A detailed analysis of kernel parameters in Gaussian process-based optimization

Hossein Mohammadi, Rodolphe Le Riche, Eric Touboul

A detailed analysis of kernel parameters in Gaussian process-based optimization

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

[1] Ecole Nationale Supérieure des Mines de Saint-Etienne, France,
[2] CNRS LIMOS, UMR 5168, Saint-Etienne, France
{hossein.mohammadi,leriche,touboul}@emse.fr

Abstract. The global optimization of expensive-to-evaluate functions frequently occurs in many real-world applications. Among the methods developed for solving such problems, Efficient Global Optimization (EGO) is regarded as one of the state-of-the-art unconstrained continuous optimization algorithms. The most important control on the efficiency of EGO is the Gaussian process covariance function which must be chosen together with the objective function. Traditionally, a parameterized family of covariance functions is considered whose parameters are learned by maximum likelihood or cross-validation. In this paper, we theoretically and empirically analyze the effect of length-scale covariance parameters and nugget on the design of experiments generated by EGO and the associated optimization performance.

Keywords: Continuous global optimization; EGO; Gaussian processes.

1 Introduction

We wish to find the global minimum of a function f, $\min_{\mathbf{x}\in\mathcal{S}} f(\mathbf{x})$, where the search space $\mathcal{S} = [LB, UB]^d$ is a compact subset of $\mathbb{R}^d$. We assume that f is an expensive-to-compute black-box function. In this situation, optimization can only be attempted at a low number of function evaluations. The Efficient Global Optimization (EGO) algorithm [4] has become a 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.

The way the kriging model is learned from data points is essential to the EGO performance as the kernel determines the set of functions processed by the algorithm to make optimization decisions. Several methods alternative to cross-validation or ML have been proposed to tune the kernel parameters. For example, a fully Bayesian approach is used in [2]. In [4], the process of estimating parameters and searching for the optimum are combined together through a likelihood which encompasses a targeted objective. In [7], the bounds on the parameter values are changing with the iterations following an a priori schedule. The existing methods for learning kernel parameters are complex so that the basic phenomena taking place in the optimization when tuning the kernel cannot

be clearly observed. This paper allows to more deeply understand the influence of kriging parameters on the efficiency of EGO by studying the convergence of EGO with fixed parameters on a unimodal and a multimodal function. The effect of nugget is also investigated.

2 Kriging model summary

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

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

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

Here, μ and σ^2 are the process 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, $\mathbf{r}_i = \mathrm{Cor}(Y(\mathbf{x}), Y(\mathbf{x}^i))$, and $\mathbf{R}$ is an $n \times n$ correlation matrix between sample points, $\mathbf{R}_{ij} = \mathrm{Cor}(Y(\mathbf{x}^i), Y(\mathbf{x}^j))$. The covariance function (i.e., kernel) used in this paper is the isotropic Mateŕn 5/2 function defined as [5]

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

in which the parameter $\theta > 0$ is called *characteristic length-scale* and controls the correlation strength between pairs of response values. The smaller θ, the least two response values at given points are correlated, and vice versa.
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. The process mean and variance are estimated by ML [5], In this paper, the process mean and variance are estimated by the following ML closed-form expressions [5],

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

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

At any point $\mathbf{x}$ in $\mathcal{S}$, the improvement is defined as $I(\mathbf{x}) = \max(0, f_{min} - Y(\mathbf{x}) \mid Y(\mathbf{X}) = \mathbf{y})$ 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 the best observed function value. The expected improvement can be calculated analytically as

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

where Φ and ϕ denote the cumulative distribution function and probability density function of the standard normal distribution, respectively. EI($\mathbf{x}$) is null at data points, positive everywhere else. It is increasing when the kriging variance increases and when the kriging mean decreases. The first term in Eq. (6) is dominated by the contribution of kriging mean to the improvement while the second term is dominated by the contribution of kriging variance. The EGO algorithm consists in the sequential maximization of EI, $\mathbf{x}^{n+1} = \arg\max_{x \in \mathcal{S}} EI(\mathbf{x})$ 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 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 $UB = 5$. Fig. 4 illustrates the kriging models on the Ackley test function (defined below) in 1 dimension and the associated EIs for small and large length-scales.

When θ is small, there is a low correlation between response values so that data points have an influence on the process only in their immediate neighborhood. As $\theta \to 0$ and 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 flat function: when $\mathbf{x}$ is away from $\mathbf{x}^i$, $EI(\mathbf{x}) \approx (f_{min} - \hat{\mu})\Phi\left(\frac{f_{min}-\hat{\mu}}{\hat{\sigma}}\right) + \hat{\sigma}\phi\left(\frac{f_{min}-\hat{\mu}}{\hat{\sigma}}\right)$. Also, $\hat{\mu} = \frac{\sum_{i=1}^{n} y^i}{n}$ and $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y^i - \hat{\mu})^2}{n}$ since $\mathbf{R}$ becomes an identity matrix in Equation (2). In this case, the next infill sample is always in the immediate vicinity of the design point with the lowest function value (if there are many points with lowest observation, the most isolated one). Irrespectively of the function being optimized, the set of design points created by EGO with small θ, has characteristically repeated samples near the best observed points (cf. example at the bottom left of Fig. 4).

When the length-scale is small, the observations have low range influence. In the limit case, one can assume that in a vicinity of ith design point the correlation between y^i and other observations is zero. Let $B_\epsilon((\mathbf{x}^i, y^i)) = \left\{\mathbf{x} \in \mathcal{S} : \ \|\mathbf{x} - \mathbf{x}^i\| \leq \epsilon\right\}$ be the immediate vicinity of data point $(\mathbf{x}^i, y^i)$. If the amount of correlation between y^i and $Y(\mathbf{x})$ for $\mathbf{x} \in B_\epsilon((\mathbf{x}^i, y^i))$, is denoted by $r \in (0, 1)$, the kriging mean

and variance and the EI can be expressed by r in this region as follows:

$$m(r(\mathbf{x})) = \mu + r(1)^{-1}(f_{min} - \mu) = \mu(1-r) + r f_{min}, \tag{7}$$

$$s^2(r(\mathbf{x})) = \sigma^2\left(1 - r(1)^{-1} r\right) = \sigma^2(1-r^2), \tag{8}$$

$$EI(r(\mathbf{x})) = (1-r)(f_{min}-\mu)\Phi\left(\frac{f_{min}-\mu}{\sigma}\sqrt{\frac{1-r}{1+r}}\right) + \sigma\sqrt{1-r^2}\,\phi\left(\frac{f_{min}-\mu}{\sigma}\sqrt{\frac{1-r}{1+r}}\right). \tag{9}$$

By dividing both sides of Equation 9 by σ and introducing the new variable A that equals $\frac{f_{min}-\mu}{\sigma} \leq 0$, the normalized expected improvement EI/σ, reads

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

Note that when $r \to 0$, $EI/\sigma = A\Phi(A) + \phi(A)$. Fig. 1 illustrates the case that EI is explained as $EI(r(\mathbf{x}))$ when the length-scale of the kriging model is 10^{-12}. Also, EI/σ is plotted with three different values of A.

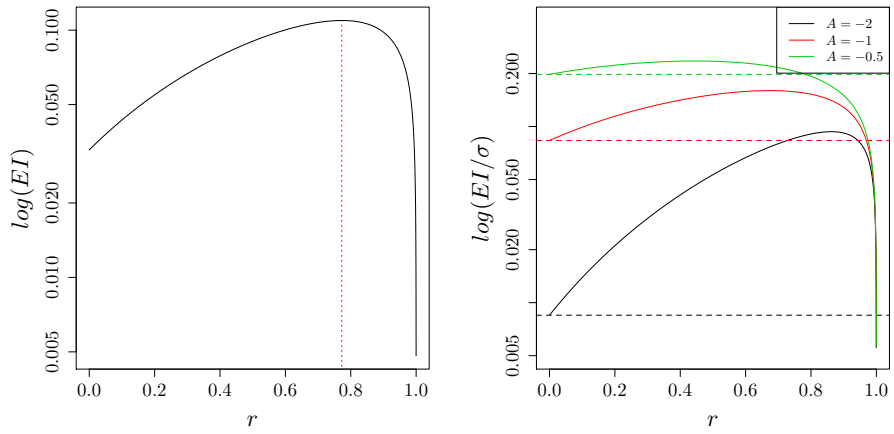

Fig. 1: EI as a function of $r \in (0,1)$ for $\theta = 10^{-12}$. The next infill sample will be taken where $r \simeq 0.77$.

An example is provided in Fig. 2: when θ tends to zero, the next EGO iterate is in the neighborhood of an already observed point with minimal objective function value. Because in practice θ has a positive finite value, after some iterations the next EGO iterate will jump from a design point with lowest objective function value to the neighborhood of the next observed point striking the best compromise between low objective and low sampling density.

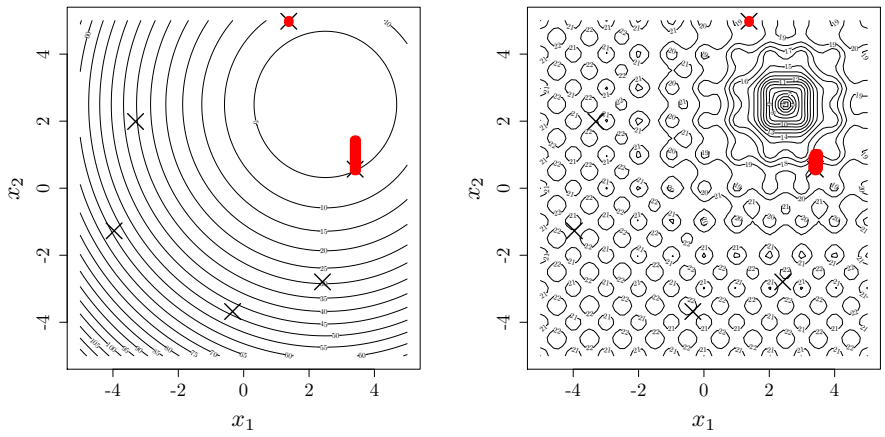

Fig. 2: $2D$ illustration of search points (bullets) obtained by EGO with small length-scale ($\theta = 0.01$) on the Sphere (left) and Ackley (right) functions.

On the contrary, when $\theta \to \infty$, it can be proved that the term $\mathbf{r}(\mathbf{x})^\top \mathbf{R}^{-1}\mathbf{r}(\mathbf{x})$ in Equation (2) tends to 1. Because $\mathbf{r}(\mathbf{x})$ and $\mathbf{R}$ become a vector and matrix of ones. The matrix $\mathbf{R}$ has only one non-zero eigenvalue that equals n, the matrix size [1]. The corresponding eigenvector is $\mathbf{v} = \frac{\sqrt{n}}{n}(1, \ldots, 1)^\top$. To invert such a non-invertible matrix, we use *Moore-Penrose pseudoinverse* [6] denoted by $\mathbf{R}^\dagger$ and defined as

$$\mathbf{R}^\dagger = [\mathbf{v}\ \mathbf{W}] \begin{bmatrix} \frac{1}{n} & \mathbf{0}_{1\times(n-1)} \\ \mathbf{0}_{(n-1)\times 1} & \mathbf{0}_{(n-1)\times(n-1)} \end{bmatrix} [\mathbf{v}\ \mathbf{W}]^\top, \tag{11}$$

in which $\mathbf{W}$ contains the $n-1$ eigenvectors associated with the zero eigenvalues. Now, it is easy to show that $\mathbf{r}(\mathbf{x})^\top \mathbf{R}^{-1}\mathbf{r}(\mathbf{x}) = (1, \ldots, 1)\mathbf{R}^\dagger(1, \ldots, 1)^\top = 1$. As a result, the kriging variance, $s^2(\mathbf{x})$, becomes zero. In this case, $EI(\mathbf{x}) \to f_{min} - m(\mathbf{x})$ and the EGO search degenerates to an iterative minimization and updating of the kriging mean $m(\mathbf{x})$.

Minimizing kriging mean does not define a valid global optimization scheme 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 $\mathbf{x}^{n+1} = \arg\min_{\mathbf{x}\in\mathcal{S}} m(\mathbf{x})$, 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. The DoE created by EGO with large θ can vary greatly depending on the function and the initial DoE. On the one hand, 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 not be the global optimum, see Fig. 3.

On the other hand, if $m(\mathbf{x}^{n+1})$ is different from $f(\mathbf{x}^{n+1})$, the kriging mean changes a lot between iterations because new observations have long range in-

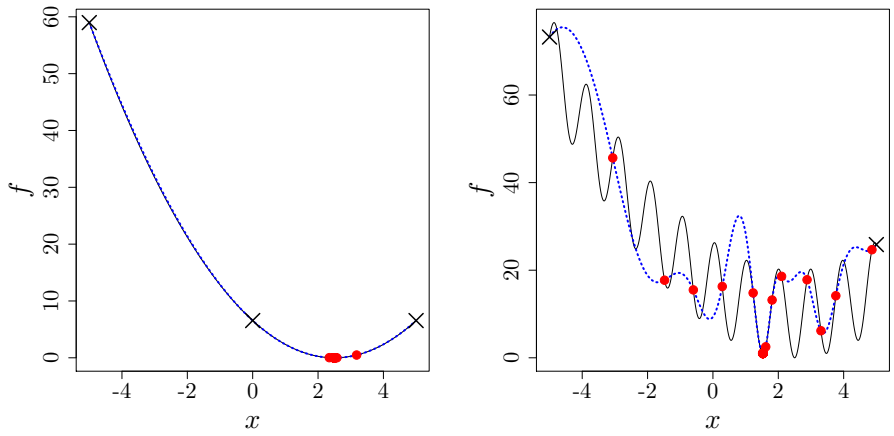

Fig. 3: DoE created by EGO with $\theta = 100$. The global minimum is located at 2.5.

fluence. The kriging mean overshoots observations in both upper and lower directions (cf. the dotted blue curve in the upper left plot of Fig. 4). The resulting DoE is more space-filling than the DoE of small length scales. An example of such DoE is provided at the bottom right of Fig. 4.

In the sequel, the efficiency of EGO with different fixed length-scales is compared with the standard EGO whose length-scale is learned by ML. Tests are carried out on two isotropic functions, the unimodal sphere and the highly multimodal Ackley functions:

$$f_{\text{sphere}}(\mathbf{x}) = \sum_{i=1}^{d} (x_i)^2$$

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

Both functions have 0 as minimal value. Each optimization is repeated 5 times on 5 dimensional instances of the problems, $d = 5$. The initial DoE is fixed and has size $3 \times d$. The search length is $70 \times d$. To allow comparisons of the results, the functions are scaled (multiplied) by $\frac{2}{f_{DoE}^{max} - f_{DoE}^{min}}$, where f_{DoE}^{min} and f_{DoE}^{max} are the smallest and the largest value of function f in the initial DoE.

Fig. 5 shows the results of the comparison in terms of median objective functions. The θ values belong to the set $\{0.01, 0.1, 1, 5, 10, 20\}$. On both test functions, the algorithm does not converge quickly towards the minimum when $\theta = 0.01$ or $\theta = 0.1$ because it focuses on the neighborhoods of the best points found early in the search. On the Sphere function, EGOs with large length-scales, $\theta = 20$ or $\theta = 10$, have performances equivalent to that of the standard EGO. Indeed, the Sphere function is very smooth and, as can be seen on the rightmost plot of Fig. 5, ML estimates of θ are equal to 20 (the upper bound of the ML) rapidly after a few iterations. With the multimodal Ackley function,

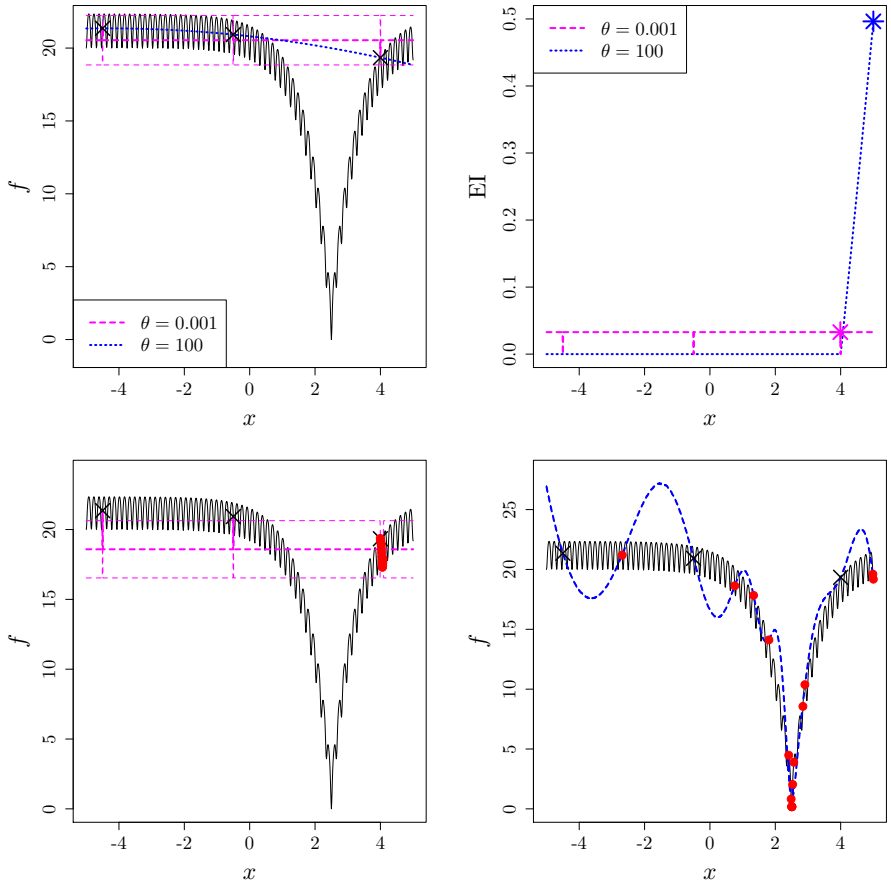

Fig. 4: Ackley function (black solid line) approximated by a kriging model (mean ± std. deviation, thick/thin lines) with $\theta = 0.001$ (dashed pink) and $\theta = 100$ (dotted blue). The crosses are the initial DoE. Top, right: EIs at iteration 1 with the stars indicating the EI maximums. Bottom, red bullets: DoEs created by EGO after 20 iterations with $\theta = 0.001$ (left) and $\theta = 100$ (right).

the best fixed θ is equal to 1. It temporarily outperforms the standard EGO at the beginning of the search (until about 70 evaluations) but then ML allows decreasing the θ's until about 0.5 (see rightmost plot) and fine tuning the search in the already located high performance region.

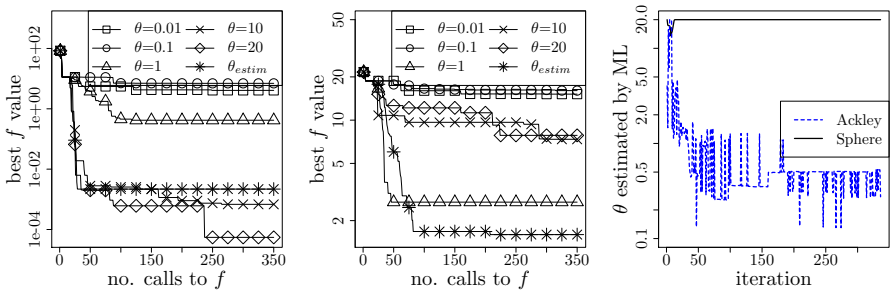

Fig. 5: Median of the best objective function vs. number of calls of standard EGO and EGO with different fixed length-scales on the Sphere (left) and the Ackley (middle) functions, $d = 5$. Right: evolution of θ learned by ML in standard EGO.

4 Effect of nugget on EGO convergence

To investigate the effect of nugget on EGO, we carry out the same test protocol as above but the length-scales are set by ML and two scenarios are considered: 1) the 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). Fig. 6 shows the results. For both test 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 nugget, which is 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 the uncertainty $s^2(\mathbf{x})$ everywhere including in the immediate vicinity of data points, where it would be null without nugget, 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 0 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 performance. In this case, the global optimum is not accurately located after

$70 \times d$ evaluations of f, there is no need to allow through nugget an accumulation of points near the best observation.

Note that on both functions, ML estimation of nugget is not a good strategy.

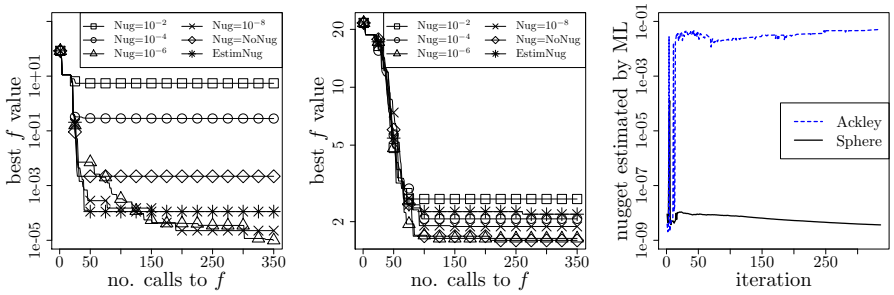

Fig. 6: Median of the best objective function vs. number of calls to f for EGO with different nugget values on the Sphere (left) and Ackley (middle) functions in dimension 5. Right: ML estimated nugget, τ^2, vs. number of calls to f.

5 Concluding remarks

To sum up, this paper carefully explains the DoEs generated by EGO with fixed length-scale and nugget. In terms of performance, ML estimation of the length-scales is a good choice but ML estimation of nugget is not recommended (a fixed small nugget value should be preferred). As a perspective, EGO strategies starting with a large fixed length-scale and then decreasing it while keeping a small amount of nugget should be efficient while avoiding ML estimations which require $O(n^3)$ computations [3].

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

References

1. Andrianakis, I., Challenor, P.G.: The effect of the nugget on Gaussian process emulators of computer models. Computational Statistics & Data Analysis 56(12), 4215–4228 (2012)
2. Benassi, R., Bect, J., Vazquez, E.: Robust gaussian process-based global optimization using a fully bayesian expected improvement criterion. In: Coello, C. (ed.) Learning and Intelligent Optimization, Lecture Notes in Computer Science, vol. 6683, pp. 176–190. Springer Berlin Heidelberg (2011)

3. Cressie, N., Johannesson, G.: Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(1), 209–226 (2008)
4. Jones, D.R.: A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization 21, 345–383 (2001)
5. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning, The MIT Press (2005)
6. Strang, G.: Linear Algebra and Its Applications. Brooks Cole (1988), http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20{\&}path=ASIN/0155510053
7. Wang, Z., Zoghi, M., Hutter, F., Matheson, D., de Freitas, N.: Bayesian optimization in high dimensions via random embeddings. In: International Joint Conferences on Artificial Intelligence (IJCAI) (2013)