

Convergence properties of the expected improvement algorithm with fixed mean and covariance functions

Emmanuel Vazquez and Julien Bect

SUPELEC, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France
email: {firstname}.{lastname}@supelec.fr

Abstract

This paper deals with the convergence of the expected improvement algorithm, a popular global optimization algorithm based on a Gaussian process model of the function to be optimized. The first result is that under some mild hypotheses on the covariance function k of the Gaussian process, the expected improvement algorithm produces a dense sequence of evaluation points in the search domain, when the function to be optimized is in the reproducing kernel Hilbert space generated by k . The second result states that the density property also holds for \mathbb{P} -almost all continuous functions, where \mathbb{P} is the (prior) probability distribution induced by the Gaussian process.

Key words: Bayesian optimization, computer experiments, Gaussian process, global optimization, sequential design, RKHS

62M20; 62K20; 60G15; 60G25; 46E22; 90C99

1. Introduction

Global optimization is the task of finding the global optima of a real valued function using the results of some pointwise evaluations, which can be chosen sequentially, or in batches, when parallelization is possible. The function to be optimized is generally called *objective function*. In the field of design and analysis of computer experiments, as pioneered by Sacks et al. (1989) and Currin et al. (1991), the objective function—typically an expensive-to-evaluate numerical model of some physical phenomenon—is seen as a sample path of a stochastic process. The stochastic model captures prior knowledge about the objective function and makes it possible to infer the position of the global optima before evaluating the function. This Bayesian decision-theoretic point of view has been largely explored during the 70's and the 80's by the Vilnius school of global optimization led by J. Mockus (see Mockus et al., 1978; Mockus, 1989; Törn and Zilinskas, 1989; Zilinskas, 1992, and references therein).

In this paper, we consider the *expected improvement* (EI) algorithm, a popular optimization algorithm proposed by J. Mockus in the 70's and brought to the field of computer experiments by D.R. Jones, M. Schonlau and W.J. Welch (Schonlau and Welch, 1996; Schonlau, 1997; Schonlau et al., 1997; Jones et al., 1998). Let \mathbb{X} be a compact subset of \mathbb{R}^d , $d \geq 1$, and let ξ be a real valued Gaussian process with parameter $x \in \mathbb{X}$. Our goal is to maximize a given objective function, which is assumed to be a sample path of ξ . The EI algorithm is a sequential planning strategy that constructs a sequence $(x_n)_{n \in \mathbb{N}} \in \mathbb{X}^{\mathbb{N}}$ in such a way that each evaluation point x_n is a function of the previous evaluation points x_i , $i < n$, and the corresponding values of the objective function. Let $M_n = \xi(x_1) \vee \dots \vee \xi(x_n)$ be the observed maximum at step n ; then, a new evaluation point x_{n+1} is chosen in order to maximize the quantity

$$\rho_n(x) := \mathbb{E} [(\xi(x) - M_n)_+ \mid \xi(x_1), \dots, \xi(x_n)] \quad (1)$$

where $z_+ = z \vee 0$. Note that this is equivalent to choosing the evaluation point x_{n+1} that maximizes $\mathbb{E} [M_n \vee \xi(x) \mid \xi(x_1), \dots, \xi(x_n)]$ with respect to x . The function $\rho_n(x)$, which is called the expected improvement at x , is the conditional mean excess of $\xi(x)$ above the current maximum M_n . It is well-known that the expected improvement has a closed-form expression, which can be written using the kriging predictor and its variance (see, e.g., Jones et al., 1998).

This paper addresses the convergence of the EI algorithm, under the assumption that ξ is a Gaussian process with zero mean and known covariance. (Our results still apply if some parameters of the covariance function—for instance,

the range and regularity parameters of a Matérn covariance function—are estimated using a first batch of evaluations and held fixed afterward.) It is easily seen that a global optimization algorithm converges for all continuous functions if and only if the sequence of evaluation points produced by the algorithm is dense for all continuous functions (Törn and Zilinskas, 1989, Theorem 1.3). In the case of the EI algorithm, this property was proved by Locatelli (1997), with $d = 1$, $\mathbb{X} = [0, 1]$ and ξ a Brownian motion. Mockus (1989, Section 4.2) claims a much more general convergence result, but his proof unfortunately contains a severe technical gap¹.

The main contribution of this paper is a couple of convergence results for the EI algorithm. The first result (Theorem 6) states that the sequence of evaluation points is dense in the search domain provided that the objective function belongs to the reproducing kernel Hilbert space \mathcal{H} attached to ξ , under a non-degeneracy assumption on the covariance function that we call the No-Empty-Ball (NEB) property. This convergence result is quite natural from the point of view of interpolation theory. The second result (Theorem 7) states that the density property also holds for \mathbb{P} -almost all continuous functions, where \mathbb{P} is the (prior) probability distribution of the Gaussian process ξ .

The paper is outlined as follows. Section 2 introduces our framework, notations and standing assumptions. Section 3 describes the EI algorithm in greater details and states the main results of the paper. Section 4 provides a sufficient condition for the NEB property, in the case of a stationary covariance function. Section 5 contains the proof of the main theorems. Finally, Section 6 gives our conclusions and discusses future work.

2. Preliminaries

2.1. Framework and standing assumptions

The central mathematical object in global optimization theory is the objective function $\omega : \mathbb{X} \rightarrow \mathbb{R}$, defined on some *search space* \mathbb{X} . A deterministic search strategy can therefore be seen as a mapping \underline{X} from the set $\Omega = \mathbb{R}^{\mathbb{X}}$ to the set $\mathbb{X}^{\mathbb{N}}$ of all sequences in \mathbb{X} ,

$$\underline{X}(\omega) := (X_1(\omega), X_2(\omega), \dots), \quad (2)$$

with the property that, for all $n \geq 1$, $X_{n+1}(\omega)$ depends only on the first n evaluations $\omega(X_1(\omega)), \dots, \omega(X_n(\omega))$. Assuming measurability of the X_n s with respect to the product σ -algebra \mathcal{A} on Ω (i.e. the σ -algebra generated by cylinder sets), this can be reformulated in the language of probability theory—although there is no probability measure involved yet. Indeed, let

$$\xi : \mathbb{X} \times \Omega \rightarrow \mathbb{R}, \quad (x, \omega) \mapsto \xi(x, \omega) := \omega(x), \quad (3)$$

denote the canonical process on the path space (Ω, \mathcal{A}) . Then, the above search strategy \underline{X} can be seen as a *random* sequence in \mathbb{X} , with the property that X_{n+1} is \mathcal{F}_n -measurable, where \mathcal{F}_n is the σ -algebra generated by $\xi(X_1), \dots, \xi(X_n)$. It must be stressed that, despite the lexical shift, we are still dealing with *deterministic* algorithms: randomness only comes from the fact that we are now considering the objective function $\xi(\cdot, \omega) = \omega$ as a random element in Ω .

In the Bayesian approach to global optimization, prior information on the objective function is taken into account under the form of a probability measure \mathbb{P} on (Ω, \mathcal{A}) , which amounts to specifying the probability distribution of the stochastic process ξ . This prior information is then updated at each step of the search, through the computation of the conditional distribution $\mathbb{P}\{\cdot | \mathcal{F}_n\}$. For practical reasons, only Gaussian process priors have been considered in the literature: in this case, the prior is completely specified by the mean $m(x)$ and the covariance function $k(x, x')$, and the process ξ remains Gaussian under the conditional distributions $\mathbb{P}\{\cdot | \mathcal{F}_n\}$, $n \geq 1$. Throughout the paper we shall make the following standing assumptions:

Assumption 1.

- i) \mathbb{X} is a compact subset of \mathbb{R}^d , for some $d \geq 1$,
- ii) ξ is a centered Gaussian process under \mathbb{P} ,
- iii) the covariance function k is continuous and positive definite.

¹More precisely, the arguments given on page 45 fail to prove the key result claimed in Lemma 4.2.2, i.e. the density of the sequence of evaluation points.

Let $\mathcal{H} \subset \Omega$ denote the Reproducing Kernel Hilbert Space (RKHS) that is canonically attached to ξ (also known as the Cameron-Martin space of ξ ; see, e.g., Bogachev, 1998). Assumption 1.iii entails that \mathcal{H} is a space of continuous functions. We shall denote by $(\cdot, \cdot)_{\mathcal{H}}$ the inner product of \mathcal{H} and by $\|\cdot\|_{\mathcal{H}}$ the corresponding norm. It is worth noting that $\mathbb{P}(\mathcal{H}) = 0$ (see, e.g., Lukic and Beder, 2001, Driscoll's theorem). We shall comment on this fact with respect to our convergence result in Section 3.

Remark 2. Unless otherwise specified (see Section 4), it is not assumed that the covariance k is stationary. To the best of our knowledge, however, most practical applications of the EI algorithm have used stationary covariances to model the objective function prior to any evaluation.

2.2. Linear prediction and the No-Empty-Ball property

For $n \geq 1$, $\underline{x}_n = (x_1, \dots, x_n) \in \mathbb{X}^n$ and $x \in \mathbb{X}$, we denote by $\widehat{\xi}_n(x; \underline{x}_n)$ the conditional expectation of $\xi(x)$ given $\xi(x_1), \xi(x_2), \dots, \xi(x_n)$. Since ξ is a centered Gaussian process, the conditional expectation is also the best linear predictor in $L^2(\Omega, \mathcal{A}, \mathbb{P})$, and therefore can be written as

$$\widehat{\xi}_n(x, \omega; \underline{x}_n) = \sum_{i=1}^n \lambda_n^i(x; \underline{x}_n) \xi(x_i, \omega). \quad (4)$$

Let $\sigma_n^2(x; \underline{x}_n)$ denote the mean-square prediction error, i.e.

$$\sigma_n^2(x; \underline{x}_n) := \mathbb{E} \left[\left(\xi(x) - \widehat{\xi}_n(x; \underline{x}_n) \right)^2 \right]. \quad (5)$$

(Recall that, since ξ is a Gaussian process, the error of prediction is independent of the σ -algebra generated by the $\xi(x_i)$ s, $1 \leq i \leq n$; see, e.g., Chilès and Delfiner (1999), Section 3.3.4.)

Definition 3. We shall say that the Gaussian process ξ —or, equivalently, the covariance function k —has the No-Empty-Ball (NEB) property if, for all sequences $(x_n)_{n \geq 1}$ in \mathbb{X} and all $y \in \mathbb{X}$, the following assertions are equivalent:

i) y is an adherent point of the set $\{x_n, n \geq 1\}$,

ii) $\sigma_n^2(y; \underline{x}_n) \rightarrow 0$ when $n \rightarrow +\infty$.

Since k is assumed continuous, (i) always implies (iii) in Definition 3. The NEB property is therefore equivalent to the assertion that, if the prediction error at y goes to zero, then there can be no “empty ball” centered at y (i.e. for all $\epsilon > 0$, there exists $n \geq 1$ such that $|y - x_n| < \epsilon$)—hence its name. A sufficient condition for the NEB property will be given in Section 4. To the best of our knowledge, finding necessary and sufficient condition for the NEB property is an open problem.

2.3. Simplified notations

Since the notations introduced in (4) and (5) would rapidly become cumbersome in the next sections, the following simplified notations will be used:

$$\widehat{\xi}_n(x, \omega) := \widehat{\xi}_n(x, \omega; \underline{X}_n(\omega)), \quad (6)$$

$$\sigma_n^2(x, \omega) := \mathbb{E} \left[\left(\xi(x) - \widehat{\xi}_n(x; \underline{X}_n) \right)^2 \middle| \mathcal{F}_n \right] (\omega) = \sigma_n^2(x; \underline{X}_n(\omega)), \quad (7)$$

with $\underline{X}_n = (X_1, \dots, X_n)$. Remark that $\sigma_n^2(x, \omega)$ is a stochastic process indexed by \mathbb{X} . The second equality in (7) follows from the fact that $\widehat{\xi}_n(x; \cdot)$ is continuous for all $x \in \mathbb{X}$.

3. Main results

In this paper, we shall consider a generalization of the EI criterion. Define

$$\rho_n(x) = \gamma\left(\widehat{\xi}_n(x) - M_n, \sigma_n^2(x)\right), \quad (8)$$

where the function $\gamma : \mathbb{R} \times [0; +\infty) \rightarrow [0; +\infty)$ satisfies the following requirements:

$$\begin{aligned} R_1 : & \quad \gamma \text{ is continuous,} \\ R_2 : & \quad \forall z \leq 0, \gamma(z, 0) = 0, \\ R_3 : & \quad \forall z \in \mathbb{R}, \forall s > 0, \gamma(z, s) > 0. \end{aligned} \quad (9)$$

The corresponding optimization algorithm can then be written as:

$$\begin{cases} X_1 & = x_{\text{init}} \in \mathbb{X}, \\ X_{n+1} & = \arg \max_{x \in \mathbb{X}} \rho_n(x). \end{cases} \quad (10)$$

Remark 4. It is well-known (Schonlau and Welch, 1996) that the EI criterion defined by (1) can be rewritten under the form (8). More precisely, let Φ denote the Gaussian cumulative distribution function. Then (8) holds for the EI criterion with

$$\gamma(z, s) = \begin{cases} \sqrt{s} \Phi' \left(\frac{z}{\sqrt{s}} \right) + z \Phi \left(\frac{z}{\sqrt{s}} \right) & \text{if } s > 0, \\ \max(z, 0) & \text{if } s = 0. \end{cases} \quad (11)$$

In fact, equation (8) with γ thus defined should be taken as the true definition of the EI criterion. Indeed, the exact mathematical meaning of “ $\rho_n(x) := \mathbb{E}[(\xi(x) - M_n)_+ | \mathcal{F}_n]$ ” has to be specified, since, for each x , the conditional expectation is only defined up to a \mathbb{P} -negligible subset of Ω .

Remark 5. The criterion $x \mapsto \rho_n(x)$ is continuous, but there is no guarantee that the maximizer over \mathbb{X} will be unique. Therefore, a more rigorous statement of the iterative part of (10) would be: $X_{n+1} \in \arg \max_{x \in \mathbb{X}} \rho_n(x)$. In this way, instead of a single algorithm, we encapsulate the family of all algorithms that choose (measurably) X_{n+1} among the maximizers of ρ_n . General measurable selection theorems (see, e.g., Molchanov, 2005) ensure that such an algorithm does exist.

The first result of this paper is the following density theorem:

Theorem 6. Assume that the covariance function k has the NEB property. Then, for all $x_{\text{init}} \in \mathbb{X}$ and all $\omega \in \mathcal{H}$, the sequence $(X_n(\omega))_{n \geq 1}$ generated by (10) is dense in \mathbb{X} .

The fact that Theorem 6 is stated for objective functions in the RKHS \mathcal{H} calls for some comments. From the point of view of interpolation theory, it is indeed quite natural that an algorithm built on the best interpolants $\widehat{\xi}_n(\cdot, \omega)$ in an RKHS \mathcal{H} should be provably working, using the tools of RHKS theory, only when ω is in this very space. From the probabilistic point of view, however, the event $\{\xi(\cdot) \in \mathcal{H}\}$ almost never happens according to Driscoll’s theorem (Lukic and Beder, 2001). The second result of this paper states that the result of Theorem 6 also holds \mathbb{P} -almost surely in Ω .

Theorem 7. Assume that the covariance function k has the NEB property. Then, for all $x_{\text{init}} \in \mathbb{X}$, the sequence $(X_n)_{n \geq 1}$ generated by (10) is \mathbb{P} -almost surely dense in \mathbb{X} .

It is still an important open question to determine whether the algorithm converges for *all* continuous functions, as claimed in Mockus (1989). Another interesting open problem would be to determine whether the NEB assumption can be relaxed.

Remark 8. We have assumed for the sake of simplicity that the optimization algorithm starts after a single evaluation performed at $X_1 = x_{\text{init}}$. In practice, especially when some parameters of the covariance need to be estimated, the algorithm starts with an initial design of several evaluations $x_{\text{init}}^1, \dots, x_{\text{init}}^{n_0}$. This is equivalent to saying that \mathcal{F}_1 is the σ -algebra generated by $\xi(x_{\text{init}}^1), \dots, \xi(x_{\text{init}}^{n_0})$. The proofs of Theorem 6 and Theorem 7 carry over without modification.

4. A sufficient condition for the NEB property

4.1. Statement of the result

In this section we shall prove that the following assumption is a sufficient condition for the NEB property:

Assumption 9. *The process ξ is stationary and has spectral density S , with the property that S^{-1} has at most polynomial growth.*

In other words, Assumption 9 means that there exist $C > 0$ and $r \in \mathbb{N} \setminus \{0\}$ such that $S(u)(1 + |u|^r) \geq C$, almost everywhere on \mathbb{R}^d . This assumption prevents k from being *too regular*. In particular, the so-called *Gaussian* covariance,

$$k(x, y) = \sigma^2 e^{-\alpha \|x-y\|^2}, \quad \sigma > 0, \alpha > 0, \quad (12)$$

does not satisfy Assumption 9. However, we are still allowed to consider a large class of covariances. For instance, the exponential covariances

$$k(x, y) = \sigma^2 e^{-\alpha \|x-y\|^s}, \quad \sigma > 0, \alpha > 0, 0 < s < 2, \quad (13)$$

the class of Matérn covariances (see, e.g., Stein, 1999), and their anisotropic versions, all satisfy Assumption 9. The main result of this section is:

Proposition 10. *Let $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ be two sequences in \mathbb{X} . Assume that the sequence (y_n) is convergent, and denote by y^* its limit. Then each of the following conditions implies the next one:*

- i) y^* is an adherent point of the set $\{x_n, n \geq 1\}$,
- ii) $\sigma_n^2(y_n; \underline{x}_n) \rightarrow 0$ when $n \rightarrow \infty$,
- iii) $\widehat{\xi}_n(y_n, \omega; \underline{x}_n) \rightarrow \xi(y^*, \omega)$ when $n \rightarrow \infty$, for all $\omega \in \mathcal{H}$.

Moreover, under Assumption 9, the three conditions are equivalent and therefore ξ has the NEB property.

Remark 11. As already observed, the Gaussian covariance does not satisfy Assumption 9. In fact, it is known that the Gaussian covariance does not even have the NEB property (Vazquez and Bect, 2010).

4.2. Consequence of Assumption 9 in terms of RKHS

Let \mathcal{H}' denote the RKHS associated to k on \mathbb{R}^d . It is well-known (Aronszajn, 1950, Section 1.5) that \mathcal{H} embeds isometrically into \mathcal{H}' and that, for all $\omega \in \mathcal{H}'$, the orthogonal projection of ω onto \mathcal{H} is simply its restriction to \mathbb{X} .

Under Assumption 9, \mathcal{H}' contains the Sobolev space $H^{\frac{r}{2}}(\mathbb{R}^d)$, and the injection is continuous. Indeed, denoting by $\widehat{\omega}$ the Fourier transform of $\omega \in \mathcal{H}'$, we have

$$\int (1 + |u|^r) |\widehat{\omega}(u)|^2 du \geq C \int S(u)^{-1} |\widehat{\omega}(u)|^2 du = \|\omega\|_{\mathcal{H}'}^2.$$

A useful consequence is that \mathcal{H}' contains the space $C_c^\infty(\mathbb{R}^d)$ of all compactly supported infinitely differentiable functions on \mathbb{R}^d , for any r . In particular, k is a *universal kernel* on \mathbb{X} in the sense of Steinwart (2001), which means that \mathcal{H} is dense in the Banach space $C(\mathbb{X})$ of all continuous functions on \mathbb{X} .

4.3. Proof of Proposition 10

(i) \Rightarrow (ii). Assume that $y^* \notin \{x_n, n \geq 1\}$ (otherwise the result holds trivially). Let (x_{ϕ_k}) be a subsequence of (x_n) converging to y^* and let $\psi_n = \max\{\phi_k; \phi_k \leq n\}$. Then,

$$\sigma_n^2(y_n; \underline{x}_n) = \text{var}[\xi(y_n) - \widehat{\xi}_n(y_n; \underline{x}_n)] \leq \text{var}[\xi(y_n) - \xi(x_{\psi_n})].$$

Since $\psi_n \rightarrow \infty$, it follows from the continuity of k that

$$\text{var}[\xi(y_n) - \xi(x_{\psi_n})] = k(y_n, y_n) + k(x_{\psi_n}, x_{\psi_n}) - 2k(x_{\psi_n}, y_n) \rightarrow 0.$$

(ii) \Rightarrow (iii). Using the Cauchy-Schwarz inequality in \mathcal{H} , we have

$$\left| \xi(y_n, \omega) - \widehat{\xi}_n(y_n, \omega; \underline{x}_n) \right| \leq \sigma_n(y_n; \underline{x}_n) \|\omega\|_{\mathcal{H}}$$

Therefore

$$\begin{aligned} \left| \xi(y^*, \omega) - \widehat{\xi}_n(y_n, \omega; \underline{x}_n) \right| &\leq \left| \xi(y^*, \omega) - \xi(y_n, \omega) \right| + \left| \xi(y_n, \omega) - \widehat{\xi}_n(y_n; \underline{x}_n) \right| \\ &\leq \left| \omega(y^*) - \omega(y_n) \right| + \sigma_n(y_n; \underline{x}_n) \|\omega\|_{\mathcal{H}} \rightarrow 0, \end{aligned}$$

since ω is continuous.

Under Assumption 9, (iii) \Rightarrow (i). Suppose (i) is false. Then, there exists a neighborhood U of y^* in \mathbb{R}^d that does not intersect $\{x_n, n \geq 1\}$. Besides, it follows from Assumption 9 that there exists $\omega \in \mathcal{H}$ such that $\text{supp } \omega \subset U$ and $\omega(y^*) > 0$ (where $\text{supp } \omega$ denotes the support of ω). Then, $\widehat{\xi}_n(y^*, \omega; \underline{x}_n) = 0$ for all n , whereas $\xi(y^*, \omega) = \omega(y^*) \neq 0$. Therefore (iii) does not hold. \square

5. Proofs of the main theorems

5.1. Proof of Theorem 6

Let $v_n = \sup_{x \in \mathbb{X}} \rho_n(x)$, where ρ_n is the criterion defined by equation (8). Note that, for all $n \geq 1$,

$$v_n = \rho_n(X_{n+1}) = \gamma\left(\widehat{\xi}_n(X_{n+1}) - M_n, \sigma_n^2(X_{n+1})\right).$$

Our proof of Theorem 6 will be based on the following result (which does not require the NEB property):

Lemma 12. For all $\omega \in \mathcal{H}$, $\liminf_{n \rightarrow \infty} v_n(\omega) = 0$.

Proof. Fix $\omega \in \mathcal{H}$. For all $n \geq 1$, set $x_n = X_n(\omega)$, $s_n = \sigma_n^2(x_{n+1}, \omega)$ and $z_n = \widehat{\xi}_n(x_{n+1}, \omega) - M_n(\omega)$, so that $v_n(\omega) = \gamma(z_n, s_n)$. Let y^* be a cluster point of the sequence (x_n) and let (x_{ϕ_n}) be any subsequence converging to y^* : we are going to prove that $v_{\phi_{n-1}}(\omega) \rightarrow 0$. It follows from Proposition 10, (i) \Rightarrow (iii), that $\widehat{\xi}_{\phi_{n-1}}(x_{\phi_n}, \omega) \rightarrow \omega(y^*)$. Moreover, $(M_{\phi_{n-1}}(\omega))$ is a bounded increasing sequence, with the property that $M_{\phi_{n-1}}(\omega) \geq M_{\phi_{n-1}}(\omega) \geq \omega(x_{\phi_{n-1}}) \rightarrow \omega(y^*)$. Therefore $(z_{\phi_{n-1}})$ has a finite limit, such that

$$\lim_{n \rightarrow \infty} z_{\phi_{n-1}} = \lim_{n \rightarrow \infty} \widehat{\xi}_{\phi_{n-1}}(x_{\phi_n}, \omega) - \lim_{n \rightarrow \infty} M_{\phi_{n-1}}(\omega) \leq 0.$$

By Proposition 10, (i) \Rightarrow (ii), we also know that $s_{\phi_{n-1}} = \sigma_{\phi_{n-1}}^2(x_{\phi_n}, \omega) \rightarrow 0$. Therefore, using (R_1) and (R_2) ,

$$v_{\phi_{n-1}}(\omega) = \gamma(z_{\phi_{n-1}}, s_{\phi_{n-1}}) \rightarrow \gamma\left(\lim_{n \rightarrow \infty} z_{\phi_{n-1}}, 0\right) = 0.$$

This completes the proof of Lemma 12. \square

Proof of Theorem 6. Now fix $\omega \in \mathcal{H}$, and suppose that $\{X_n(\omega), n \geq 1\}$ is not dense in \mathbb{X} . Then there exist a point $y^* \in \mathbb{X}$ that is not adherent to $\{X_n(\omega), n \geq 1\}$. This implies, by the NEB property, that

$$\inf_{n \geq 1} \sigma_n^2(y^*, \omega) > 0.$$

Besides, using the Cauchy-Schwarz inequality in \mathcal{H} , we observe that the sequence $(\widehat{\xi}_n(y^*, \omega))$ is bounded. Indeed, we have

$$\left| \widehat{\xi}_n(y^*, \omega) - \omega(y^*) \right|^2 \leq \sigma_n^2(y^*, \omega) \|\omega\|_{\mathcal{H}}^2 \leq k(y^*, y^*) \|\omega\|_{\mathcal{H}}^2.$$

The sequence $(M_n(\omega))$ is also obviously bounded by $\|\omega\|_{\infty}$. Therefore, we obtain as a consequence of (R_1) and (R_3) that

$$\rho_n(y^*, \omega) \geq \inf_{k \geq 1} \gamma(\widehat{\xi}_k(y^*, \omega) - M_k(\omega), \sigma_k^2(y^*, \omega)) > 0.$$

This is a contradiction with Lemma 12, since $v_n(\omega) = \max_{x \in \mathbb{X}} \rho_n(x, \omega)$. The proof is thus complete. \square

5.2. Proof of Theorem 7

In essence, the structure of the proof of Theorem 7 is the same as that of Theorem 6. The first step is to obtain an almost sure version of Lemma 12.

Lemma 13. $\liminf_{n \rightarrow \infty} v_n = 0$ almost surely.

Proof. Let $D_n = \min_{1 \leq i \leq n} |X_{n+1} - X_i|$ be the distance of X_{n+1} to the set of all previous evaluation points. Define $T_k = \min\{n \geq 1; D_n \leq r_k\}$, with (r_k) a sequence of positive numbers such that $\lim r_k = 0$. Note that each T_k is finite, since the set \mathbb{X} is compact, and is an (\mathcal{F}_n) -stopping time since the sequence (D_n) is (\mathcal{F}_n) -adapted.

The first step is to see that, as in the proof of Proposition 10,

$$\sigma_{T_k}^2(X_{T_k+1}) \leq \eta_k := \sup_{|x-y| \leq r_k} k(x, x) + k(y, y) - 2k(x, y) \xrightarrow[k \rightarrow \infty]{} 0. \quad (14)$$

Note that $(X_{T_k+1})_k$ does not necessarily converge.

The next step is to prove that $\widehat{\xi}_{T_k}(X_{T_k+1}) - \xi(X_{T_k+1})$ converges to zero almost surely, for a suitable choice of the sequence (r_k) . First, using that T_k is a stopping time, we have:

$$\begin{aligned} \mathbb{E} \left[\left(\widehat{\xi}_{T_k}(X_{T_k+1}) - \xi(X_{T_k+1}) \right)^2 \right] &= \mathbb{E} \left[\sum_{n \geq 1} \mathbf{1}_{T_k=n} \left(\widehat{\xi}_n(X_{n+1}) - \xi(X_{n+1}) \right)^2 \right] \\ &= \sum_{n \geq 1} \mathbb{E} \left[\mathbf{1}_{T_k=n} \mathbb{E} \left[\left(\widehat{\xi}_n(X_{n+1}) - \xi(X_{n+1}) \right)^2 \mid \mathcal{F}_n \right] \right] \\ &= \mathbb{E} \left[\sum_{n \geq 1} \mathbf{1}_{T_k=n} \sigma_n^2(X_{n+1}) \right] \\ &= \mathbb{E} \left[\sigma_{T_k}^2(X_{T_k+1}) \right] \leq \eta_k. \end{aligned}$$

Then, for each $\varepsilon > 0$, it follows from Markov's inequality that

$$\mathbb{P} \left\{ \left(\widehat{\xi}_{T_k}(X_{T_k+1}) - \xi(X_{T_k+1}) \right)^2 > \varepsilon \right\} \leq \eta_k / \varepsilon.$$

Choosing r_k such that, for instance, $\eta_k = 1/k^2$, ensures that $\widehat{\xi}_{T_k}(X_{T_k+1}) - \xi(X_{T_k+1})$ converges to zero almost surely. Therefore, the sequence $(\widehat{\xi}_{T_k}(X_{T_k+1}))$ is almost surely bounded. Moreover,

$$\begin{aligned} \limsup_{k \rightarrow \infty} \widehat{\xi}_{T_k}(X_{T_k+1}) - M_{T_k} &= \limsup_{k \rightarrow \infty} \widehat{\xi}_{T_k}(X_{T_k+1}) - M_{T_k+1} \\ &\leq \lim_{k \rightarrow \infty} \widehat{\xi}_{T_k}(X_{T_k+1}) - \xi(X_{T_k+1}) = 0 \quad \text{a.s.}, \end{aligned} \quad (15)$$

where we have used the fact that (M_n) is convergent.

Finally, using (R_1) and (R_2) , the fact that $(\widehat{\xi}_{T_k}(X_{T_{k+1}}) - M_{T_k})$ is almost surely bounded, (14) and (15), we conclude that

$$\nu_{T_k} = \gamma(\widehat{\xi}_{T_k}(X_{T_{k+1}}) - M_{T_k}, \sigma_{T_k}^2(X_{T_{k+1}})) \xrightarrow[k \rightarrow \infty]{} 0 \quad \text{a.s.} \quad \square$$

Proof of Theorem 7. Fix $x \in \mathbb{X}$ and define the event $A_x \in \mathcal{A}$ by

$$A_x = \{x \text{ is not an adherent point of the set } \{X_n, n \geq 1\}\}.$$

Then $\inf_{n \geq 1} \sigma_n^2(x) > 0$ on A_x by the NEB property. Moreover, the martingale $\widehat{\xi}_n(x) = \mathbb{E}[\xi(x) | \mathcal{F}_n]$ is bounded in L^2 since $\mathbb{E} \widehat{\xi}_n(x)^2 \leq k(x, x) < +\infty$, and thus converges almost surely and in L^2 to a random variable $\widehat{\xi}_\infty(x)$ (see, e.g., Williams, 1991). As a consequence, the event

$$B_x := \{(\widehat{\xi}_n(x) - M_n) \text{ is bounded}\}$$

has probability one, since (M_n) is also convergent. Therefore, we obtain by (R_1) and (R_3) that, on $A_x \cap B_x$,

$$\nu_n \geq \rho_n(x) \geq \inf_{k \geq 1} \gamma(\widehat{\xi}_k(x) - M_k, \sigma_k^2(x)) > 0.$$

Since $\mathbb{P}(B_x) = 1$, it follows from Lemma 13 that $\mathbb{P}(A_x) = 0$.

Finally, let $\tilde{\mathbb{X}}$ be a countable dense subset of \mathbb{X} and let $\Omega_0 = \Omega \setminus \bigcup_{x \in \tilde{\mathbb{X}}} A_x$. Then $\mathbb{P}(\Omega_0) = 1$ and it is straightforward to see that for each $\omega \in \Omega_0$, the set $\{X_n(\omega), n \geq 1\}$ is dense in \mathbb{X} . \square

6. Discussion

Since Jones et al. (1998), the expected improvement (EI) algorithm has become a very popular algorithm to optimize an expensive-to-evaluate function. Such functions are often encountered in industrial problems, where the function value may be the output of a complex computer simulation, or the result of costly measurements on prototypes. A body of empirical studies, based on optimization test-beds and real applications, have shown that the EI algorithm can lead to significant evaluation savings over traditional optimization methods (see, e.g. Jones, 2001; Huang et al., 2006; Forrester et al., 2008). Yet, making use of an optimization algorithm without knowing its convergence properties is not satisfying, not only theoretically, but also from a practical viewpoint. Indeed, if it turned out that the EI algorithm could not get arbitrarily close to a global optimizer when the number of function evaluations increases, using this algorithm on a restricted budget of function evaluations would hardly be justified.

In this paper, we have provided two important results. The first one is that the EI improvement algorithm behaves consistently provided that the objective function belongs to the reproducing kernel Hilbert space (RKHS) attached to ξ , under a non-degeneracy assumption on the covariance function that we have called the No-Empty-Ball (NEB) property. This result is obviously interesting from a theoretical viewpoint; it is less so in practice because one seldom knows in advance whether the objective function belongs to a given RKHS. The second main result of this paper, which states that convergence also takes place for \mathbb{P} -almost all continuous functions, where \mathbb{P} is the (prior) probability distribution of the Gaussian process ξ , is what really matters from a practical point of view.

These results constitute a first step toward a deeper understanding of global optimization algorithms based on the EI criterion, or more generally on criteria satisfying (9). Possible directions for future research include the derivation of pathwise or average convergence rates, the convergence of the algorithm when some parameters of the covariance are re-estimated after each new evaluation, and the extension—possibly under more restrictive assumptions—of our convergence results to all continuous functions.

References

- Aronszajn, N., 1950. Theory of reproducing kernels. *Trans. Amer. Math. Soc.* 68, 337–404.
 Bogachev, V. I., 1998. *Gaussian Measures*. Vol. 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.
 Chilès, J.-P., Delfiner, P., 1999. *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York.
 Currin, C., Mitchell, T., Morris, M., Ylvisaker, D., 1991. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Amer. Statist. Assoc.*, 953–963.

- Forrester, A., Söbester, A., Keane, A., 2008. Engineering Design via Surrogate Modelling. Wiler, Chichester.
- Huang, D., Allen, T., Notz, W., Zeng, N., 2006. Global optimization of stochastic black-box systems via sequential kriging meta-models. *J. Global Optim.* 34, 441–466.
- Jones, D., 2001. A taxonomy of global optimization methods based on response surfaces. *J. Global Optim.* 21, 345–383.
- Jones, D. R., Schonlau, M., Welch, W. J., 1998. Efficient global optimization of expensive black-box functions. *J. Global Optim.* 13, 455–492.
- Locatelli, M., 1997. Bayesian algorithms for one-dimensional global optimization. *J. Global Optim.* 10, 57–76.
- Lukic, M. N., Beder, J. H., 2001. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Trans. Amer. Math. Soc.* 353 (10), 3945–3969.
- Mockus, J., 1989. Bayesian approach to Global Optimization: Theory and Applications. Kluwer Acad. Publ., Dordrecht-Boston-London.
- Mockus, J., Tiesis, V., Zilinskas, A., 1978. The application of Bayesian methods for seeking the extremum. In: Dixon, L., Szego, G. (Eds.), *Towards Global Optimization*. Vol. 2. North Holland, New York, pp. 117–129.
- Molchanov, I. S., 2005. *Theory of Random Sets*. Springer, London.
- Sacks, J., Welch, W. J., Mitchell, T. J., Wynn, H. P., 1989. Design and analysis of computer experiments. *Statist. Sci.* 4 (4), 409–435.
- Schonlau, M., 1997. Computer experiments and global optimization. Ph.D. thesis, University of Waterloo, Waterloo, Ontario, Canada.
- Schonlau, M., Welch, W. J., 1996. Global optimization with nonparametric function fitting. In: *Proceedings of the ASA, Section on Physical and Engineering Sciences*. Amer. Statist. Assoc., pp. 183–186.
- Schonlau, M., Welch, W. J., Jones, D. R., 1997. A data analytic approach to bayesian global optimization. In: *Proceedings of the ASA, Section on Physical and Engineering Sciences*. Amer. Statist. Assoc., pp. 186–191.
- Stein, M. L., 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York.
- Steinwart, I., 2001. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.* 2, 67–93.
- Törn, A., Zilinskas, A., 1989. *Global Optimization*. Springer, Berlin.
- Vazquez, E., Bect, J., 2010. Pointwise consistency of the kriging predictor with known mean and covariance functions. In: Giovagnoli, A., Atkinson, A. C., Torsney, B. (Eds.), *mODa 9 – Advances in Model-Oriented Design and Analysis*. Springer, to be published.
- Williams, D., 1991. *Probability with Martingales*. Cambridge University Press, Cambridge.
- Zilinskas, A., 1992. A review of statistical models for global optimization. *J. Global Optim.* 2, 145–153.