Sequential search based on kriging: convergence analysis of some algorithms

Vazquez, Emmanuel

Bect, Julien

SUPELEC, Gif-sur-Yvette, France
e-mail: emmanuel.vazquez@supelec.fr, julien.bect@supelec.fr

1 Introduction

Let $\mathcal{F}$ be a set of real-valued functions on a set $\mathbb{X}$ and let $S : \mathcal{F} \to \mathcal{G}$ be an arbitrary mapping. We consider the problem of making inference about $S(f)$, with $f \in \mathcal{F}$ unknown, from a finite set of pointwise evaluations of f. We are mainly interested in the problems of approximation and optimization. Formally, a deterministic algorithm to infer a quantity of interest $S(f)$ from a set of n evaluations of f is a pair $\left(\underline{X}_n, \widehat{S}_n\right)$ consisting of a deterministic *search strategy*

$$\underline{X}_n : f \mapsto \underline{X}_n(f) = (X_1(f), X_2(f), \ldots, X_n(f)) \in \mathbb{X}^n \,,$$

and a mapping $\widehat{S}_n : \mathcal{F} \to \mathcal{G}$, such that:

a) $X_1(f) = x_1$, for some arbitrary $x_1 \in \mathbb{X}$

b) For all $1 \leq i < n$, $X_{i+1}(f)$ depends measurably on $\mathcal{I}_i(f)$, where $\mathcal{I}_i = ((X_1, Z_1), \ldots, (X_i, Z_i))$, and $Z_i(f) = f(X_i(f))$, $1 \leq i \leq n$.

c) There exists a measurable function ϕ_n such that $\widehat{S}_n = \phi_n \circ \mathcal{I}_n$.

The algorithm $\left(\underline{X}_n, \widehat{S}_n\right)$ describes a sequence of decisions, made from an increasing amount of information: for each $i = 1, \ldots, n-1$, the algorithm uses information $\mathcal{I}_i(f)$ to choose the next evaluation point $X_{i+1}(f)$. The estimator $\widehat{S}_n(f)$ of $S(f)$ is the terminal decision. We shall denote by $\mathcal{A}_n$ the class of all strategies $\underline{X}_n$ that query sequentially n evaluations of f and also define the subclass $\mathcal{A}_n^0 \subset \mathcal{A}_n$ of non-adaptive strategies, that is, the class of all strategies such that the X_is do not depend on f.

A classical approach to study the performance of a sequential strategy is to consider the worst error of estimation on some class of functions $\mathcal{F}$

$$\epsilon_{\text{worstcase}}(\underline{X}_n) := \sup_{f \in \mathcal{F}} L(S(f), \widehat{S}_n(f)) \,,$$

where L is a loss function. There are many results dealing with the problems of function approximation and optimization in the worst case setting. Two noticeable results concern convex and symmetric classes of bounded functions. For such classes, from a worst-case point of view, any strategy will behave similarly for the problem of global optimization and that of function approximation. Moreover the use of adaptive methods can not be justified by a worst case analysis (see, e.g., Novak, 1988, Propositions 1.3.2 and 1.3.3). These results, combined with the fact that most optimization algorithms are adaptive, lead to think that the worst-case setting may not be the most appropriate framework to assess the performance of a search algorithm in practice. Indeed, it would be also important, in practice, to know whether the loss $L(S(f), \widehat{S}_n(f))$ is close to, or on the contrary much smaller than $\epsilon_{\text{worstcase}}$, for "typical" functions $f \in \mathcal{F}$ not corresponding to worst cases. To address this question, a classical approach is to adopt a Bayesian point of view.

In this paper, we consider methods where f is seen as a sample path of a real-valued random process ξ defined on some probability space $(\Omega, \mathcal{B}, \mathsf{P}_0)$ with parameter in $\mathbb{X}$. Then, $\underline{X}_n(\xi)$ is a random

sequence in $\mathbb{X}$, with the property that $X_{n+1}(\xi)$ is measurable with respect to the σ-algebra generated by $\xi(X_1(\xi))$, ..., $\xi(X_n(\xi))$. From a Bayesian decision-theoretic point of view, the random process represents prior knowledge about f and makes it possible to infer a quantity of interest before evaluating the function. This point of view has been widely explored in the domain of optimization and computer experiments. Under this setting, the performance of a given strategy $\underline{X}_n$ can be assessed by studying the average loss

$$\epsilon_{\text{average}}(\underline{X}_n) := \mathsf{E}\, L(S(\xi), \widehat{S}_n(\xi))\,.$$

How much does adaption help on the average, and is it possible to derive rates of decay for errors in average? In this article, we shall make a brief review of results concerning average error bounds of Bayesian search methods based on a random process prior.

This article has three parts. The precise assumptions about ξ are given in Section 2. Section 3 deals with the problem of function approximation, while Section 4 deals with the problem of optimization.

2 Framework

Let ξ be a random process defined on a probability space $(\Omega, \mathcal{B}, \mathsf{P}_0)$, with parameter $x \in \mathbb{R}^d$. Assume moreover that ξ has a zero mean and a continuous covariance function. The kriging predictor of $\xi(x)$, based on the observations $\xi(X_i(\xi))$, $i = 1, \ldots, n$, is the orthogonal projection

$$\widehat{\xi}_n(x) \;:=\; \sum_{i=1}^{n} \lambda^i(x; \underline{X}_n(\xi))\, \xi(X_i(\xi)) \tag{1}$$

of $\xi(x)$ onto $\text{span}\{\xi(X_i(\xi)), i = 1, \ldots, n\}$ in $L^2(\Omega, \mathcal{B}, \mathsf{P}_0)$. At step $n \geq 1$, given evaluation points $\underline{X}_n(\xi)$, the kriging coefficients $\lambda^i(x; \underline{X}_n(\xi))$ can be obtained by solving a system of linear equations (see, e.g., Chilès and Delfiner, 1999). Note that for any sample path $f = \xi(\omega, \cdot)$, $\omega \in \Omega$, the value $\widehat{\xi}_n(\omega, x)$ is a function of $\mathcal{I}_n(f)$ only.

The mean-square error (MSE) of estimation at a fixed point $x \in \mathbb{R}^d$ will be denoted by

$$\sigma_n^2(x) := \mathsf{E}\{(\xi(x) - \widehat{\xi}(x; \underline{X}_n(\xi)))^2\}\,.$$

It is generally not possible to compute $\sigma_n^2(x)$ when $\underline{X}_n$ is an adaptive strategy.

Regularity assumptions. Assume that there exists $\Phi : \mathbb{R}^d \to \mathbb{R}$ such that $k(x, y) = \Phi(x - y)$, which is in $L^2(\mathbb{R}^d)$ and has a Fourier transform

$$\tilde{\Phi}(u) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \Phi(x) e^{i(x,u)} dx$$

that satisfies

$$c_1(1 + \|u\|_2^2)^{-s} \leq \tilde{\Phi}(u) \leq c_2(1 + \|u\|_2^2)^{-s}\,, \quad u \in \mathbb{R}^d\,, \tag{2}$$

with $s > d/2$ and constants $0 < c_1 \leq c_2$. Note that the Matérn covariance with regularity parameter ν (see, e.g., Stein, 1999) satisfies such a regularity assumption, with $s = \nu + d/2$. Tensor-product covariance functions, however, never satisfy such a condition (see Ritter, 2000, chapter 7, for some results in this case).

Let $\mathcal{H}$ be the RKHS of functions generated by k. Denote by $(\cdot\,, \cdot)_{\mathcal{H}}$ the inner product of $\mathcal{H}$, and by $\|\cdot\|_{\mathcal{H}}$ the corresponding norm. It is well known (see, e.g. Wendland, 2005) that $\mathcal{H}$ is the Sobolev space

$$W_2^s(\mathbb{R}^d) = \left\{ f \in L^2(\mathbb{R}^d);\ \tilde{f}(\cdot)(1 + \|\cdot\|_2^2)^{s/2} \in L^2(\mathbb{R}^d) \right\}$$

due to the following result.

Proposition 1. *$\mathcal{H} \subset L^2(\mathbb{R}^d)$ and $\forall f \in \mathcal{H}$,*

$$\|f\|_{\mathcal{H}}^2 = \int_{\mathbb{R}^d} |\tilde{f}(u)|^2 \, \tilde{\Phi}(u)^{-1} \, du \, .$$

$\|f\|_{\mathcal{H}}^2$ is equivalent to the Sobolev norm

$$\|f\|_{W_2^s(\mathbb{R}^d)}^2 = \|\tilde{f}(\,\cdot\,) \left(1 + \|\cdot\|_2^2\right)^{s/2} \|_{L^2(\mathbb{R}^d)}$$

3 Approximation

We first consider the problem of approximation, with the point of view exposed in Section 2. Using the notations introduced above, the problem of approximation corresponds to considering operators S and $\widehat{S}_n$ defined by $S(\xi) := \xi_{|\mathbb{X}}$ and $\widehat{S}_n(\xi) := \widehat{\xi}_{n\,|\mathbb{X}}$, with $\mathbb{X} \subset \mathbb{R}^d$ a compact domain with non-empty interior. For the design of computer experiments, classical criteria for assessing the quality of a strategy $\underline{X}_n \in \mathcal{A}_n$ for the approximation problem are the maximum mean-square error (MMSE)

$$\epsilon_{\text{MMSE}}(\underline{X}_n) \; := \; \sup_{x \in \mathbb{X}} \mathsf{E}\left(\left(\xi(x) - \widehat{\xi}_n(x)\right)^2 \right) \; = \; \sup_{x \in \mathbb{X}} \sigma_n^2(x)$$

and the integrated mean-square error (IMSE)

$$\epsilon_{\text{IMSE}}(\underline{X}_n) \; := \; \mathsf{E}\left(\|\xi - \widehat{\xi}_n\|_{L^2(\mathbb{X},\mu)}^2 \right) \; = \; \int_{\mathbb{X}} \sigma_n(x)^2 \, \mu(\mathrm{d}x)$$

(see, e.g., Sacks et al., 1989; Currin et al., 1991; Welch et al., 1992; Santner et al., 2003). These criteria correspond to G-optimality and I-optimality in the theory of (parametric) optimal design.

As mentioned earlier, computing $\sigma_n^2(x)$ is usually not possible in the case of adaptive sampling strategies, even for a Gaussian process. From a theoretical point of view, however, it is important to know if adaptive strategies can improve upon non-adaptive strategies for the approximation problem.

Proposition 2. *Assume that ξ is a Gaussian process. Then adaptivity does not help for the approximation problem, with respect to either the MMSE or the IMSE criterion.*

Proof. For any adaptive strategy $\underline{X}_n$, it can be proved by induction (using the fact that X_{i+1} only depends on $\mathcal{I}_i$) that, for each $x \in \mathbb{X}$,

$$\sigma_n^2(x) \; = \; \mathsf{E}\left(\sigma^2(x; X_1(\xi), \ldots, X_n(\xi)) \right) \, , \tag{3}$$

where $\sigma^2(x; x_1, \ldots, x_n)$, $x_1, \ldots, x_n \in \mathbb{X}$, denotes the MSE at x of the non-adaptive strategy that selects the points $x_1, \ldots, x_n$. Therefore, for each $x \in \mathbb{X}$,

$$\sigma_n^2(x) \; \geq \; \min_{x_1, \ldots, x_n \in \mathbb{X}} \sigma^2(x; x_1, \, \ldots, \, x_n) \, ,$$

which proves the claim in the case of the MMSE criterion. Similarly, integrating (3) yields

$$\begin{aligned} \int_{\mathbb{X}} \sigma_n^2 \, d\mu \; &= \; \mathsf{E}\left\{ \int_{\mathbb{X}} \sigma^2(x; \underline{X}_n(\xi)) \, \mu(\mathrm{d}x) \right\} \\ &\geq \; \min_{x_1, \ldots, x_n \in \mathbb{X}} \int_{\mathbb{X}} \sigma^2(x; \, x_1, \, \ldots, \, x_n) \, \mu(\mathrm{d}x) \, , \end{aligned}$$

which proves the claim in the case of the IMSE criterion. □

In the case of the IMSE criterion, Proposition 2 can be seen as a special case of a general result about linear problems (see, e.g., Ritter, 2000, Chapter 7). The following proposition establishes a connection between the MMSE criterion and the worst-case L^∞-error of approximation in the unit ball of $\mathcal{H}$, which will be useful to establish the optimal rate for IMSE- and MMSE-optimal designs.

Proposition 3. *Let $\mathcal{H}_1$ denote the unit ball of $\mathcal{H}$. For any non-adaptive strategy $\underline{X}_n \in \mathcal{A}_n^0$, the MMSE criterion equals the squared worst-case L^∞-error of approximation in $\mathcal{H}_1$ using $\widehat{S}_n$:*

$$\epsilon_{\text{MMSE}}(\underline{X}_n) = \left(\sup_{f \in \mathcal{H}_1} \|S(f) - \widehat{S}_n(f)\|_{L^\infty(\mathbb{X})} \right)^2 .$$

Proof. Let $\underline{X}_n \in \mathcal{A}_n^0$ be a non-adaptive strategy such that $X_i(\xi) = x_i$, $i = 1, \ldots, n$, for some arbitrary x_is in $\mathbb{X}$. Denote by $\lambda_i(x) = \lambda_i(x; \underline{X}_n(\xi))$ the corresponding kriging coefficients (which do not depend on ξ). Using the fact that the mapping $\xi(x) \mapsto k(x, \cdot)$ extends linearly to an isometry from $\overline{\text{span}}\{\xi(y),\, y \in \mathbb{R}^d\}$ to $\mathcal{H}$, we have for all $x \in \mathbb{X}$

$$\begin{aligned}
\sigma_n(x) &= \left\|\xi(x) - \widehat{\xi}_n(x)\right\|_{L^2(\Omega, \mathcal{B}, \mathrm{P}_0)} \\
&= \left\|k(x, \cdot) - \textstyle\sum_i \lambda^i(x)\, k(x_i, \cdot)\right\|_{\mathcal{H}} \\
&= \sup_{f \in \mathcal{H}_1} \left(f\,,\, k(x, \cdot) - \textstyle\sum_i \lambda^i(x)\, k(x_i, \cdot) \right)_{\mathcal{H}} . \\
&= \sup_{f \in \mathcal{H}_1} (f - \widehat{S}_n f)(x) .
\end{aligned}$$

Thus,

$$\sup_{x \in \mathbb{X}} \sigma_n(x) \;=\; \sup_{f \in \mathcal{H}_1} \sup_{x \in \mathbb{X}} (f - \widehat{S}_n f)(x) \;=\; \sup_{f \in \mathcal{H}_1} \left\| f - \widehat{S}_n f \right\|_{L^\infty(\mathbb{X})} .$$

□

The following proposition summarizes known results concerning the optimal rate of decay in the class of non-adaptive strategies for both the IMSE criterion and the MMSE criterion. Note that, by Proposition 2, this rate is also the optimal rate of decay in the class of all adaptive strategies if ξ is a Gaussian process.

Proposition 4. *Assume that ξ has a continuous covariance function satisfying the regularity assumptions of Section 2, and let $\nu = s - d/2 > 0$. Then there exists $C_1 > 0$ such that, for any $\underline{X}_n \in \mathcal{A}_n^0$,*

$$C_1\, n^{-2\nu/d} \;\le\; \epsilon_{\text{IMSE}}(\underline{X}_n) \;\le\; \mu(\mathbb{X})\, \epsilon_{\text{MMSE}}(\underline{X}_n) \tag{4}$$

Moreover, if $\mathbb{X}$ has a Lipschitz boundary and satisfies an interior cone condition, then there exists $C_2 > 0$ such that

$$\inf_{\underline{X}_n \in \mathcal{A}_n^0} \epsilon_{\text{IMSE}}(\underline{X}_n) \;\le\; \mu(\mathbb{X}) \inf_{\underline{X}_n \in \mathcal{A}_n^0} \epsilon_{\text{MMSE}}(\underline{X}_n) \;\le\; C_2\, n^{-2\nu/d} . \tag{5}$$

The optimal rate of decay is therefore $n^{-2\nu/d}$ for both criteria.

Proof. It is proved in (Ritter, 2000, Chapter 7, Proposition 8) that there exists $C_1 > 0$ such that $\epsilon_{\text{IMSE}}(\underline{X}_n) \;\ge\; C_1\, n^{-2\nu/d}$ in the case where $\mathbb{X} = [0; 1]^d$. This readily proves the lower bound (4) since any $\mathbb{X}$ with non-empty interior contains an hypercube on which Ritter's result holds.

If $\mathbb{X}$ is a bounded Lipschitz domain satisfying an interior cone condition, then (Narcowich et al., 2005, Proposition 3.2) there exists $c_1 > 0$ such that $\|S(f) - \widehat{S}_n(f)\|_{L^\infty(\mathbb{X})} \le c_1 h_n^{s-d/2} \|S(f)\|_{W_2^s(\mathbb{X})}$ for

all $f \in \mathcal{H}$, where $h_n = \sup_{x\in\mathbb{X}} \min_{i\in\{1,\ldots,n\}} \|x - X_i(f)\|_2$ is the fill distance of the non-adaptive strategy $\underline{X}_n$ in $\mathbb{X}$. Therefore

$$\|S(f) - \widehat{S}_n(f)\|_{L^\infty(\mathbb{X})} \;\le\; c_1 h_n^\nu \, \|S(f)\|_{W_2^s(\mathbb{X})} \;\le\; c_1 h_n^\nu \, \|f\|_{W_2^s(\mathbb{R}^d)} \;\le\; c_2 h_n^\nu \, \|f\|_{\mathcal{H}}$$

for some $c_2 > 0$, using the equivalence of the Sobolev $W_2^s(\mathbb{R}^d)$ norm with the RKHS norm (see Section 2). Considering any non-adaptive space-filling strategy $\underline{X}_n$ with a fill distance $h_n = O(n^{-1/d})$ yields

$$\inf_{\underline{X}_n \in \mathcal{A}_n^0} \sup_{f\in\mathcal{H}_1} \left\| f - \widehat{S}_n f \right\|_{L^\infty(\mathbb{X})} \;\le\; c_3 \, n^{-\nu/d}$$

for some $c_3 > 0$ and the upper-bound (5) then follows from Proposition 3. □

Finding a non-adaptive MMSE-optimal design is a difficult non-convex optimization problem in nd dimensions. Instead of addressing directly such a high-dimensional global optimization problem, we can use the classical sequential non-adaptive greedy strategy $\underline{X}_n(\,\cdot\,) = (x_1, \ldots, x_n) \in \mathbb{X}^n$ defined by

$$x_{i+1} \;=\; \operatorname*{argmax}_{x\in\mathbb{X}} \sigma^2\,(x; x_1, \ldots, x_i)\;, \quad 1 \le i < n\,. \tag{6}$$

Of course, the strategy is suboptimal but it only involves simpler optimization problems in d dimensions and has the advantage that it can be stopped at any time. Following Binev et al. (2010), it can be established that this greedy strategy is rate optimal.

Proposition 5. *Assume that ξ has a continuous covariance function satisfying the regularity assumptions of Section 2, and let $\nu = s - d/2 > 0$. Let $\underline{X}_n$ be the sequential strategy defined by (6). Then,*

$$\epsilon_{\text{MMSE}}(\underline{X}_n) = \mathrm{O}(n^{2\nu/d})\,.$$

Proof. Theorem 3.1 in Binev et al. (2010), applied to the compact subset $\{\xi(x), x \in \mathbb{X}\}$ in $L^2(\Omega, \mathcal{B}, \mathsf{P}_0)$, states that the greedy algorithm (6) preserves polynomial rates of decay. The result follows from Proposition 4. □

4 Optimization

In this section, we consider the problem of global optimization on a compact domain $\mathbb{X} \subset \mathbb{R}^d$, which corresponds formally to operators S and $\widehat{S}_n$ defined by $S(\xi) = \sup_{x\in\mathbb{X}} \xi(x)$ and $\widehat{S}_n(\xi) = \max_{i\in 1,\ldots,n} \xi(X_i(\xi))$.

In a Bayesian setting, a classical criterion to assess the performance of an optimization procedure is the average error

$$\epsilon_{\overline{\text{OPT}}}(\underline{X}_n) := \mathsf{E}(S(\xi) - \widehat{S}_n(\xi))\,.$$

Although it may be not possible in the context of this article to make a comprehensive review of known results concerning the average case in the Gaussian case, it can be safely said however that such results are scarce and specific.

In fact, most available results about the average-case error concern the one-dimensional Wiener process ξ on the interval $[0, 1]$. Under this setting, Ritter (1990) shows that the average error of the best non-adaptive optimization procedure decreases at rate $n^{-1/2}$ (extensions of this result for

non-adaptive algorithms and the r-fold Wiener measure can be found in Wasilkowski, 1992). Under the same assumptions for ξ, Calvin (1997) derives the exact limiting distribution of the error of a particular adaptive algorithm, which suggests that adaptivity does yield a better average error for the optimization problem—the result is that, for any $0 < \delta < 1$, it is possible to find an adaptive strategy such that $n^{(1-\delta)}(S(\xi) - \widehat{S}_n(\xi))$ converges in distribution.

A theoretical result concerning the optimal average-error criterion for less restrictive Gaussian priors is also available. If the covariance of a Gaussian process ξ is α-Hölder continuous, then Grünewälder et al. (2010) show that a space filling strategy $\underline{X}_n$ achieves

$$\epsilon_{\overline{\mathrm{OPT}}}(\underline{X}_n) = \mathrm{O}(n^{-\alpha/(2d)}(\log n)^{1/2})\,. \tag{7}$$

Thus, under the assumptions of Section 2, for a Matérn covariance with regularity parameter ν, the rate of the optimal average error of estimation of the optimum is less than $n^{-\nu/d}(\log n)^{1/2}$ (since a Matérn covariance is α-Hölder continuous with $\alpha = 2\nu$). Note that this bound is not sharp in general since the optimal non-adaptive rate is $n^{-1/2}$ for the Brownian motion on $[0; 1]$, the covariance function of which is α-Hölder continuous with $\alpha = 1$.

In view of these results, we can safely say that characterizing the average behavior of adaptive sequential optimization algorithms is still an open (and apparently difficult) problem. At present, the only way to draw useful conclusions about the interest of a particular optimization algorithm is to resort to numerical simulations.

In the following paragraphs, we shall illustrate the kind of results that can be expected from such empirical studies. Benassi et al. (2011) provide an empirical comparison between four optimization algorithms. The first algorithm is a non-adaptive space-filling strategy. The second algorithm assumes a Gaussian prior about the objective function and use the expected improvement (EI) sampling criterion (Mockus et al., 1978) for choosing the evaluation points. In practice however, it is often difficult to choose a Gaussian prior before any evaluation is made. As a result, the covariance function of ξ is usually chosen in some parametric class of positive definite functions, the value of the parameters assumed to be unknown. The third algorithm compared in Benassi et al. (2011) is a fully Bayesian algorithm (FBA), which is used to deal with uncertain parameters of the covariance of ξ. The fourth strategy is the popular efficient global optimization (EGO) algorithm introduced by Jones et al. (1998), which assumes a Gaussian process prior and takes a plug-in approach to deal with the uncertain parameters of the covariance.

In order to compare the four optimization strategies, Benassi et al. (2011) build several testbeds $\mathcal{T}_k$, $k = 1, 2, \ldots$, of functions $f_{k,l}$, $l = 1, \ldots, L$, corresponding to sample paths of a Gaussian process, with zero-mean and a Matérn covariance function, simulated on a set of $q = 600$ points in $[0, 1]^d$ generated using a Latin hypercube sampling (LHS), with different values for d and for the parameters of the covariance. Here, we present the results obtained for two testbeds in dimension 1 and 4 (the actual parameters are provided in Table 1).

Figures 1 and 2 show the average errors and also the distributions of the error of estimation of the global optimum. These empirical results show that the EI strategy performs much better in average than the space-filling strategy. Large errors are also less frequent with the EI strategy. Moreover, we can also assess the cost of estimating the parameters of the covariance. EGO and FBA have very similar average performances. In fact, both of them perform almost as well, in this experiments, as the EI strategy, where the true parameters are assumed to be known. Comparing the tails of complementary cumulative distribution function of the error $Sf - \widehat{S}_n f$ shows, however, that using a fully Bayesian approach brings a reduction of the occurrence of large errors with respect to the EGO algorithm. In other words, the fully Bayesian approach appears to be statistically more robust than the plug-in approach, while retaining the same average performance. Empirical studies such as the

Parameter \ Testbed	$\mathcal{T}_1$	$\mathcal{T}_2$
Dimension d	1	4
Number of sample paths L	20000	20000
Variance σ^2	1.0	1.0
Regularity ν	2.5	2.5
Scale $\beta = (\beta_1, \ldots, \beta_d)$	0.1	(0.7, 0.7, 0.7, 0.7)

Table 1: Parameters used for building the testbeds of Gaussian-process sample-paths. The Gaussian process has a zero-mean and a isotropic Matérn covariance function $k_{[\nu,\sigma^2,\rho]} : (x, y) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto \sigma^2 \kappa_\nu(\|x - y\|/\rho)$ with $\kappa_\nu(h) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(2\nu^{1/2}h\right)^\nu \mathcal{K}_\nu \left(2\nu^{1/2}h\right)$, $h \in \mathbb{R}$, where Γ is the Gamma function, $\mathcal{K}_\nu$ is the modified Bessel function of the second kind, and ν, σ^2 ρ are strictly positive scalar parameters (see Stein, 1999).

one presented here are therefore very useful from a practical point of view, since they make it possible to obtain fine and sound performance assessments of any strategy with a reasonable computational cost.

References

R. Benassi, J. Bect, and E. Vazquez. Robust gaussian process-based global optimization using a fully bayesian expected improvement criterion. In *Proceedings of fifth Learning and Intelligent Optimization Conference (LION 5)*, Rome, 2011.

P. Binev, A. Cohen, W. Dahmen, R. DeVore, G; Petrova, and P. Wojtaszczyk. *Convergence Rates for Greedy Algorithms in Reduced Basis Methods*, volume IGPM Report 310. RWTH Aachen, 2010.

J.M. Calvin. Average performance of a class of adaptive algorithms for global optimization. *The Annals of Applied Probability*, 7(3):711–730, 1997.

J.-P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York, 1999.

C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Amer. Statist. Assoc.*, pages 953–963, 1991.

S. Grünewälder, J.Y. Audibert, M. Opper, and J. Shawe-Taylor. Regret bounds for gaussian process bandit problems. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

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

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*, volume 2, pages 117–129, North Holland, New York, 1978.

F. J. Narcowich, J. D. Ward, and H. Wendland. Sololev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. *Math. Comp.*, 74:743–763, 2005.

E. Novak. *Deterministic and stochastic error bounds in numerical analysis*, volume 1349 of *Lecture Notes in Mathematics*. Springer-Verlag, 1988.

K. Ritter. Approximation and optimization on the wiener space. *Journal of Complexity*, 6(4):337–364, 1990.

K. Ritter. *Average-case analysis of numerical problems*, volume 1733 of *Lecture Notes in Mathematics*. Springer Verlag, 2000.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statist. Sci.*, 4(4):409–435, 1989.

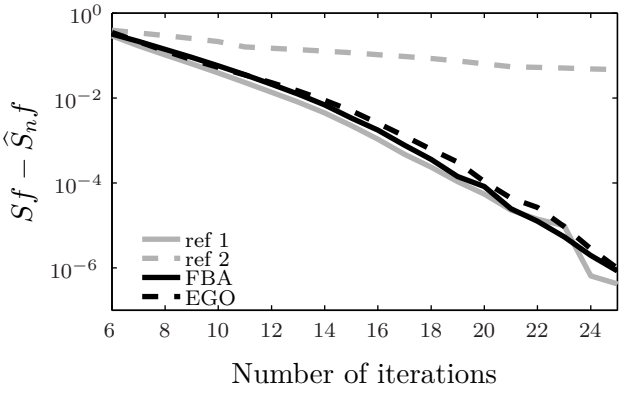
T. J. Santner, B. J. Williams, and W. I. Notz. *The Design and Analysis of Computer Experiments*. Springer, 2003.

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

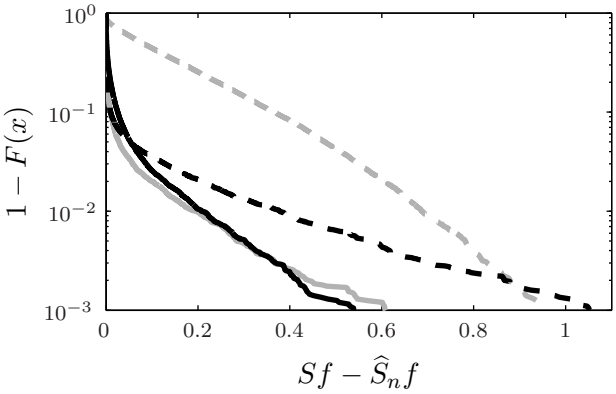
G.W. Wasilkowski. On average complexity of global optimization problems. *Mathematical programming*, 57(1):313–324, 1992.

W. J. Welch, R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris. Screening, predicting, and computer experiments. *Technometrics*, 34(1):15–25, 1992.

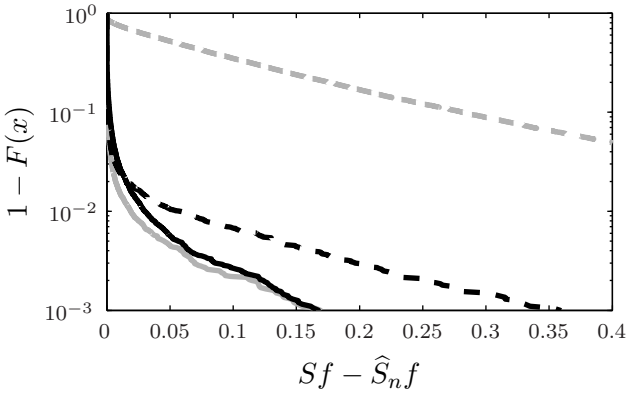
H. Wendland. *Scattered Data Approximation*. Monographs on Applied and Computational Mathematics. Cambridge Univ. Press, Cambridge, 2005.



(a) Average error to the maximum

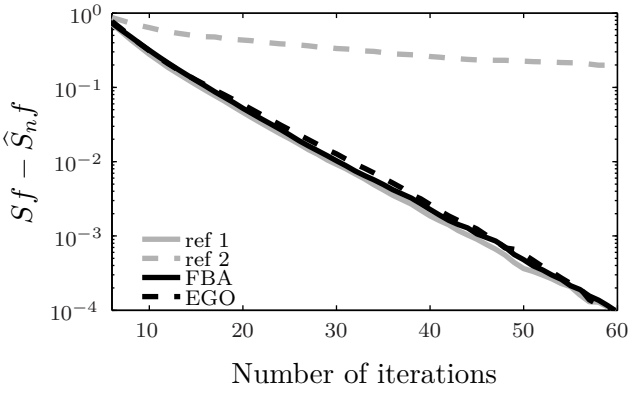


(b) Distribution of errors at iteration 13

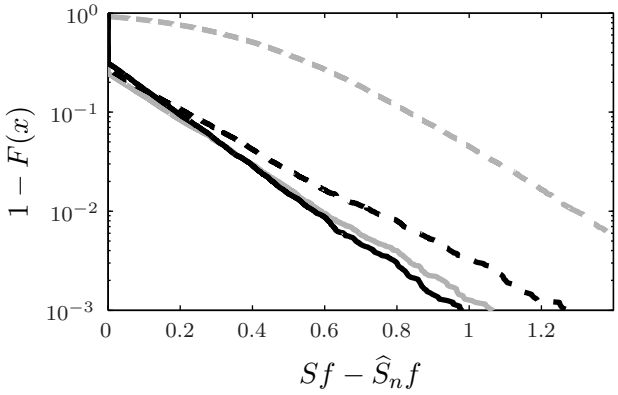


(c) Distribution of errors at iteration 16

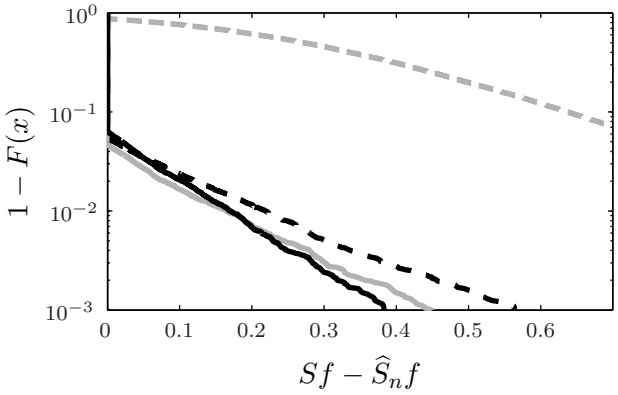
Figure 1: Average results and error distributions for testbed $\mathcal{T}_1$, for FBA (solid black line), EGO (dashed black line), the EI with the parameters used to generate sample paths (solid gray line), the space-filling strategy (dashed gray line). More precisely, (a) represents the average approximation error as a function of the number of evaluation points. In (b) and (c), $F(x)$ stands for the cumulative distribution function of the approximation error. We plot $1-F(x)$ in logarithmic scale in order to analyze the behavior of the tail of the distribution (big errors with small probabilities of occurrence). Small values for $1 - F(x)$ mean better results.



(a) Average error to the maximum



(b) Distribution of errors at iteration 20



(c) Distribution of errors at iteration 34

Figure 2: Average results and distribution of errors for testbed $\mathcal{T}_2$. See Figure 1 for details.