Bayesian optimization for parameter identification on a small simulation budget

Julien Villemonteix [*] **, Emmanuel Vazquez** [*] **and Eric Walter** [**]

[*] *Department of Signal and Electronic Systems, SUPELEC.*
91192 Gif-sur-Yvette, France; e-mail: emmanuel.vazquez@supelec.fr.
[**] *Laboratoire des Signaux et Systèmes, CNRS, SUPELEC, Univ Paris-Sud.*
91192 Gif-sur-Yvette, France.

Abstract — Bayesian optimization uses a probabilistic model of the objective function to guide the search for the optimum. It is particularly interesting for the optimization of expensive-to-evaluate functions. For the last decade, it has been increasingly used for industrial optimization problems and especially for numerical design involving complex computer simulations. We feel that Bayesian optimization should be considered with attention by anyone who has to identify the parameters of a model based on a very limited number of model simulations because of model complexity. In this paper, we wish to describe, as simply as possible, how Bayesian optimization can be used in parameter identification and to present a new application. We concentrate on two algorithms, namely EGO (for Efficient Global Optimization) and IAGO (for Informational Approach to Global Optimization), and describe how they can be used for parameter identification when the budget for evaluating the cost function is severely limited. Some open questions that must be addressed for theoretical and practical reasons are indicated.

Keywords: Gaussian processes, global optimization, Kriging

1. INTRODUCTION

The identification of a parametric model (or parameter estimation) is usually carried out by minimizing some cost function $f(\mathbf{x})$ that quantifies the difference between experimental data and results of model simulation, where $\mathbf{x} \in \mathbb{X}$ is the parameter vector of the model. When this cannot be done analytically, minimizing f by conventional iterative methods, possibly combined with multistart to try to escape local minima, generally requires many model simulations. We assume in this paper that the number of model simulations (or function evaluations) effectively achievable is severely limited by either time or cost.

It then becomes essential to look for optimization methods that use the information available as efficiently as possible. Such methods often use an approximation based on the results of past 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.

During the last decade, surrogate approximations relying on a probabilistic model of the function to be optimized have received particular attention [Huang, 2005, Jones, 2001, Villemonteix et al., 2008b]. The field of optimization techniques that rely on such models is known as *Bayesian Optimization* [Zilinskas, 1992]. The interest of a Bayesian approach to the optimization of expensive-to-evaluate functions has already been discussed extensively (especially in Jones [2001]) and confirmed by many applications to industrial design problems (see, e.g., Huang and Allen [2005], Villemonteix et al. [2008a]). The context of restricted evaluation budget indeed makes most algorithms inefficient. Since the mere simulation of complex knowledge-based parametric models often requires an heavy computational effort and the usual iterative algorithms for nonlinear parameter estimation rely on a large number of such simulations, the potential of Bayesian optimization in parameter estimation must be investigated.

Our first results in this context were presented in Villemonteix et al. [2007], but practical details were reduced to a minimum, and no real-life application was proposed. In this paper, we wish to describe, as simply as possible, how Bayesian optimization can be used in parameter identification and to present an application. We shall concentrate on two algorithms, both based on Gaussian processes and Kriging, namely the reference algorithm EGO (for Efficient Global Optimization [Jones et al., 1998]) and a very recent alternative IAGO (for Informational Approach to Global Optimization [Villemonteix et al., 2008a,b]).

In Section 2, we recall the principles behind EGO and IAGO. We give some details about these algorithms in Section 3, present an application in Section 4, and finally discuss some open questions in Section 5.

2. GAUSSIAN PROCESSES FOR OPTIMIZATION

Bayesian optimization is based on two main ideas. The first one is to model the cost function $f(\cdot)$ by a random process $F(\cdot)$, here assumed to be Gaussian with mean function $m(\cdot)$ and covariance function $k(\cdot,\cdot)$ (in what follows, we shall simply call them mean and covariance). This means in particular that for $\mathbb{S}_n = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subset \mathbb{X}$ a set of evaluation points, the vector of evaluation results $\mathbf{f}_n = [f(\mathbf{x}_1), ..., f(\mathbf{x}_n)]^\mathsf{T}$ is viewed as a realization of the random vector $\mathbf{F}_n = [F(\mathbf{x}_1), ..., F(\mathbf{x}_n)]^\mathsf{T}$. The second idea is to choose the evaluation points iteratively based on what has been learned through the previous evaluation. If n evaluations have been performed at the points in $\mathbb{S}_n$, the $(n+1)$-th point is chosen by optimizing a sampling criterion $J(\mathbf{x}, \mathbb{S}_n, \mathbf{f}_n, F)$ that measures the interest of an additional eval-

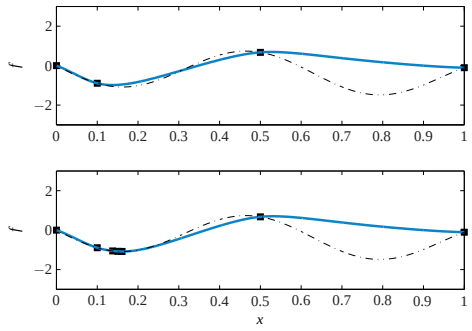

Figure 1. Naive approach to optimization based on a surrogate approximation. *(Top)* Surrogate approximation $\hat{f}$ (bold line) of f (dash-dotted line) based on available evaluation results (squares). *(Bottom)* Surrogate approximation obtained after four iterations of the algorithm that evaluates f at a minimizer of $\hat{f}$. A local minimizer has been identified, but the global one is ignored.

uation at $\mathbf{x}$, given the results $\mathbf{f}_n$ of past evaluations and the Gaussian model F. To choose the $(n+1)$-th evaluation point, one has to solve

$$\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}\in\mathbb{X}} J(\mathbf{x}, \mathbb{S}_n, \mathbf{f}_n, F). \tag{1}$$

In summary, Bayesian optimization replaces the optimization of an expensive-to-evaluate function by a series of optimizations of a cheap criterion. This criterion quantifies the interest of any additional evaluation, and reflects our concern for a careful spending of the evaluation budget.

In this section, we shall discuss how to build a surrogate approximation using Kriging, before giving some details about two possible sampling criteria build from this surrogate approximation. But first, we shall illustrate what we are trying to achieve. Assume that some surrogate approximation $\hat{f}(\mathbf{x})$ of $f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{X}$ has been built. A simple sampling criterion would then be the minimization of $\hat{f}(\mathbf{x})$. The optimization algorithm that ensues might converge towards a local optimum (cf. Fig. 1) and is completely dependent on the initial sampling scheme. To make search more global, one should take into account the uncertainty associated with the surrogate approximation, and this is exactly what Kriging allows us to do.

2.1 An introduction to Kriging

Assume that the mean m of F is a finite linear combination of known functions, which can be written as $m(\cdot) = \boldsymbol{\beta}^\mathsf{T}\mathbf{p}(\cdot)$, with $\mathbf{p}$ a vector of known functions, and $\boldsymbol{\beta}$ a vector of fixed but *unknown* coefficients. The theory of Kriging then addresses the construction, for all $\mathbf{x} \in \mathbb{X}$, of the *best unbiased linear predictor* (BLUP) $\hat{F}(\mathbf{x})$ of F based on the random vector $\mathbf{F}_n$. As a linear predictor, $\hat{F}$ can be written as

$$\hat{F}(\mathbf{x}) = \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T}\mathbf{F}_n. \tag{2}$$

Using a Lagrangian formulation, one can show that, for the predictor to be a BLUP, the vector of Kriging coefficients $\boldsymbol{\lambda}(\mathbf{x})$ must satisfy

$$\begin{pmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^\mathsf{T} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}(\mathbf{x}) \\ \boldsymbol{\mu}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{k}(\mathbf{x}) \\ \mathbf{p}(\mathbf{x}) \end{pmatrix}, \tag{3}$$

with

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

the covariance matrix of the evaluation results,

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

the vector of covariances between $F(\mathbf{x})$ and $\mathbf{F}_n$,

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

the regression matrix, and $\boldsymbol{\mu}(\mathbf{x})$ a vector of Lagrange coefficients.

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

$$\hat{\sigma}^2(\mathbf{x}) = \mathbb{E}\left[F(\mathbf{x}) - \hat{F}(\mathbf{x})\right]^2 = k(\mathbf{x},\mathbf{x}) - \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T}\mathbf{k}(\mathbf{x}) - \mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\mu}(\mathbf{x}), \tag{4}$$

Usually, however, the covariance of F is not known a priori. It may then be chosen among a parametric family (for example, the Matérn covariance class [Stein, 1999]), with the parameters estimated by maximum likelihood (we shall discuss this point in Section 5). Given the evaluation results $\mathbf{f}_n$, the prediction of $f(\mathbf{x})$ becomes $\hat{f}(\mathbf{x}) = \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T}\mathbf{f}_n$ which can be used as a surrogate approximation for f, while $\hat{\sigma}^2(\mathbf{x})$ gives us an explicit characterisation of the uncertainty associated with this prediction. The conditional distribution of $F(\mathbf{x})$ is indeed Gaussian with mean $\hat{f}(\mathbf{x})$ and variance $\hat{\sigma}^2(\mathbf{x})$. Fig. 2(a) presents an example of prediction by Kriging. In what follows we describe two sampling criteria that use Kriging to balance local and global searches.

2.2 Maximizing the expected improvement

Sampling at a maximizer of the *expected improvement* (EI) was initially proposed by Mockus et al. [1978]. This strategy has been the object of most publications in the field in the last few years [Jones, 2001, Jones et al., 1998] and has been used for industrial applications (see, e.g., Huang [2005]).

A trivial estimator of the minimum of f after n evaluations is $M_n = \min_{\mathbf{x}_i\in\mathbb{S}_n} F(\mathbf{x}_i)$. The EI strategy chooses as the next evaluation point $\mathbf{x}_{n+1}$, a minimizer of $\mathbb{E}[\min(M_n, F(\mathbf{x}))|\mathcal{F}_n]$, the expected value of this estimator *after* the evaluation at $\mathbf{x}$ and given the previous evaluation results $\mathcal{F}_n = \{\mathbf{F}_n = \mathbf{f}_n\}$. This amount to maximizing the expected improvement

$$\mathrm{EI}(\mathbf{x}) = \mathbb{E}[\max(M_n - F(\mathbf{x}), 0)|\mathcal{F}_n], \tag{5}$$

which represents the average excursion of $F(\mathbf{x})$ under the current estimate of the minimum. The EI has a closed-form expression [Jones, 2001], which involves both the Kriging prediction $\hat{f}$ and the variance $\hat{\sigma}^2$ of its error

$$\mathrm{EI}(\mathbf{x}) = \hat{\sigma}(\mathbf{x})\Phi'(u) + u\Phi(u), \tag{6}$$

with $u = (\mathbf{x} - m_n)/\hat{\sigma}(\mathbf{x})$, $m_n = \min_{\mathbf{x}_i\in\mathbb{S}_n} f(\mathbf{x}_i)$ and $\Phi(\cdot)$ the Gaussian cumulative density function.

As can be seen on Fig. 2(c), the EI criterion favors sampling where the prediction is small, but also where the uncertainty on this prediction is large.

2.3 Minimizing the conditional entropy of the global minimizers

Where EI concentrates on the global *minimum*, we proposed in Villemonteix et al. [2008b] to concentrate rather on the global *minimizers*, which is of particular importance in a context of parameter estimation. Instead of considering an estimator of the minimum, we estimate *the probability distribution* of the global minimizers. Let $\mathbb{X}_\mathrm{d}$ be a finite subset of $\mathbb{X}$ and

denote by $\mathbf{X}^*$ a minimizer of F over $\mathbb{X}_\mathrm{d}$. The probability distribution $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n) = \mathrm{P}(\mathbf{X}^* = \cdot|\mathcal{F}_n)$ of the random vector $\mathbf{X}^*$ given all previous evaluation results can then be approximated using *conditional simulations* of F [Villemonteix et al., 2008b]. Conditional simulations are simulations of F that interpolate all available evaluation results (see, e.g., [Chilès and Delfiner, 1999] for details on how to generate them).

Assuming that $\mathbb{X}_\mathrm{d}$ is a suitable representation of $\mathbb{X}$ (we shall discuss this idea in Section 3.2), $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ represents what has been learned (through evaluations) and assumed (through the Gaussian model) about the minimizers. The main idea is then to quantify the uncertainty left on the location of these minimizers by the *entropy* of $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$. The progress made in finding a solution to the optimization problem is thus summarized by a scalar.

We then suggest to sample where the largest uncertainty reduction is expected, by minimizing the *conditional entropy of the global minimizers* (CEM)

$$H_n(\mathbf{x}) = \int_{y\in\mathbb{R}} p_{F(\mathbf{x})}(y|\mathcal{F}_n) H(\mathbf{X}^*|\mathcal{F}_n, F(\mathbf{x}) = y)dy, \qquad (7)$$

with $p_{F(\mathbf{x})}(\cdot|\mathcal{F}_n)$ the conditional distribution of the evaluation result $F(\mathbf{x})$ at $\mathbf{x}$ (Gaussian with mean and variance obtained by Kriging) and

$$H(\mathbf{X}^*|\mathcal{F}_n, F(\mathbf{x}) = y) = -\sum_{\mathbf{u}\in\mathbb{X}_\mathrm{d}} P_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n, F(\mathbf{x}) = y) \log_2(P_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n, F(\mathbf{x}) = y)) \qquad (8)$$

the entropy of $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n, F(\mathbf{x}) = y)$, the distribution of $\mathbf{X}^*$ given $\mathcal{F}_n$ and $\{F(\mathbf{x}) = y\}$. Practical details on the evaluation of (7) and particularly the approximation of $p_{F(\mathbf{x})}(\cdot|\mathcal{F}_n)$ are discussed in Villemonteix et al. [2008b].On Fig. 2(c) and 2(d), we can compare, on the same simple example, the EI and CEM sampling criteria. Their difference appears particularly clearly near the origin. EI samples at $x = 0$, clearly not the most informative choice with half of its neighborhood outside search space.

3. USING IAGO AND EGO

In the previous section, we have described briefly the principles of two sampling criteria for optimization. Both have been especially conceived for a careful spending of the evaluation budget. In this section, we shall discuss how to insert these criteria into optimization algorithms, and give some recommendations regarding their practical use.

3.1 From the sampling criterion to the optimization algorithm

We have seen in Section 2 that the principle of Bayesian optimization was iteratively to choose points at which the cost function should be evaluated, by optimizing a sampling criteria J. If we use, as here, a Gaussian model and Kriging to build the surrogate approximation, the resulting optimization algorithm looks like Algorithm 1. First, f has to be sampled on a small initial design in order to initialize the Kriging predictor, and possibly to allow an estimation of the covariance parameters. Then, the Kriging predictor is computed, and an additional point is selected to be the location of the next evaluation of f, by optimizing the sampling criterion J, based on this predictor. The covariance function may then be updated (for example by re-estimating its parameters), the model is re-computed, and the process of choosing new points continues until some stopping condition is met. The optimization algorithms based on EI and CEM have been named respectively EGO and IAGO. The advantages of IAGO over EGO will be discussed in Section 5. However, it must be stated right away that the computational complexity of IAGO is significantly higher than that of EGO (cf. Villemonteix et al. [2008b]). The complexity of computing EI is $O(n^3)$, while for CEM it is $O(n^2N)$, with N the size of $\mathbb{X}_\mathrm{d}$. In practice, IAGO should only be used for simulation times that are significantly larger than the time needed to minimize the CEM sampling criterion (a few minutes on a personal computer for problems such as the one described in Section 4). Fig. 3 presents an application of IAGO to the sine function already used for Fig. 2. After twelve iterations, the three global minimizers have been identified as shown by the conditional distribution (right-hand part of Fig. 3).

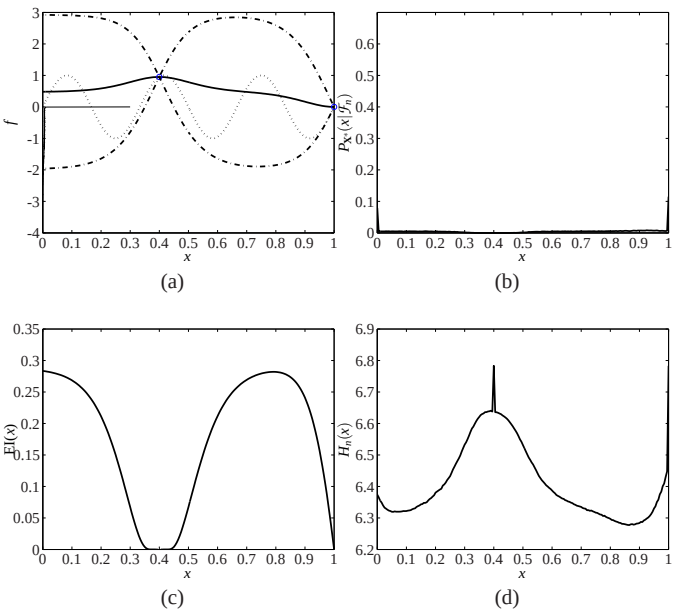

Figure 2. *(a)* Example of prediction by Kriging (bold line) based on two evaluation results (circles) of a supposedly unknown function (dotted line). The dash-dotted lines delimit 95% intervals for this prediction, and the thin line is the conditional distribution of the minimum. *(b)* Corresponding distribution of the global minimizers. *(c)* Expected improvement corresponding to the prediction. The next evaluation point would be $x = 0$ (the value of EI at $x = 0.8$ is slightly lower). *(d)* Corresponding CEM, which would lead to sample at $x = 0.87$.

Algorithm 1. Bayesian optimization based on Kriging and the sampling criterion J (for example EI or CEM)

1: Evaluate f on an initial design ▷ cf. Section
2: **while** the stopping condition is not satisfied, **do** ▷ cf. Section 3.3
3: Choose a covariance function ▷ cf. Section 5
4: Compute the Kriging prediction based on all available evaluation results
5: Evaluate f at $\arg\mathrm{opt}_{\mathbf{x}\in\mathbb{X}} J(\mathbf{x}, \mathbb{S}_n, \mathbf{f}_n, F)$.
6: **end while**

In the rest of Section 3, we shall discuss how to optimize EI and CEM in practice, and the stopping condition. The covariance choice will be discussed in Section 5, along with the initial design as we wish to insist on the difficulties still left to deal with in practice.

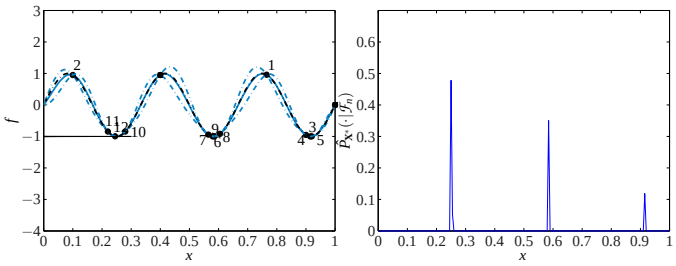

Figure 3. Twelve iterations of IAGO on a sine function. *(Left)* Kriging prediction (bold line) based on three initial evaluation results (squares) and on the twelve additional evaluations chosen by IAGO (circles). The dotted line is the sine function to be optimized (supposedly unknown). The thin horizontal line is the conditional distribution of the global minimum after twelve iterations of IAGO, and its support is almost reduced to the true global minimum. *(Right)* Conditional distribution of the global minimizers after twelve iterations of IAGO.

3.2 Optimization of the sampling criterion

Optimizing the sampling criterion is an important practical difficulty of Bayesian optimization, as there may be many local optima. However, one has to bear in mind that the evaluation of the sampling criterion does not require any evaluation of the cost function f, and that there is no need for exact optimization since the sampling criterion is only used to determine the next evaluation point. A small error will have little influence on the final estimation of the minimum of f and of the corresponding minimizers.

At the beginning of the optimization procedure, we have very little *a priori* on the potential location of the minimizers. However, as the number of evaluations increases, it becomes obvious that certain areas of search space do not have any interest for an additional evaluation as they stand no chance of containing a global minimizer. Therefore, we propose to use $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ as an auxiliary distribution for the optimization of the sampling criterion. For IAGO, this does not require any additional computation as this auxiliary distribution is already computed. The idea is to sample a *set of candidate points* from $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ and to compute exhaustively the values taken by the sampling criterion over this set. An evaluation is then performed at the most interesting point; $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ is recomputed and the set of candidate points re-sampled.

Note that, to compute $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$, a finite sub-set $\mathbb{X}_\mathrm{d}$ of $\mathbb{X}$ is needed. This set should represent the support of the probability density of the global minimizer of F over $\mathbb{X}$. If it stays fixed as the number of evaluation increases, the support of this density will dwindle, as well as the capacity of $\mathbb{X}_\mathrm{d}$ to describe this support. That is why we propose to re-sample $\mathbb{X}_\mathrm{d}$ after each new evaluation using $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ (estimated for example with a Gaussian kernel estimator [Chib and Greenberg, 1995]).

In summary, we can use $\mathbb{X}_\mathrm{d}$ as the set of candidate points and re-sample it to improve the precision of the optimization of the sampling criterion, and to maintain a good representation of the support of the probability density of the minimizers of F over $\mathbb{X}$. The resulting optimization algorithm is summarized by Algorithm 2 (with either EI or CEM as the sampling criterion).

Algorithm 2. Bayesian optimization based on Kriging and the sampling criterion J (for example EI or CEM), when the criterion is computed extensively over a set of candidate points re-sampled after each new evaluation.

Evaluate f on an initial design
Choose an initial grid $\mathbb{X}_\mathrm{d}$
while the stopping condition is not satisfied, **do**
    Choose a covariance function
    Compute the Kriging prediction based on all available evaluation results
    Estimate the conditional distribution of $\mathbf{X}^*$
    Re-sample $\mathbb{X}_\mathrm{d}$ given the new conditional distribution
    Evaluate J exhaustively over $\mathbb{X}_\mathrm{d}$
    Evaluate f at $\arg\mathrm{opt}_{\mathbf{x}\in\mathbb{X}_\mathrm{d}} J(\mathbf{x}, \mathbb{S}_n, \mathbf{f}_n, F)$.
end while

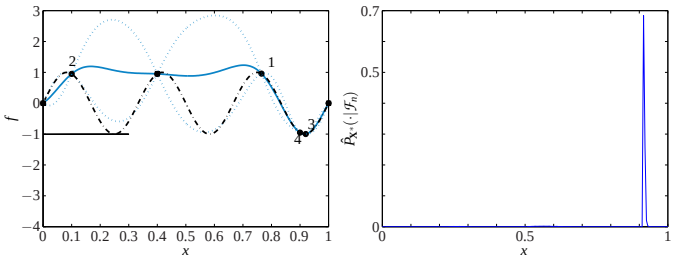

Figure 4. Optimization of a sine function with IAGO using (9) as the stopping condition, with $\sigma_\mathrm{stop} = 0.1$. Conventions are as in Fig. 3.

3.3 Stopping condition

Usually a simple stopping condition is available, namely exhaustion of the evaluation budget. It is indeed very common in practice to have exhausted the evaluation budget while still having a very poor estimation of the minimizers. Other stopping conditions have been proposed for either EGO [Jones et al., 1998] or IAGO [Villemonteix et al., 2008b]. They may be used when no budget has been specified beforehand. In this paper, we propose a new stopping condition, which is easier to use in practice. Let F^* be the minimum of F and let $\mathrm{var}(F^*|\mathcal{F}_n)$ be the variance of F^* given all past evaluation results. This quantity represent the uncertainty left on the estimation of the minimum. We then suggest to stop the algorithm when

$$\sqrt{\mathrm{var}(F^*)|\mathcal{F}_n} < \sigma_\mathrm{stop}, \tag{9}$$

where σ_stop is a positive threshold to be chosen by the user to reflect the desired precision on the solution. The conditional distribution of F^* (an example of which is presented on Fig. 2(a)) can be approximated along with this variance using conditional simulations. Fig. 4 shows the impact of the new stopping condition on the example of Fig. 3.

4. AN INDUSTRIAL APPLICATION

The tuning of a parametric controller based on a performance criterion quantifying deviation form some ideal response is formally equivalent to parameter identification. We had to optimize the 32 parameters of the control law of an electrical power steering device (EPSD). This was based on data generated by a Matlab/Simulink model of the device interacting with a model of the car on a driving test case characteristic of the typical use of the vehicle. The response of the car to this test case can be visualised in the torque – steering wheel angle plane (cf. Fig. 5). Its quality was quantified by a cost function f reflecting the distance between the response and the ideal one (as specified

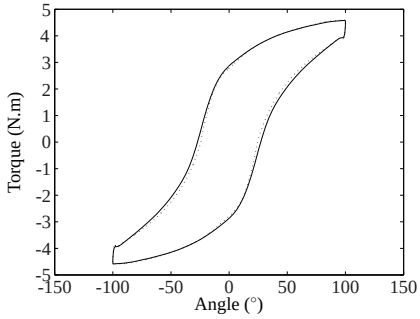

Figure 5. Response of an EPSD to a driving test case in the torque – steering wheel angle plane. The dotted line is obtained by simulation for the control laws obtained after optimization of 32 parameters using EGO. The continuous line represents the ideal response according to experts in the field.

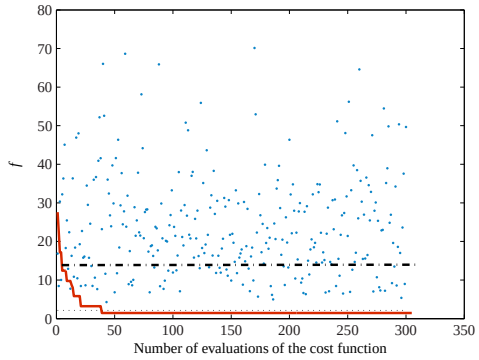

Figure 6. Application of EGO to the identification of the parameters of the control law of an EPSD. The bold line represents the best evaluation result obtained by EGO. The dash-dotted line represents the solution found by the Nelder-Mead Simplex method after 300 iterations. The points represent the results of the 300 evaluations on the LHS Sample. Finally, the dotted line represents the result of the evaluation at the minimizer of a polynomial model build from the 300 evaluation results of the LHS sample.

by experts in the field). Given the simulation time (10 minutes) for the test case and industrial needs, the minimization of f had to be carried out in less than 100 evaluations.

We dealt with this problem using EGO (preferred here to IAGO because of the relatively short simulation time) and re-estimated the parameters of a Matérn covariance after each evaluation, using maximum likelihood estimation. We compared our results with those of two other optimization approaches. The first was to use the Nelder-Mead Simplex, and the second was exhaustive computation over a Latin Hyper Square (LHS) sample [McKay et al., 1979] with 300 points. EGO turned out to be largely superior to the other approaches (see Fig. 6). It found in 45 evaluations of the cost function a better solution than that obtained via LHS sampling in 300 evaluations. Using the Nelder-Mead simplex was clearly the least efficient of the three approaches.

5. DISCUSSION

Choosing an algorithm. The interest of a given optimization algorithm can be tested in many ways. If convergence rates can be derived for a general class of functions, it becomes easy to

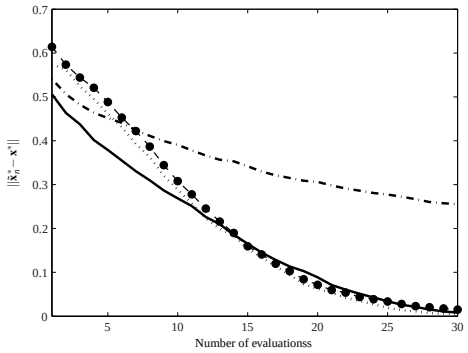

Figure 7. Mean of the estimation errors $||\tilde{\mathbf{x}}_n^* - \mathbf{x}^*||$ over 1000 sample paths as a function of the number of evaluations of f, for IAGO (bold line) EGO (dotted line) and a random search (with a uniform distribution, dash-dotted line), when confronted to a irregular Gaussian process indexed on $[0,1]^2$. The standard deviation of the estimation of the mean is always less than 0.02.

assert the interest of the algorithm compared to those of its competitors for which convergence rates have been obtained. When this is not the case, one may resort to large sets of test functions (see, e.g., Floudas et al. [1999]). An opinion can then be forged out of the precision obtained in the location of the minimum and minimizers on each test function, as well as the number of evaluations required. The practical interest of the latter approach can actually be questioned. If the comparison is conducted seriously, no method should turn out to be better than all others on every test case. A natural question is then, how can we deduce from the tests conducted which method to use *a priori*?

In the context of Bayesian optimization, we think that a more satisfying approach can be used to compare algorithms. For example, to compare EGO and IAGO, it seems natural to use sample paths of the Gaussian model behind both algorithms. By doing so, we can obtain empirical convergence results such as the one presented on Fig. 7. To obtain this result, 1000 sample paths of a Gaussian process (with $\mathbb{X} = [0,1]^2$) were generated and IAGO and EGO were applied to each of them. After each iteration and for each sample path, $||\tilde{\mathbf{x}}_n^* - \mathbf{x}^*||$ was computed, where $\tilde{\mathbf{x}}_n^*$ is an estimator of $\mathbf{x}^*$, obtained as a value of the parameter vector associated with the cost

$$\tilde{m}_n = \min(m_{n-1}, f(\arg\max_{\mathbf{x}\in\mathbb{X}_d} P_{\mathbf{X}^*}(\mathbf{x}|\mathcal{F}_n))).$$

The vector $\tilde{\mathbf{x}}_n^*$ is thus the solution that would have been retained in practice with a budget of n evaluation.

On Fig. 7, IAGO appears to be significantly quicker to converge than EGO. The drawbacks of the use of EGO are evidenced on Fig. 8, where IAGO and EGO are applied on one of the sample paths used to obtain the results of Fig. 7. It appears that the EI stays stuck on a local minimum. This result has been obtained on a *single* Gaussian process. In Villemonteix et al. [2008a], we noted that, if the test process is the same as the model underlying IAGO and EGO, IAGO seemed regularly to outperform EGO.

This approach could clearly be extended to any type of global optimization method. We could, for example, create a set of test random processes, carefully chosen to represent several types of regularity or other characteristics.

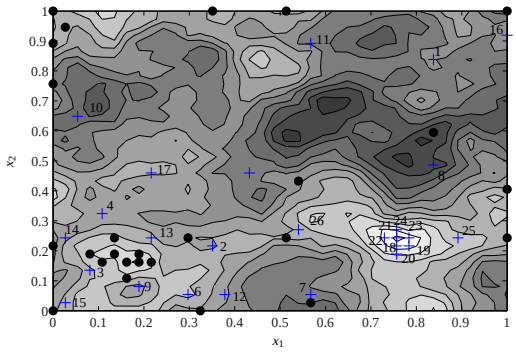

Figure 8. Minimization of one of the conditional sampled paths used to obtain the results of Fig. 7. The circles indicate the evaluation locations chosen by EGO, the crosses these chosen by IAGO. The order in which the evaluations are chosen is also indicated for IAGO. The global minimizer is at $\mathbf{x} = (0.77, 0.25)^\mathsf{T}$.

Choosing a covariance. The choice of a covariance is a central question for Kriging prediction and, by way of consequence, for Bayesian optimization. Generally, the covariance is chosen in a parametrized class, and its parameters are estimated from available data. The method used for this estimation depends on the field (variogram fitting in geostatistics, cross-validation in machine learning or maximum likelihood in statistics [Villemonteix et al., 2008b]), but in a context of small evaluation budget, these methods may not be applicable.

[Jones et al., 1998] propose to initialize the EGO algorithm by an LHS sample of size $10d$, with d the dimension of search space $\mathbb{X}$. By doing so, the algorithm gets information on the promising areas in search space, but most of all a first estimation of the parameters can be carried out. This idea is appealing if d is small, but as d increases, the size of search space grows exponentially, and it soon becomes very hard to estimate the parameters of the covariance. Besides, if we consider the available evaluation budget for the application in Section 4, the size of the initial sample would have exceeded the total evaluation budget.

It finally appears that classical approaches for tuning the covariance do not apply in our context [1]. Now, EGO or IAGO may perform very poorly if the covariance is inadequate (see, e.g., Jones [2001] for examples). Some preliminary results, not presented here, indicate however that when choosing the covariance inside the Matérn covariance class, it is possible to ensure a satisfactory behavior of IAGO or EGO by limiting search space for the covariance parameters. There even seems to be fixed covariances that ensure a satisfactory behavior of these algorithms over a large class of functions.

For the example of Section 4, we estimated the covariance parameters using maximum likelihood over a very restricted search space. In future work, we feel that a Bayesian approach would be best suited here to guide the estimation of the parameters of the covariance.

[1] Except when evaluations results are available from a problem close to the one considered. In such a case the parameters can be estimated directly.

6. CONCLUSIONS

The main goal of this paper was to present EGO and IAGO, two Bayesian optimisation algorithms dedicated to problems with very restricted simulation budgets to the parameter identification community. We described the principles shared by the two algorithms and discussed their use in practice. The interest of Bayesian optimization for parameter identification was evidenced by an industrial application.

Comparison using the Gaussian process model behind EGO and IAGO indicates a clear superiority for IAGO. However, its computational complexity restricts its potential applications to the optimization of functions that require a large amount of computer time to be evaluated.

REFERENCES

S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *Amer. Statist.*, 49(4):327–335, 1995.

J.P. Chilès and P. Delfiner. *Geostatistics, Modeling Spatial Uncertainty*. John Wiley & Sons, Inc, New York, 1999.

C.A. Floudas, P.M. Pardalos, C. Adjiman, W.R. Esposito, Z.H. Gümüs, S.T. Harding, J.L. Klepeis, C.A. Meyer, and C.A. Schweiger. *Handbook of Test Problems in Local and Global Optimization*. Nonconvex Optimization and Its Applications , Vol. 33. Kluwer Academic Publishers, 1999.

D. Huang. *Experimental Planning and Sequential Kriging Optimization Using Variable Fidelity Data*. PhD thesis, Ohio State University, 2005.

D. Huang and T. Allen. Design and analysis of variable fidelity experimentation applied to engine valve heat treatment process design. *Journal of Royal Statistics, Series C.*, 54, Part 2: 443–463, 2005.

D.R. Jones. A taxonomy of global optimization methods based on response surfaces. *J. Global Optim.*, 21:345–383, 2001.

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

M.D. McKay, W.J. Conover, and R.J. Beckman. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21: 239–245, 1979.

J. Mockus, V. Tiesis, and A. Zilinskas. The application of Bayesian methods for seeking the extremum. In L.C.W. Dixon and G.P. Szego, editors, *Towards Global Optimization 2*, pages 117–129, North Holland, New York, 1978.

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

J. Villemonteix, E. Vazquez, and É. Walter. Identification of expensive-to-simulate parametric models using kriging and stepwise uncertainty reduction. In *46th IEEE Conference on Decision and Control*, pages 5505–5510, New Orleans (USA), December 12-14, 2007.

J. Villemonteix, E. Vazquez, M. Sidorkiewicz, and É. Walter. Global optimization of expensive-to-evaluate functions: an empirical comparison of two sampling criteria. *To appear in the Mykonos special issue of the J. Global Optim.*, 2008a.

J. Villemonteix, E. Vazquez, and É. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *To appear in J. Global Optim.*, 2008b.

A. Zilinskas. A review of statistical models for global optimization. *J. Global Optim.*, 2:145–153, 1992.