present here and use it as reference in Section IV.

In [6], the improvement expected from an additional evaluation of f at $\mathbf{x}$ given the past evaluations in $\mathbf{f}_\mathbb{S}$ is expressed as

$$\mathrm{EI}(\mathbf{x}) = \mathbb{E}\left[I(\mathbf{x})|\mathbf{F}_\mathbb{S} = \boldsymbol{f}_\mathbb{S}\right], \tag{6}$$

with

$$\mathrm{I}(\mathbf{x}) = \begin{cases} 0 & \text{if} \quad F(\mathbf{x}) \geq \hat{f}_{min} \\ \hat{f}_{min} - F(\mathbf{x}) & \text{otherwise} \end{cases},$$

and $\hat{f}_{min}$ the best value of f yet obtained. Using integration by part, one can easily rewrite (6) as

$$\mathrm{EI}(\mathbf{x}) = \hat{\sigma}(\mathbf{x})\left[u\Phi(u) + \Phi'(u)\right], \tag{7}$$

with

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

and Φ the normal cumulative distribution. The new evaluation point is then chosen as a global maximizer of $\mathrm{EI}(\mathbf{x})$.

Besides EI, all commonly used criteria aim at answering the same question: What is the most likely position of $\mathbf{x}^*$? They *implicitly* seek a likely value for the optimum location, and then assume it to be a suitable location for an additional evaluation of f. By contrast, our main contribution will be the *explicit* characterization (through $\hat{p}_{\mathbf{X}^*_\mathbb{G}|\mathbf{f}_\mathbb{S}}$) of the uncertainty on the minimizers stemming from the lack of information on the function. We shall also see that a more pertinent problem can in fact be solved: Where should the evaluation be carried out optimally to improve knowledge on the global minimizers?

B. Stepwise uncertainty reduction

In [12], conditional entropy has been introduced to measure the information gain to be provided on the minimizers by an additional evaluation. In active learning, this is part of the *Stepwise Uncertainty Reduction* (SUR) strategy [3], which chooses the point that potentially brings the largest reduction in entropy (seen as a measure of uncertainty). To apply the SUR principle to global optimization, the IAGO strategy evaluates this gain at $\mathbf{x}$ by using Kriging to generate the necessary conditional simulations for the approximation of the distribution of the minimizers conditionally to past evaluations and to a possible evaluation at $\mathbf{x}$. This approach, is relatively expensive but, as detailed in [12], the same set of sample paths can be used throughout the procedure which makes the algorithm applicable (see [11] for an example in the automotive industry). Let us present the IAGO algorithm in more detail.

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

$$H(U) = -\sum_u \mathrm{P}(U=u)\log_2 \mathrm{P}(U=u).$$

$H(U)$ quantifies the spread of the distribution of U, and decreases as this distribution gets more peaked.

Similarly, the conditional entropy [2] of U given a discrete random variable V and an event $\mathcal{B}$ is

$$H(U|\mathcal{B},V) = \sum_v \mathrm{P}(V=v|\mathcal{B})H(U|\mathcal{B},V=v), \tag{8}$$

with

$$\begin{aligned} H(U|\mathcal{B},V=v) = &-\sum_u \mathrm{P}(U=u|\mathcal{B},V=v) \\ &\log_2 \mathrm{P}(U=u|\mathcal{B},V=v), \end{aligned} \tag{9}$$

the conditional entropy of U given $\mathcal{B}$ and $\{V=v\}$.

For our optimisation problem to be fully solved, there should not remain any uncertainty on $\mathbf{x}^*$. Therefore, we would like to ensure that $H(\mathbf{X}^*_\mathbb{G}|\mathbf{F}_\mathbb{S}=\mathbf{f}_\mathbb{S})=0$. The idea of the IAGO strategy is then iteratively to ensure a one-step optimal reduction of $H(\mathbf{X}^*_\mathbb{G}|\mathbf{F}_\mathbb{S}=\mathbf{f}_\mathbb{S})$ given what is known of the system. In other words, $\mathbf{x}'$ is chosen as a new evaluation point if it minimizes $H_\mathbb{S}(\mathbf{x})$ the conditional entropy of $\mathbf{X}^*_\mathbb{G}$ given all past evaluations and $F_\mathrm{Q}(\mathbf{x})$, a discrete version of $F(\mathbf{x})$, obtained by quantization at levels $f_1,\ldots,f_M$ (the quantization is necessary for the computation of conditional entropy). By using (8) we can then write

$$\begin{aligned} H_\mathbb{S}(\mathbf{x}) = &\sum_{i=1}^{M} \mathrm{P}(F_\mathrm{Q}(\mathbf{x})=f_i|\mathbf{F}_\mathbb{S}=\mathbf{f}_\mathbb{S}) \\ &H(\mathbf{X}^*_\mathbb{G}|\mathbf{F}_\mathbb{S}=\mathbf{f}_\mathbb{S},F_\mathrm{Q}](\mathbf{x})=f_i), \end{aligned} \tag{10}$$

with

$$\begin{aligned} H(\mathbf{X}^*_\mathbb{G}|\mathbf{F}_\mathbb{S}=\mathbf{f}_\mathbb{S},F_\mathrm{Q}(\mathbf{x})=f_i) = \\ -\sum_{\mathbf{u}\in\mathbb{G}} p_{\mathbf{X}^*_\mathbb{G}|\mathbf{f}_\mathbb{S},f_i}(\mathbf{u})\log_2 p_{\mathbf{X}^*_\mathbb{G}|\mathbf{f}_\mathbb{S},f_i}(\mathbf{u}), \end{aligned} \tag{11}$$

and

$$p_{\mathbf{X}^*_\mathbb{G}|\mathbf{f}_\mathbb{S},f_i}(\mathbf{u}) = \mathrm{P}(\mathbf{X}^*=\mathbf{u}|\mathbf{F}_\mathbb{S}=\mathbf{f}_\mathbb{S},F_\mathrm{Q}(\mathbf{x})=f_i),$$

computed using conditional simulations.

The criterion $H_{\mathbb{S}}$ thus takes into account the conditional statistical properties of F and particularly the covariance of the model to choose a one-step optimal evaluation point. By contrast, the EI criterion depends only on the conditional mean and variance of F at the design point considered (and this is actually true for most standard strategies).

C. Computational issues

Our algorithm is similar in spirit to a particular strategy for Kriging-based optimization known as *Efficient Global Optimization* (EGO [6]). EGO starts with a small initial set of evaluations of f, estimates the parameters of the covariance (see [12] and the reference therein for details on this subject) 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 covariance 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 information obtained to the prior on the function, but the SUR strategy is used instead of the maximization of EI. Another specific feature of our algorithm is that we advocate the use of the Matèrn covariance [9] and of maximum likelihood estimation for the parameters.

Stopping criterion: 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 the current minimum of the Kriging interpolation than a given tolerance threshold. This stopping criterion is well suited here, since the estimation of the repartition function of $F(\mathbf{X}^*)$ can be carried out using conditional simulations in exactly the same fashion as for the estimation of $\hat{p}_{\mathbf{X}^*_{\mathbb{G}}|\mathbf{f}_{\mathbb{S}}}$.

Computational burden: The IAGO algorithm involves the minimization of the conditional entropy $H_{\mathbb{S}}(\mathbf{x})$ over a set of candidate evaluations points. We propose to solve this optimization problem using points in $\mathbb{G}$ as candidate points, and sampling $\mathbb{G}$ with $\hat{p}_{\mathbf{X}^*_{\mathbb{G}}|\mathbf{f}_{\mathbb{S}}}$ as prior. By doing so, areas of the design space where the density is sufficiently small are ignored as they are not likely to be of interest for the reduction of entropy.

As detailed earlier, the computation of $H_{\mathbb{S}}(\mathbf{x})$ requires the use of conditional simulations of F over $\mathbb{G}$. This can be done in $\mathrm{O}(N)$ operations (cf. [12]), with N the cardinal of $\mathbb{G}$. Choosing a new evaluation point for f therefore requires $\mathrm{O}(N^2)$ opérations.

Given this complexity, trying to cover parameter space while keeping the same accuracy as dimension increases leads to an exponential increase in computational burden. In a context of expensive function evaluation, however, the objective is less to specify exactly all global minimizers (which could be too demanding in function evaluations anyway), than to use available information to efficiently reduce the likely areas for the location of these minimizers. N can therefore be kept relatively small (in [11] we used 1000 points for a 6 dimension parameter space). Besides, as $\mathbb{G}$ is re-sampled after every evaluation of f, the number of candidate points effectively explored is considerably larger than N. Lastly, the result obtained can be trusted to be a consistent choice within this set of candidate points, in regard of what has been learned (observations) and assumed (covariance of the GP model) about f. Anyhow, the computation of $H_{\mathbb{S}}(\mathbf{x})$ only involves the surrogate approximation. Computational burden is therefore a minor issue as long as it stays small in comparison with the computational burden of an evaluation of f.

IV. Example

A typical example of identification for which the IAGO method is particularly relevant is the estimation of the few physical parameters of a knowledge-based model described by partial differential equations with complex boundary conditions. We chose, however, to consider a much simpler illustrative problem, for three reasons. First, it is possible to briefly give enough details to allow the reader to use it to compare the performance of the IAGO approach with those other methods not considered here. Second, nothing is lost by considering such an example, as the methodology would be strictly the same for a more expensive to simulate model. Last, it will turn out that this example is not so easy to solve and demonstrates the superiority of our approach over more conventional ones.

We thus consider a deceptively simple two compartmental model. Its state vector $\mathbf{q} = [q_1, q_2]^{\mathsf{T}}$ corresponds to the amounts of material in two compartments, which are governed by the evolution equations

$$\left\{ \begin{array}{rcl} \dot{q}_1 & = & -(x_1+x_3)q_1 + x_2 q_2, \\ \dot{q}_2 & = & x_1 q_1 - x_2 q_2. \end{array} \right. \tag{12}$$

At time $t = 0$, a unit injection of material takes place in compartment 1, so $\mathbf{q}(0) = (1,0)^{\mathsf{T}}$. Measurement $y(t_i)$ of the quantity of material in Compartment 2 are collected at time t_i, $i = 1, ..., 15$.

For this simulated example, a noise-free vector of measurements $\mathbf{y}$ is generated using the ODE solver of Matlab with a parameter vector $\mathbf{x}_0 = (0.6, 0.15, 0.35)^{\mathsf{T}}$. The optimization is then carried out over $[0,1]^3$ using the quadratic cost function

$$f(\mathbf{x}) = \sum_{i=1}^{15} (q_2(\mathbf{x}, t_i) - y(t_i))^2.$$

This example is actually difficult for two reasons. First, as suggest by the level sets of Figure 2 (thin lines), the zones where f is small are relatively large in proportion to the size of the search space. Second, the model parameters are not uniquely identifiable, as the values of p_2 and p_3 can be exchanged without modifying the system output [14]. So there are two global minimizers of f, namely $\mathbf{x}_0$ but also $\mathbf{x}_1 = (0.6, 0.35, 0.15)^{\mathsf{T}}$.

With the IAGO algorithm, after 40 evaluations of f, the zones where the approximate density of the minimizers is non-zero are consistent with the 0.3-level set of f (cf.

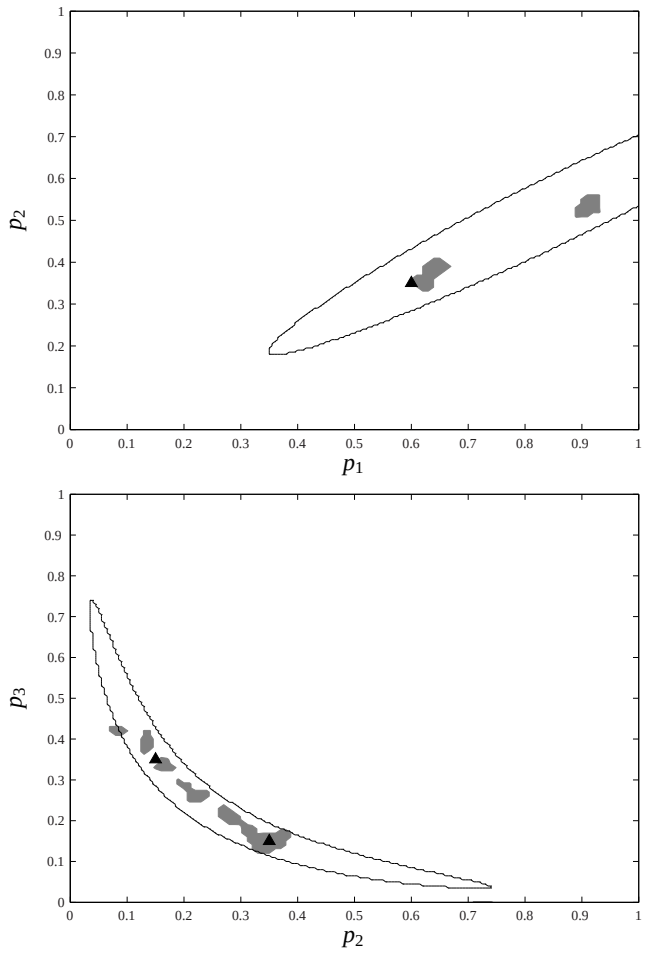

Fig. 2. Cross sections of the density of the minimizers in the ($p_3 = 0.15$) plane (*Top*) and ($p_1 = 0.6$) plane (*bottom*) estimated after 40 evaluations of f using the IAGO algorithms. The points where the estimated density is non-zero are contained by the dark areas. The black curve is the 0.03-level set of the cost function. The true global minimizers $\mathbf{x}_0$ and $\mathbf{x}_1$ are indicated by triangles.

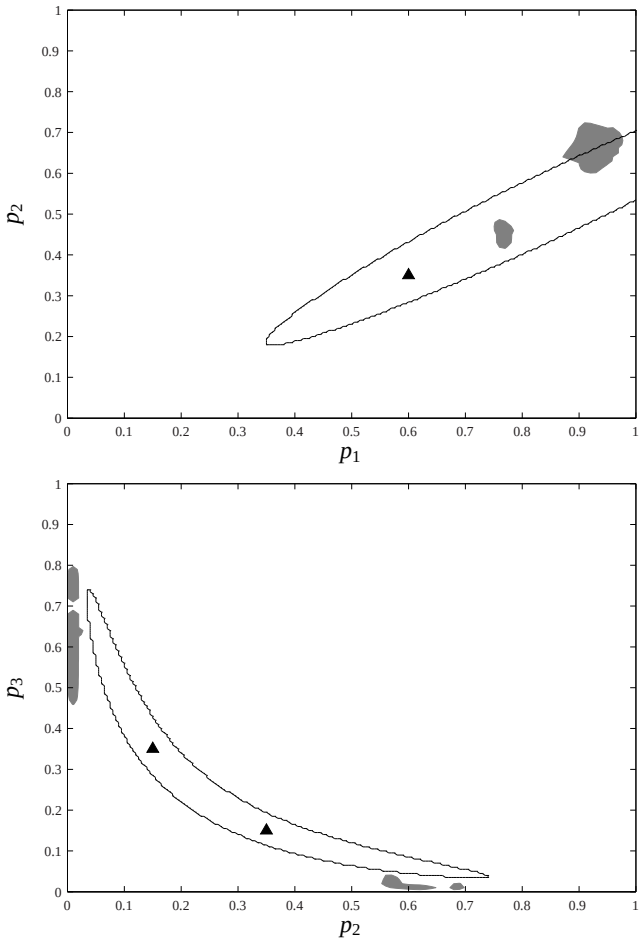

Fig. 3. Cross sections of the density of the minimizers in the ($p_3 = 0.15$) plan (*Top*) and ($p_1 = 0.6$) plan (*bottom*) estimated after 40 evaluations of f using the EGO algorithm. The graphic conventions are the same as those of Figure 2.

Figure 2). Both $\mathbf{x}_0$ and $\mathbf{x}_1$ are within high probability zones for the global minimizer. By comparison, with the EGO algorithm, after 40 evaluations (cf. Figure 3), the approximate density of the minimizers misses both $\mathbf{x}_0$ and $\mathbf{x}_1$. In terms of convergence rates, IAGO performs well on this example (see Table IV), as both minimizers are found with 0.01 precision after 80 evaluations of f. In comparison, the EGO algorithm has only identified $\mathbf{x}_1$, and it takes an average of 160 evaluations to the Nelder-Meald simplex to reach this precision for one of the global minimizers, while entirely missing the other.

V. Conclusions

In this paper, we have presented the IAGO algorithm as an efficient way of handling parameter identification when confronted with, possibly non-uniquely identifiable, expensive-to-evaluate parametric models. The approach, as others before it, uses Kriging to provide a surrogate approximation of the cost function. However, to the best of our

TABLE I

Results after 40 and 80 function evaluations.[a]

Algorithm	Nelder-Meald	EGO	IAGO
Estimation error for the minimizers after 40 evaluations	0.44	0.269 0.090	**0.063** **0.025**
Estimation error for the minimum after 40 evaluations	0.135	10^{-2} 10^{-3}	$\mathbf{10^{-3}}$ $\mathbf{10^{-3}}$
Estimation error for the minimizers after 80 evaluations	0.35	0.399 0.011	**0.011** **0.011**
Estimation error for the minimum after 80 evaluations	5.10^{-2}	10^{-2} 10^{-4}	$\mathbf{10^{-7}}$ $\mathbf{10^{-5}}$

[a]For EGO and IAGO, two results are given, corresponding to the two global minimizers. For the Nelder Mead simplex, a single result is presented, as it is a local method. The local search is repeated for 100 different starting points, and the average precision is presented (each time, the most favorable minimizer is chosen). The estimation error, is either the Euclidean distance between the estimated minimizer and a true one, or the estimation of the minimum (as the true minimum is zero).

knowledge, no other method has used Kriging to compute the density of the minimizers explicitly, which allows, at each iteration of the search, to perform an evaluation at the point that is most likely to reduce the uncertainty on the position of the minimum. As evidenced by the example, the evaluations savings offered by the IAGO algorithm can be significant in comparison with the widespread Nelder-Meald simplex algorithm, but also in comparison with the EGO algorithm, a standard procedure in Kriging-based optimization. The method is particularly well suited to the identification of the parameters of knowledge-based models, which are often very expensive to simulate.

REFERENCES

[1] J.P. Chilès and P. Delfiner. *Geostatistics, Modeling Spatial Uncertainty*. John Willey & Sons, Inc, New York, 1999.

[2] T. M. Cover and A. T. Joy. *Elements of Information Theory*. John Willey & Sons, Inc, New York, 1991.

[3] D. Geman and B. Jedynak. An active testing model for tracking roads in satellite images. Technical Report 2757, Institut National de Recherche en Informatique et en Automatique (INRIA), December 1995.

[4] D. Huang, T. Allen, W. Notz, and N. Zeng. Global optimization of stochastic black-box systems via sequential Kriging meta-models. *Journal of Global Optimization*, 34:441–466, 2006.

[5] D.R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21:345–383, 2001.

[6] D.R. Jones, M. Schonlau, and J. William. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

[7] G. Matheron. Principles of geostatistics. *Economic Geology*, 58:1246–1266, 1963.

[8] A.J. Smola. *Learning with Kernels*. PhD thesis, Technische Universität Berlin, 1998.

[9] M.L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New-York, 1999.

[10] Emmanuel Vazquez. *Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et application*. PhD thesis, Université Paris Sud, UFR Scientifique d'Orsay, 2005.

[11] J. Villemonteix, E. Vazquez, M. Sidorkiewicz, and E. Walter. Gradient-based IAGO strategy for the global optimization of expensive-to-evaluate functions and application to intake-port design. Accepted for the conference on Advances in Global Optimization: Theory and Applications, to be help in June 2007.

[12] J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Submited to the Journal of Global Optimization*, 2006.

[13] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, volume 6, pages 69–87, Boston, 1998. MIT Press.

[14] E. Walter and L. Pronzato. *Identification of Parametric Models from Experimental Data*. Springer-Verlag, London, 1997.