An analysis of covariance parameters in Gaussian Process-based optimization

Hossein Mohammadi, Rodolphe Le Riche, Xavier Bay, Eric Touboul

An analysis of covariance parameters in Gaussian process-based optimization

Hossein Mohammadi[1,*], Rodolphe Le Riche[2], Xavier Bay[2] and Eric Touboul[2]

[1] *College of Engineering, Mathematics and Physical Sciences, University of Exeter, Prince of Wales Road, EX4 4SB Exeter, UK*
E-mail: ⟨H.Mohammadi@exeter.ac.uk⟩

[2] *Ecole des Mines de Saint Etienne, Institut H. Fayol, 158 Cours Fauriel, 42023 Saint-Etienne, France*
E-mail: ⟨{leriche, bay, touboul}@emse.fr⟩

Abstract. The need for globally optimizing expensive-to-evaluate functions frequently occurs in many real-world applications. Among the methods developed for solving such problems, the Efficient Global Optimization (EGO) is regarded as one of the state-of-the-art unconstrained continuous optimization algorithms. The surrogate model used in EGO is a Gaussian process (GP) conditional on data points. The most important control on the efficiency of the EGO algorithm is the GP covariance function (or kernel), which is taken as a parameterized function. In this paper, we theoretically and empirically analyze the effect of the covariance parameters, the so-called "characteristic length scale" and "nugget", on EGO performance. More precisely, we analyze the EGO algorithm with fixed covariance parameters and compare them to the standard setting where they are statistically estimated. The limit behavior of EGO with very small or very large characteristic length scales is identified. Experiments show that a "small" nugget should be preferred to its maximum likelihood estimate. Overall, this study contributes to a better theoretical and practical understanding of a key optimization algorithm.

Keywords: Covariance kernel, EGO, Gaussian process, global optimization

Received: September 23, 2016; accepted: June 8, 2018; available online: July 24, 2018

DOI: 10.17535/crorr.2018.0001

1. Introduction

We wish to find the global minimum of a function f, $\min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x})$, where the search space $\mathcal{D} = [LB, UB]^d$ is a compact subset of $\mathbb{R}^d$. We assume that f is an expensive-to-compute black-box function. Subsequently, optimization can only be attempted for a low number of function evaluations. The Efficient Global Optimization (EGO) algorithm [3, 4] has become standard for optimizing such expensive unconstrained continuous problems. Its efficiency stems from an embedded conditional Gaussian process (GP), also known as kriging, which acts as a surrogate for the objective function. Certainly, other surrogate techniques can be employed instead of GPs. For example, [9] proposes a variant of EGO in which a quadratic regression model serves as a surrogate. However, it is shown by some of their examples that the standard EGO performs better than this variant.

A kriging model is described principally by the associated kernel that determines the set of possible functions processed by the algorithm to make optimization decisions. Several alternative methods to cross-validation or maximum likelihood (ML) have been suggested to tune the

[*] Corresponding author.

kernel parameters. For example, a fully Bayesian approach is used in [1]. In [4], the process of estimating parameters and searching for the optimum are combined through a likelihood which encompasses a targeted objective. In [8], the bounds on the parameter values change within iterations following an a priori schedule. In our view, the existing methods induce interactions between kernel learning at each iteration and the optimization dynamics that are still difficult to understand. The goal of this study is to more deeply understand the influence of the kernel parameters on the efficiency of EGO by studying the convergence of EGO with fixed parameters on both a unimodal and multimodal function. In addition, the effect of the "nugget" term is investigated.

2. Kriging model summary

Let $\mathbf{X} = \{\mathbf{x}^i\}_{i=1}^n$ be a set of n design points and $\mathbf{y} = \{y_i = f(\mathbf{x}^i)\}_{i=1}^n$ the associated function values at $\mathbf{X}$. Suppose the observations are a realization of a stationary GP, $Y(\mathbf{x})$. The kriging model is the GP conditional on the observations, $Y(\mathbf{x}) \mid Y(\mathbf{x}^1) = y_1, \dots, Y(\mathbf{x}^n) = y_n$. The GP prediction (simple kriging mean) and variance of prediction (simple kriging variance) at a generic point $\mathbf{x}$ are

$$m(\mathbf{x}) = \mu + \mathbf{r}(\mathbf{x})^\top \mathbf{R}^{-1}(\mathbf{y} - \mu\mathbf{1}), \tag{1}$$

$$s^2(\mathbf{x}) = \sigma^2 \left(1 - \mathbf{r}(\mathbf{x})^\top \mathbf{R}^{-1}\mathbf{r}(\mathbf{x})\right). \tag{2}$$

Here, μ and σ^2 are the constant prior mean and variance, $\mathbf{1}$ is a $n \times 1$ vector of ones, $\mathbf{r}(\mathbf{x})$ is the vector of correlations between point $\mathbf{x}$ and the n sample points,
$\mathbf{r}(\mathbf{x}) = [\mathrm{Corr}(Y(\mathbf{x}), Y(\mathbf{x}^1)), \dots, \mathrm{Corr}(Y(\mathbf{x}), Y(\mathbf{x}^n))]^\top$, and $\mathbf{R}$ is the correlation matrix between sample points of general term $\mathbf{R}_{ij} = \mathrm{Corr}(Y(\mathbf{x}^i), Y(\mathbf{x}^j))$. The covariance kernel mostly used here is the isotropic Matérn 5/2 function defined as [7]

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \mathrm{Corr}(Y(\mathbf{x}), Y(\mathbf{x}')) = \sigma^2 \left(1 + \tfrac{\sqrt{5}\|\mathbf{x}-\mathbf{x}'\|}{\theta} + \tfrac{5\|\mathbf{x}-\mathbf{x}'\|^2}{3\theta^2}\right) \exp\left(-\tfrac{\sqrt{5}\|\mathbf{x}-\mathbf{x}'\|}{\theta}\right), \tag{3}$$

where the scalar parameter $\theta > 0$ is the *characteristic length scale* and controls the correlation strength between pairs of response values. The smaller the length scale θ, the least any two response values at given points are correlated, and vice versa (see Figure 1). When a nugget τ^2, is added to the model, the covariance function becomes

$$k_{\tau^2}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') + \tau^2 \delta(\mathbf{x}, \mathbf{x}'), \tag{4}$$

where $\delta(.,.)$ is the Kronecker's delta. Adding a nugget to the model means that the observations are perturbed by an additive Gaussian noise $\mathcal{N}(0, \tau^2)$. The nugget also increases kriging variance throughout the search domain since, beside the changes in the covariance matrix $\mathbf{R}$, the term σ^2 becomes $\sigma^2 + \tau^2$ in Equation (2).

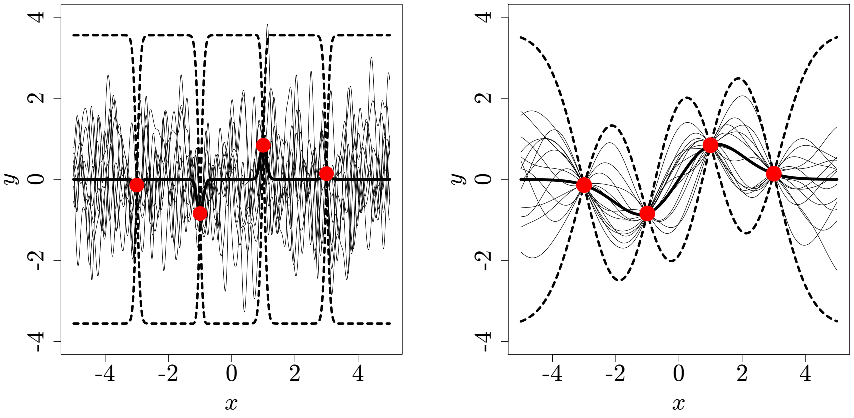
Classically here, the prior mean and variance, without the nugget, are estimated by the following ML closed-form expressions [7],

$$\hat{\mu} = \frac{\mathbf{1}^\top \mathbf{R}^{-1}\mathbf{y}}{\mathbf{1}^\top \mathbf{R}^{-1}\mathbf{1}} \quad , \quad \hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{1}\hat{\mu})^\top \mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu})}{n} \ , \tag{5}$$

so that the only kernel parameters left are θ and τ^2.

At any point $\mathbf{x} \in \mathcal{D}$, the improvement is defined as the random variable $I(\mathbf{x}) = \max(0, f_{min} - Y(\mathbf{x}))$ where f_{min} is the best objective function value observed so far. The improvement is the random excursion of the process at any point below f_{min}. The expected improvement (EI) can be calculated analytically as

$$EI(\mathbf{x}) = \begin{cases} (f_{min} - m(\mathbf{x}))\Phi\left(\frac{f_{min} - m(\mathbf{x})}{s(\mathbf{x})}\right) + s(\mathbf{x})\phi\left(\frac{f_{min} - m(\mathbf{x})}{s(\mathbf{x})}\right) & \text{if } s(\mathbf{x}) > 0 \\ 0 & \text{if } s(\mathbf{x}) = 0 \ , \end{cases} \tag{6}$$



(a) $\theta = 0.1$ (b) $\theta = 1$

Figure 1: *Kriging mean (thick solid line) along with the 95% confidence intervals (thick dashed lines), i.e.,* $m(\boldsymbol{x}) \pm 1.96 s(\boldsymbol{x})$. *The thin lines are the sample paths of the GP. As* θ *changes, the class of possible functions considered for the optimization decision changes.*

in which Φ and ϕ are the cumulative distribution function and probability density function of the standard normal distribution respectively. $EI(\mathbf{x})$ is positive everywhere in $\mathcal{D}$. It is increasing when the kriging variance increases (at a fixed kriging mean) and when the kriging mean decreases (at a fixed kriging variance). The first term in Equation (6) is dominated by the contribution of the kriging mean to the improvement while the second term is dominated by the contribution of the kriging variance. The EGO algorithm is the sequential maximization of the EI, $\mathbf{x}^{n+1} \in \arg\max_{x\in\mathcal{D}} EI(\mathbf{x})$ followed by the updating of the kriging model with $\mathbf{X} \cup \{\mathbf{x}^{n+1}\}$ and the associated responses $\mathbf{y}$.

3. EGO with fixed length scale

We start by discussing the behavior of EGO with two different fixed length scales (small and large). The magnitude of length scale is measured with respect to the longest possible distance in the search space, $Dist_{max}$, which in our d-dimensional box search space is equal to $(UB - LB)\sqrt{d}$. θ is large if it is close to or larger than $Dist_{max}$ and vice versa. Here, $LB = -5$ and $UB = 5$. For the sake of clarity and brevity, illustrations are given on only two isotropic functions, the unimodal sphere and the highly multimodal Ackley functions. They demand two radically different behaviors from the optimization algorithm and are defined as

$$f_{\text{Sphere}}(\mathbf{x}) = \sum_{i=1}^{d} x_i^2, \tag{7}$$

$$f_{\text{Ackley}}(\mathbf{x}) = -20\exp\left(-0.2\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(2\pi x_i)\right) + 20 - \exp(1). \tag{8}$$

Figure 3 illustrates the kriging models on Ackley for small and large length scales. More details can be found in [6].

3.1. EGO with small characteristic length scale

When θ is small, there is a low correlation between response values so that data points influence the process only in their immediate neighborhood. As θ tends to 0 and at points away from the data points, the kriging mean and variance of Equations (1) and (2) turn into the constants μ and σ^2 respectively. Thus the EI becomes a constant flat function when $\mathbf{x}$ is away from $\mathbf{x}^i$s, $EI(\mathbf{x}) \to EI^{\text{asymp}} := (f_{min} - \hat{\mu})\Phi\left(\frac{f_{min}-\hat{\mu}}{\hat{\sigma}}\right) + \hat{\sigma}\phi\left(\frac{f_{min}-\hat{\mu}}{\hat{\sigma}}\right)$, where $\hat{\mu} \to \frac{\sum_{i=1}^{n} y_i}{n}$ and $\hat{\sigma}^2 \to \frac{\sum_{i=1}^{n} (y_i-\hat{\mu})^2}{n}$ since the correlation matrix $\mathbf{R}$ in Equation (5) tends to $\mathbf{I}$, the identity matrix.

Proposition 1 (EGO iterates for small length scale)**.** *Without loss of generality, we assume that the best observed point is unique. As the characteristic length scale of the kernel tends to zero, the EGO iterates are located in a shrinking neighborhood of the best observed point.*

This proposition is now further explained. Irrespectively of the function being optimized and the current design of experiments (DoE), provided that f_{min} is uniquely defined, the set of design points created by EGO with small θ has characteristically repeated samples near f_{min}. When the length scale is small, the observations have a low range of influence. In the limit case, one can assume that in the vicinity of the ith design point the correlation between $Y(\mathbf{x}^i)$ and the other observations is zero, i.e. $\text{Corr}(Y(\mathbf{x}^i), Y(\mathbf{x}^j)) \to 0\ ,\ 1 \le j \le n\ ,\ j \ne i$, so that $\mathbf{R} \to \mathbf{I}$. Let $\mathbf{x}$ be in the neighborhood of $\mathbf{x}^i$, $B_\epsilon(\mathbf{x}^i) = \left\{\mathbf{x} \in \mathcal{D} :\ \|\mathbf{x} - \mathbf{x}^i\| \le \epsilon\right\}$, for a sufficiently small ϵ and away from the other points of the DoE $j \ne i$ so that the correlation vector tends to $\mathbf{r}(\mathbf{x}) \to [0, \ldots, 0, r, 0, \ldots 0]$ where $r = \text{Corr}(Y(\mathbf{x}), Y(\mathbf{x}^i))$. In this situation, the kriging mean and variance can be fully expressed in terms of the correlation r (a scalar in $[0, 1]$):

$$m(r) = \hat{\mu} + r(y_i - \hat{\mu}) = \hat{\mu}(1 - r) + r y_i \quad , \quad s^2(r) = \hat{\sigma}^2(1 - r^2), \tag{9}$$

The above equations show that among the points of the DoE, the EI will be the largest near the best observed point as, for any given r, the variance will be the same and the mean will be the lowest. By setting $y_i = f_{min}$ in Equation (9), dividing the equation by $\hat{\sigma}$ and introducing the new variable $A := \frac{f_{min}-\hat{\mu}}{\hat{\sigma}}$, the normalized EI (Equation (6)) in the vicinity of the best observed point reads,

$$EI(r)/\hat{\sigma} = (1 - r)A\Phi\left(A\sqrt{\frac{1-r}{1+r}}\right) + \sqrt{1-r^2}\phi\left(A\sqrt{\frac{1-r}{1+r}}\right). \tag{10}$$

The normalized EI is handy in that, for small length scale, it sums up what happens for all objective functions, DoEs and kernels in terms of only two scalars, i.e. the correlation r and A. Notice that $A \le 0$ because $f_{min} \le y_i\ ,\ \forall i$. Instances of the normalized EI are plotted for a set of As in $[-2, -0.001]$ in the left of Figure 2. The value of normalized EI when $r \to 0^+$ is the value of EI as $\mathbf{x}$ moves away from data points. The maximum of EI (equivalently $EI/\hat{\sigma}$) is reached at $r^\star$, which is strictly larger than 0, and thus in the neighborhood of the best observed point.

3.2. EGO with large characteristic length scale

Proposition 2 (EGO iterates for large length scale)**.** *As the characteristic length scale of kernels goes to infinity, the EGO algorithm degenerates into the sequential minimization of the kriging mean, $m(.)$.*

Partial proof: As the length scale, θ, goes to infinity, the kriging variance vanishes everywhere, $\lim_{\theta\to\infty} s^2(\mathbf{x}) = 0$, with the implication on the EI of Equation (6) that $\lim_{\theta\to\infty} EI(\mathbf{x}) = f_{min} - m(\mathbf{x})$ in regions where $m(\mathbf{x}) < f_{min}$. To save space, we do not prove here that such regions exist in general. The intuition is that because $m(\mathbf{x})$ is interpolating, and stiff for large θs, it overshoots the best point value. Now we just need to prove that kriging variance $s^2(\mathbf{x})$ tends to zero when $\theta \to \infty$. This is established under general conditions by the following

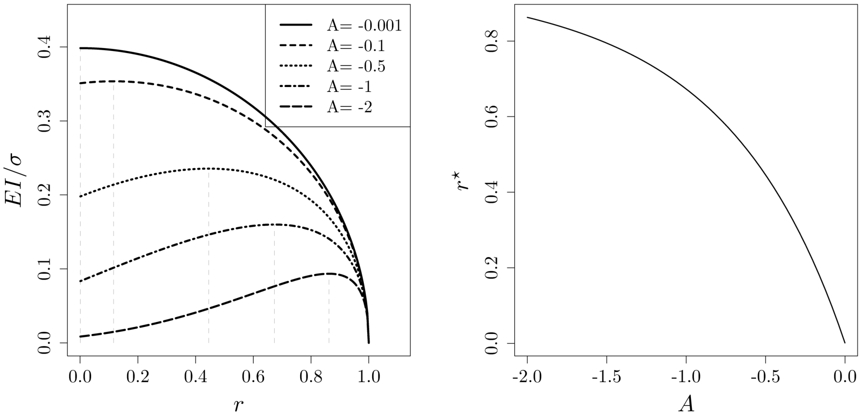


Figure 2: *Left: Normalized EI as a function of $r \in [0,1]$ in the vicinity of the sample point with the lowest function value for a small length scale. Right: location of the next EGO iterate ($r^\star$ where EI is maximized) as a function of A.*

Lemma, where the stationary GP $Y(\mathbf{x})_{\mathbf{x}\in\mathcal{D}}$ is assumed to be centered, i.e. $\mathbb{E}[Y(\mathbf{x})] = 0$, and square-integrable. Dealing with a centered GP is a standard assumption used in simple kriging which is recovered in general by removing the average from the process.

Lemma 1. *If $\exists\, i\ ,\ i = 1,\dots,n$, such that $\langle Y(\mathbf{x}), Y(\mathbf{x}^i)\rangle \to \sigma^2$ as $\theta \to \infty$, then $s^2(\mathbf{x}) \to 0$.*

Proof. We define $\mathcal{S}$ to be the subspace spanned by the $Y(\mathbf{x}^i)$s: $\mathcal{S} = span\left(Y(\mathbf{x}^1),\dots,Y(\mathbf{x}^n)\right)$. Let $\mathcal{P}$ be the orthogonal projection operator onto $\mathcal{S}$. The kriging variance at an arbitrary location is

$$s^2(\mathbf{x}) = \|Y(\mathbf{x}) - \mathcal{P}\left(Y(\mathbf{x})\right)\|^2 = \|Y(\mathbf{x})\|^2 - \|\mathcal{P}(Y(\mathbf{x}))\|^2\ . \tag{11}$$

Any vector in $\mathcal{S}$ can be represented by a linear combination of $Y(\mathbf{x}^1),\dots,Y(\mathbf{x}^n)$, including $\mathcal{P}\left(Y(\mathbf{x})\right)$, which is written as

$$\mathcal{P}(Y(\mathbf{x})) = \sum_{j=1}^{n} \beta_j Y(\mathbf{x}^j) = \boldsymbol{\beta}^\top \mathbf{y}, \tag{12}$$

Because the projection error is perpendicular to the projection plane, the β_js are solutions of $\forall\ i\ ,\ \left\langle Y(\mathbf{x}) - \sum_{j=1}^{n}\beta_j Y(\mathbf{x}^j), Y(\mathbf{x}^i)\right\rangle = 0$, which implies

$$\forall\ i\ ,\ \langle Y(\mathbf{x}), Y(\mathbf{x}^i)\rangle = \left\langle \sum_{j=1}^{n} \beta_j Y(\mathbf{x}^j), Y(\mathbf{x}^i)\right\rangle = \sum_{j=1}^{n} \beta_j \left\langle Y(\mathbf{x}^j), Y(\mathbf{x}^i)\right\rangle. \tag{13}$$

In the Euclidean space, the following definitions hold for any two $\mathbf{x}$ and $\mathbf{x}' \in \mathcal{D}$: $\langle Y(\mathbf{x}), Y(\mathbf{x}')\rangle = \mathbb{E}\left[Y(\mathbf{x})Y(\mathbf{x}')\right] = \mathrm{Cov}\left(Y(\mathbf{x}), Y(\mathbf{x}')\right) = \sigma^2 \mathrm{Corr}\left(Y(\mathbf{x}), Y(\mathbf{x}')\right)$, where the term $\mathrm{Corr}\left(Y(\mathbf{x}), Y(\mathbf{x}')\right)$ is (geometrically) the cosine of the angle between $Y(\mathbf{x})$ and $Y(\mathbf{x}')$. Thus, one can rewrite Equation (13) as

$$\mathbb{E}\left[Y(\mathbf{x})Y(\mathbf{x}^i)\right] = \sum_{j=1}^{n} \beta_j \mathbb{E}\left[Y(\mathbf{x}^j)Y(\mathbf{x}^i)\right] \quad \text{or} \quad \sigma^2 \mathrm{Corr}\left(Y(\mathbf{x}), \mathbf{y}\right) = \sigma^2 \mathrm{Corr}\left(\mathbf{y}, \mathbf{y}\right)\boldsymbol{\beta}\ ,$$

$$\text{i.e.,} \qquad \mathbf{r}(\mathbf{x}) = \mathbf{R}\boldsymbol{\beta} \ \Rightarrow\ \boldsymbol{\beta}^\top = \mathbf{r}^\top(\mathbf{x})\mathbf{R}^{-1}. \tag{14}$$

Summing up, Equations (12 – 14) result in $\mathcal{P}(Y(\mathbf{x})) = \mathbf{r}^\top(\mathbf{x})\mathbf{R}^{-1}\mathbf{y}$, which is an important property: the kriging prediction at any location $\mathbf{x} \in \mathcal{D}$ is the orthogonal projection of $Y(\mathbf{x})$ onto $\mathcal{S}$.

In the sequel, Equation (11) is rewritten by replacing the terms $\|Y(\mathbf{x})\|^2$ and $\|\mathcal{P}(Y(\mathbf{x}))\|^2$ by σ^2 and $\mathbb{E}\left[\mathcal{P}(Y(\mathbf{x}))\mathcal{P}(Y(\mathbf{x}))\right]$ as follows

$$\begin{aligned}\sigma^2 - \mathbb{E}\left[\mathbf{r}^\top(\mathbf{x})\mathbf{R}^{-1}\mathbf{y}\mathbf{y}^\top\mathbf{R}^{-1}\mathbf{r}(\mathbf{x})\right] &= \sigma^2 - \mathbf{r}^\top(\mathbf{x})\mathbf{R}^{-1}\underbrace{\mathbb{E}\left[\mathbf{y}\mathbf{y}^\top\right]}_{\sigma^2\mathbf{R}}\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}) \\ &= \sigma^2 - \left(1 - \mathbf{r}^\top(\mathbf{x})\mathbf{R}^{-1}\mathbf{r}(\mathbf{x})\right) = s^2(\mathbf{x}). \qquad (15)\end{aligned}$$

We can now see why the kriging variance tends to zero when $\theta \to \infty$. Remember that $\langle Y(\mathbf{x}), Y(\mathbf{x}')\rangle = \sigma^2 \mathrm{Corr}\left(Y(\mathbf{x}), Y(\mathbf{x}')\right)$. When $\theta \to \infty$, the correlation between $Y(\mathbf{x})$ and $Y(\mathbf{x}')$ is 1 and $\langle Y(\mathbf{x}), Y(\mathbf{x}')\rangle$ becomes σ^2. Apply this result to $\mathbf{x}' \equiv \mathbf{x}^i$:

$$\begin{aligned}&s^2(\mathbf{x}) = \|Y(\mathbf{x})\|^2 - \|\mathcal{P}(Y(\mathbf{x}))\|^2 \leq \left\|Y(\mathbf{x}) - Y(\mathbf{x}^i)\right\|^2 = \\ &\underbrace{\|Y(\mathbf{x})\|^2}_{\sigma^2} + \underbrace{\left\|Y(\mathbf{x}^i)\right\|^2}_{\sigma^2} - 2\underbrace{\left\langle Y(\mathbf{x}), Y(\mathbf{x}^i)\right\rangle}_{\sigma^2} = 0, \qquad (16)\end{aligned}$$

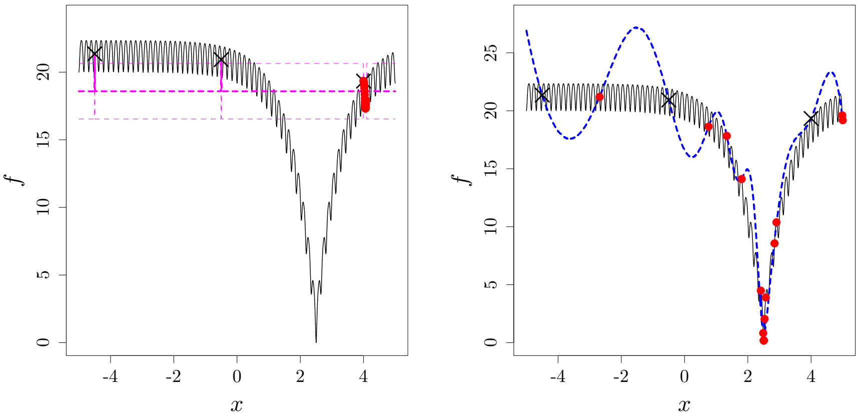
and the proof is complete. □

Minimizing the kriging mean does not define a valid global optimization scheme for two reasons. Firstly, because premature convergence occurs as soon as the minimum of $m(\mathbf{x})$ coincides with an observation of the true function [4]; when $m(\mathbf{x}^{n+1}) = f(\mathbf{x}^{n+1})$ where $\mathbf{x}^{n+1} = \arg\min_{\mathbf{x}\in\mathcal{D}} m(\mathbf{x})$, the EGO iterations with large θ stop producing new points; however, $\mathbf{x}^{n+1} \cup \mathbf{X}$ may not even contain a local optimum of f. Secondly, it should be remembered that the kriging mean discussed here stems from large length scale, which may not allow an accurate prediction of the objective function considered. It suits a function like the sphere with a Matérn kernel, but it is not appropriate for a multimodal function like Ackley.

The DoE created by EGO with large θ can vary greatly depending on the function and the initial DoE. So, if the function is regular and well predicted by $m(.)$ around $\mathbf{x}^{n+1}$, like the sphere function, the kriging mean rapidly converges to the true function and points are accumulated in this region which may or may not be the global optimum. Conversely, if $m(\mathbf{x}^{n+1})$ is different from $f(\mathbf{x}^{n+1})$, the kriging mean changes a lot between iterations, which can be understood as a manifestation of Runge's phenomenon because new observations have a long-range influence. The kriging mean overshoots observations in both upper and lower directions (cf. the dotted curve in the upper right plot of Figure 3). The resulting DoE is more space-filling than the DoE of small length scale. We end this section by further specifying the asymptotic behavior of the kriging mean as $\theta \to \infty$. Our results are based on 1-dimensional observations and are summarized as follows.

Conjecture 1 (Asymptotic behavior of kriging mean for large length scales)**.**
In 1 dimension with n data points, if the covariance function is $n-1$ times differentiable, then the kriging mean tends to the interpolating Lagrange polynomial [2] as the length scale, θ, tends to infinity.

For example, Figure 4 shows processes with Matérn 3/2, Matérn 5/2 and square exponential kernels which are once, twice and infinitely differentiable respectively. Herein, the length scale is 300, which is very large with respect to the distance between the sample locations, and the number of data points is $n = 3$. As a result, the kriging means of the processes with Matérn 5/2 and square exponential kernels are identical to the Lagrange polynomial, but not when the kernel is Matérn 3/2. Note also that, since a cubic spline [5] is made of third degree polynomials, it is the same as the Lagrange polynomial when the number of data points is $n \leq 4$.



(a) $\theta = 0.001$

(b) $\theta = 100$

Figure 3: *Ackley function (solid line) approximated by a kriging model (mean ± std. deviation, thick/thin lines) with* $\theta = 0.001$ *(dashed) and* $\theta = 100$ *(dotted). The crosses are the initial DoE. Bullets: DoEs created by EGO after 20 iterations.*

3.3. Comparison of EGO with fixed and adapted length scale

In this section, the efficiency of EGO with different fixed length scales is compared with the standard EGO whose length scale is determined by ML. Tests are carried out on the previously defined functions, Equations (7) and (8) respectively. Each optimization is repeated five times on 5-dimensional instances of the problems, $d = 5$. The initial DoE is fixed and has size $3 \times d$. The total budget is $70 \times d$. To compare results adequately, the functions are scaled (multiplied) by $\frac{2}{f^{max}_{DoE} - f^{min}_{DoE}}$, where f^{min}_{DoE} and f^{max}_{DoE} are the smallest and the largest value of f in the initial DoE. Figure 5 shows the results in terms of median objective functions. The medians are significant even when accounting for the spread in the results of the experiments. The θ values belong to the set $\{0.01, 0.1, 1, 5, 10, 20\}$. On both functions, EGO does not converge quickly towards the minimum when $\theta = 0.01$ or $\theta = 0.1$ because, as explained in Section 3, it focuses on the neighborhoods of the best points found early in the search. On the sphere function, EGOs with large length scale, i.e. 10 or 20, have performances equivalent to that of the standard EGO. Indeed, the sphere function is very smooth and, as can be seen on the rightmost picture of Figure 5, ML estimates of θ are rapidly equal to 20 (the upper bound of the ML) after a few iterations. With the Ackley function, the best fixed θ is 1. It temporarily outperforms the standard EGO at the beginning of the search (until about 70 evaluations) but then ML makes it possible to decrease the θ until about 0.5 (see rightmost plot) and to fine-tune the search in the already located high performance region.

4. Effect of a nugget on EGO convergence

To investigate the effect of a nugget on EGO, the same test protocol as above is conducted but the length scales are set by ML and two scenarios are considered: 1) nugget τ^2 is estimated by ML, 2) a fixed nugget is taken from the set $\tau^2 \in \{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, 0\}$ ($\tau^2 = 0$ means no nugget). Figure 6 shows the results. For both functions, when the nugget value is large (10^{-2} or 10^{-4} or ML estimated on Ackley), EGO exhibits the worst performances: it does not converge faster and stops further from the optimum. The reason is that a large nugget deteriorates the interpolation quality of a kriging model when observations are not noisy like here. On the sphere function, EGO rapidly locates the area of the optimum but the EI without a nugget, which is

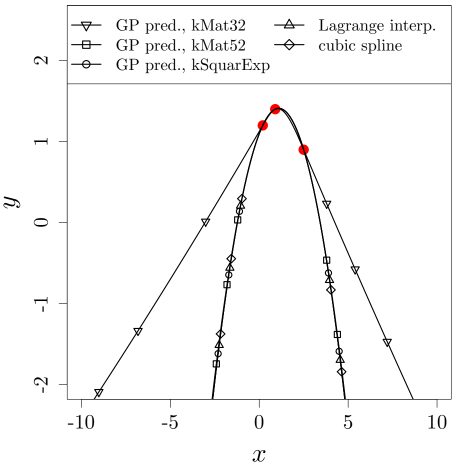
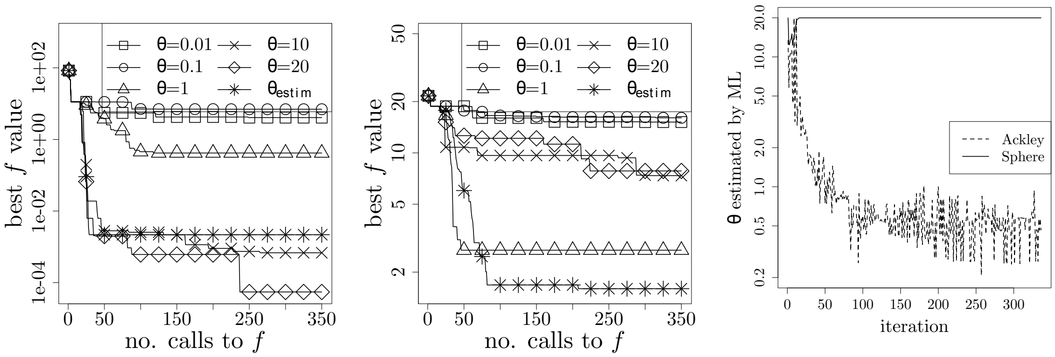


Figure 4: *Comparison of kriging predictions for Matérn 3/2, Matérn 5/2 and square exponential kernels with the interpolating Lagrange polynomial and the cubic spline.* $n = 3$, $\theta = 300$.



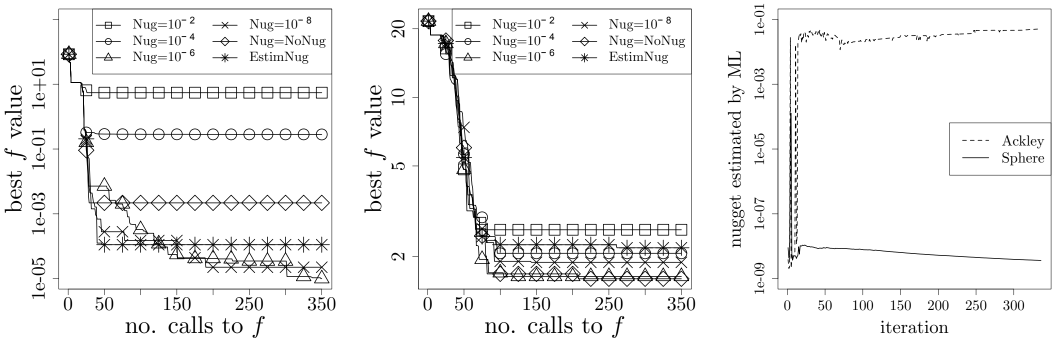
(a) *Sphere function*

(b) *Ackley function*

(c) *Evolution of* $\hat{\theta}_{ML}$ *in standard EGO*

Figure 5: *Median of the best objective function vs. number of calls of standard EGO and EGO with different fixed length scale in dimension* 5.

null at data points, pushes the search away from it. However, a nugget value equal to 10^{-6} or 10^{-8} hardly slows down convergence and significantly improves the accuracy with which the optimum is found. Indeed, by increasing $s^2(\mathbf{x})$ everywhere including in the immediate vicinity of data points, where it would be null without a nugget, the nugget increases the EI there and allows a higher concentration of EGO iterates near the best observed point. The nugget learned by ML on the sphere tends to zero which, as just explained, is not the best setting for optimization. On Ackley, besides large nugget values ($\tau^2 \geq 10^{-4}$) which significantly degrade the EGO search, values ranging from $\tau^2 = 0$ to 10^{-6} do not notably affect efficiency. In this case, the global optimum is not accurately located after $70 \times d$ evaluations of f and there is no need to allow an accumulation of points near the best observation through the nugget. Note that on both functions, when considering the best point found so far, ML estimation of the nugget is not a good strategy.



(a) *Sphere function* (b) *Ackley function* (c) *Nugget estimated by ML*

Figure 6: *Median of the best objective function vs. number of calls to f for EGO with different nugget values in dimension* 5.

5. Conclusion

This paper provides a careful analysis of the effect of length scale and a nugget on EGO iterations. Based on our tests, ML estimation of the length scale is a good choice but ML estimation of the nugget is not recommended and a fixed small nugget value is preferred. As a perspective, our experiments suggest that EGO strategies starting with a large fixed length scale, which is then decreased while keeping a small amount of nugget, should be efficient and will also eliminate the need for ML estimations, which require $O(n^3)$ computations.

Acknowledgements

The authors would like to acknowledge support from the French national research agency (ANR) within the Modèles Numériques project "NUMerical Black-Box Optimizers" (NumBBO).

References

[1] Benassi, R., Bect, J. and Vazquez, E. (2011). Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion. In Battiti, R., Kvasov, D. E. and Sergeyev, Y. D. (Eds.) Learning and Intelligent Optimization (pp. 176-190). Springer International Publishing.

[2] Berrut, J.-P. and Trefethen, L. N. (2004). Barycentric lagrange interpolation. SIAM REVIEW, 46(4), 501-517.

[3] Jones, D. R., Schonlau, M. and Welch, W. J (1998). Effcient global optimization of expensive black-box functions. Journal of Global Optimization, 13(4), 455-492.

[4] Jones, D. R. A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization, 21, 345-383.

[5] Schumaker, L. L. (1981). Spline Functions: Basic Theory. New York: Wiley-Interscience.

[6] Mohammadi, H., Le Riche, R. and Touboul, E. (2015). A detailed analysis of kernel parameters in Gaussian process-based optimization. Ecole Nationale Supérieure des Mines. Technical report, HAL report no. hal-01246677, LIMOS.

[7] Rasmussen, C. E. and Williams, C. K. I. (2005). Gaussian Processes for Machine Learning. (Adaptive Computation and Machine Learning). The MIT Press.

[8] Wang, Z., Zoghi, M., Hutter, F., Matheson, D. and de Freitas, N. (2013). Bayesian optimization in high dimensions via random embeddings. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. https://www.ijcai.org/Proceedings/13/Papers/263.pdf [Accessed 23/9/2016]

[9] Weihs, C., Herbrandt, S., Bauer, N., Friedrichs, K. and Horn, D. (2016). Effcient Global Optimization: Motivation, Variations and Applications. Universitätsbibliothek Dortmund. Discussion Paper No. SFB823. https://publikationen.bibliothek.kit.edu/1000065446/4076602 [Accessed 23/9/2016]