Kriging is well-suited to parallelize optimization

David Ginsbourger, Rodolphe Le Riche, and Laurent Carraro

1 Introduction

1.1 Motivations: efficient optimization algorithms for expensive computer experiments

Beyond both estalished frameworks of derivative-based descent and stochastic search algorithms, the rise of expensive optimization problems creates the need for new specific approaches and procedures. The word "expensive" —which refers to price and/or time issues— implies severely restricted budgets in terms of objective function evaluations. Such limitations contrast with the computational burden typically associated with stochastic search techniques, like genetic algorithms. Furthermore, the latter evaluations provide no differential information in a majority of expensive optimization problems, whether the objective function originate from physical or from simulated experiments. Hence there exists a strong motivation for developing derivative-free algorithms, with a particular focus on their optimization performances in a drastically limited number of evaluations. Investigating and implementing adequate strategies constitute a contemporary challenge at the interface between Applied Mathematics and Computational Intelligence, especially when it comes to reducing optimization durations by efficiently taking advantage of parallel computation facilities.

The primary aim of this chapter is to address parallelization issues for the optimization of expensive-to-evaluate simulators, such as increasingly encountered in engineering applications like car crash tests, nuclear safety, or reservoir forecasting. More specifically, the work presented here takes place in the frame of metamodel-based design of computer experiments, in the sense of [42]. Even though the results and discussions might be extended to a more general scope, we restrict ourself here for clarity to single-objective opti-

David Ginsbourger
Département 3MI, Ecole Nationale Supérieure des Mines, 158 cours Fauriel, Saint-Etienne, France
e-mail: ginsbourger@emse.fr

Rodolphe Le Riche
CNRS (UMR 5146) and Département 3MI, Ecole Nationale Supérieure des Mines, 158 cours Fauriel, Saint-Etienne, France
e-mail: leriche@emse.fr

Laurent Carraro
Département 3MI, Ecole Nationale Supérieure des Mines, 158 cours Fauriel, Saint-Etienne, France
e-mail: carraro@emse.fr

mization problems for deterministic codes. The simulator is seen as black-box function y with d-dimensional vector of inputs and scalar output, the latter being often obtained as combination of several responses. *Metamodels*, also called *surrogate models*, are simplified representations of y. They can be used for predicting values of y outside the initial design, or visualizing the influence of each variable on y [27, 43]. They may also guide further sampling decisions for various purposes, such as refining the exploration of the input space in preferential zones or optimizing the function y [22]. Classical surrogates include radial basis functions [37], interpolation splines [52], neural nets [8] (*deterministic* metamodels), or linear and non-linear regression [2], and Kriging [7] (*probabilistic* metamodels). We concentrate here on the advantages of probabilistic metamodels for parallel exploration and optimization, with a particular focus on the virtues of Kriging.

1.2 Where Computational Intelligence and Kriging meet

Computational intelligence (CI) methods share, in various proportions, four features:

An history going from experiments to theory: CI methods very often originate from empirical computing experiments, in particular from experiments that mimick natural processes (e.g., neural networks [4], ant colony optimization [5], simulated annealing [25]). Later on, as researchers use and analyze them, theory develops and their mathematical content grows. A good example is provided by the evolutionary algorithms [9] which have progressively mixed the genetic metaphor and stochastic optimization theory.

An indirect problem representation: In standard evolutionary optimization methods, knowledge about the cost function takes the indirect form of a set of well-performing points, known as "current population". Such set of points is an implicit, partial, representation of a function. In fuzzy methods, the probability density functions of the uncertain variables are averaged out. Such indirect representations enable to work with few mathematical assumptions and have broadened the range of applicability of CI methods.

Parallelized decision process: Most CI approaches are inherently parallel. For example, the evolutionary or particle swarm optimization [24] methods process sets of points in parallel. Neural networks have a internal parallel structure. Today, parallelism is crucial for taking advantage of the increasingly distributed computing capacity. The parallel decision making possibilities are related to the indirect problem representations (through set of points, distributions) and to the use of randomness in the decision process.

Heuristics: Implicit problem representations and the empirical genesis of the CI methods rarely allow mathematical proofs of the methods properties. Most CI methods are thus *heuristics.*

Kriging has recently gained popularity among several research communities related to CI, ranging from *Data Mining* [16] and *Bayesian Statistics* [34, 48] to *Machine Learning* [39], where it is linked to *Gaussian Process Regression* [53] and *Kernel Method* [12]. Recent works [17, 30, 31] illustrate the practical relevance of Kriging to approximate computer codes in application areas such as aerospace engineering or materials science. Indeed, probabilistic metamodels like Kriging seem to be particularly adapted for the optimization of black-box functions, as analyzed and illustrated in the excellent article [20]. The current Chapter is devoted to the optimization of black-box functions using a kriging metamodel [14, 22, 49, 51]. Let us now stress some essential relationships between Kriging and CI by revisiting the above list of features.

A history from field studies to mathematical statistics: Kriging comes from the earth sciences [29, 33], and has been progressively developed since the 1950's along with the discipline called *geostatistics* [23, 32]. Originally aimed at estimating natural ressources in mining applications, it has later been adapted to address very general interpolation and approximation problems [42, 43]. The word "kriging" comes

from the name of a mining engineer, Prof. Daniel G. Krige, who was a pioneer in the application of mathematical statistics to the study of new gold mines using a limited number of boreholes [29].

An indirect representation of the problem: As will be detailed later in the text, the kriging metamodel has a powerful interpretation in terms of stochastic process conditionned by observed data points. The optimized functions are thus indirectly represented by stochastic processes.

Parallelized decision process: The central contribution of this chapter is to propose tools enabling parallelized versions of state-of-the art kriging-based optimization algorithms.

Heuristics: Although the methods discussed here are mathematically founded on the multipoints expected improvement, the maximization of this criterion is not mathematically tractable beyond a few dimensions. In the last part of the chapter, it is replaced by the "kriging believer" and the "constant liar" heuristics.

Through their indirect problem representation, their parallelism and their heuristical nature, the kriging-based optimization methods presented hereafter are Computational Intelligence methods.

1.3 Towards Kriging-based parallel optimization: summary of obtained results and outline of the chapter

This chapter is a follow-up to [14]. It proposes metamodel-based optimization criteria and related algorithms that are well-suited to parallelization since they yield several points at each iteration. The simulations associated with these points can be distributed on different processors, which helps performing the optimization when the simulations are calculation intensive. The algorithms are derived from a multi-points optimization criterion, the *multi-points* or *q-points expected improvement* (q-EI). In particular, an analytic expression is derived for the 2-EI, and consistent statistical estimates relying on Monte-Carlo methods are provided for the general case. All calculations are performed in the framework of Gaussian processes (**GP**). Two classes of heuristic strategies, the *Kriging Believer* (**KB**) and *Constant Liar* (**CL**), are subsequently introduced to obtain approximately q-EI-optimal designs. The latter strategies are tested and compared on a classical test case, where the *Constant Liar* appears to constitute a legitimate heuristic optimizer of the q-EI criterion. Without too much loss of generality, the probabilistic metamodel considered is Ordinary Kriging (**OK**, see eqs. 1,2,35), like in the founder work [22] introducing the now famous **EGO** algorithm. In order to make this document self-contained, non-specialist readers may find an overview of existing criteria for kriging-based sequential optimization in the next pages, as well as a short but dense introduction to GP and OK in the body of the chapter, with complements in appendix. The outline of the chapter is as follows:

- Section 2 (*Background in Kriging for Sequential Optimization*) recalls the OK equations, with a focus on the joint conditional distributions associated with this probabilistic metamodel. A progressive introduction to kriging-based criteria for sequential optimization is then proposed, culminating with the presentation of the EGO algorithm and its obvious limitations in a context of distributed computing.
- Section 3 (*The Multi-points Expected Improvement*) consists in the presentation of the q-EI criterion — continuing the work initiated in [47]—, its explicit calculation when $q = 2$, and the derivation of estimates of the latter criterion in the general case, relying on Monte-Carlo simulations of gaussian vectors.
- Section 4 (*Approximated q-EI maximization*) introduces two heuristic strategies, KB and CL, to circumvent the computational complexity of a direct q-EI maximization. These strategies are tested on a classical test-case, and CL is found to be a very promizing competitor for approximated q-EI maximization

- Section 5 (*Towards Kriging-based Parallel Optimization: Conclusion and Perspectives*) gives a summary of obtained results as well as some related practical recommendations, and finally suggests what the authors think are perspectives of research to address the most urgently in order to extend this work.
- The appendix 6 is a short but dense introduction to GP for machine learning, with an emphasis on the foundations of both Simple Kriging and Ordinary Kriging by GP conditionning.

Some notations: $y : \mathbf{x} \in D \subset \mathbb{R}^d \rightarrow y(\mathbf{x}) \in \mathbb{R}$ refers to the objective function, where $d \in \mathbb{N}\backslash\{0\}$ is the number of input variables and D is the set in which the inputs vary, most of the time assumed to be a compact and connex [1] subset of $\mathbb{R}^d$. At first, y is known at a *Design of Experiments* $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$, where $n \in \mathbb{N}$ is the number of initial runs or experiments, and each $\mathbf{x}^i$ $(1 \leq i \leq n)$ is hence a d-dimensional vector $(\mathbf{x}^i_1, \ldots, \mathbf{x}^i_d)$. We denote by $\mathbf{Y} = \{y(\mathbf{x}^1), ..., y(\mathbf{x}^n)\}$ the set of observations made by evaluating y at the points of $\mathbf{X}$. The data $(\mathbf{X}, \mathbf{Y})$ provides information on which is initially based the metamodeling of y, with an accuracy that depends on n, the geometry of $\mathbf{X}$, and the regularity of y. The OK mean predictor and prediction variance are denoted by the functions $m_{OK}(.)$ and $s^2_{OK}(.)$. The random process implicitely underlying OK is denoted by $Y(.)$, in accordance with the notations of eq. (35) presented in appendix. The symbol "|" is used for conditioning, together with the classical symbols for probability and expectation, respectively $\mathbb{P}$ and $\mathbb{E}$.

2 Background in Kriging for Sequential Optimization

2.1 The Ordinary Kriging metamodel and its Gaussian Process interpretation

OK is the most popular Kriging metamodel, simultaneously due to its great versatility and applicability. It provides a mean predictor of spatial phenomena, with a quantification of the expected prediction accuracy at each site. A full derivation of the OK mean predictor and variance in a GP setting is proposed in the appendix. The corresponding OK mean and variance functions are given by the following formulae:

$$m_{OK}(\mathbf{x}) = \left[\mathbf{c}(\mathbf{x}) + \left(\frac{1 - \mathbf{c}(\mathbf{x})^T \Sigma^{-1} \mathbf{1}_n}{\mathbf{1}_n^T \Sigma^{-1} \mathbf{1}_n}\right) \mathbf{1}_n\right]^T \Sigma^{-1} \mathbf{Y}, \tag{1}$$

$$s^2_{OK}(\mathbf{x}) = \sigma^2 - \mathbf{c}(\mathbf{x})^T \Sigma^{-1} \mathbf{c}(\mathbf{x}) + \frac{(1 - \mathbf{1}_n^T \Sigma^{-1} \mathbf{c}(\mathbf{x}))^2}{\mathbf{1}_n^T \Sigma^{-1} \mathbf{1}_n}, \tag{2}$$

where $\mathbf{c}(x) := \left(c(Y(\mathbf{x}), Y(\mathbf{x}^1)), ..., c(Y(\mathbf{x}), Y(\mathbf{x}^n))\right)^T$, and Σ and σ^2 are defined following the assumptions[2] and notations given in appendix 6. Classical properties of OK include that $\forall i \in [1, n]$ $m_{OK}(\mathbf{x}^i) = y(\mathbf{x}^i)$ and $s^2_{OK}(\mathbf{x}^i) = 0$, therefore $[Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}]$ is interpolating. Note that $[Y(\mathbf{x}^a)|Y(\mathbf{X}) = \mathbf{Y}]$ and $[Y(\mathbf{x}^b)|Y(\mathbf{X}) = \mathbf{Y}]$ are dependent random variables, where $\mathbf{x}^a$ and $\mathbf{x}^b$ are arbitrary points of D, as we will develop later.

The OK metamodel of the Branin-Hoo function (Cf. eq. 25) is plotted on fig. 2.1. The OK interpolation (upper middle) is based only on 9 observations. Even if the shape is reasonably respected (lower middle),

[1] Connexity is sometimes untenable in practical applications, see e.g. [46] for a treatment of disconnected feasible regions.

[2] An extension to covariance non-stationary processes [35] is straightforward but beyond the scope of the present work.

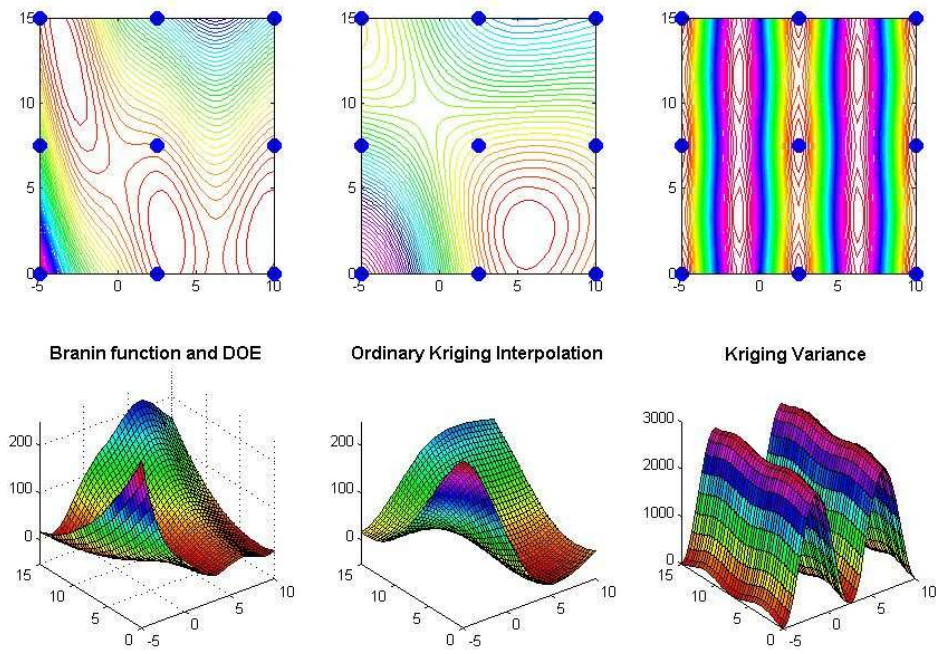

Fig. 1 Ordinary Kriging of the Branin-Hoo function (function, Kriging mean value and variance, from left to right). The design of experiments is a 3×3 factorial design. The covariance is an anisotropic squared exponential with parameters estimated by gaussian likelihood maximization [7].

the contour of the mean shows an artificial optimal zone (upper middle, around the point $(6,2)$). In other respects, the variance is not depending on the observations[3] (eq. 2). Note its particular shape, due to the anisotropy of the covariance kernel estimated by likelihood maximization. In modern interpretations [39], deriving OK equations is often based on the assumption that y is a realization of a random process Y with unknown constant mean and known covariance (see [1] or [12] for a review of classical covariance kernels). Here we follow the derivation of 6.4, which has the advantage of delivering a gaussian posterior distribution:

$$[Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}(m_{OK}(\mathbf{x}), s^2_{OK}(\mathbf{x})) \tag{3}$$

Note that both a structure selection and a parametric estimation are made in practice: one often chooses a generalized exponential kernel with plugged-in maximum likelihood covariance hyperparameters, i.e. without taking the estimation variance into account [22]. This issue is sometimes addressed using a full bayesian treatment, as can be found in [43], or more recently in [15, 34, 39]. Rephrasing 3, under the latter GP as-

[3] phenomenon known as homoskedasticity of the Kriging variance with respect to the observations [7]

sumptions, the random variable $Y(\mathbf{x})$ knowing the values of $\{y(\mathbf{x^1}),...,y(\mathbf{x^n})\}$ follows a gaussian distribution which mean and variance are respectively $\mathbb{E}[Y(\mathbf{x})|Y(\mathbf{X})=\mathbf{Y}]=m_{OK}(\mathbf{x})$ and $Var[Y(\mathbf{x})|Y(\mathbf{X})=\mathbf{Y}]=s^2_{OK}(\mathbf{x})$. In fact, as shown in appendix (Cf. eq. 38), one can even get much more than these marginal conditional distributions; $Y(.)|Y(\mathbf{X})=\mathbf{Y}$ constitutes a random process which is itself gaussian, and as such completely characterized by its conditional mean, m_{OK}, and conditional covariance kernel c_{OK} explicited herunder:

$$[Y(.)|Y(\mathbf{X})=\mathbf{Y}] \sim GP(m_{OK}(.),c_{OK}(.,.)), \tag{4}$$

$$\text{where } c_{OK}(\mathbf{x},\mathbf{x}')=c(\mathbf{x}-\mathbf{x}')-\mathbf{c}(\mathbf{x})^T\boldsymbol{\Sigma}^{-1}\mathbf{c}(\mathbf{x}')+\sigma^2\left[\frac{(1-\mathbf{1}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{c}(\mathbf{x}))(1-\mathbf{1}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{c}(\mathbf{x}'))}{\mathbf{1}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_n}\right]. \tag{5}$$

This new kernel c_{OK} is not stationary, even if c is. In other respects, the knowledge of m_{OK} and c_{OK} is the first step to performing conditional simulations of Y knowing the observations $Y(\mathbf{X})=\mathbf{Y}$, which is easily feasible at any new finite design of experiments, whatever the dimension of inputs. This will enable the computation of any multi-points sampling criterion, such as proposed in the forthcoming section about parallelization.

2.2 Kriging-based optimization criteria

GP metamodels [39, 53] such as OK has been used for optimization (minimization, by default). There is a detailed review of optimization methods relying on a metamodel in [44, 45] or [20]. The latter analyzes why directly optimizing a deterministic metamodel (like a spline, a polynomial, or the kriging mean) is dangerous, and does not even necessarily lead to a local optimum. Kriging-based sequential optimization strategies (as developed in [22], and commented in [20]) address the issue of converging to non (locally) optimal points, by taking the kriging variance term into account (hence encouraging the algorithms to explore outside the already visited zones). Such algorithms produce one point at each iteration that maximizes a figure of merit based upon $[Y(\mathbf{x})|Y(\mathbf{X})=\mathbf{Y}]$. In essence, the criteria balance kriging mean prediction and uncertainty.

2.2.1 Visiting the point with most promizing mean: minizing m_{OK}

When approximating y by m_{OK}, it might seem natural to hope that minimizing m_{OK} instead of y brings satisfying results. However, a function and its approximation (m_{OK} or other) can show substantial differences in terms of optimal values and optimizers. More specifically, depending on the kind of covariance kernel used in OK, the minimizer of m_{OK} is susceptible to lie at (or near to) the design point with minimal y value. Taking the geometry of the design of experiments and space-filling considerations into account within exploration criteria then makes sense. The Kriging variance can be of providential help for this purpose.

2.2.2 Visiting the point with highest uncertainty: maximizing s_{OK}

A fundamental mistake of minimizing m_{OK} is that no account is done of the uncertainty associated with it. At the extreme inverse, it is possible to define the next optimization iterate as the least known point in D,

$$\mathbf{x}' = argmax_{\mathbf{x}\in D}\, s_{OK}(\mathbf{x}) \tag{6}$$

This procedure defines a series of $\mathbf{x}'$s which will fill the space D and hence ultimately locate a global optimum. Yet, since no use is made of previously obtained $\mathbf{Y}$ information —look at formula 2 for s^2_{OK}—, there is no bias in favor of high performance regions. Maximizing the uncertainty is inefficient in practice.

2.2.3 Compromizing between m_{OK} and s_{OK}

The most general formulation for compromizing between the exploitation of previous simulations brought by m_{OK} and the exploration based on s_{OK} is the multicriteria problem

$$\begin{cases} \min_{\mathbf{x}\in D} m_{OK}(\mathbf{x}) \\ \max_{\mathbf{x}\in D} s_{OK}(\mathbf{x}) \end{cases} \tag{7}$$

Let $\mathscr{P}$ denote the Pareto set of solutions [4]. Finding one (or many) elements in $\mathscr{P}$ remains a difficult problem since $\mathscr{P}$ typically contains an infinite number of points. A comparable approach called *direct*, although not based on OK, is described in [21]: the metamodel is piecewise linear and the uncertainty measure is a distance to already known points. The space D is discretized and the Pareto optimal set defines areas where discretization is refined. The method becomes computationally expensive as the number of iterations and dimensions increase. Note that [3] proposes several parallelized versions of *direct*.

2.2.4 Maximizing the probability of improvement

Among the numerous criteria presented in [20], the probability of getting an improvement of the function with respect to the past evaluations seems to be one of the most fundamental. This function is defined for every $\mathbf{x} \in D$ as the probability for the random variable $Y(\mathbf{x})$ to be below the currently known minimum $min(\mathbf{Y}) = \min\{y(\mathbf{x}^1), ..., y(\mathbf{x}^n)\}$ conditional on the observations at the design of experiments:

$$PI(\mathbf{x}) := P\left(Y(\mathbf{x}) \leq \min(Y(\mathbf{X}))|Y(\mathbf{X}) = \mathbf{Y}\right) \tag{8}$$

$$= \mathbb{E}\left[\mathbf{1}_{Y(\mathbf{x})\leq \min(Y(\mathbf{X}))}|Y(\mathbf{X}) = \mathbf{Y}\right] = \Phi\left(\frac{\min(\mathbf{Y}) - m_{OK}(\mathbf{x})}{s_{OK}(\mathbf{x})}\right), \tag{9}$$

where Φ is the gaussian cumulative distribution function, and the last equality follows 3. The threshold $\min(\mathbf{Y})$ is sometimes replaced by some arbitrary target $T \in \mathbb{R}$, as evokated in [38]. PI is known to provide a very local search whenever the value of T is equal or close to $min(\mathbf{Y})$. Taking several T's is a remedy proposed by [20] to force global exploration. Of course, this new degree of freedom is also one more parameter to fit. In other respects, PI has also been succesfully used as pre-selection criterion in GP-assisted evolution strategies [49], where it was pointed out that PI is performant but has a tendency to sample in unexplored areas. We argue that the chosen covariance structure plays a capital role in such matters, depending whether the kriging mean is overshooting the observations or not. The next presented criterion, the *expected improvement*, is less sensitive to such issues since it explicitly integrates both kriging mean and variance.

[4] Definition of the Pareto front of $(s_{OK}, -m_{OK})$: $\forall x \in \mathscr{P}, \nexists\, z \in D : (m_{OK}(z) < m_{OK}(x)$ and $s_{OK}(z) \geq s_{OK}(x))$ or $(m_{OK}(z) \leq m_{OK}(x)$ and $s_{OK}(z) > s_{OK}(x))$

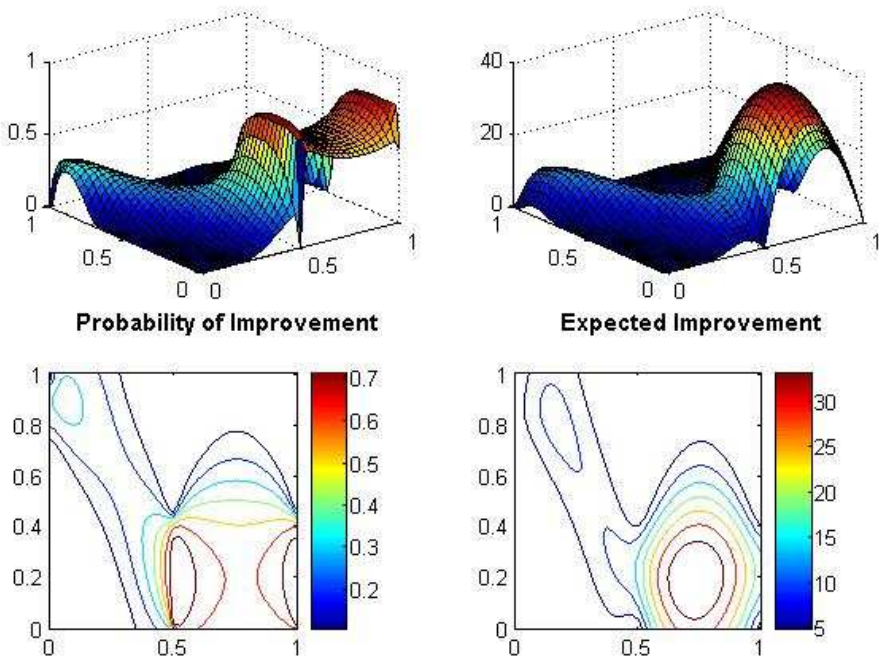

Fig. 2 PI and EI surfaces of the Branin-Hoo function (same design of experiments, Kriging model, and covariance parameters as in fig. (2.1)). Maximizing PI leads to sample near the good points (associated with low observations) whereas maximizing EI leads here to sample between the good points. By construction, both criteria are null at the design of experiments, but the probability of improvement is very close to $\frac{1}{2}$ in a neighborhood of the point(s) where the function takes its current minimum.

2.2.5 Maximizing the expected improvement

An alternative solution is to maximize the *expected improvement* (EI),

$$EI(\mathbf{x}) := \mathbb{E}[(\min(Y(\mathbf{X}) - Y(\mathbf{x}))^+ | Y(\mathbf{X}) = \mathbf{Y}] = \mathbb{E}[\max\{0, \min(Y(\mathbf{X}) - Y(\mathbf{x})\} | Y(\mathbf{X}) = \mathbf{Y}], \qquad (10)$$

that additionally takes into account the magnitude of the improvements. EI measures how much improvement is expected by sampling at $\mathbf{x}$. *In fine*, the improvement will be 0 if $y(\mathbf{x})$ is above $min(\mathbf{Y})$ and $min(\mathbf{Y}) - y(\mathbf{x})$ else. Knowing the conditional distribution of $Y(\mathbf{x})$, it is straightforward to calculate EI in closed form:

$$EI(\mathbf{x}) = (min(\mathbf{Y}) - m_{OK}(\mathbf{x}))\Phi\left(\frac{min(\mathbf{Y}) - m_{OK}(\mathbf{x})}{s_{OK}(\mathbf{x})}\right) + s_{OK}(\mathbf{x})\phi\left(\frac{min(\mathbf{Y}) - m_{OK}(\mathbf{x})}{s_{OK}(\mathbf{x})}\right), \qquad (11)$$

where ϕ stands for the probability density function of the standard normal law $\mathcal{N}(0,1)$.

Proof of 11: $EI(\mathbf{x}) = \mathbb{E}[(min(\mathbf{Y}) - Y(\mathbf{x}))_{\ Y(\mathbf{x}) \leq min(\mathbf{Y})} | Y(\mathbf{X}) = \mathbf{Y}]$

$$= \int_{-\infty}^{min(\mathbf{Y})} (min(\mathbf{Y}) - t)\, f_{\mathcal{N}(m_{KO}(\mathbf{x}), s^2_{KO}(\mathbf{x}))}(t)dt = \int_{-\infty}^{\frac{min(\mathbf{Y}) - m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}} (min(\mathbf{Y}) - m_{KO}(\mathbf{x}) - s_{KO}(\mathbf{x}) \times u)\, f_{\mathcal{N}(0,1)}(u)du$$

$$= (min(\mathbf{Y}) - m_{KO}(\mathbf{x})) \int_{-\infty}^{\frac{min(\mathbf{Y}) - m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}} f_{\mathcal{N}(0,1)}(u)du - s_{KO}(\mathbf{x}) \int_{-\infty}^{\frac{min(\mathbf{Y}) - m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}} u \times f_{\mathcal{N}(0,1)}(u)du$$

$$= (min(\mathbf{Y}) - m_{KO}(\mathbf{x}))\, \Phi\left(\frac{min(\mathbf{Y}) - m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}\right) + s_{KO}(\mathbf{x})\phi\left(\frac{min(\mathbf{Y}) - m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}\right)$$

EI represents a trade-off between promising and uncertain zones. This criterion has important properties for sequential exploration: it is null at the already visited sites, and positive everywhere else with a magnitude that is increasing with the Kriging variance and with the decreasing Kriging mean (EI maximizers are indeed part of the Pareto front of $(s_{OK}, -m_{OK})$). Such features are usually demanded from global optimization procedures (see [21] for instance). EI and the probability of improvement are compared in fig. (2).

2.2.6 The *Stepwise Uncertainty Reduction* (SUR) strategy

SUR has been introduced in [11] and extended to global optimization in [50, 51]. By modeling y using the process Y's conditional law $Y(\mathbf{x})|\mathbf{Y}$, it is possible to define $\mathbf{x}^*|\mathbf{Y}$, the conditional law of Y's global minimizer $\mathbf{x}^*$, and its density $p_{\mathbf{x}^*|\mathbf{Y}}(\mathbf{x})$. The uncertainty about the location of $\mathbf{x}^*$ is measured as the entropy of $p_{\mathbf{x}^*|\mathbf{Y}}(\mathbf{x})$, $H(\mathbf{x}^*|\mathbf{Y})$. $H(\mathbf{x}^*|\mathbf{Y})$ diminishes as the distribution of $\mathbf{x}^*|\mathbf{Y}$ gets more peaked. Conceptually, the SUR strategy for global optimization chooses as next iterate the point that specifies the most the location of the optimum,

$$\mathbf{x}' = argmin_{\mathbf{x} \in D} H(\mathbf{x}^*|\mathbf{Y}, Y(\mathbf{x})) \tag{12}$$

In practice, $p_{\mathbf{x}^*|\mathbf{Y}}(\mathbf{x})$ is estimated by Monte-Carlo sampling of $Y(\mathbf{x})|\mathbf{Y}$ at a finite number of locations in D, which may become a problem in high dimensional D's as the number of locations must geometrically increase with d to properly fill the space. The SUR criterion is different in nature from the criteria presented so far in that it does not maximize an immediate (i.e. at the next iteration) payoff but rather lays the foundation of a delayed payoff by gaining a more global knowledge on Y (reduce the entropy of its optima). The multi-points EI criterion we are focusing on in the present chapter also uses a delayed payoff measure.

2.2.7 The *Efficient Global Optimization* (EGO) algorithm

EGO [22] relies on the EI criterion. Starting with an initial Design $\mathbf{X}$ (typically a Latin Hypercube), EGO sequentially visits the current global maximizer of EI (say the first visited one if there is more than one global maximizer) and updates the OK metamodel at each iteration, including hyperparameters re-estimation:

1. Evaluate y at $\mathbf{X}$, set $\mathbf{Y} = y(\mathbf{X})$ and estimate covariance parameters of Y by MLE (Maximum Likelihood Estimation)
2. While stopping criterion not met
   a. Compute $\mathbf{x}' = argmax_{\mathbf{x} \in D} EI(\mathbf{x})$, set $\mathbf{X} = \mathbf{X} \cup \{\mathbf{x}'\}$ and $\mathbf{Y} = \mathbf{Y} \cup \{y(\mathbf{x}')\}$
   b. Re-estimate covariance parameters by MLE

After having been developed in [22, 47], EGO has inspired contemporary works in optimization of expensive-to-evaluate functions. For instance, [19] exposes some EGO-based methods for the optimization of noisy black-box functions like stochastic simulators. [18] focuses on multiple numerical simulators with different levels of fidelity, and introduces the so-called *augmented EI* criterion, integrating possible heterogeneity in the simulation times. Moreover, [26] proposes an adaptation to multi-objective optimization, [17] proposes an original multi-objective adaptation of EGO for physical experiments, and [28] focuses on robust criteria for multiobjective constrained optimization with applications to laminating processes.

In all, one major drawback of the EGO-like algorithms discussed so far is that they do not allow parallel evaluations of y, which is desirable for costly simulators (e.g. a crash-test simulation run typically lasts 24 hours). This was already pointed out in [47], where the multi-points EI was defined but not further developed. Here we continue this work by expliciting the latter multi-points EI (q-EI), and by proposing two classes of heuristics strategies meant to approximatly optimize the q-EI, and hence (almost) simultaneously deliver an arbitrary number of points without intermediate evaluations of y. In particular, we analytically derive the 2-EI, and explain in detail how to take advantage of statistical interpretations of Kriging to consistently compute q-EI by simulation when $q > 2$, which happens to provide quite a general template for desiging Kriging-based parallel evaluation strategies dedicated to optimization or other purposes.

3 The Multi-points Expected Improvement (q-EI) Criterion

The main objective of the present work is to analyze and then approximately optimize a global optimization criterion, the q-EI, that yields q points. Since q-EI is an extension of EI, all derivations are performed within the framework of OK. Such criterion is the first step towards a parallelized version of the EGO algorithm [22]. It also departs, like the SUR criterion, from other criteria that look for an immediate payoff. We now propose a progressive construction of the q-EI, by coming back to the random variable *improvement*.

Both criteria of PI and EI that we have previously recalled share indeed the feature of being conditional expectations of quantities involving the *improvement*. The *improvement* brought by sampling at some $\mathbf{x} \in D$ is indeed defined by $I(\mathbf{x}) := (min(Y(\mathbf{X})) - Y(\mathbf{x}))^+$, and is positive whenever the value sampled at $\mathbf{x}$, $Y(\mathbf{x})$, is below the current minimum $min(Y(\mathbf{X}))$. Now, if we sample Y at q new locations $\mathbf{x}^{n+1}, \ldots, \mathbf{x}^{n+q} \in D$ simultaneously, it seems quite natural to define the joint —or *multipoints*— improvement as follows:

$$\begin{aligned} \forall \mathbf{x}^{n+1}, \ldots, \mathbf{x}^{n+q} \in D,\ I(\mathbf{x}^{n+1}, \ldots, \mathbf{x}^{n+q}) &:= \max\left(I(\mathbf{x}^{n+1}), \ldots, I(\mathbf{x}^{n+q})\right) \\ &= \max\left((min(Y(\mathbf{X})) - Y(\mathbf{x}^{n+1}))^+, \ldots, (min(Y(\mathbf{X})) - Y(\mathbf{x}^{n+q}))^+\right) \\ &= \left(\min(Y(\mathbf{X})) - \min(Y(\mathbf{x}^{n+1}), \ldots, Y(\mathbf{x}^{n+q}))\right)^+, \end{aligned} \tag{13}$$

where we used the fact that $\forall a, b, c \in \mathbb{R},\ \max((a-b)^+, (a-c)^+) = (a-b)^+$ if $b \leq c$ and $(a-c)^+$ else. The way of unifying the q criteria of (1-point) improvements used in eq. 13 deserves to be called *elitist*: one juges the quality of the set of q-points as a function only of the one that performs the best. This is to be compared for instance to the weighted sums of criteria encountered in many political science applications.

The q-points EI criterion (as already defined but not developed in [47] under the name "q-step EI") is then straightforwardly defined as conditional expectation of the improvement brought by the q considered points:

$$\begin{aligned} EI(\mathbf{x}^{n+1},...,\mathbf{x}^{n+q}) :&= \mathbb{E}[\max\{(min(Y(\mathbf{X})) - Y(\mathbf{x}^{n+1}))^+,...,(min(\mathbf{Y}) - Y(\mathbf{x}^{n+q}))^+\}|Y(\mathbf{X}) = \mathbf{Y}] \\ &= \mathbb{E}\left[(\min(Y(\mathbf{X})) - \min\left(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+q})\right))^+|Y(\mathbf{X}) = \mathbf{Y}\right] \\ &= \mathbb{E}\left[(\min(\mathbf{Y}) - \min\left(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+q})\right))^+|Y(\mathbf{X}) = \mathbf{Y}\right] \end{aligned} \tag{14}$$

Hence, the q-EI may be seen as the regular EI applied to the random variable $\min(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+q}))$. We thus have to deal with a minimum of dependent random variables. Fortunately, eq. 4 provides us with the exact joint distribution of the q unknown responses conditional on the observations:

$$[(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+q}))|Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}((m_{OK}(\mathbf{x}^{n+1}),...,m_{OK}(\mathbf{x}^{n+q})),S_q) \tag{15}$$

where the elements of the conditional covariance matrix S_q are $(S_q)_{i,j} = c_{OK}(\mathbf{x}^{n+i},\mathbf{x}^{n+j})$ (See eq. 5). We now propose two different ways to evaluate the criterion eq. 14, depending whether $q = 2$ or $q \geq 3$.

3.1 Analytical calculation of 2-EI

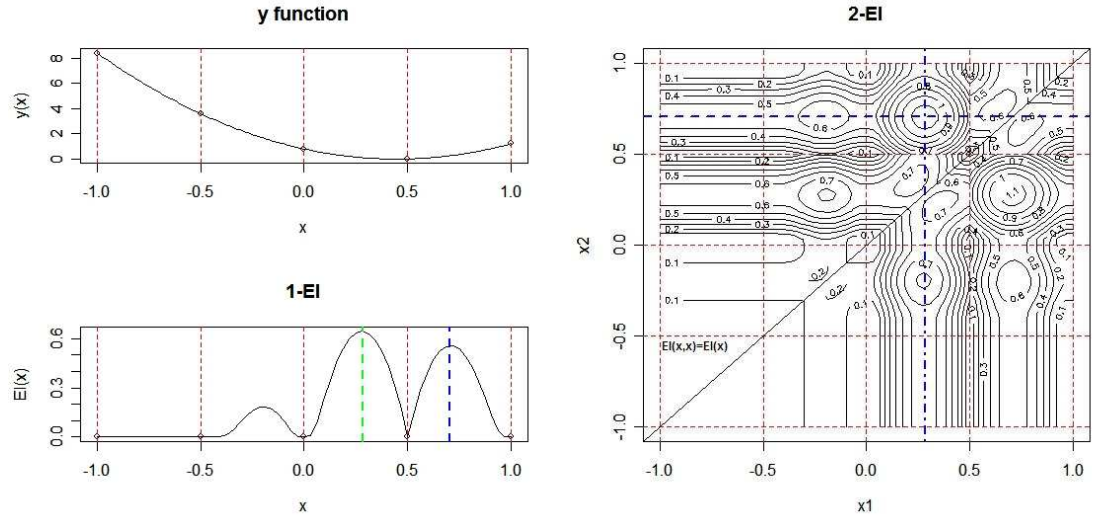

Fig. 3 1-EI (lower left) and 2-EI (right) functions associated with a monodimensional quadratic function ($y(x) = 4 \times (x - 0.45)^2$ known at $\mathbf{X} = \{-1, -0.5, 0, 0.5, 1\}$. The OK metamodel has here a cubic covariance with parameters $\sigma^2 = 10,$ scale $= 0.9$).

We first focus on the calculation of the 2-EI associated with two arbitrary points $\mathbf{x}^{n+1}, \mathbf{x}^{n+2} \in D$, defined as

$$EI(\mathbf{x}^{n+1}, \mathbf{x}^{n+2}) := \mathbb{E}[(\min(Y(\mathbf{X})) - min(Y(\mathbf{x}^{n+1}), Y(\mathbf{x}^{n+2})))^+ | Y(\mathbf{X}) = \mathbf{Y}],$$

Let us remark that in reformulating the positive part function, the expression above can also be written:

$$EI(\mathbf{x}^{n+1}, \mathbf{x}^{n+2}) = \mathbb{E}[(\min(Y(\mathbf{X})) - min(Y(\mathbf{x}^{n+1}), Y(\mathbf{x}^{n+2})))\mathbf{1}_{min(Y(\mathbf{x}^{n+1}), Y(\mathbf{x}^{n+2})) \leq \min(\mathbf{Y})} | Y(\mathbf{X}) = \mathbf{Y}].$$

We will now show that the 2-EI can be developed as a sum of two 1-EI's, plus a correction term involving 1- and 2-dimensional gaussian cumulative distributions.

Before all, some classical results of conditional calculus allow us to precise the dependence between $Y(\mathbf{x}^{n+1})$ and $Y(\mathbf{x}^{n+2})$ conditional on $Y(\mathbf{X}) = \mathbf{Y}$, and to fix some additional notations. $\forall i, j \in \{1,2\}$ $(i \neq j)$, we note:

$$\begin{cases} m_i := m_{KO}(\mathbf{x}^i) = \mathbb{E}[Y(\mathbf{x}^{n+i}) | Y(\mathbf{X}) = \mathbf{Y}], \\ \sigma_i := s_{KO}(\mathbf{x}^{n+i}) = \sqrt{Var[Y(\mathbf{x}^{n+i}) | Y(\mathbf{X}) = \mathbf{Y}]}, \\ c_{1,2} := \rho_{1,2}\sigma_1\sigma_2 := cov[Y(\mathbf{x}^{n+1}), Y(\mathbf{x}^{n+2}) | Y(\mathbf{X}) = \mathbf{Y}], \\ m_{i|j} = \mathbb{E}[Y(\mathbf{x}^{n+i}) | Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{x}^{n+j}))] = m_i + c_{1,2}\sigma_i^{-2}(Y(\mathbf{x}^{n+j}) - m_j), \\ \sigma^2_{i|j} = \sigma_i^2 - c_{1,2}\sigma_j^{-2} = \sigma_i^2(1 - \rho_{12}^2). \end{cases} \tag{16}$$

At this stage we are in position to compute $EI(\mathbf{x}^{n+1}, \mathbf{x}^{n+2})$ in four steps. From now on, we replace the complete notation $Y(\mathbf{x}^{n+i})$ by Y_i and forget the conditioning on $Y(\mathbf{X}) = \mathbf{Y}$ for the sake of clarity.

Step 1.

$$\begin{aligned} EI(\mathbf{x}^{n+1}, \mathbf{x}^{n+2}) &= \mathbb{E}[(min(\mathbf{Y}) - min(Y_1, Y_2))\mathbf{1}_{min(Y_1,Y_2) \leq min(\mathbf{Y})}] \\ &= \mathbb{E}[(min(\mathbf{Y}) - min(Y_1, Y_2))\mathbf{1}_{min(Y_1,Y_2) \leq min(\mathbf{Y})}(\mathbf{1}_{Y_1 \leq Y_2} + \mathbf{1}_{Y_2 \leq Y_1})] \\ &= \mathbb{E}[(min(\mathbf{Y}) - Y_1)\mathbf{1}_{Y_1 \leq min(\mathbf{Y})}\mathbf{1}_{Y_1 \leq Y_2}] + E[(min(\mathbf{Y}) - Y_2)\mathbf{1}_{Y_2 \leq min(\mathbf{Y})}\mathbf{1}_{Y_2 \leq Y_1}] \end{aligned}$$

Since both terms of the last sum are similar (up to a permutation between $\mathbf{x}^{n+1}$ and $\mathbf{x}^{n+2}$), we will restrict our attention to the first one. Using $\mathbf{1}_{Y_1 \leq Y_2} = 1 - \mathbf{1}_{Y_2 \leq Y_1}$ [5] , we get:

$$\begin{aligned} \mathbb{E}[(min(\mathbf{Y}) - Y_1)\mathbf{1}_{Y_1 \leq min(\mathbf{Y})}\mathbf{1}_{Y_1 \leq Y_2}] &= \mathbb{E}[(min(\mathbf{Y}) - Y_1)\mathbf{1}_{Y_1 \leq min(\mathbf{Y})}(1 - \mathbf{1}_{Y_2 \leq Y_1})] \\ &= EI(\mathbf{x}^{n+1}) - \mathbb{E}[(min(\mathbf{Y}) - Y_1)\mathbf{1}_{Y_1 \leq min(\mathbf{Y})}\mathbf{1}_{Y_2 \leq Y_1}] \\ &= EI(\mathbf{x}^{n+1}) + B(\mathbf{x}^{n+1}, \mathbf{x}^{n+2}) \end{aligned}$$

where $B(\mathbf{x}^{n+1}, \mathbf{x}^{n+2}) = E[(Y_1 - min(\mathbf{Y}))\mathbf{1}_{Y_1 \leq min(\mathbf{Y})}\mathbf{1}_{Y_2 \leq Y_1}]$. Informally, $B(\mathbf{x}^{n+1}, \mathbf{x}^{n+2})$ is the opposite of the improvement brought by Y_1 when $Y_2 \leq Y_1$ and hence that doesn't contribute to the 2-points improvement. Our aim in the next steps will be to give an explicit expression for $B(\mathbf{x}^{n+1}, \mathbf{x}^{n+2})$.

Step 2.

[5] This expression should be noted $1 - \mathbf{1}_{Y_2 < Y_1}$, but since we work with continous random variables, it sufficies that their correlation is $\neq 1$ for the expression to be exact ($\{Y_1 = Y_2\}$) is then neglectable). We implicitely do this assumption in the following.

$$B(\mathbf{x}^{n+1},\mathbf{x}^{n+2}) = \mathbb{E}[Y_1\mathbf{1}_{Y_1\leq min(\mathbf{Y})}\mathbf{1}_{Y_2\leq Y_1}] - min(\mathbf{Y})\mathbb{E}[\mathbf{1}_{Y_1\leq min(\mathbf{Y})}\mathbf{1}_{Y_2\leq Y_1}]$$

At this point, it is worth noticing that $Y_1 \overset{\mathscr{L}}{=} m_1 + \sigma_1 N_1$ (always conditional on $Y(\mathbf{X}) = \mathbf{Y}$) with $N_1 \sim \mathscr{N}(0,1)$. Substituing this decomposition in the last expression of $B(\mathbf{x}^{n+1},\mathbf{x}^{n+2})$ delivers:

$$B(\mathbf{x}^{n+1},\mathbf{x}^{n+2}) = \sigma_1\mathbb{E}[N_1\mathbf{1}_{Y_1\leq min(\mathbf{Y})}\mathbf{1}_{Y_2\leq Y_1}] + (m_1 - min(\mathbf{Y}))\mathbb{E}[\mathbf{1}_{Y_1\leq min(\mathbf{Y})}\mathbf{1}_{Y_2\leq Y_1}]$$

The two terms of this sum require some attention. We compute them in detail in the two next steps.

Step 3. Using a key property of conditional calculus [6], we obtain

$$\mathbb{E}[N_1\mathbf{1}_{Y_1\leq min(\mathbf{Y})}\mathbf{1}_{Y_2\leq Y_1}] = \mathbb{E}[N_1\mathbf{1}_{Y_1\leq min(\mathbf{Y})}\mathbb{E}[\mathbf{1}_{Y_2\leq Y_1}|Y_1]],$$

and the fact that $Y_2|Y_1 \sim \mathscr{N}(m_{2|1}(Y_1), s^2_{2|1}(Y_1))$ (all conditional on the observations) leads to the following:

$$\mathbb{E}[\mathbf{1}_{Y_2\leq Y_1}|Y_1] = \Phi\left(\frac{Y_1 - m_{2|1}}{s_{2|1}}\right) = \Phi\left(\frac{Y_1 - m_2 - \frac{c_{1,2}}{\sigma_1^2}(Y_1 - m_1)}{\sigma_2\sqrt{1-\rho_{12}^2}}\right)$$

Back to the main term and using again the normal decomposition of Y_1, we get:

$$\mathbb{E}\left[N_1\mathbf{1}_{Y_1\leq min(\mathbf{Y})}1_{Y_2\leq Y_1}\right] = \left[N_1 1_{N_1\leq\frac{min(\mathbf{Y})-m_1}{\sigma_1}}\,\Phi\left(\frac{m_1 - m_2 + (\sigma_1 - \rho_{12}\sigma_2)N_1}{\sigma_2\sqrt{1-\rho_{12}^2}}\right)\right] = \mathbb{E}\left[N_1\mathbf{1}_{N_1\leq\gamma_1}\Phi(\alpha_1 N_1 + \beta_1)\right]$$

$$\text{where } \gamma_1 = \frac{min(\mathbf{Y}) - m_1}{\sigma_1},\ \beta_1 = \frac{m_1 - m_2}{\sigma_2\sqrt{1-\rho_{12}^2}} \text{ and } \alpha_1 = \frac{\sigma_1 - \rho_{12}\sigma_2}{\sigma_2\sqrt{1-\rho_{12}^2}} \tag{17}$$

$\mathbb{E}[N_1\mathbf{1}_{N_1\leq\gamma_1}\Phi(\alpha_1 N_1 + \beta_1)]$ can be computed applying an integration by parts:

$$\int_{-\infty}^{\gamma_1} u\phi(u)\Phi(\alpha_1 u + \beta_1)du = -\phi(\gamma_1)\Phi(\alpha_1\gamma_1 + \beta_1) + \frac{\alpha_1}{2\pi}\int_{-\infty}^{\gamma_1} e^{\frac{-u^2-(\alpha_1 u+\beta_1)^2}{2}}du$$

And since $u^2 + (\alpha_1 u + \beta_1)^2 = \left(\sqrt{(1+\alpha_1^2)}u + \frac{\alpha_1\beta_1}{\sqrt{1+\alpha_1^2}}\right)^2 + \frac{\beta_1^2}{1+\alpha_1^2}$, the last integral reduces to:

$$\sqrt{2\pi}\phi\left(\sqrt{\frac{\beta_1^2}{1+\alpha_1^2}}\right)\int_{-\infty}^{\gamma_1} e^{\frac{-\left(\sqrt{(1+\alpha_1^2)}u + \frac{\alpha_1\beta_1}{\sqrt{1+\alpha_1^2}}\right)^2}{2}}du = \frac{2\pi\phi\left(\sqrt{\frac{\beta_1^2}{1+\alpha_1^2}}\right)}{\sqrt{(1+\alpha_1^2)}}\int_{-\infty}^{\sqrt{(1+\alpha_1^2)}\gamma_1 + \frac{\alpha_1\beta_1}{\sqrt{1+\alpha_1^2}}}\frac{e^{\frac{-v^2}{2}}}{\sqrt{2\pi}}dv$$

We conclude in using the definition of the cumulative distribution function:

[6] For all function ϕ in $\mathscr{L}^2(\mathbb{R},\mathbb{R})$, $E[X\phi(Y)] = E[E[X|Y]\phi(Y)]$

$$\mathbb{E}[N_1 \mathbf{1}_{Y_1 \leq min(\mathbf{Y})} 1_{Y_2 \leq Y_1}] = -\phi(\gamma_1)\Phi(\alpha_1\gamma_1+\beta_1) + \frac{\alpha_1 \phi\left(\sqrt{\frac{\beta_1^2}{1+\alpha_1^2}}\right)}{\sqrt{(1+\alpha_1^2)}} \Phi\left(\sqrt{(1+\alpha_1^2)}\gamma_1 + \frac{\alpha_1\beta_1}{\sqrt{1+\alpha_1^2}}\right)$$

Step 4. We then compute the term $\mathbb{E}[\mathbf{1}_{Y_1 \leq min(\mathbf{Y})}\mathbf{1}_{Y_2 \leq Y_1}] = E[\mathbf{1}_{X \leq min(\mathbf{Y})}\mathbf{1}_{Z \leq 0}]$, where $(X,Z) := (Y_1, Y_2 - Y_1)$ follows a 2-dimensional gaussian distribution with expectation $M = (m_1, m_2 - m_1)$, and covariance matrix $\Gamma := \begin{pmatrix} \sigma_1^2 & c_{1,2} - \sigma_1^2 \\ c_{1,2} - \sigma_1^2 & \sigma_2^2 + \sigma_1^2 - 2c_{1,2} \end{pmatrix}$. The final results rely on the fact that: $\mathbb{E}[\mathbf{1}_{X \leq min(\mathbf{Y})}\mathbf{1}_{Z \leq 0}] = CDF(M, \Gamma)(min(\mathbf{Y}), 0)$, where CDF stands for the bi-gaussian cumulative distribution function:

$$EI(\mathbf{x}^1, \mathbf{x}^2) = EI(\mathbf{x}^1) + EI(\mathbf{x}^2) + B(\mathbf{x}^1, \mathbf{x}^2) + B(\mathbf{x}^2, \mathbf{x}^1) \tag{18}$$

$$\text{where } \begin{cases} B(\mathbf{x}^1, \mathbf{x}^2) = (m_{OK}(\mathbf{x}^1) - min(\mathbf{Y}))\delta(\mathbf{x}^1, \mathbf{x}^2) + \sigma_{OK}(\mathbf{x}^1)\varepsilon(\mathbf{x}^1, \mathbf{x}^2) \\ \varepsilon(\mathbf{x}^1, \mathbf{x}^2) = \alpha_1 \phi\left(\frac{|\beta_1|}{\sqrt{(1+\alpha_1^2)}}\right) \Phi\left((1+\alpha_1^2)^{\frac{1}{2}}\left(\gamma + \frac{\alpha_1\beta_1}{1+\alpha_1^2}\right)\right) - \phi(\gamma)\Phi(\alpha_1\gamma + \beta_1) \\ \delta(\mathbf{x}^1, \mathbf{x}^2) = CDF(\Gamma)\begin{pmatrix} min(\mathbf{Y}) - m_1 \\ m_1 - m_2 \end{pmatrix} \end{cases}$$

Figure 3.1 represents the 1-EI and the 2-EI contour plots associated with a deterministic polynomial function known at 5 points. 1-EI advises here to sample between the "good points" of $\mathbf{X}$. The 2-EI contour illustrates some general properties: 2-EI is symmetric and its diagonal equals 1-EI, what can be easily seen by coming back to the definitions. Roughly said, 2-EI is high whenever the 2 points have high 1-EI and are reasonably distant from another (precisely, in the sense of the metric used in OK). Additionally, maximizing 2-EI selects here the two best local optima of 1-EI ($x_1 = 0.3$ and $x_2 = 0.7$). This is not a general fact. The next example illustrates for instance how 2-EI maximization can yield two points located around (but different from) 1-EI's global optimum whenever 1-EI has one single peak of great magnitude (see fig. 4).

3.2 q-EI computation by Monte Carlo Simulations

Extrapolating the calculation of 2-EI to the general case gives complex expressions depending on q-dimensional gaussian cdf's. Hence, it seems that the computation of q-EI when q is large would have to rely on numerical multivariate integral approximation techniques anyway. Therefore, directly evaluating q-EI by Monte-Carlo Simulation makes sense. Thanks to eq. 15, the random vector $(Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+q}))$ can be simulated conitional on $Y(\mathbf{X}) = \mathbf{Y}$ using a decomposition (e.g. Mahalanobis) of the covariance matrix S_q:

$$\forall k \in [1, n_{sim}],\ M_k = (m_{OK}(\mathbf{x}^{n+1}), ..., m_{OK}(\mathbf{x}^{n+q})) + [S_q^{\frac{1}{2}} N_k]^T, N_k \sim \mathcal{N}(\mathbf{0}_q, \mathbf{I}_q)\ i.i.d. \tag{19}$$

Computing the conditional expectation of any function (not necessarily linearly) of the conditioned random vector $(Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+q}))$ knowing $Y(\mathbf{X}) = \mathbf{Y}$ can then be done in averaging the images of the simulated

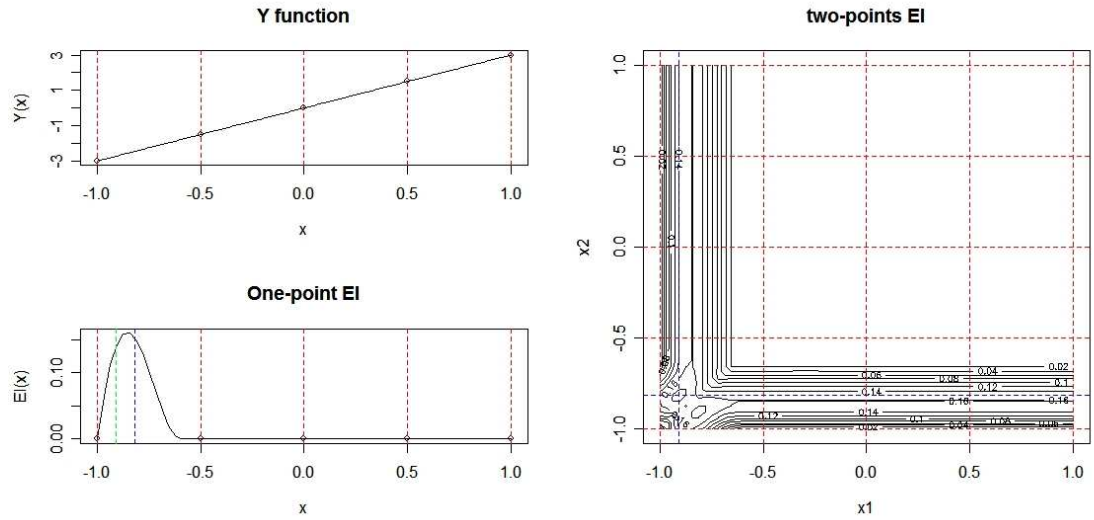

Fig. 4 1-point EI (lower left) and 2-points EI (right) functions associated with a monodimensional linear function ($y(x) = 3 \times x$) known at $\mathbf{X} = \{-1, -0.5, 0, 0.5, 1\}$. The OK metamodel has here a cubic covariance with parameters $\sigma^2 = 10,$ scale $= 1.4$).

vectors by the considered function:

function Q-$\mathbb{E}$I($\mathbf{X}$, $\mathbf{Y}$, $\mathbf{X}^{new}$)
  $L = \text{chol}(Var[Y(\mathbf{X}^{new})|Y(\mathbf{X}) = \mathbf{Y}])$ ▷ Decomposition of S_q (Cholesky, Mahalanobis, etc.)
  for $i \leftarrow 1, n_{sim}$ **do**
    $N \sim \mathcal{N}(0, I_q)$ ▷ Drawing a standard gaussian vector N at random
    $M_i = m_{OK}(\mathbf{X}^{new}) + LN$ ▷ Simulating Y at $\mathbf{X}^{new}$ conditional on $Y(\mathbf{X}) = \mathbf{Y}$
    $qI_{sim}(i)$=$[\min(\mathbf{Y}) - \min(M_i)]^+$ ▷ Simulating the improvement at $\mathbf{X}^{new}$
  end for
  $qEI_{sim} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} qI_{sim}(i)$ ▷ Estimation of the q-points Expected Improvement
end function

A straightforward application of the Law of Large Numbers (LLN) yields indeed

$$qEI_{sim} = \sum_{i=1}^{n_{sim}} \frac{[\min(\mathbf{Y}) - \min(M_i)]^+}{n_{sim}} \xrightarrow[n_{sim} \to +\infty]{} EI(\mathbf{x}^1, ..., \mathbf{x}^q) \text{ a.s.}, \tag{20}$$

and the Central Limit Theorem (CLT) can finally be used to control the precision of the Monte Carlo approximation as a function of n_{sim} (see [40] for details concerning the variance estimation):

$$\sqrt{n_{sim}}\left(\frac{qEI_{sim}-EI(\mathbf{x}^1,...,\mathbf{x}^q)}{\sqrt{Var[I(\mathbf{x}^1,...,\mathbf{x}^q)]}}\right)\xrightarrow[n_{sim}\to+\infty]{}\mathcal{N}(0,1)\text{ in law.} \tag{21}$$

4 Approximated q-EI maximization

The multi-points criterion that we have presented in the last section can potentially be used to deliver an additional design of experiments in one step through the resolution of the optimization problem

$$(\mathbf{x}'^{n+1},\mathbf{x}'^{n+2},...,\mathbf{x}'^{n+q}) = argmax_{\mathbf{X}'\in D^q}[EI(\mathbf{X}')] \tag{22}$$

However, the computation of q-EI becomes intensive as q increases. Moreover, the optimization problem (22) is of dimension $d\times q$, and with a noisy and derivative-free objective function in the case where the criterion is estimated by Monte-Carlo. Here we try to find pseudo-sequential greedy strategies that approach the result of problem 22 while avoiding its numerical cost, hence circumventing the curse of dimensionality.

4.1 A first greedy strategy to build a q-points design with the 1-point EI

Instead of searching for the globally optimal vector $(\mathbf{x}'^{n+1},\mathbf{x}'^{n+2},...,\mathbf{x}'^{n+q})$, an intuitive way of replacing it by a sequential approach is the following: first look for the next best single point $\mathbf{x}^{n+1}=argmax_{\mathbf{x}\in D}EI(\mathbf{x})$, then feed the model and look for $\mathbf{x}^{n+2}=argmax_{\mathbf{x}\in D}EI(x)$, and so on. Of course, the value $y(\mathbf{x}^{n+1})$ is not known at the second step (else we would be in a real sequential algorithm, like EGO). Nevertheless, we dispose of two pieces of information: the site $\mathbf{x}^{n+1}$ is assumed to have already been visited at the previous iteration, and $[Y(\mathbf{x}^{n+1})|\mathbf{Y}=Y(\mathbf{X})]$ has a known distribution. More precisely, the latter is $[Y(\mathbf{x}^{n+1})|Y(\mathbf{X})=\mathbf{Y}]\sim\mathcal{N}(m_{OK}(\mathbf{x}^{n+1}),s^2_{OK}(\mathbf{x}^{n+1}))$. Hence, the second site $\mathbf{x}^{n+2}$ can be computed as:

$$\mathbf{x}^{n+2}=argmax_{\mathbf{x}\in D}\mathbb{E}\left[\mathbb{E}\left[(Y(\mathbf{x})-\min(Y(\mathbf{X})))^+|Y(\mathbf{X})=\mathbf{Y},Y(\mathbf{x}^{n+1})\right]\right], \tag{23}$$

and the same procedure can be applied iteratively to deliver q points, computing $\forall j\in[1,q-1]$:

$$\mathbf{x}^{n+j+1}=argmax_{\mathbf{x}\in D}\int_{\mathbf{u}\in\mathbb{R}^j}\left[\mathbb{E}\left[(Y(\mathbf{x})-\min(Y(\mathbf{X})))^+|Y(\mathbf{X})=\mathbf{Y},Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+j-1})\right]\right]f_{Y(\mathbf{X}^{1:j})|Y(\mathbf{X})=\mathbf{Y}}(\mathbf{u})d\mathbf{u}, \tag{24}$$

where $f_{Y(\mathbf{X}^{1:j})|Y(\mathbf{X})=\mathbf{Y}}$ is the multivariate gaussian density of the OK conditional distrubtion at $(\mathbf{x}^{n+1},...,\mathbf{x}^{n+j})$. Although eq. 24 is a sequentialized version of the q-points expected improvement maximization, it doesn't completely fulfill our objectives. There is still a multivariate gaussian density to integrate, which seems to be a typical curse in such problems dealing with dependent random vectors. We now present two classes of heuristic strategies meant to circumvent the computational complexity encountered in (24).

4.2 The Kriging Believer (KB) and Constant Liar (CL) strategies

Lying to escape intractable calculations: starting from the principle of (24), we propose to weaken the conditional knowledge taken into account at each iteration. This very elementary idea inspired two heuristic strategies that we expose and test in the next two subsections: the *Kriging Believer* and the *Constant Liar*.

4.2.1 Believing the OK predictor: the KB heuristic strategy

The *Kriging Believer* strategy replaces the conditional knowledge about the responses at the sites chosen within the last iterations by deterministic values equal to the expectation of the kriging predictor. Keeping the same notations as previously, the strategy can be summed up as follows:

Algorithm 1 The Kriging Believer algorithm: a first approximate solution of the multipoints problem $(\mathbf{x}'^{n+1}, \mathbf{x}'^{n+2}, ..., \mathbf{x}'^{n+q}) = argmax_{\mathbf{X}' \in D^q}[EI(\mathbf{X}')]$

1: **function** KB($\mathbf{X}$, $\mathbf{Y}$, q)
2: **for** $i \leftarrow 1, q$ **do**
3: $\mathbf{x}^{n+i} = argmax_{\mathbf{x} \in D} EI(\mathbf{x})$
4: $m_{OK}(\mathbf{x}^{n+i}) = \mathbb{E}[Y(\mathbf{x}^{n+i}) | Y(\mathbf{X}) = \mathbf{Y}]$
5: $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^{n+i}\}$
6: $\mathbf{Y} = \mathbf{Y} \bigcup \{m_{OK}(\mathbf{x}^{n+i})\}$
7: **end for**
8: **end function**

This sequential strategy delivers a q-points design and is computationally affordable since it relies on the analytically known EI, optimized in d dimensions. However, there is a risk of failure, since believing an OK predictor that overshoots the observed data may lead to a sequence that gets trapped in a non-optimal region for many iterations (see 4.3). We now propose a second strategy that reduces this risk.

4.2.2 Updating the OK metamodel with fake observations: the CL heuristic strategy

Let us now consider a sequential strategy in which the metamodel is updated (still without hyperparameter re-estimation) at each iteration with a value L exogenously fixed by the user, here called a "lie". The strategy referred to as the *Constant Liar* consists in lying with the same value L at every iteration: maximize EI (i.e. find $\mathbf{x}^{n+1}$), actualize the model as if $y(\mathbf{x}^{n+1}) = L$, and so on always with the same $L \in \mathbb{R}$:
The effect of L on the performance of the resulting optimizer is investigated in the next section. L should logically be determined on the basis of the values taken by y at $\mathbf{X}$. Three values, $min\{\mathbf{Y}\}$, $mean\{\mathbf{Y}\}$, and $max\{\mathbf{Y}\}$ are considered here. The larger L is, the more explorative the algorithm will be, and vice versa.

4.3 Empirical comparisons with the Branin-Hoo function

The four optimization strategies presented in the last section are now compared on the the Branin-Hoo function which is a classical test-case in global optimization [22, 38, 47]:

Algorithm 2 The Constant Liar algorithm: another approximate solution of the multipoints problem $(\mathbf{x}'^{n+1}, \mathbf{x}'^{n+2}, ..., \mathbf{x}'^{n+q}) = argmax_{\mathbf{X}' \in D^q}[EI(\mathbf{X}')]$

1: **function** CL($\mathbf{X}$, $\mathbf{Y}$, L, q)
2: **for** $i \leftarrow 1, q$ **do**
3: $\mathbf{x}^{n+i} = argmax_{\mathbf{x} \in D} EI(\mathbf{x})$
4: $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^{n+i}\}$
5: $\mathbf{Y} = \mathbf{Y} \bigcup \{L\}$
6: **end for**
7: **end function**

$$\begin{cases} y_{BH}(x_1, x_2) = (x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6)^2 + 10(1 - \frac{1}{8\pi})cos(x_1) + 10 \\ x_1 \in [-5, 10],\ x_2 \in [0, 15] \end{cases} \tag{25}$$

y_{BH} has three global minimizers $(-3.14, 12.27)$, $(3.14, 2.27)$, $(9.42, 2.47)$, and the global minimum is approximately equal to 0.4. The variables are normalized by the transformation $x_1' = \frac{x_1+5}{15}$ and $x_2' = \frac{x_2}{15}$. The initial design of experiments is a 3×3 complete factorial design $\mathbf{X}_9$ (see 5), thus $\mathbf{Y} = y_{BH}(\mathbf{X}_9)$. Ordinary Kriging is applied with a stationary, anisotropic, gaussian covariance function

$$\forall h = (h_1, h_2) \in \mathbb{R}^2,\ c(h_1, h_2) = \sigma^2 e^{-\theta_1 h_1^2 - \theta_2 h_2^2} \tag{26}$$

where the parameters (θ_1, θ_2) are fixed to their Maximum Likelihood Estimate $(5.27, 0.26)$, and σ^2 is estimated within kriging, as an implicit function of (θ_1, θ_2) (like in [22]). We build a 10-points optimization design with each strategy, and additionally estimated by Monte Carlo simulations ($n_{sim} = 10^4$) the PI and EI values brought by the q first points of each strategy (here $q \in \{2, 6, 10\}$). The results are gathered in Tab. 4.3.

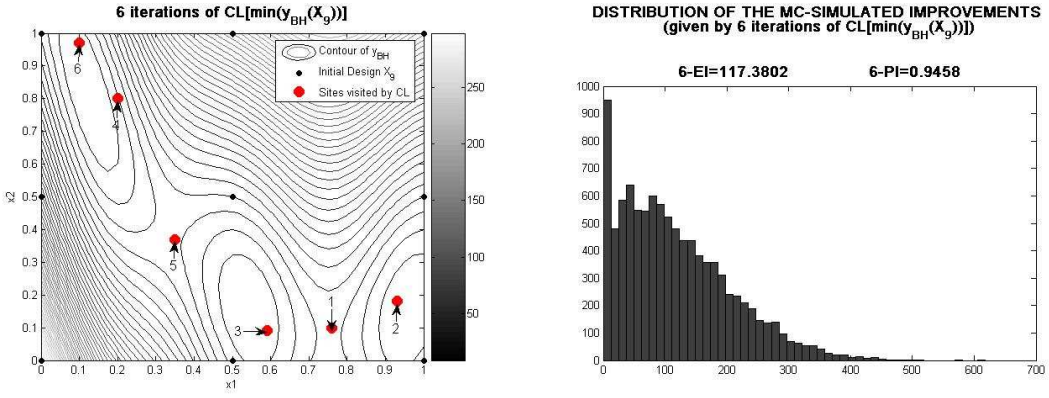

Fig. 5 (Left) contour of the Branin-Hoo function with the design $\mathbf{X}_9$ (small black points) and the 6 first points given by the heuristic strategy CL[min($y_{BH}(\mathbf{X}_9)$)] (large bullets). (Right) Histogram of 10^4 Monte Carlo simulated values of the improvement brought by the 6-points CL[min($y_{BH}(\mathbf{X}_9)$)] strategy. The corresponding estimates of 6-points PI and EI are given above.

The four strategies (KB and the three variants of CL) gave clearly different designs and optimization performances. In the first case, *Constant Liar* (CL) sequences behaved as if the already visited points generated a repulsion, with a magnitude increasing with L. The tested values $L = \max(\mathbf{Y})$ and $L = mean(\mathbf{Y})$ forced the exploration designs to fill the space by avoiding $\mathbf{X_9}$. Both strategies provided space-filling, exploratory designs with high probabilities of improvement (10-PI near 100%) and promising q-EI values (see Table 1). *In fine*, they brought respective actual improvements of 7.86 and 6.25.

Of all the tested strategies, CL[min($\mathbf{Y}$)] gave here the best results. In 6 iterations, it visited the three locally optimal zones of y_{BH}. In 10 iterations, it gave the best actual improvement among the considered strategies, which is furthermore in agreement with the 10-points EI values simulated by Monte-Carlo. It seems in fact that the soft repulsion when $L = \min(\mathbf{Y})$ is the right tuning for the optimization of the Branin-Hoo function, with the initial design $\mathbf{X}_9$.

In the second case, the KB has yielded here disappointing results. All the points (except one) were clustered around the first visited point $\mathbf{x}^{n+1}$ (the same as in *CL*, by construction). This can be explained by the exaggeratedly low prediction given by Kriging at this very point: the mean predictor overshoots the data (because of the Gaussian covariance), and the expected improvement becomes abusively large in the neighborhood of $\mathbf{x}^{n+1}$. Then $\mathbf{x}^{n+2}$ is then chosen near $\mathbf{x}^{n+1}$, and so on. The algorithm gets temporarily trapped at the first visited point. KB behaves in the same way as *CL* would do with a constant L below $\min(\mathbf{Y})$. As can be seen in Table 1 (last column), the phenomenon is visible on both the q-PI and q-EI criteria: they remain almost constant when q increases. This illustrates in particular how q-points criteria can help in rejecting unappropriate strategies.

	CL[$min(\mathbf{Y})$]	CL[$mean(\mathbf{Y})$]	CL[$max(\mathbf{Y})$]	KB
PI (first 2 points)	87.7%	87%	88.9%	65%
EI (first 2 points)	114.3	114	113.5	82.9
PI (first 6 points)	94.6%	95.5%	92.7%	65.5%
EI (first 6 points)	117.4	115.6	115.1	85.2
PI (first 10 points)	99.8%	99.9%	99.9%	66.5%
EI (first 10 points)	122.6	118.4	117	85.86
Improvement (first 6 points)	7.4	6.25	7.86	0
Improvement (first 10 points)	8.37	6.25	7.86	0

Table 1 Multipoints PI, EI, and actual improvements for the 2, 6, and 10 first iterations of the heuristic strategies CL[min($\mathbf{Y}$)], CL[$mean(\mathbf{Y})$], CL[max($\mathbf{Y}$)], and Kriging Believer (here $\min(\mathbf{Y}) = \min(y_{BH}(\mathbf{X}_9))$). $q-PI$ and $q-EI$ are evaluated by Monte-Carlo simulations (Eq. (20), $n_{sim} = 10^4$).

In other respects, the results shown in Tab. 4.3 highlight a major drawback of the q-PI criterion. When q increases, the PI associated with all 3 CL strategies quickly converges to 100%, such that it is not possible to discriminate between the good and the very good designs. The q-EI is a more selective measure thanks to taking the magnitude of possible improvements into account. Nevertheless, q-EI overevaluates the improvement associated with all designs considered here. This effect (already pointed out in [47]) can be explained by considering both the high value of σ^2 estimated from $\mathbf{Y}$ and the small difference between the minimal value reached at $\mathbf{X}_9$ (9.5) and the actual minimum of y_{BH} (0.4).

We finally compared CL[min], CL[max], latin hypercubes (LHS) and uniform random designs (UNIF) in terms of q-EI values, with $q \in [1,10]$. For every $q \in [1,10]$, we sampled 2000 q-elements designs of each type (LHS and UNIF) and compared the obtained empirical distributions of q-points Expected Improvement

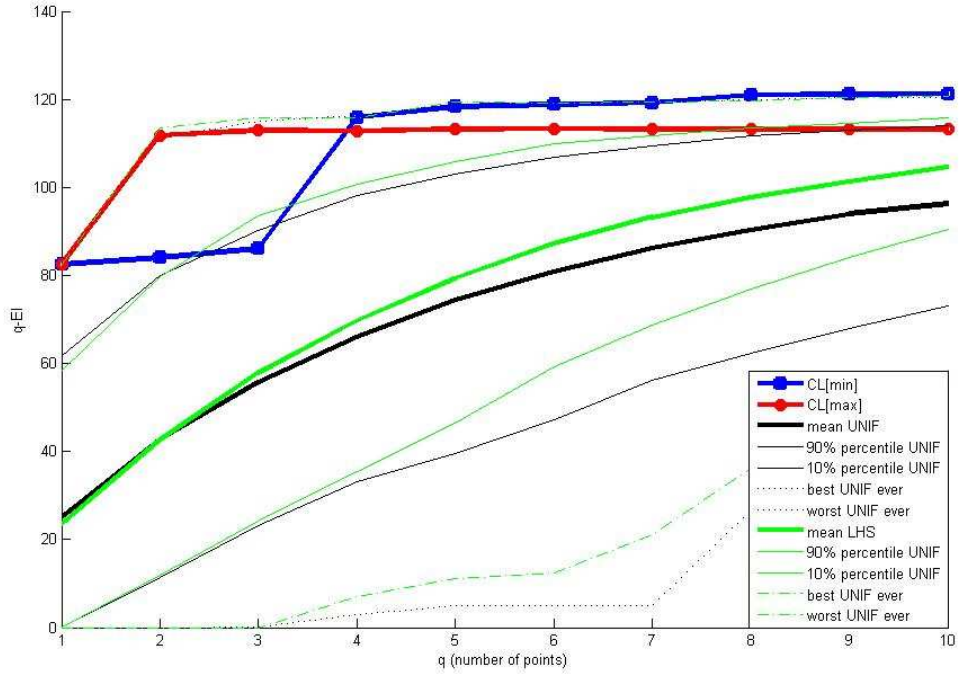

Fig. 6 Comparaison of the q-EI associated with the q first points ($q \in [1,10]$) given by the constant liar strategies (min and max), 2000 q-points designs uniformly drawn for every q, and 2000 q-points LHS designs taken at random for every q.

to the q-points Expected Improvement estimates associated with the q first points of both CL strategies.

As can be seen on fig. 6, CL[max] (light bullets) and CL[min] (dark squares) offer very good q-EI results compared to random designs, especially for small values of q. By definition, the two of them start with the 1-EI global maximizer, which ensures a q-EI at least equal to 83 for all $q \geq 1$. Both associated q-EI series then seem to converge to threshold values, almost reached for $q \geq 2$ by CL[max] (which dominates CL[min] when $q = 2$ and $q = 3$) and for $q \geq 4$ by CL[min] (which dominates CL[max] for all $4 \leq q \leq 10$). The random designs have less promizing q-EI expected values. Their q-EI distributions are quite dispersed, which can be seen for instance by looking at the $10\% - 90\%$ interpercentiles represented on fig. 6 by thin full lines (respectively dark and light for UNIF and LHS designs). Note in particular that the q-EI distribution of the LHS designs seem globally better than the one of the uniform designs. Interestingly, the best designs ever found among the UNIF designs (dark dotted lines) and among the LHS designs (light dotted lines) almost match with CL[max] when $q \in \{2,3\}$ and CL[min] when $4 \leq q \leq 10$. We haven't yet observed a design sampled at random that clearly provides better q-EI values than the proposed heuristic strategies.

5 Towards Kriging-based Parallel Optimization: Conclusion and Perspectives

Optimization problems with objective functions steming from expensive computer simulations strongly motivates the use of data-driven simplified mathematical representations of the simulator, or *metamodels*. An increasing number of optimization algorithms developed for such problems rely on metamodels, competing with and/or complementing population-based Computational Intelligence methods. A representative example is given by the EGO algorithm [22], a sequential black-box optimization procedure, which has gained popularity during the last decade and inspired numerous recent works in the field [10, 17, 18, 19, 20, 26, 28, 36, 44, 50]. EGO relies on a Kriging-based criterion, the expected improvement (EI), accounting for the exploration-exploitation trade-off[7]. The latter algorithm unfortunately produces only one point at each iteration, which prevents to take advantage of parallel computation facilities. In the present work, we came back to the interpretation of Kriging in terms of Gaussian Process[39] in order to propose a framework for Kriging-based parallel optimization, and to prepare the work for parallel variants of EGO.

The probabilistic nature of the Kriging metamodel allowed us to calculate the joint probability distribution associated with the predictions at any set of points, upon which we could rediscover (see [47]) and characterize a criterion named here *multi-points expected improvement*, or q-EI. The q-EI criterion makes it possible to get an evaluation of the "optimization potential" given by any set of q new experiments. An analytical derivation of 2-EI was performed, providing a good example of how to manipulate joint Kriging distributions for choosing additional designs of experiments, and enabling us to shed more light on the nature of the q-EI thanks to selected figures. For the computation of q-EI in the general case, an alternative computation method relying on Monte-Carlo simulations was proposed. As pointed out in illustrated in the chapter, Monte-Carlo simulations offer indeed the opportunity to evaluate the q-EI associated with any given design of experiment, whatever its size n, and whatever the dimension of inputs d. However, deriving q-EI-optimal designs on the basis of such estimates is not straightforward, and crucially depending on both n and d. Hence some greedy alternative problems were considered: four heuristic strategies, the "Kriging Believer" and three "Constant Liars" have been proposed and compared that aim at maximizing q-EI while being numerically tractable. It has been verified in the frame of a classical test case that the CL strategies provide q-EI vales comparable with the best Latin Hypercubes and uniform designs of experiments found by simulation. This simple application illustrated a central practical conclusion of this work: considering a set of candidate designs of experiments, provided for instance by heuristic strategies, it is always possible —whatever n and d— to evaluate and rank them using estimates of q-EI or related criteria, thanks to conditional Monte-Carlo simulation.

Perspectives include of course the development of synchronous parallel EGO variants delivering a set of q points at each iteration. The tools presented in the chapter may constitute bricks of these algorithms, as it has very recently been illustrated on a succesful 6-dimensional test-case in the thesis [13]. An R package covering that subject is in an advanced stage of preparation and should be released soon [41]. On a longer term, the scope of the work presented in this chapter, and not only its modest original contributions, could be broaden. If the considered methods could seem essentially restricted to the Ordinary Kriging metamodel and concern the use of an optimization criterion meant to obtain q points in parallel, several degrees of freedom can be played on in order to address more general problems. First, any probabilistic metamodel potentially providing joint distributions could do well (regression models, smoothing splines, etc.). Second, the final goal of the new generated design might be to improve the global accuracy of the metamodel,

[7] Other computational intelligence optimizers, e.g. evolutionary algorithms [9], address the exploration/exploitation trade-off implicitely through the choice of parameters such as the population size and the mutation probability.

to learn a quantile, to fill the space, etc : the work done here with the q-EI and associate strategies is just a particular case of what one can do with the flexibility offered by probabilistic metamodels and all possible decision-theoretic criteria. To finish with two challenging issues of Computationnal Intelligence, the following perspectives seem particularly relevant at both sides of the interface with this work:

- CI methods are needed to maximize the q-EI criterion, which inputs live in a $(n \times d)$-dimensional space, and which evaluation is noisy, with tunable fidelity depending on the chosen n_{sim} values,
- q-EI and related criteria are now at disposal to help pre-selecting good points in metamodel-assisted evolution strategies, in the flavour of [10].

Aknowledgements: This work was conducted within the frame of the DICE (Deep Inside Computer Experiments) Consortium between ARMINES, Renault, EDF, IRSN, ONERA, and Total S.A. The authors wish to thank X. Bay, R. T. Haftka, B. Smarslok, Y. Richet, O. Roustant, and V. Picheny for their help and rich comments. Special thanks to the R project people [6] for developing and spreading such a useful freeware. David moved to Neuchâtel University (Switzerland) for a postdoc, and he gratefully thanks the Mathematics Institute and the Hydrogeology Department for letting him spend time on the revision of the present chapter.

6 Appendix

6.1 Gaussian Processes for Machine Learning

A real random process $(Y(\mathbf{x}))_{\mathbf{x}\in D}$ is defined as a *Gaussian Process* (GP) whenever all its finite-dimensional distributions are gaussian. Consequently, for all $n \in \mathbb{N}$ and for all set $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$ of n points of D, there exists a vector $\mathbf{m} \in \mathbf{R}^n$ and a symmetric positive semi-definite matrix $\Sigma \in \mathscr{M}_n(\mathbb{R})$ such that $(Y(\mathbf{x}^1), ..., Y(\mathbf{x}^n))$ is a gaussian Vector, following a multigaussian probability distribution $\mathscr{N}(\mathbf{m}, \Sigma)$. More specifically, for all $i \in [1,n]$, $Y(\mathbf{x}^i) \sim \mathscr{N}(\mathbb{E}[Y(\mathbf{x}^i)], Var[Y(\mathbf{x}^i)])$ where $\mathbb{E}[Y(\mathbf{x}^i)]$ is the ith coordinate of $\mathbf{m}$ and $Var[Y(\mathbf{x}^i)]$ is the ith diagonal term of Σ. Furthermore, all couples $(Y(\mathbf{x}^i), Y(\mathbf{x}^j))$ $i, j \in [1,n], i \neq j$ are multigaussian with a covariance $Cov[Y(\mathbf{x}^i), Y(\mathbf{x}^j)]$ equal to the non-diagonal term of Σ indexed by i and j.
A Random Process Y is said to be *first order stationary* if its mean is a constant, i.e. if $\exists \mu \in \mathbb{R} |\ \forall \mathbf{x} \in D,\ \mathbb{E}[Y(\mathbf{x})] = \mu$. Y is said to be *second order stationary* if it is first order stationary and if there exists furthermore a function of positive type, $c : D - D \longrightarrow \mathbb{R}$, such that for all pairs $(\mathbf{x}, \mathbf{x}') \in D^2$, $Cov[Y(\mathbf{x}), Y(\mathbf{x}')] = c(\mathbf{x} - \mathbf{x}')$. We then have the following expression for the covariance matrix of the observations at $\mathbf{X}$:

$$\Sigma := (Cov[Y(\mathbf{x}_i), Y(\mathbf{x}_j)])_{i,j\in[1,n]} = (c(\mathbf{x}_i - \mathbf{x}_j))_{i,j\in[1,n]} = \begin{pmatrix} \sigma^2 & c(\mathbf{x}_1 - \mathbf{x}_2) & ... & c(\mathbf{x}_1 - \mathbf{x}_n) \\ c(\mathbf{x}_2 - \mathbf{x}_1) & \sigma^2 & ... & c(\mathbf{x}_2 - \mathbf{x}_n) \\ ... & ... & ... & ... \\ c(\mathbf{x}_n - \mathbf{x}_1) & c(\mathbf{x}_n - \mathbf{x}_2) & ... & \sigma^2 \end{pmatrix} \tag{27}$$

where $\sigma^2 := c(0)$. Second order stationary processes are sometimes called *weakly stationary*. A major feature of GPs is that their *weak stationarity* is equivalent to *strong stationarity*: if Y is a weakly stationary GP, the law of probability of the random variable $Y(\mathbf{x})$ doesn't depend on $\mathbf{x}$, and the joint distribution of $(Y(\mathbf{x}^1), ..., Y(\mathbf{x}^n))$ is the same as the distribution of $(Y(\mathbf{x}^1 + \mathbf{h}), ..., Y(\mathbf{x}^n + \mathbf{h}))$ whatever the set of points $\{\mathbf{x}^1, ..., \mathbf{x}^n\} \in D^n$ and the vector $\mathbf{h} \in \mathbb{R}^n$ such that $\{\mathbf{x}^1 + \mathbf{h}, ..., \mathbf{x}^n + \mathbf{h}\} \in D^n$. To sum up, a stationary GP is entirely defined by its mean μ and its covariance function $c(.)$. The classical framework of Kriging for Computer Experiments is to make predictions of a costly simulator y at a new set of sites $\mathbf{X}_{new} = \{\mathbf{x}^{n+1}, ..., \mathbf{x}^{n+q}\}$ (most of the time, $q = 1$), on the basis of the collected observations at the initial design $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$, and under the assumption that y is one realization of a stationary GP Y with known covariance function c (in theory). Simple Kriging (SK) assumes a known mean, $\mu \in \mathbb{R}$. In Ordinary Kriging (OK), μ is estimated.

6.2 Conditioning Gaussian Vectors

Let us consider a centered Gaussian vector $V = (V_1, V_2)$ with covariance matrix

$$\Sigma_V = \mathbb{E}[VV^T] = \begin{pmatrix} \Sigma_{V_1} & \Sigma_{cross}^T \\ \Sigma_{cross} & \Sigma_{V_2} \end{pmatrix} \tag{28}$$

Key properties of Gaussian vectors include that the orthogonal projection of a Gaussian vector onto a linear subspace is still a Gaussian vector, and that the orthogonality of two subvectors V_1, V_2 of a Gaussian vector V (i.e. $\Sigma_{cross} = \mathbb{E}[V_2 V_1^T] = 0$) is equivalent to their independence. We now express the conditional expectation $\mathbb{E}[V_1|V_2]$. $\mathbb{E}[V_1|V_2]$ is by definition such that $V_1 - \mathbb{E}[V_1|V_2]$ is independent of V_2. $\mathbb{E}[V_1|V_2]$ is thus fully characterized as orthogonal projection on the vector space spanned by the components of V_2, solving the so called *normal equations*:

$$\mathbb{E}[(V_1 - \mathbb{E}[V_1|V_2])V_2^T] = 0 \tag{29}$$

Assuming linearity of $\mathbb{E}[V_1|V_2]$ in V_2, i.e. $\mathbb{E}[V_1|V_2] = AV_2$ ($A \in \mathcal{M}_n(\mathbb{R})$), a straightforward development of (eq.29) gives the matrix equation $\Sigma_{cross}^T = A\Sigma_{V_2}$, and hence $\Sigma_{cross}^T \Sigma_{V_2}^{-1} V_2$ is a suitable solution provided Σ_{V_2} is full ranked[8]. We conclude that

$$\mathbb{E}[V_1|V_2] = \Sigma_{cross}^T \Sigma_{V_2}^{-1} V_2 \tag{30}$$

by uniqueness of the orthogonal projection onto a closed linear subspace in a Hilbert space. Using the independence between $(V_1 - \mathbb{E}[V_1|V_2])$ and V_2, one can calculate the conditional covariance matrix $\Sigma_{V_1|V_2}$:

$$\begin{aligned} \Sigma_{V_1|V_2} &= \mathbb{E}[(V_1 - \mathbb{E}[V_1|V_2])(V_1 - \mathbb{E}[V_1|V_2])^T|V_2] = \mathbb{E}[(V_1 - AV_2)(V_1 - AV_2)^T] \\ &= \Sigma_{V_1} - A\Sigma_{cross} - \Sigma_{cross}^T A^T + A\Sigma_{V_2} A^T = \Sigma_{V_1} - \Sigma_{cross}^T \Sigma_{V_2}^{-1} \Sigma_{cross} \end{aligned} \tag{31}$$

Now consider the case of a non-centered random vector $V = (V_1, V_2)$ with mean $m = (m_1, m_2)$. The conditional distribution $V_1|V_2$ can be obtained by coming back to the centered random vector $V - m$. We then find that $\mathbb{E}[V_1 - m_1|V_2 - m_2] = \Sigma_{cross}^T \Sigma_{V_2}^{-1}(V_2 - m_2)$ and hence $\mathbb{E}[V_1|V_2] = m_1 + \Sigma_{cross}^T \Sigma_{V_2}^{-1}(V_2 - m_2)$.

6.3 Simple Kriging Equations

Let us come back to our metamodeling problem and assume that y is one realization of a Gaussian Process Y, defined as follows:

$$\begin{cases} Y(\mathbf{x}) = \mu + \varepsilon(\mathbf{x}) \\ \varepsilon(\mathbf{x}) \text{ centered stationary GP with covariance function } c(.) \end{cases} \tag{32}$$

where $\mu \in \mathbb{R}$ is known. Now say that Y has already been observed at n locations $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$ ($Y(\mathbf{X}) = \mathbf{Y}$) and that we wish to predict Y a q new locations $\mathbf{X}_{new} = \{\mathbf{x}^{n+1}, ..., \mathbf{x}^{n+q}\}$. Since $(Y(\mathbf{x}^1), ..., Y(\mathbf{x}^n), Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+q}))$ is a Gaussian Vector with mean $\mu \mathbf{1}_{n+q}$ and covariance matrix

$$\Sigma_{tot} = \begin{pmatrix} \Sigma & \Sigma_{cross}^T \\ \Sigma_{cross} & \Sigma_{new} \end{pmatrix} = \begin{pmatrix} \sigma^2 & c(\mathbf{x}_1 - \mathbf{x}_2) & \dots & c(\mathbf{x}_1 - \mathbf{x}_{n+q}) \\ c(\mathbf{x}_2 - \mathbf{x}_1) & \sigma^2 & \dots & c(\mathbf{x}_2 - \mathbf{x}_{n+q}) \\ \dots & \dots & \dots & \dots \\ c(\mathbf{x}_{n+q} - \mathbf{x}_1) & c(\mathbf{x}_{n+q} - \mathbf{x}_2) & \dots & \sigma^2 \end{pmatrix} \tag{33}$$

We can directly apply eq. (30) and eq. (31) to derive the Simple Kriging Equations:

$$[Y(\mathbf{X}_{new})|Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}(m_{SK}(\mathbf{X}_{new}), \Sigma_{SK}(\mathbf{X}_{new})) \tag{34}$$

[8] If Σ_{V_2} is not invertible, the equation holds in replacing $\Sigma_{V_2}^{-1}$ by the pseudo-inverse $\Sigma_{V_2}^{\dagger}$.

with $m_{SK}(\mathbf{X}_{new}) = \mathbb{E}[Y(\mathbf{X}_{new})|Y(\mathbf{X}) = \mathbf{Y}] = \mu\mathbf{1}_q + \Sigma_{cross}^T\Sigma^{-1}(\mathbf{Y} - \mu\mathbf{1}_q)$ and $\Sigma_{SK}(\mathbf{X}_{new}) = \Sigma_{new} - \Sigma_{cross}^T\Sigma^{-1}\Sigma_{cross}$. When $q = 1$, $\Sigma_{cross} = \mathbf{c}(\mathbf{x}^{n+1}) = Cov[Y(\mathbf{x}^{n+1}), Y(\mathbf{X})]$ and the covariance matrix reduces to $s_{SK}^2(\mathbf{x}) = \sigma^2 - \mathbf{c}(\mathbf{x}^{n+1})^T\Sigma^{-1}\mathbf{c}(\mathbf{x}^{n+1})$, which is called the *Kriging Variance*. When μ is constant but not known in advance, it is not mathematically correct to sequentially estimate μ and plug in the estimate in the Simple Kriging equations. Ordinary Kriging addresses this issue.

6.4 Ordinary Kriging Equations

Compared to Simple Kriging, Ordinary Kriging (OK) is used when the mean of the underlying random process is constant and unknown. We give here a derivation of OK in a Bayesian framework, assuming that μ has an improper uniform prior distribution $\mu \sim \mathscr{U}(\mathbb{R})$. y is thus seen as a realization of a random process Y, defined as the sum of μ and a centered GP [9]:

$$\begin{cases} Y(\mathbf{x}) = \mu + \varepsilon(\mathbf{x}) \\ \varepsilon(\mathbf{x}) \text{ centered stationary GP with covariance function } c(.) \\ \mu \sim \mathscr{U}(\mathbb{R}) \text{ (prior), independent of } \varepsilon \end{cases} \tag{35}$$

Note that conditioning with respect to μ actually provides SK equations. Letting μ vary, we aim to find the law of $[Y(\mathbf{X}_{new})|Y(\mathbf{X}) = \mathbf{Y}]$. Starting with $[Y(\mathbf{X}) = \mathbf{Y}|\mu] \sim \mathscr{N}(\mu\mathbf{1}_n, \Sigma)$, we get μ's posterior distribution:

$$[\mu|Y(\mathbf{X}) = \mathbf{Y}] \sim \mathscr{N}\left(\hat{\mu}, \sigma_\mu^2\right) = \mathscr{N}\left(\frac{\mathbf{1}^T\Sigma^{-1}\mathbf{Y}}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}, \frac{1}{\mathbf{1}_q^T\Sigma^{-1}\mathbf{1}_q}\right) \text{ (posterior)} \tag{36}$$

We can re-write the SK equations $[Y(\mathbf{X}_{new})|Y(\mathbf{X}) = \mathbf{Y}, \mu] \sim \mathscr{N}(m_{SK}(\mathbf{X}_{new}), \Sigma_{SK}(\mathbf{X}_{new}))$. Now it is very useful to notice that the random vector $(Y(\mathbf{X}_{new}), \mu)$ is Gaussian conditional on $Y(\mathbf{X}) = \mathbf{Y}$. [10]. It follows that $[Y(\mathbf{X}_{new})|Y(\mathbf{X}) = \mathbf{Y}]$ is Gaussian, and its mean and covariance matrix can finally be calculated with the help of classical conditional calculus results. Hence using $m_{OK}(\mathbf{X}_{new}) = \mathbb{E}[Y(\mathbf{X}_{new})|Y(\mathbf{X}) = \mathbf{Y}] = \mathbb{E}_\mu[\mathbb{E}[Y(\mathbf{X}_{new})|Y(\mathbf{X}) = \mathbf{Y}, \mu]]$, we find that $m_{OK}(\mathbf{X}_{new}) = \hat{\mu} + \Sigma_{cross}^T\Sigma^{-1}(\mathbf{Y} - \hat{\mu}\mathbf{1}_n)$. Similarly, $\Sigma_{OK}(\mathbf{X}_{new})$ can be obtained using that $Cov[A,B] = Cov[\mathbb{E}[A|C], \mathbb{E}[B|C]] + \mathbb{E}[Cov[A,B|C]]$ for all random variables A,B, C such that all terms exist. We then get for all couples of points $(\mathbf{x}^{n+i}, \mathbf{x}^{n+j})$ $(i, j \in [1,q])$:

$$\begin{aligned} &Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})|Y(\mathbf{X}) = \mathbf{Y}] \\ =&\mathbb{E}\left[Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})|Y(\mathbf{X}) = \mathbf{Y}, \mu]\right] + Cov\left[\mathbb{E}[Y(\mathbf{x}^{n+i})|Y(\mathbf{X}) = \mathbf{Y}, \mu], \mathbb{E}[Y(\mathbf{x}^{n+j})|Y(\mathbf{X}) = \mathbf{Y}, \mu]\right]. \end{aligned} \tag{37}$$

The left term $Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})|Y(\mathbf{X}) = \mathbf{Y}, \mu]$ is the conditional covariance under the Simple Kriging Model. The right term is the covariance between $\mu + \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}(\mathbf{Y} - \mu\mathbf{1}_q)$ and $\mu + \mathbf{c}(\mathbf{x}^{n+j})^T\Sigma^{-1}(\mathbf{Y} - \mu\mathbf{1}_q)$ conditional on the observations $Y(\mathbf{X}) = \mathbf{Y}$. Using eq. 36, we finally obtain:

$$\begin{aligned} &Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})|Y(\mathbf{X}) = \mathbf{Y}] \\ =&Cov_{SK}[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})|Y(\mathbf{X}) = \mathbf{Y}] \\ +&Cov[\mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}(\mathbf{Y}) + \mu(1 + \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}\mathbf{1}_q), \mathbf{c}(\mathbf{x}^{n+j})^T\Sigma^{-1}(\mathbf{Y}) + \mu(1 + \mathbf{c}(\mathbf{x}^{n+j})^T\Sigma^{-1}\mathbf{1}_q)] \\ =&c(\mathbf{x}^{n+i} - \mathbf{x}^{n+j}) - \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}\mathbf{c}(\mathbf{x}^{n+j}) + \frac{(1 + \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}\mathbf{1}_q)(1 + \mathbf{c}(\mathbf{x}^{n+j})^T\Sigma^{-1}\mathbf{1}_q)}{\mathbf{1}_q^T\Sigma^{-1}\mathbf{1}_q}. \end{aligned} \tag{38}$$

[9] The resulting random process Y is not Gaussian.

[10] which can be proved by considering its Fourier transform.

References

1. Abrahamsen P (1997) A review of gaussian random fields and correlation functions, second edition. Tech. Rep. 917, Norwegian Computing Center, Olso
2. Antoniadis A, Berruyer J, Carmona R (1992) Régression non linéaire et applications. Economica, Paris
3. Baker C, Watson LT, Grossman B, Mason WH, Haftka RT (2001) Parallel global aircraft configuration design space exploration. Practical parallel computing pp 79–96
4. Bishop C (1995) Neural Networks for Pattern Recognition. Oxford Univ. Press
5. Blum C (2005) Ant colony optimization : introduction and recent trends. Physics of life review 2:353–373
6. development Core Team R (2006) R: A language and environment for statistical computing. URL `http:www.R-project.org`
7. Cressie N (1993) Statistics for spatial data. Wiley series in probability and mathematical statistics
8. Dreyfus G, Martinez JM (2002) Réseaux de neurones. Eyrolles
9. Eiben A, Smith J (2003) Introduction to Evolutionary Computing. Springer Verlag
10. Emmerich M, Giannakoglou K, Naujoks B (2006) Single-and multiobjective optimization assisted by gaussian random field metamodels. IEEE Transactions on Evolutionary Computation 10(4):421439
11. Geman D, Jedynak B (December 1995) An active testing model for tracking roads in satellite images. Tech. rep., Institut National de Recherches en Informatique et Automatique (INRIA)
12. Genton M (2001) Classes of kernels for machine learning: A statistics perspective. Journal of Machine Learning Research 2:299–312
13. Ginsbourger D (2009) Multiples métamodèles pour l'approximation et l'optimisation de fonctions numériques multivariables. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne
14. Ginsbourger D, Le Riche R, Carraro L (2007) A multipoints criterion for parallel global optimization of deterministic computer experiments. In: Non-Convex Programming 07
15. Goria S (2004) Evaluation d'un projet minier: approche bayésienne et options réelles. PhD thesis, Ecole des Mines de Paris
16. Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. Springer
17. Henkenjohann N, Göbel R, Kleiner M, Kunert J (2005) An adaptive sequential procedure for efficient optimization of the sheet metal spinning process. Qual Reliab Engng Int 21:439–455
18. Huang D, Allen T, Notz W, Miller R (2006) Sequential kriging optimization using multiple fidelity evaluations. Sructural and Multidisciplinary Optimization 32:369–382
19. Huang D, Allen T, Notz W, Zheng N (2006) Global optimization of stochastic black-box systems via sequential kriging meta-models. Journal of Global Optimization 34:441–466
20. Jones D (2001) A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization 21:345–383
21. Jones D, Pertunen C, Stuckman B (1993) Lipschitzian optimization without the lipschitz constant. Journal of Optimization Theory and Application 79(1)
22. Jones D, Schonlau M, Welch W (1998) Efficient global optimization of expensive black-box functions. Journal of Global Optimization 13:455–492
23. Journel A (1988) Fundamentals of geostatistics in five lessons. Tech. rep., Stanford Center for Reservoir Forecasting
24. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: IEEE Intl. Conf. on Neural Networks, vol 4, pp 1942–1948
25. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680

26. Knowles J (2005) Parego: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. IEEE transactions on evolutionnary computation
27. Koehler J, Owen A (1996) Computer experiments. Tech. rep., Department of Statistics, Stanford University
28. Kracker H (2006) Methoden zur analyse von computerexperimenten mit anwendung auf die hochdruckblechumformung. Master's thesis, Dortmund University
29. Krige D (1951) A statistical approach to some basic mine valuation problems on the witwatersrand. J of the Chem, Metal and Mining Soc of South Africa 52 (6):119139
30. Martin J, Simpson T (2004) A monte carlo simulation of the kriging model. In: 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Albany, NY, August 30 - September 2, AIAA, AIAA-2004-4483.
31. Martin J, Simpson T (2005) Use of kriging models to approximate deterministic computer models. AIAA Journal 43 (4):853–863
32. Matheron G (1963) Principles of geostatistics. Economic Geology 58:1246–1266
33. Matheron G (1970) La théorie des variables régionalisées et ses applications. Tech. rep., Centre de Morphologie Mathématique de Fontainebleau, Ecole Nationale Supérieure des Mines de Paris
34. O'Hagan A (2006) Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering and System Safety 91(91):1290–1300
35. Paciorek C (2003) Nonstationary gaussian processes for regression and spatial modelling. PhD thesis, Carnegie Mellon University
36. Ponweiser W, Wagner T, Biermann D, Vincze M (2008) Multiobjective optimization on a limited budget of evaluations using model-assisted S-metric selection. In: Rudolph G, Jansen T, Lucas S, Poloni C, Beume N (eds) Parallel Problem Solving from Nature PPSN X. 10th International Conference Dortmund, Germany, September 13-17, Springer, Lecture Notes in Computer Science, vol 5199, pp 784–794
37. Praveen C, Duvigneau R (2007) Radial basis functions and kriging metamodels for aerodynamic optimization. Tech. rep., INRIA
38. Queipo N, Verde A, Pintos S, Haftka R (2006) Assessing the value of another cycle in surrogate-based optimization. In: 11th Multidisciplinary Analysis and Optimization Conference, AIAA
39. Rasmussen C, Williams K (2006) Gaussian Processes for Machine Learning. M.I.T. Press
40. Ripley B (1987) Stochastic Simulation. John Wiley and Sons, New York
41. Roustant O, Ginsbourger D, Deville Y (2009) The DiceKriging package: kriging-based metamodeling and optimization for computer experiments, the UseR! Conference, Agrocampus-Ouest, Rennes, France
42. Sacks J, Welch W, Mitchell T, Wynn H (1989) Design and analysis of computer experiments. Statistical Science 4(4):409–435
43. Santner T, Williams B, Notz W (2003) The Design and Analysis of Computer Experiments. Springer
44. Sasena M (2002) Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations. PhD thesis, University of Michigan
45. Sasena MJ, Papalambros P, Goovaerts P (2002) Exploration of metamodeling sampling criteria for constrained global optimization. Journal of Engineering Optimization
46. Sasena MJ, Papalambros PY, Goovaerts P (2002) Global optimization of problems with disconnected feasible regions via surrogate modeling. In: Proceedings of the 9th AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Atlanta, GA, 5573
47. Schonlau M (1997) Computer experiments and global optimization. PhD thesis, University of Waterloo
48. Schonlau M, Welch W, Jones D (1997) A data-analytic approach to bayesian global optimization. In: Proceedings of the A.S.A.

49. Ulmer H, Streichert F, Zell A (2003) Evolution strategies assisted by gaussian processes with improved pre-selection criterion. Tech. rep., Center for Bioinformatics Tuebingen (ZBIT)
50. Villemonteix J (2008) Optimisation de fonctions coûteuses: Modèles gaussiens pour une utilisation efficace du budget d' évaluations : théorie et pratique industrielle. PhD thesis, Université Paris-sud XI, Faculté des Sciences dOrsay
51. Villemonteix J, Vazquez E, Walter E (2009) An informational approach to the global optimization of expensive-to-evaluate functions. Journal of Global Optimization 44(4):509–534
52. Wahba G (1990) Spline Models for Observational Data. Siam
53. Williams C, Rasmussen C (1996) Gaussian processes for regression. Advances in Neural Information Processing Systems 8