Global optimization based on noisy evaluations: an empirical study of two statistical approaches

Emmanuel Vazquez, Julien Villemonteix, Maryan Sidorkiewicz, Eric Walter

Global optimization based on noisy evaluations: an empirical study of two statistical approaches

Emmanuel Vazquez [a], **Julien Villemonteix** [b], **Maryan Sidorkiewicz** [b] **and Éric Walter** [c]

a. SUPELEC, 91192 Gif-sur-Yvette, France
b. RENAULT S.A., France
c. Laboratoire des Signaux et Systèmes, CNRS, SUPELEC, UNIV. PARIS-SUD, 91192 Gif-sur-Yvette, France

E-mail: emmanuel.vazquez@supelec.fr

Abstract. The optimization of the output of complex computer codes has often to be achieved with a small budget of evaluations. Algorithms dedicated to such problems have been developed and compared, such as the Expected Improvement algorithm (EI) or the Informational Approach to Global Optimization (IAGO). However, the influence of noisy evaluation results on the outcome of these comparisons has often been neglected, despite its frequent appearance in industrial problems. In this paper, empirical convergence rates for EI and IAGO are compared when an additive noise corrupts the result of an evaluation. IAGO appears more efficient than EI and various modifications of EI designed to deal with noisy evaluations.
Keywords. Global optimization; computer simulations; kriging; Gaussian process; noisy evaluations.

1. Context and objectives

Computer simulations are often used to optimize a product or a process, i.e., to find the best feasible values for design parameters. These optimization problems have specific difficulties due to the very nature of the function to be optimized (thereafter called objective function). First, the derivatives of the objective function cannot generally be obtained, which prevents the use of gradient-based optimization methods. Second, computer simulations are often very time consuming (several hours for one run is common). The optimization must therefore be carried out with a *restricted budget of function evaluations*, generally excluding stochastic search algorithms, such as simulated annealing. Third, the objective functions encountered hardly ever turn out to be convex, discouraging the use of local methods. Fourth, evaluations of the objective function may be corrupted by noise. This noise might for instance stem from propagation of round-off errors, bad conditioning or slow convergence of iterative schemes. Figure 1 illustrates a problem of convergence in a Navier-Stokes simulation. Other examples can be found in [1] in the context of aeronautics.

In this paper, we empirically compare two global optimization algorithms, namely the Expected Improvement (EI) algorithm[1] [2] and the Informational Approach to Global

[1] In the literature, EI refers to a criterion that has to be optimized to choose the location of the next evaluation. Here, to simplify the presentation, EI stands for the optimization algorithm based on this criterion.

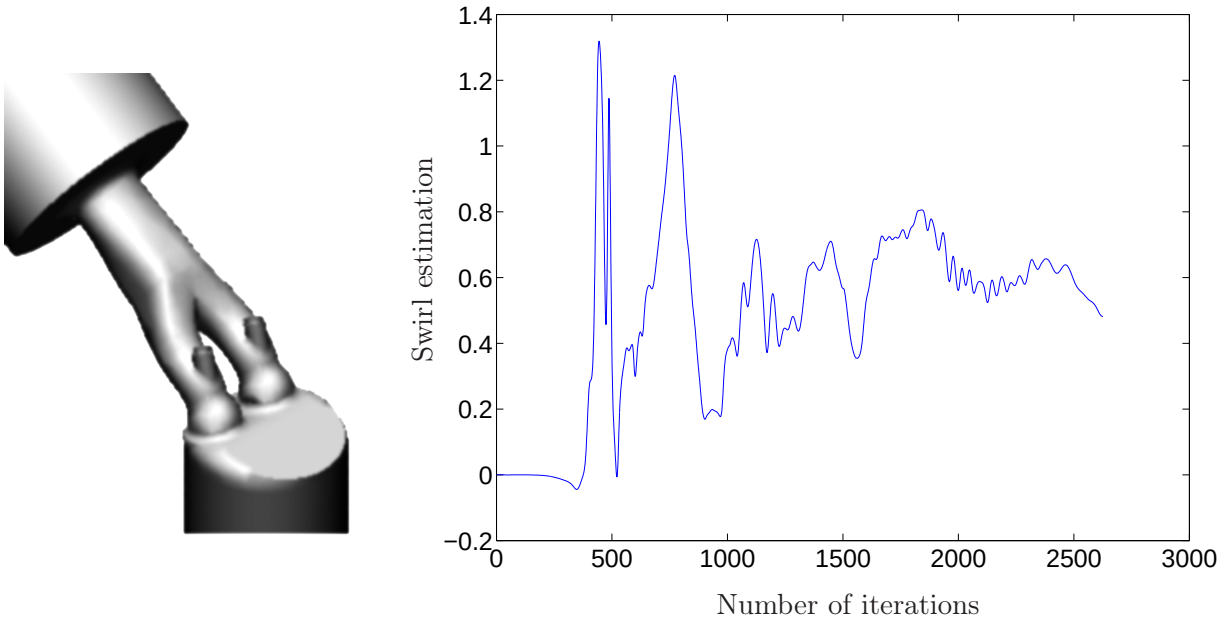

Figure 1: Simulation of airflow in an intake port of a Renault engine. This Navier-Stokes problem is solved iteratively by a finite element method. Left: 3D representation of an intake port. Right: value of the *swirl*, which quantifies the level of turbulence in the combustion chamber, as a function of the number of iterations of the solving procedure. Due to limitation of computer resources, the solver must be stopped before the stabilization of the solution.

Optimization (IAGO) strategy recently proposed in [3], *when the evaluations are noisy*. Both algorithms are based on a stochastic model of the objective function and aim at using a few evaluations of the objective function as efficiently as possible to infer the location of optimizers. A comparison in the noise-free case is presented in [4], however, no comparison between the two methods have yet been carried out when the evaluations are noisy. A brief introduction to statistical approaches for global optimization in the noise-free case is presented in Section 2. Section 3 describes the modifications necessary to deal with noisy evaluations. Empirical convergence rates are provided in Section 4 for both algorithms.

2. Noise-free GP-based global optimization

Global optimization algorithms based on statistical models have recently received particular attention [5–7]. These algorithms use a statistical model to build a prediction of the objective function (also called *surrogate model*). In this section, it is assumed that there is no noise, i.e., that the evaluations results correspond exactly to the values of the objective function. Using a prediction allows one to infer the most likely locations of the optimum given the past evaluations and to choose the next evaluation points accordingly. In principle, the statistical model need not be of any particular kind but the use of Gaussian random processes (GPs) turns out to be both easy and very efficient in practice. In this section, we shall recall some basic facts about Kriging, a well-known linear prediction method in the context of GPs, before very briefly presenting the EI and IAGO algorithms.

2.1. Kriging basics

Let ξ be a GP indexed by a parameter space $\mathbb{X} \subset \mathbb{R}^d$ with mean $m(x)$, $x \in \mathbb{X}$, and covariance function $k(x, y)$, $(x, y) \in \mathbb{X}^2$. Assume that $m(x)$ can be written as a linear parametric function $m(x) = b^{\mathsf{T}} p(x)$, where $p(x)$ is the q-dimensional vector of all monomials of degree less than

or equal to $l \in \mathbb{N}$, and $b \in \mathbb{R}^q$ is a vector of fixed but *unknown* coefficients. The theory of (universal) kriging (see, e.g., [8]) is concerned with the construction of the *best linear unbiased predictor* (BLUP) of ξ based on a finite set of pointwise observations of the process. For $x \in \mathbb{X}$, $\underline{x}_n = (x_1, \ldots, x_n) \in \mathbb{X}^n$, $n \geq 1$, denote by $\widehat{\xi}(x; \underline{x}_n)$ a linear predictor of $\xi(x)$ based on $\xi(x_1), \ldots, \xi(x_n)$, which can be written as

$$\widehat{\xi}(x; \underline{x}_n) = \lambda(x; \underline{x}_n)^{\mathsf{T}} \underline{\xi}_n \,, \tag{1}$$

with $\underline{\xi}_n = (\xi(x_1), \ldots, \xi(x_n))^{\mathsf{T}}$ and $\lambda(x; \underline{x}_n)$ a vector of weights $\lambda_i(x; \underline{x}_n)$, $i = 1, \ldots, n$. A fundamental result of kriging theory is that the BLUP is the linear projection of $\xi(x)$ onto $\mathrm{span}\{\xi(x_i), i \leq n\}$ orthogonally to the space of functions $\mathcal{P} := \{b^{\mathsf{T}} p(x); b \in \mathbb{R}^q\}$ and in such way that the norm of the prediction error is minimum, which leads to express the vector of kriging coefficients $\lambda(x; \underline{x}_n)$ as the solution of the linear system of equations

$$\begin{pmatrix} k(\underline{x}_n, \underline{x}_n) & P^{\mathsf{T}} \\ P & 0 \end{pmatrix} \begin{pmatrix} \lambda(x; \underline{x}_n) \\ \alpha(x; \underline{x}_n) \end{pmatrix} = \begin{pmatrix} k(x, \underline{x}_n) \\ p(x) \end{pmatrix} , \tag{2}$$

where $k(\underline{x}_n, \underline{x}_n)$ is the $n \times n$ matrix of covariances $k(x_i, x_j)$, P is a $q \times n$ matrix with entries ${x_j}^i$ for $j = 1, \ldots, n$ and multi-indexes $i = (i_1, \ldots, i_d)$ such that $|i| := i_1 + \cdots + i_d \leq l$, $\alpha(x; \underline{x}_n)$ is a vector of Lagrange coefficients, $k(x, \underline{x}_n)$ is a vector of size n with entries $k(x, x_i)$ and $p(x)$ is a vector of size q with entries x^i, i such that $|i| \leq l$.

The variance of the kriging error, or *kriging variance*, is given by

$$\widehat{\sigma}^2(x; \underline{x}_n) := \mathrm{var}[\xi(x) - \widehat{\xi}(x; \underline{x}_n)] = k(x, x) - \lambda(x; \underline{x}_n)^{\mathsf{T}} k(x, \underline{x}_n) - \alpha(x; \underline{x}_n)^{\mathsf{T}} p(x) \,. \tag{3}$$

The knowledge of this variance makes it possible to derive confidence intervals for the predictions.

2.2. The Expected Improvement algorithm

The EI algorithm [2] is a well-known optimization method based on GP modeling [9]. The objective function $f : \mathbb{X} \to \mathbb{R}$ (to be maximized, say) is modeled by a GP ξ. Let $M_n = \max_{i=1,\ldots,n} \xi(x_i)$ be the maximum observed at step n. The EI strategy chooses an evaluation point x_{n+1} that maximizes the quantity

$$\mathrm{E}[M_{n+1} \mid \underline{\xi}_n] = \mathrm{E}\left[\max\left(\xi(x_{n+1}), M_n\right) \mid \underline{\xi}_n\right] = M_n + \mathrm{E}\left[\max\left(\xi(x_{n+1}) - M_n, 0\right) \mid \underline{\xi}_n\right] \tag{4}$$

(we have used the fact that $\mathrm{E}[M_n \mid \underline{\xi}_n] = M_n$). The function $\rho_n(x) := \mathrm{E}\left[\max\left(\xi(x) - M_n, 0\right) \mid \underline{\xi}_n\right]$ is called the *expected improvement* at x. This quantity is always positive and represents the average excursion of $\xi(x)$ above the current maximum M_n. The expected improvement has a closed-form expression, based on the kriging predictor:

$$\rho_n(x) \;=\; \begin{cases} \widehat{\sigma}(x; \underline{x}_n)\, \Phi'\left(\frac{u}{\widehat{\sigma}(x; \underline{x}_n)}\right) + u\, \Phi\left(\frac{u}{\widehat{\sigma}(x; \underline{x}_n)}\right) & \text{if } \widehat{\sigma} > 0, \\ \max(u, 0) & \text{if } \widehat{\sigma} = 0 \,, \end{cases} \tag{5}$$

with $u = \widehat{\xi}(x; \underline{x}_n) - M_n$, and where Φ denotes the Gaussian cumulative density function.

Practical issues. A first practical issue is the choice of a covariance function for ξ. Generally, this covariance is chosen inside a parametrized class of covariances, for instance Matérn covariances [10] or polynomial covariances [11]. The parameters can be estimated using restricted maximum likelihood estimation (REML) [10; 12; 13]. A second practical issue is the maximization of $\rho_n(x)$. Because evaluating $\rho_n(x)$ is cheap, a very large number of evaluation points can be chosen. Thus, a convenient solution is to restrict the search of the maximum on a *finite subset* $\mathbb{X}_d$ of $\mathbb{X}$ and to compute $\rho_n(x)$ extensively over $\mathbb{X}_d$. This approach is acceptable provided that $\mathbb{X}_d$ ensures a regular filling of $\mathbb{X}$. This can for example be achieved by using a regular lattice (in low dimension) or a Latin hypercube sampling (LHS) [14] (in high dimension).

2.3. The IAGO algorithm

The authors have recently introduced an *Informational Approach to Global Optimization* (IAGO) based on GP modeling [3]. IAGO provides a choice of evaluation point that is one-step optimal in terms of reduction of the uncertainty on the maximizer location. It is based on two main ideas.

Let $\mathbb{X}_d$ be a finite subset of $\mathbb{X}$, as above, and denote by $X^* \in \mathbb{X}_d$ a global maximizer of ξ on $\mathbb{X}_d$. Note that X^* is a random variable. The first idea of the IAGO strategy is to estimate the probability distribution $P_{X^*|\underline{\xi}_n} : \mathbb{X}_d \to [0,1]$ of X^* conditioned on the observations $\xi(x_i)$, $i = 1, \ldots, n$. This estimation can be carried out using *conditional simulations* of ξ [3]. A conditional simulation of ξ is simply the generation of a sample path of ξ conditioned on the observations. Generating a conditional sample path is straightforward using the kriging predictor (see [3; 8] for an insight into how conditional simulations are generated). In the noise-free setting, conditional simulations interpolate the observations (see Figure 2).

The second idea is to consider the entropy of $P_{X^*|\underline{\xi}_n}$ as a measure of uncertainty on the maximizer location and then to select a new observation point $x_{n+1} \in \mathbb{X}_d$ that will, *in mean*, maximize the decrease of the entropy. Formally, the IAGO strategy is defined as

$$x_{n+1} = \arg\min_{x \in \mathbb{X}_d} \mathrm{E}[H(X^*; \underline{\xi}_n, \xi(x)) \mid \underline{\xi}_n] \,, \tag{6}$$

$$= \arg\min_{x \in \mathbb{X}_d} \int_{z \in \mathbb{R}} H(X^*; \underline{\xi}_n, \xi(x) = z) p_{\xi(x)|\underline{\xi}_n}(z) dz \,, \tag{7}$$

where $H(X^*; \underline{\xi}_n, \xi(x))$ stands for the entropy of X^* conditioned on the vector of observations $\underline{\xi}_n$ and the candidate observation $\xi(x)$, which can be written as

$$H(X^*; \underline{\xi}_n, \xi(x)) = -\sum_{y \in \mathbb{X}_d} P_{X^*|\underline{\xi}_n, \xi(x)}(y) \log P_{X^*|\underline{\xi}_n, \xi(x)}(y) \,,$$

and $p_{\xi(x)|\underline{\xi}_n}$ denotes the density of the candidate observation conditioned on $\underline{\xi}_n$.

Practical issues regarding the IAGO strategy are discussed in [3].

3. Optimization with noisy evaluations

The literature on the optimization of computer models generally assumes that the evaluations are noise-free. However this may be too-idealized a view, as illustrated by the example of Figure 1. Evaluation errors usually stem from a trade-off between the precision of the numerical model and computer resources. The problem to be addressed in this section is the optimization of an objective function $f : \mathbb{X} \to \mathbb{R}$ from noisy evaluations

$$f_i^{\mathrm{obs}} = f(x_i) + \varepsilon_i \,,$$

where, for all i, $\varepsilon_i \in \mathbb{R}$ represents an additive evaluation error.

3.1. Kriging with noisy observations

Assume that the f_i^{obs}s are sample values of the random variables $\xi_i^{\mathrm{obs}} = \xi(x_i) + N_i$, $i = 1, \ldots, n$, where the N_is are Gaussian random variables with zero-mean and known covariance matrix K_N. (If a parametrized covariance is chosen for the noise, its parameters can be estimated by maximum likelihood together with those of the covariance of ξ). When, as generally assumed, the noise is white, $K_N = \sigma_N^2 I_n$, where I_n is the $n \times n$ identity matrix.

The BLUP $\widehat{\xi}(x; \underline{x}_n)$ of $\xi(x)$ is the linear projection of $\xi(x)$ onto $\mathrm{span}\{\xi_i^{\mathrm{obs}}, i \leq n\}$ orthogonally to the space of functions $\mathcal{P} := \{b^{\mathsf{T}} p(x); b \in \mathbb{R}^q\}$ and in such way that the norm of the prediction error is minimum. Thus,

$$\widehat{\xi}(x; \underline{x}_n) = \lambda(x; \underline{x}_n)^{\mathsf{T}} \underline{\xi}_n^{\mathrm{obs}} \,,$$

with $\underline{\xi}_n^{\text{obs}} = (\xi_1^{\text{obs}}, \ldots, \xi_n^{\text{obs}})^{\mathsf{T}}$. The vector $\lambda(x;\underline{x}_n) = (\lambda_1(x;\underline{x}_n), \ldots, \lambda_n(x;\underline{x}_n))^{\mathsf{T}}$ is obtained by solving the system

$$\begin{pmatrix} k(\underline{x}_n,\underline{x}_n) + K_N & P^{\mathsf{T}} \\ P & 0 \end{pmatrix} \begin{pmatrix} \lambda(x;\underline{x}_n) \\ \alpha(x;\underline{x}_n) \end{pmatrix} = \begin{pmatrix} k(x,\underline{x}_n) \\ p(x) \end{pmatrix} . \tag{8}$$

As in Section 2.1, the variance of the prediction is given by (3).

In the next two sections, we discuss the specific changes to be made to the EI and IAGO algorithms to deal with noisy evaluations.

3.2. EI with noisy evaluations

In principle, the EI algorithm can be used in the case of noisy evaluations without modification. An iteration of the EI algorithm then writes

$$x_{n+1} = \arg\max_{x\in\mathbb{X}_d} \rho_n(x)\,,$$

with $\rho_n(x)$ again defined by (5), where M_n is replaced by $\max_{i\in\{1,\ldots,n\}} \xi_i^{\text{obs}}$, and where $\widehat{\xi}(x;\underline{x}_n)$ and $\widehat{\sigma}(x;\underline{x}_n)$ are the kriging predictor and the kriging variance obtained from (8).

However, $M_n = \max_{i\in\{1,\ldots,n\}} \xi_i^{\text{obs}}$ no longer converges to the maximum of ξ when the number of observations grows (M_n actually tends to exceed the maximum of ξ). Other choices for the estimator of the maximum could be considered. For instance, we could choose $M_n = \max_{x\in\mathbb{X}_d} \widehat{\xi}(x;\underline{x}_n)$, i.e., we would consider the excursions (the improvement) above the maximum of the predictor instead of the maximum of the noisy observations (this modification will be referred to as EIm). Another approach, called Augmented Expected Improvement (AEI), is proposed in [7]. The AEI is an empirical modification of the classical EI criterion, which was shown to perform better on several test problems. To the best of our knowledge, there are no other modifications of the EI criterion designed specifically to cope with noisy evaluations[2]. In Section 4, they will be compared, with the results obtained with IAGO.

3.3. IAGO with noisy evaluations

As for the EI algorithm, the IAGO algorithm can be used without modification in the case of noisy evaluations. Note that sample paths conditioned on noisy observations have to be generated, which is again straightforward using the noisy version of kriging. Figure 2 illustrates conditional simulation in the case of noisy evaluations.

4. Numerical experiments

We feel that the comparison of EI and IAGO should be based on a Monte-Carlo approach rather than on a small set of classical test functions [4]. Indeed, the Monte-Carlo approach makes it possible to estimate convergence rates for an entire class of functions, as opposed to a few particular functions.

The proposed comparison methodology uses sample paths of a GP with a given covariance function as objective functions, and observes the resulting mean convergence rates using EI, EIm, AEI and IAGO in the case of noisy evaluations. We use the *same* covariance function in the optimization algorithms, and therefore compare the algorithms on the class of functions they are assumed to optimize. To do so, sample paths of a GP with a Matérn covariance are generated over a regular lattice $\mathbb{X}_d \subset [0,1]^2$. Using the parametrization in [10, p. 50], the Matérn covariance can be written as

$$k(x,y) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}h}{\rho}\right)^{\nu} \mathcal{K}_\nu \left(\frac{2\nu^{1/2}h}{\rho}\right) , \tag{9}$$

[2] In [1], the authors propose to optimize with the EI algorithm an approximate version of the objective function when the evaluations are noisy. We feel that it is not relevant to include this approach in our comparisons.

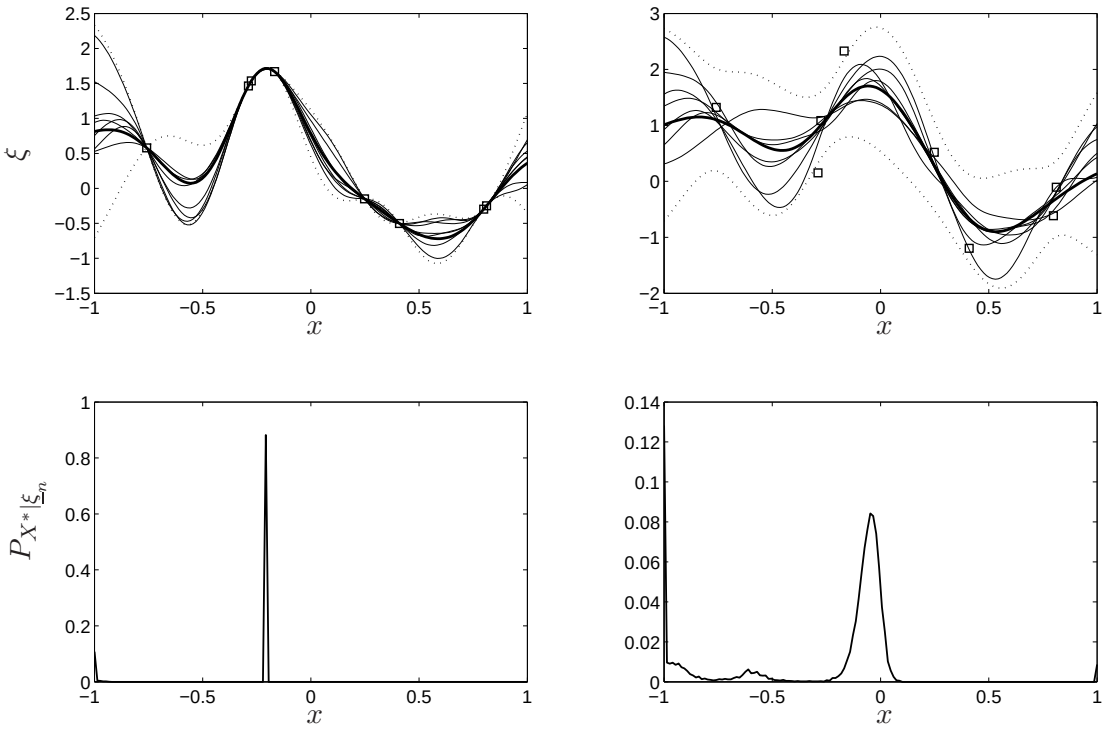

Figure 2: Conditional simulations and estimates of the maximizer distribution. Top: irregularly sampled evaluations of $f(x)$ (squares) without noise (left) and with noise (right), kriging prediction (bold solid line), 95% confidence intervals of the prediction (dotted lines), conditional sample paths (thin solid line). Bottom : estimates of $P_{X^*|\underline{\xi}_n}$ without noise (left) and with noise (right).

where $h = \|x - y\|$, $\mathcal{K}_\nu$ is the modified Bessel function of the second kind, $\nu > 0$ controls the regularity (the differentiability) of the covariance at the origin, $\sigma^2 > 0$ is a variance parameter, and $\rho > 0$ is a scale parameter.

The regularity parameter ν has a significant influence on the convergence rates in the noise-free case [4]. When ν increases, the sample paths get smoother and easier to optimize. To account for a variety of possible behaviors for the objective function, we chose to work with two regularities. A set of 500 smooth sample paths were generated using $\nu = 5$, $\rho = 0.3$ and $\sigma = 1.5$ as parameters for the Matérn covariance. The same number of irregular sample paths were generated using $\nu = 1.5$, $\rho = 0.3$ and $\sigma = 1.5$.

For each sample path, sixty optimization steps were performed using EI, EIm, AEI and IAGO. To provide a reference, we also performed a uniform random search over $\mathbb{X}_d$. Each evaluation was corrupted by additive Gaussian white noise (with zero mean and standard deviation $\sigma_N = 0.5$). We provide at each iteration step, and for each criterion, the mean distance (over all sample paths) between the true global maximum $\max_{x\in\mathbb{X}_d} \xi(x)$ and the current estimate of the maximum given by the kriging predictor $\max_{i=1,\ldots,n} \widehat{\xi}(x_i; \underline{x}_n)$. Note that this convergence measure corresponds to the EIm criterion. We also compute the mean entropy (over all sample paths) of $P_{X^*|\underline{\xi}_n}$ to account for the quality of the estimation of the maximizer.

For both comparison criteria, and both regularities, IAGO performs better, and this right from the start (see Figures 3 and 4). In fact, the regularity of sample paths has little influence on the convergence rate of one algorithm relatively to another in the case of noisy evaluations (compare Figures 3 and 4), as opposed to the noise-free case [4]. For both regularities, EI

converges faster than random search in terms of the distance to the maximum but does not bring any improvement in terms of entropy reduction. The two variants of EI do not bring any major improvement over the original version. EIm is actually outperformed not only by EI, but even by a simple random search, while the small gain offered by AEI is insufficient to better IAGO.

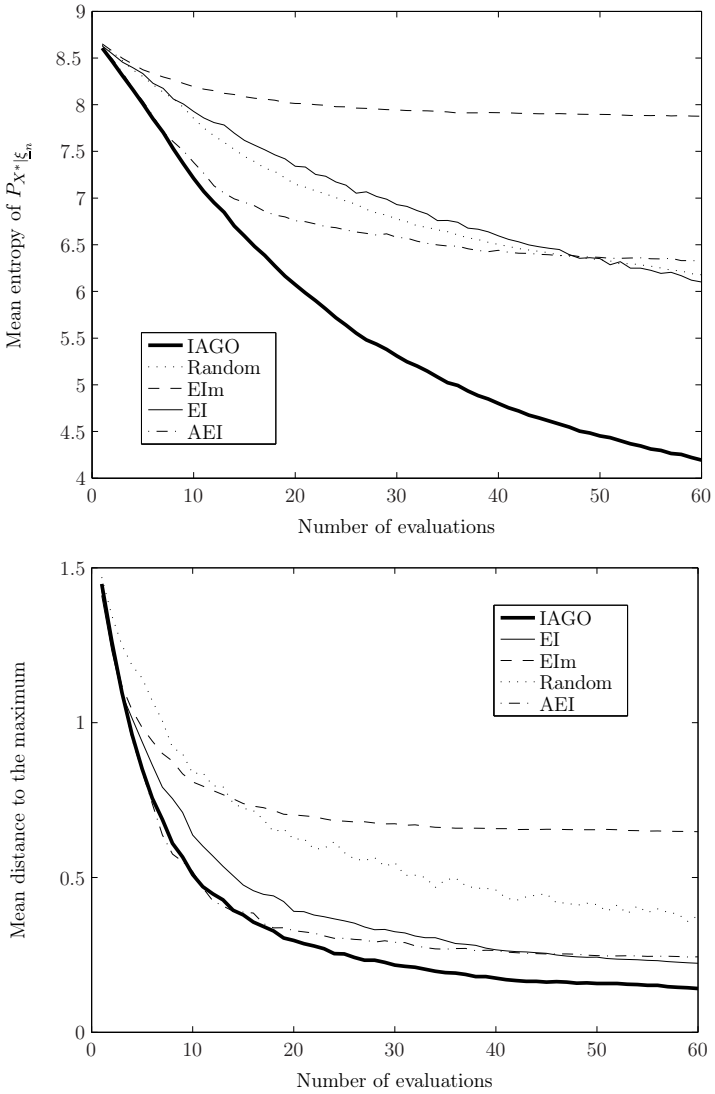

Figure 3: Convergence rates using EI, EIm, AEI, IAGO and random search, when convergence is measured by entropy of $P_{X^*|\underline{\xi}_n}$ (top), and when convergence is measured by the distance $\max_{x\in\mathbb{X}_d} \xi(x) - \max_{i=1,\ldots,n} \widehat{\xi}(x_i;\underline{x}_n)$ (bottom). The sample paths used here are smoother than those used for Figure 4 (the parameters for the Matérn covariance are $\nu = 5$, $\rho = 0.3$ and $\sigma = 1.5$). The noise standard deviation is $\sigma_N^2 = 0.5$

5. Conclusions

This paper has presented two statistical global optimization algorithms in the case of noisy evaluations of the objective function. We conducted an empirical study of their performance and found a clear superiority of IAGO over EI in the case of noisy evaluations.

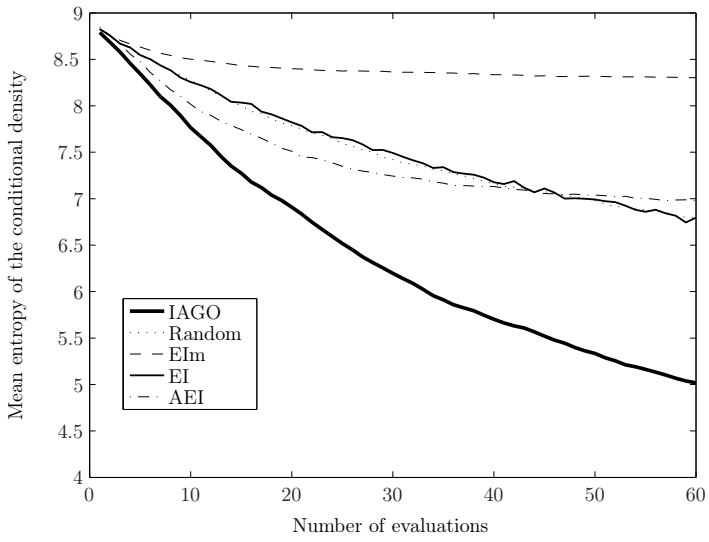

Figure 4: Convergence rates using EI, EIm, AEI, IAGO and random search, when convergence is measured by entropy of $P_{X^*|\underline{\xi}_n}$. The sample paths are simulated using the covariance (9) with $\nu = 1.5$, $\rho = 0.3$ and $\sigma = 1.5$. The standard deviation of noise is $\sigma_N^2 = 0.5$. We do not provide the convergence rates in terms of the distance to the maximum, as they are similar to those of Figure 3.

References

[1] Forrester A I J, Keane A J and Bressloff N W 2006 *AIAA Journal* **44** 2331–2339

[2] Schonlau M and Welch W 1996 *Proc. Section on Physical and Engineering Sciences,* (American Statistical Association) pp 183–186

[3] Villemonteix J, Vazquez E and Walter E 2006 An informational approach to the global optimization of expensive-to-evaluate functions. submitted to J. Global Optim. URL `http://arxiv.org/abs/cs.NA/0611143`

[4] Villemonteix J, Vazquez E, Sidorkiewicz M and Walter E 2007 Global optimization of expensive-to-evaluate functions: an empirical comparison of two sampling criteria. submitted to J. Global Optim. URL `http://hal.archives-ouvertes.fr/hal-00205120/fr/`

[5] Jones D, Schonlau M and William J 1998 *J. Global Optim.* **13** 455–492

[6] Jones D 2001 *J. Global Optim.* **21** 345–383

[7] Huang D, Allen T, Notz W and Zeng N 2006 *J. Global Optim.* **34** 441–466

[8] Chilès J P and Delfiner P 1999 *Geostatistics: Modeling Spatial Uncertainty* (New York: Wiley)

[9] Zilinskas A 1992 *J. Global Optim.* **2** 145–153

[10] Stein M L 1999 *Interpolation of Spatial Data: Some Theory for Kriging* (New York: Springer)

[11] Matheron G 1973 *Adv. Appl. Prob.* **5** 439–468

[12] Mardia K V and Marshall R J 1984 *Biometrika* **73** 135–146

[13] Vazquez E 2005 *Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications* Ph.D. thesis Univ Paris XI Orsay, France

[14] McKay M D, Beckman R J and Conover W J 1979 *Technometrics* **21** 239–245