Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion

Romain Benassi, Julien Bect, Emmanuel Vazquez

Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion

Romain Benassi, Julien Bect, and Emmanuel Vazquez

SUPELEC
Gif-sur-Yvette, France

Abstract. We consider the problem of optimizing a real-valued continuous function f, which is supposed to be expensive to evaluate and, consequently, can only be evaluated a limited number of times. This article focuses on the Bayesian approach to this problem, which consists in combining evaluation results and prior information about f in order to efficiently select new evaluation points, as long as the budget for evaluations is not exhausted.

The algorithm called efficient global optimization (EGO), proposed by Jones, Schonlau and Welch (*J. Global Optim.*, 13(4):455–492, 1998), is one of the most popular Bayesian optimization algorithms. It is based on a sampling criterion called the expected improvement (EI), which assumes a Gaussian process prior about f. In the EGO algorithm, the parameters of the covariance of the Gaussian process are estimated from the evaluation results by maximum likelihood, and these parameters are then plugged in the EI sampling criterion. However, it is well-known that this plug-in strategy can lead to very disappointing results when the evaluation results do not carry enough information about f to estimate the parameters in a satisfactory manner.

We advocate a fully Bayesian approach to this problem, and derive an analytical expression for the EI criterion in the case of Student predictive distributions. Numerical experiments show that the fully Bayesian approach makes EI-based optimization more robust while maintaining an average loss similar to that of the EGO algorithm.

1 Introduction

Let f be a continuous real-valued function defined on some compact space $\mathbb{X} \subset \mathbb{R}^d$. We consider the problem of finding the maximum of f, when f is supposed to be expensive to evaluate because one evaluation takes a long time or a large amount of resources. In this case, the optimization of f must be carried out using a limited number of evaluations. More precisely, given a budget of N evaluations of f, our objective is to choose sequentially N evaluation points $X_1, \ldots, X_N \in \mathbb{X}$ so that $\varepsilon(\underline{X}_N, f) = M - M_N$ is small, where $\underline{X}_N$ stands for $(X_1, \ldots, X_N)$, $M = \max_{x \in \mathbb{X}} f(x)$ and $M_N = f(X_1) \vee \cdots \vee f(X_N)$.

In this article, we adopt a Bayesian approach to this sequential decision problem: the unknown function f is considered as a sample path of a real-valued

random process ξ defined on some probability space $(\Omega, \mathcal{B}, \mathsf{P}_0)$ with parameter $x \in \mathbb{X}$, and a good strategy is a strategy that achieves, or gets close to, the Bayes risk $r_{\mathrm{B}} := \inf_{\underline{X}_N} \mathsf{E}_0\left(\varepsilon(\underline{X}_N, \xi)\right)$, where E_0 denotes the expectation with respect to P_0 and the infinimum is taken over the set of all sequential strategies. The reader is referred to the books [1–5] for a broader view on the field of global optimization.

It is well-known [6–12] that an optimal Bayesian optimization strategy, i.e. a strategy $\underline{X}_N^\star$ such that $\mathsf{E}_0\left(\varepsilon(\underline{X}_N^\star, \xi)\right) = r_{\mathrm{B}}$, can be formally obtained by dynamic programming. Let E_n, $n = 1, 2, \ldots$, denote the conditional expectation with respect to the σ-algebra $\mathcal{F}_n$ generated by the random variables $X_1, \xi(X_1), \ldots, X_n, \xi(X_n)$. Denote by $R_N = \mathsf{E}_N\left(\varepsilon(\underline{X}_N, \xi)\right)$ the terminal risk and define by backward induction

$$R_n = \min_{x \in \mathbb{X}} \mathsf{E}_n\big(R_{n+1} \mid X_{n+1} = x\big), \quad n = N-1, \ldots, 0. \tag{1}$$

Then, we have $R_0 = r_{\mathrm{B}}$, and the strategy $\underline{X}_N^\star$ defined by

$$X_{n+1}^\star = \operatorname*{argmin}_{x \in \mathbb{X}} \mathsf{E}_n\big(R_{n+1} \mid X_{n+1} = x\big), \quad n = 1, \ldots, N-1, \tag{2}$$

is optimal. Unfortunately, solving (1)–(2) over an horizon N of more than a few steps is not numerically tractable, for both the space of possible actions and the space of possible outcomes at each step are continuous.

A natural way of dealing with this problem is to consider a suboptimal one-step lookahead strategy; see, e.g., [13, chapter 6]. This leads to choosing each new evaluation point according to

$$\begin{aligned} X_{n+1} &= \operatorname*{argmin}_{x \in \mathbb{X}} \mathsf{E}_n\left(M - M_{n+1} \mid X_{n+1} = x\right) \\ &= \operatorname*{argmax}_{x \in \mathbb{X}} \mathsf{E}_n\left(M_{n+1} \mid X_{n+1} = x\right) \\ &= \operatorname*{argmax}_{x \in \mathbb{X}} \rho_n(x) := \mathsf{E}_n\left((\xi(X_{n+1}) - M_n)_+ \;\middle|\; X_{n+1} = x\right), \end{aligned} \tag{3}$$

where $(z)_+ = 0 \vee z$. The sampling criterion ρ_n, introduced by J. Mockus [6] and popularized through the EGO algorithm [14], is known as the *expected improvement* (EI).

When ξ is a Gaussian process, or in other words, when a Gaussian process prior is chosen for f, it is well-known that the EI can be written in closed form, with the consequence that the maximization of ρ_n can be carried out with a moderate computational effort. However, a Gaussian process prior carries a high amount of information about f and it is often difficult to elicit such a prior before any evaluation is made. As a result, the covariance function of ξ is usually assumed to belong to some parametric class of positive definite functions, the value of the parameters assumed to be unknown. In the EGO algorithm, the parameters are estimated from the evaluation results by maximum likelihood, and then plugged in the EI sampling criterion (computed for a Gaussian process with known covariance function). It has been reported [15] that this plug-in

strategy can lead to very disappointing results when the evaluation results do not carry enough information about f to estimate the parameters satisfactorily. We advocate a fully Bayesian approach to this problem, following the steps of Locatelli [9, 16] and, more recently, Osborne and co-authors [17–19].

The paper is organized as follows. Section 2 recalls the expression of the EI criterion in the case of a Gaussian process prior with known covariance function, and describes the plug-in approach used in the EGO algorithm to handle the parameters of the covariance function when it is only assumed to belong to some parametric class. Section 3 explains how a fully Bayesian approach can be adopted in this problem, in order to take into account the uncertainty on the parameters of the covariance function. Section 4 presents a new closed-form expression of the EI criterion for Student predictive densities, which arises naturally when a conjugate inverse-gamma prior is used for the variance parameter of the Gaussian process prior. Section 5 illustrates with numerical results the benefits of the fully Bayesian approach, focusing more particularly on the tail of the error distribution, i.e., on the occurrence of large errors.

Nota bene. *The analytical expression of the expected improvement for Student predictive distributions, presented in Section 4, has in fact already been obtained by Williams, Santner and Notz [20] in the special case of an improper Jeffrey prior on the variance. We warmly thank Frank Hutter for pointing out this paper to us during the LION5 conference.*

2 Efficient global optimization

2.1 The expected improvement sampling criterion for a Gaussian process

Recall that the distribution of a Gaussian process ξ is uniquely determined by its mean function $m(x) := \mathsf{E}_0(\xi(x))$, $x \in \mathbb{X}$, and its covariance function $k(x, y) := \mathsf{E}_0\left((\xi(x) - m(x))(\xi(y) - m(y))\right)$, $x, y \in \mathbb{X}$. Hereafter, we assume that the mean function is constant on $\mathbb{X}$ and write $\xi \sim \mathrm{GP}\,(m,\, k)$ to denote that ξ is a Gaussian process with mean function $m(x) = m \in \mathbb{R}$ and covariance function k.

Proposition 1. *Let k be a stationary covariance function written as $k(x, y) = \sigma^2 r(x - y)$, $x, y \in \mathbb{X}$, where $\sigma^2 > 0$ and $r(0) = 1$ (hence, r is a correlation function). Assume that $\xi \mid m \sim \mathrm{GP}\,(m,\, k)$ and $m \sim \mathcal{U}(\mathbb{R})$, where $\mathcal{U}(\mathbb{R})$ denotes the (improper) uniform distribution over $\mathbb{R}$. Then, for all $x \in \mathbb{X}$,*

$$\xi(x) \mid \mathcal{F}_n \;\sim\; \mathcal{N}\left(\widehat{\xi}_n(x),\, s_n^2(x)\right),$$

where

$$\widehat{\xi}_n(x) = \widehat{m}_n + r_n(x)^{\mathsf{T}} R_n^{-1}(\underline{\xi}_n - \widehat{m}_n \mathbb{1}_n)\,, \tag{4}$$

with

$$\begin{cases} \underline{\xi}_n = (\xi(X_1), \ldots, \xi(X_n))^\mathsf{T}\,, \\ \mathbb{1}_n = (1, \ldots, 1)^\mathsf{T} \in \mathbb{R}^n, \\ R_n \text{ the correlation matrix of } \underline{\xi}_n, \\ r_n(x) \text{ the correlation vector between } \xi(x) \text{ and } \underline{\xi}_n, \\ \widehat{m}_n = \frac{\mathbb{1}_n^\mathsf{T} R_n^{-1} \underline{\xi}_n}{\mathbb{1}_n^\mathsf{T} R_n^{-1} \mathbb{1}_n}, \text{ the weighted least squares estimate of } m, \end{cases}$$

and

$$s_n^2(x) = \sigma^2 \kappa_n^2(x)\,, \tag{5}$$

with

$$\kappa_n^2(x) = 1 - r_n(x)^\mathsf{T} R_n^{-1} r_n(x) + \frac{(1 - r_n(x)^\mathsf{T} R_n^{-1} \mathbb{1}_n)^2}{\mathbb{1}_n^\mathsf{T} R_n^{-1} \mathbb{1}_n}\,. \tag{6}$$

Proposition 2. *Under the assumptions of Proposition 1, the expected improvement can be written as*

$$\rho_n(x) = \begin{cases} s_n(x)\, \Phi'\left(\frac{\widehat{\xi}_n(x) - M_n}{s_n(x)}\right) + (\widehat{\xi}_n(x) - M_n)\, \Phi\left(\frac{\widehat{\xi}_n(x) - M_n}{s_n(x)}\right) & \text{if } s_n(x) > 0, \\ \left(\widehat{\xi}_n(x) - M_n\right)_+ & \text{if } s_n(x) = 0. \end{cases} \tag{7}$$

where Φ denotes the Gaussian cumulative distribution function.

Propositions 1 and 2 show that, given a set of evaluation points and a Gaussian prior, the EI sampling criterion can be computed with a moderate amount of resources (computing (4) at q different points in $\mathbb{X}$ involves $O(qn^2)$ operations).

However, it is rare that a user has enough information about f in order to choose an adequate covariance function k before any evaluation is made. The approach generally taken consists in choosing k in a parametrized class of covariance functions and estimating the parameters of k from the evaluation results.

2.2 Classical parametrized covariance functions

There are chiefly three classes of parametrized covariance functions in the literature of Gaussian processes for modeling computer experiments. These are the class of the so-called Gaussian covariances, the class of the exponential covariances, and that of the Matérn covariances. Using Matérn covariances makes it possible to tune the mean square differentiability of ξ, which is not the case with the exponential and Gaussian covariances.

Define $\upsilon_\nu : \mathbb{R}^+ \to \mathbb{R}^+$ such that, $\forall h \geq 0$,

$$\upsilon_\nu(h) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(2\nu^{1/2} h\right)^\nu \mathcal{K}_\nu \left(2\nu^{1/2} h\right)\,, \tag{8}$$

where Γ is the Gamma function and $\mathcal{K}_\nu$ is the modified Bessel function of the second kind of order ν. The parameter $\nu > 0$ controls regularity at the origin of υ_ν.

The anisotropic form of the Matérn covariance on $\mathbb{R}^d$ may be written as $k_\theta(x,y) = \sigma^2 r_\theta(x,y)$, with

$$r_\theta(x,y) = \upsilon_\nu\left(\sqrt{\sum_{i=1}^{d}\frac{(x_{[i]}-y_{[i]})^2}{\beta_i^2}}\right), \quad x,y\in\mathbb{R}^d, \tag{9}$$

where the positive scalar σ^2 is a variance parameter (we have $k_\theta(x,x)=\sigma^2$), $x_{[i]}, y_{[i]}$ denote the i^{th} coordinate of x and y, the positive scalars β_i represent scale or *range* parameters of the covariance, or in other words, characteristic correlation lengths, and finally $\theta = (\nu,\beta_1,\ldots,\beta_d)\in\mathbb{R}_+^{d+1}$ denotes the parameter vector of the Matérn covariance. Note that an isotropic form of the Matérn covariance is obtained by setting $\beta_1 = \ldots = \beta_d = \beta$. Then, the parameter vector of the Matérn covariance is $\theta = (\nu,\beta)\in\mathbb{R}_+^2$.

2.3 The EGO algorithm

The approach taken in the EGO (efficient global optimization) algorithm [14, 21–23] consists in estimating the unknown parameters of the covariance function by maximum likelihood, after each new evaluation. Then, the EI sampling criterion is computed using the current value of the parameters of the covariance. EGO can therefore be viewed as a plug-in approach.

Remark 1 (about maximum likelihood estimation of the parameters of a covariance function of a Gaussian process). Recall that, for $\xi\sim\mathrm{GP}(m,k_\theta)$ with $k_\theta(x,y)=\sigma^2 r_\theta(x,y)$, the likelihood of the evaluation results can be written as

$$\ell_n(\underline{\xi}_n; m,\sigma^2,\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}|R_n(\theta)|^{1/2}}\, e^{-\frac{1}{\sigma^2}\left(\underline{\xi}_n - m\mathbb{1}_n\right)^{\mathsf{T}} R_n(\theta)^{-1}\left(\underline{\xi}_n - m\mathbb{1}_n\right)}, \tag{10}$$

where $R_n(\theta)$ stands for the correlation matrix of $\underline{\xi}_n$, parametrized by θ. Note that setting to zero the partial derivatives of ℓ_n with respect to m and σ^2 yields the following maximum likelihood estimates for m and σ^2:

$$\widehat{m}(\theta) \;=\; \frac{\mathbb{1}_n^{\mathsf{T}} R_n(\theta)^{-1}\underline{\xi}_n}{\mathbb{1}_n^{\mathsf{T}} R_n(\theta)^{-1}\mathbb{1}_n}, \tag{11}$$

$$\widehat{\sigma}^2(\theta) \;=\; \frac{1}{n}\left(\underline{\xi}_n - \widehat{m}\mathbb{1}_n\right)^{\mathsf{T}} R_n(\theta)^{-1}\left(\underline{\xi}_n - \widehat{m}\mathbb{1}_n\right). \tag{12}$$

Thus the maximum likelihood estimate of θ can be obtained by maximizing the profile likelihood $\theta\mapsto\ell_n(\underline{\xi}_n;\widehat{m}(\theta),\widehat{\sigma}^2(\theta),\theta)$.

2.4 The case of deceptive functions

Deceptive functions is a term coined by D. Jones (see [15, 25]) to describe functions that appear to be "flat" based on evaluation results. In fact, any function can potentially appear to be flat depending on how it is sampled.

When the available evaluation results do not bring enough information on the objective function f to estimate the parameters of the covariance function with a reasonnable precision, the variance of the error of prediction can be severely under-estimated as depicted in Figure 1. As will be shown in Section 5.1, this can lead to very unsatisfactory behaviors of the EGO algorithm, which tends to waste lots of evalutions in local search around the current maxima (exploitation), very early in the optimization procedure, to the detriment of global search (exploration).

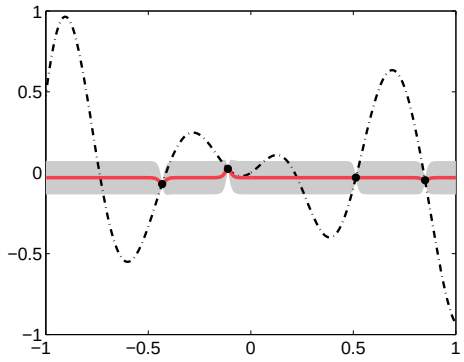


Fig. 1. Example of a deceptive sampling of a function (dashdot line). Evaluation points (black dots) are chosen such that the value of the function is around zero at these points. After having estimated the parameters of the covariance function by maximum likelihood, the prediction is very flat (solid line) and confidence intervals derived from the standard deviation of the error of prediction (gray area) are severely underestimated.

3 Fully Bayesian one-step lookahead optimization

It has been emphasized in Section 1 that the rationale behind the EI criterion is of a Bayesian decision-theoretic nature. Indeed, maximizing the EI criterion at iteration n is equivalent to minimizing the expected loss $\mathsf{E}_n\left(\max(\xi) - M_{n+1}\right)$, where the expectation is taken with respect to the value of the next evaluation, which is unknown and therefore modeled as a random variable.

In a *fully Bayesian* setting, *all* the unknown parameters of the model have to be given prior distributions. This has already been done for the unknown

mean m in Proposition 1. Let π_0 denote the prior distribution of the vector of covariance parameters $\theta' = (\sigma^2, \theta)$, and let π_n, $n = 1, \ldots, N$, denote the corresponding posterior distributions. According to Bayes' rule, the posterior distribution of $\xi(x)$ is a mixture of Gaussian distributions $\mathcal{N}\big(\widehat{\xi}_n(x;\theta'), s_n^2(x;\theta')\big)$ weighted by $\pi_n(\mathrm{d}\theta')$. The expected improvement criterion for this model can thus be written, using the tower property of conditional expectations, as

$$\begin{aligned} \mathsf{E}_n\left((\xi(x) - M_n)_+\right) &= \mathsf{E}_n\Big(\mathsf{E}_n\Big((\xi(x) - M_n)_+ \;\Big|\; \theta'\Big)\Big) \\ &= \int \rho_n(x;\theta')\,\pi_n\left(\mathrm{d}\theta'\right) . \end{aligned} \tag{13}$$

Note that the plug-in EI criterion of Section 2.3 can be seen as an approximation of the fully Bayesian criterion (13):

$$\int \rho_n(x;\theta')\,\pi_n\left(\mathrm{d}\theta'\right) \;\approx\; \rho_n(x;\widehat{\theta}'_n)\,,$$

which is justified only if the posterior distribution is concentrated enough around the MLE estimate $\widehat{\theta}'_n$. In the general case, we claim that it is safer to use the fully Bayesian criterion (13), since the corresponding expected loss integrates the uncertainty related to the fact that θ' is not exactly known. This claim will be supported by the numerical results of Section 5.

When π_0 is a finitely supported discrete distribution, the posterior distribution π_n—and therefore the integral (13)—can be computed exactly using Bayes' rule. For more general prior distribution, the integral can be approximated by stochastic techniques like MCMC sampling or SMC sampling (see [26–28] and the references therein). An alternative approach using Bayesian quadrature rules [29] has been proposed in [17–19]. In all cases, the EI criterion is approximated by an expression of the form $\sum_i w_i \rho_n(x;\theta'_i)$, which amounts to saying that π_n is approximated by the discrete distribution $\sum_i w_i \delta_{\theta'_i}$.

Remark 2. Although fully Bayesian approaches for Gaussian process models have been proposed in the literature for more than two decades (see [30, 31] and the references therein), surprisingly little has been written from this perspective in the context of Bayesian global optimization. An early attempt in this direction can be found in [9, 16], where the variance parameter of a Brownian motion is given an inverse-gamma prior and then integrated out as in (13). More recently, the fully Bayesian approach has been developed in a more general way by [17–19], but the important connection of (13) with the usual (Gaussian) EI criterion was not clearly established.

Remark 3. Discrete mixtures of Gaussian distributions and the corresponding EI criterion have also been introduced in [32] to allow for the use of several parametric classes of covariance functions, in order to provide increased robustness with respect to the choice of a particular class. The approach is not Bayesian, however, since the weights in the mixture are not posterior probabilities.

4 Student EI

Let us consider the case of a Gaussian process ξ with unknown mean m and covariance function of the form $k(x,y) = \sigma^2 r(x,y)$. We assume that m and σ^2 are independent, with m uniformly distributed on $\mathbb{R}$ (as in Proposition 1) and σ^2 following an inverse-gamma distribution with shape parameter a_0 and scale parameter b_0, hereafter denoted by $\mathrm{IG}(a_0, b_0)$. We shall prove that, in this setting, the EI criterion still has an explicit analytical expression, which is a generalization of the usual EI criterion given in Proposition 2.

First, recall that the prior chosen for σ^2 is conjugate [33]:

Proposition 3. *The conditional distribution of σ^2 given $\mathcal{F}_n$ is $\mathrm{IG}(a_n, b_n)$, with*

$$\begin{aligned} a_n &= a_0 + \frac{n-1}{2}, \\ b_n &= b_0 + \frac{1}{2}\left(\underline{\xi}_n - \widehat{m}_n \mathbb{1}_n\right)^{\mathsf{T}} R_n^{-1}\left(\underline{\xi}_n - \widehat{m}_n \mathbb{1}_n\right). \end{aligned}$$

Using this result and the fact that $\xi(x) \mid \sigma^2, \underline{\xi}_n \sim \mathcal{N}\left(0, \sigma^2 \kappa_n^2(x)\right)$, it is easy to show that the predictive distribution of $\xi(x)$ is a Student distribution. More precisely:

Proposition 4. *Let t_η denote the Student distribution with $\eta > 0$ degrees of freedom. Then, for all $x \in \mathbb{X}$,*

$$\frac{\xi(x) - \widehat{\xi}_n(x)}{\gamma_n(x)} \mid \mathcal{F}_n \sim t_{\eta_n},$$

with $\eta_n = 2a_n$, and $\gamma_n^2(x) = b_n/a_n\, \kappa_n^2(x)$.

In other words, the predictive distribution at x is a location-scale Student distribution with η_n degrees of freedom, location parameter $\widehat{\xi}_n(x)$ and scale parameter $\gamma_n(x)$. The following result is the key to our EI criterion for Student predictive distributions:

Lemma 1. *Let $T \sim t_\eta$ with $\eta > 0$. Then*

$$\mathsf{E}\big((T+u)_+\big) = \begin{cases} +\infty & \text{if } \eta \leq 1, \\ \frac{\eta+u^2}{\eta-1}\, F'_\eta(u) + u\, F_\eta(u) & \text{otherwise,} \end{cases}$$

where F_η is the cumulative distribution function of t_η.

Combining Lemma 1 and Proposition 4 finally yields an explicit expression of the EI criterion:

Theorem 1. *Under the assumptions of this section, for all $x \in \mathbb{X}$,*

$$\mathsf{E}_n\big((\xi(x) - M_n)_+\big) = \gamma_n(x)\left(\frac{\eta_n + u^2}{\eta_n - 1}\, F'_{\eta_n}(u) + u\, F_{\eta_n}(u)\right), \tag{14}$$

with $u = \big(\widehat{\xi}_n(x) - M_n\big)/\gamma_n(x)$.

It has been assumed, up to this point, that the only unknown parameter in the covariance function is the variance σ^2. More generally, assume that $k(x,y) = \sigma^2\, r(x,y;\theta)$: in this case we proceed by conditioning as in Section 3. Indeed, assume that θ is independent from $\left(m,\sigma^2\right)$ with a prior distribution π_0. Let us denote by $\tilde{\rho}_n(x;\theta) = \mathsf{E}_n\left((\xi(x) - M_n)_+ \mid \theta\right)$ the value of the EI criterion at x provided by Theorem 1 when the value of the unknown parameter is θ. Then

$$\mathsf{E}_n\left((\xi(x) - M_n)_+\right) \;=\; \mathsf{E}_n\big(\tilde{\rho}_n\left(x;\theta\right)\big) \;=\; \int \tilde{\rho}_n(x;\theta)\,\pi_n(\mathrm{d}\theta), \tag{15}$$

where π_n denotes the posterior distribution of θ after n evaluations. As explained in Section 3, the integral (15) boils down to a finite sum that can be computed exactly (using Bayes' rule) when the prior π_0 has a finite support; in the general case, approximation techniques have to be used.

5 Numerical experiments

5.1 Optimization of a deceptive function

Experiment. Consider the objective function $f : \mathbb{X} = [-1,1] \to \mathbb{R}$ defined by

$$f(x) \;=\; x\left(\sin(10x+1) + 0.1\sin(15x)\right), \quad \forall x \in \mathbb{X}\,.$$

We choose an initial set of four evaluation points with abscissas -0.43, -0.11, 0.515 and 0.85, as shown in Figure 1. Our objective is to compare the evaluation points chosen by the plug-in approach (i.e., the EGO algorithm) and those chosen by the fully Bayesian algorithm (FBA) proposed in Section 4.

In both approaches, we consider a Matérn covariance function with a known regularity parameter $\nu = 2$ (see Section 2.2). In the approach of Section 4, we choose an inverse gamma distribution $IG(0.2, 12)$ for σ^2. Since $\mathbb{X}$ has dimension one, there is only one range parameter β. To simplify the implementation of the approach proposed, we shall assume that β has a finite support distribution. More precisely, define a $\beta_{\min}$ and a $\beta_{\max}$, such that $\beta_{\min} < \beta_{\max}$, and set, for all $i = 0, \ldots, I$, $\beta_i = \beta_{\min}\left(\frac{\beta_{\max}}{\beta_{\min}}\right)^{i/I}$. We assume a uniform prior distribution over the β_is, with $\beta_{\min} = 2 \times 10^{-3}$, $\beta_{\max} = 2$ and $I = 100$.

The optimization of the two sampling criteria is performed by a Monte Carlo approach. More precisely, we generate once and for all a set of $q = 600$ candidate points uniformly distributed over $\mathbb{X}$ and the search for the maximum of each sampling criterion is carried out at each iteration by determining the value of the sampling criterion over this finite set (the same set of points is used for both criteria).

Results. Figures 2, 3 and 4 show that the standard deviation of the error of prediction is severely underestimated when using the EGO algorithm, as a result of the maximum likelihood estimation of the parameters of the covariance from

a deceptive set of evaluation points. If the uncertainty about the covariance parameters is taken into account, as explained above, the standard deviation of the error is more satisfactory. Figures 3 and 4 show that the maximum is approximated satisfactorily after only four iterations with FBA, whereas EGO needs nine more iterations before making an evaluation in the neighborhood of the maximizer. Indeed, we observe that EGO stays in the neighborhood of a local optimum for a long time, while $\mathbb{X}$ remains unexplored. This behavior is not desirable in a context of expensive-to-evaluate functions.

5.2 Comparison on sample paths of a Gaussian process

Experiment. In order to assess the performances of EGO and FBA from a statistical point of view, we study the convergence to the maximum using both algorithms on a set of sample paths of a Gaussian process.

We have built 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, due to the lack of room, we present only the results obtained for two testbeds in dimension 1 and 4 (the actual parameters are provided in Table 1).

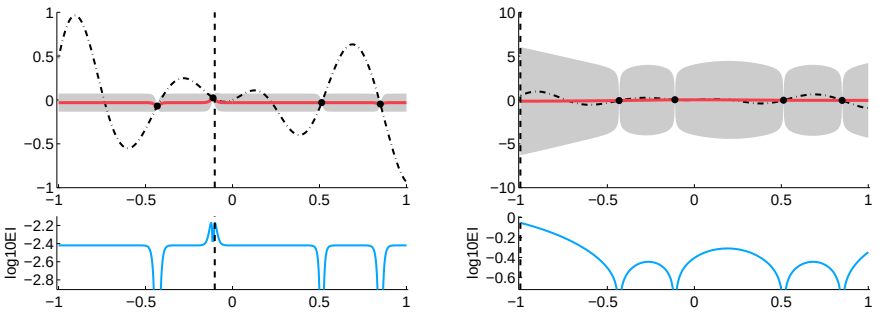
Parameter \ Testbed	$\mathcal{T}_1$	$\mathcal{T}_2$
Dimension d	1	4
Number of sample paths L	20000	20000
Variance σ^2	1.0	1.0
Regularity ν	2.5	2.5
Scale $\beta = (\beta_1, \ldots, \beta_d)$	0.1	(0.7, 0.7, 0.7, 0.7)

Table 1. Parameters used for building the testbeds of Gaussian-process sample-paths.

We shall compare the performance of EGO and FBA based on the approximation error $\varepsilon(\underline{X}_n, f_{k,l})$, $l = 1, \ldots, L$. For reference, we also provide the results obtained with two other strategies. The first strategy corresponds to using an EI criterion with the same values for the parameters of the covariance function of ξ than those used to generate the sample paths in the testbeds. In principle this strategy ought to perform very well. The second strategy corresponds to space-filling sampling, which is not necessarily a good optimization strategy.

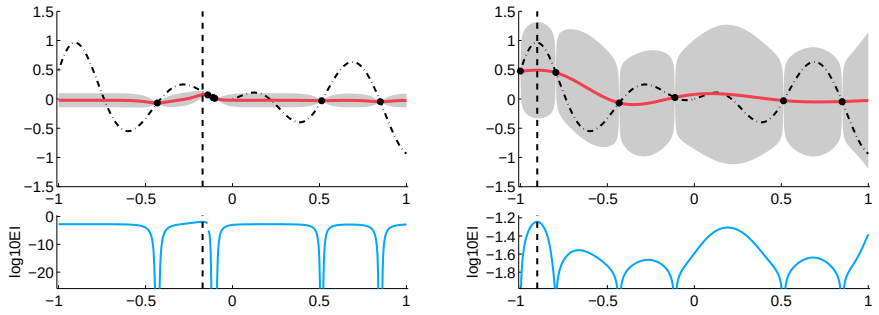
For FBA, we choose the same priors as those described in Section 5.1. More precisely, whatever be the dimension d, we choose an isotropic covariance function (with only one scale parameter) and we set $\beta_{\min} = 1/400$ and $\beta_{\max} = 2\sqrt{d}$.

Results. Figures 5(a) and 6(a) show that EGO and FBA have very similar average performances. In fact, both of them perform almost as well, in this



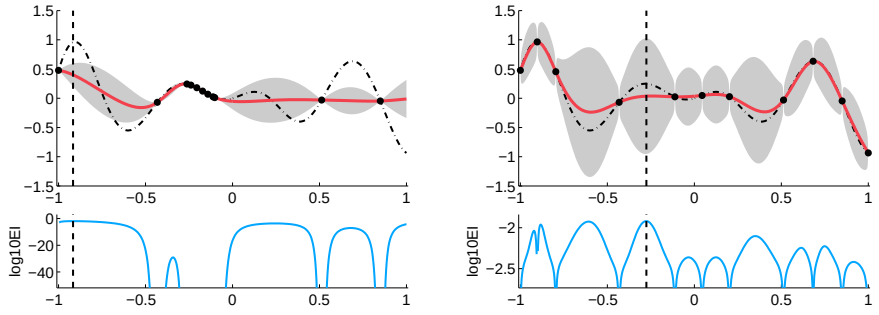
(a) parameters estimated by MLE (b) Bayesian approach for the parameters

Fig. 2. A comparison of a) EGO and b) FBA at iteration 1. Top: objective function (dashdot line), prediction (solid line), 95% confidence intervals derived from the standard deviation (gray area), sampling points (dots) and position of the next evaluation (vertical dashed line). Bottom: EI criterion.



(a) parameters estimated by MLE (b) Bayesian approach for the parameters

Fig. 3. Iteration 3 (see Figure 2 for details)



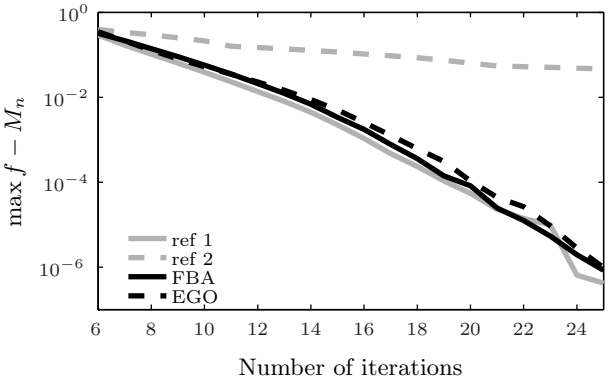
(a) parameters estimated by MLE (b) Bayesian approach for the parameters

Fig. 4. Iteration 8 (see Figure 2 for details)

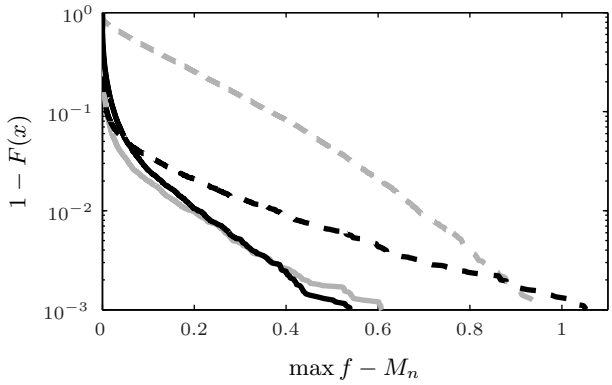
experiment, as the reference strategy where the true parameters are assumed to be known. Comparing the tails of complementary cumulative distribution function of the error $\max f - M_n$ makes it clear, however, that using a fully Bayesian approach brings a significant reduction of the occurrence of 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.

References

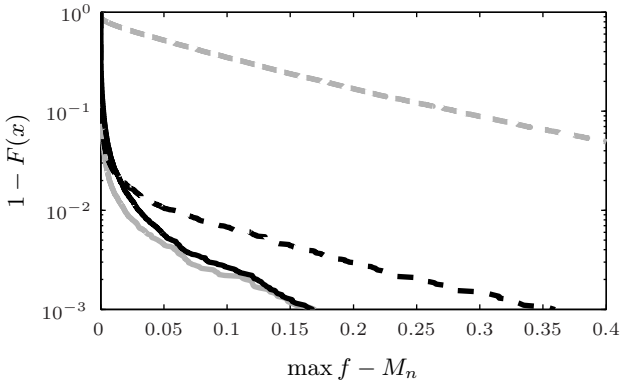
1. A. Törn and A. Zilinskas. *Global Optimization*. Springer, Berlin, 1989.
2. J. D. Pintér. *Global optimization. Continuous and Lipschitz optimization: algorithms, implementations and applications*. Springer, 1996.
3. A. Zhigljavsky and A. Zilinskas. *Stochastic global optimization*. Springer Verlag, 2007.
4. A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to derivative-free optimization*. SIAM, 2009.
5. Y. Tenne, and C. K. Goh. *Computational intelligence in optimization: applications and implementations*. Springer, 2010.
6. J. Mockus, V. Tiesis, and A. Zilinskas. The application of Bayesian methods for seeking the extremum. In L. Dixon and G. Szego, editors, *Towards Global Optimization*, volume 2, pages 117–129. Elsevier, 1978.
7. J. Mockus. *Bayesian approach to Global Optimization: Theory and Applications*. Kluwer Acad. Publ., Dordrecht-Boston-London, 1989.
8. B. Betrò. Bayesian methods in global optimization. *Journal of Global Optimization*, 1:1–14, 1991.
9. M. Locatelli and F. Schoen. An adaptive stochastic global optimization algorithm for one-dimensional functions. *Annals of Operations research*, 58(4):261–278, 1995.
10. A. Auger and O. Teytaud. Continuous lunches are free plus the design of optimal optimization algorithms. *Algorithmica*, 57(1):121–146, 2008.
11. D. Ginsbourger and R. Le Riche. Towards Gaussian process-based optimization with finite time horizon. In *mODa 9 Advances in Model-Oriented Design and Analysis*, Contribution to Statistics, 89–96, Springer, 2010.
12. S. Grünewälder, J.-Y. Audibert, M. Opper, and J. Shawe-Taylor. Regret bounds for Gaussian process bandit problems. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, volume 9 of *JMLR W&CP*, pages 273–280, 2010.
13. D. P. Bertsekas. *Dynamic programming and optimal control*. Athena Scientific Belmont, MA, 1995.
14. D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.
15. A. I. J. Forrester and D. R. Jones. Global optimization of deceptive functions with sparse sampling. In *12th AIAA/ISSMO multidisciplinary analysis and optimization conference*, 10-12 September 2008.



(a) Average error to the maximum

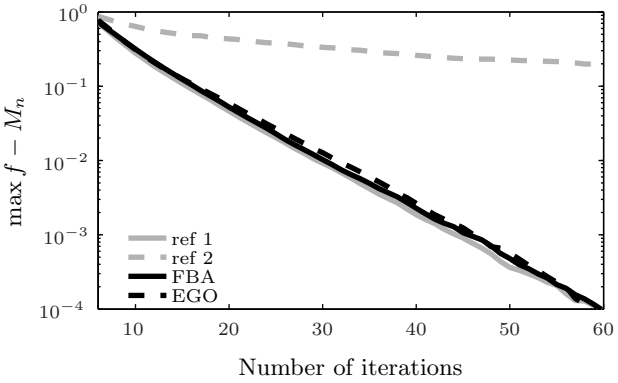


(b) Distribution of errors at iteration 13

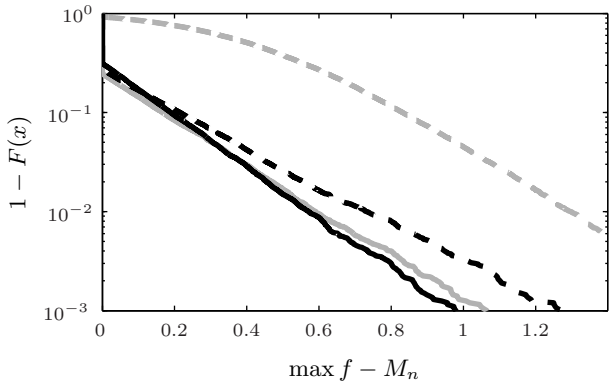


(c) Distribution of errors at iteration 16

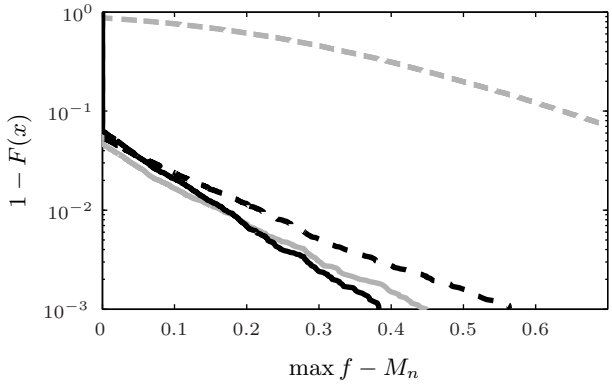
Fig. 5. Average results and error distributions for testbed $\mathcal{T}_1$, for FBA (solid black line), EGO (dashed black line), the EI with the parameters used to generate sample paths (solid gray line), the space-filling strategy (dashed gray line). More precisely, (a) represents the average approximation 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

Fig. 6. Average results and distribution of errors for testbed $\mathcal{T}_2$. See Figure 5 for details.

16. M. Locatelli. Bayesian algorithms for one-dimensional global optimization. *Journal of Global Optimization*, 10(1):57–76, 1997.

17. M. A. Osborne. *Bayesian Gaussian Processes for Sequential Prediction Optimisation and Quadrature*. PhD thesis, University of Oxford, 2010.

18. M. A. Osborne, R. Garnett, and S. J. Roberts. Gaussian processes for global optimization. In *3rd International Conference on Learning and Intelligent Optimization (LION3), online proceedings*, Trento, Italy, 2009.

19. M. A. Osborne, S. J. Roberts, A. Rogers, S. D. Ramchurn, and N. R. Jennings. Towards real-time information processing of sensor network data using computationally efficient multi-output Gaussian processes. In *Proceedings of the 7th International Conference on Information Processing in Sensor Networks*, pages 109–120. IEE Computer Society, 2008.

20. B. Williams, T. Santner, W. Notz. Sequential Design of Computer Experiments to Minimize Integrated Response Functions. *Statistica Sinica*, 10(4):1133–1152, 2000

21. M. Schonlau. *Computer experiments and global optimization*. PhD thesis, University of Waterloo, Waterloo, Ontario, Canada, 1997.

22. M. Schonlau and W. J. Welch. Global optimization with nonparametric function fitting. In *Proceedings of the ASA, Section on Physical and Engineering Sciences*, pages 183–186. Amer. Statist. Assoc., 1996.

23. M. Schonlau, W. J. Welch, and D. R. Jones. A data analytic approach to Bayesian global optimization. In *Proceedings of the ASA, Section on Physical and Engineering Sciences*, pages 186–191. Amer. Statist. Assoc., 1997.

24. A. I. J. Forrester and A. J. Keane. Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences*, 45(1-3):50–79, 2009.

25. D. R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21(4):345–383, 2001.

26. C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 2004.

27. P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

28. J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer, 2008.

29. A. O'Hagan. Bayes-Hermite quadrature. *Journal of statistical planning and inference*, 29(3):245–260, 1991.

30. A. O'Hagan. Curve Fitting and Optimal Design for Prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 40(1):1–42, 1978.

31. M. S. Handcock and M. L. Stein. A Bayesian analysis of Kriging. *Technometrics*, 35(4):403–410, 1993.

32. D. Ginsbourger, C. Helbert, and L. Carraro. Discrete mixtures of kernels for kriging-based optimization. *Quality and Reliability Engineering International*, 24:681–691, 2008.

33. A. O'Hagan. Some Bayesian numerical analysis. In *Bayesian statistics 4: proceedings of the Fourth Valencia International Meeting, April 15-20, 1991*. Oxford University Press, 1992.