A Multi-points Criterion for Deterministic Parallel Global Optimization based on Gaussian Processes

David Ginsbourger, Rodolphe Le Riche*, Laurent Carraro
Département 3MI
Ecole Nationale Supérieure des Mines
158 cours Fauriel, Saint-Etienne, France
{ginsbourger, leriche, carraro}@emse.fr

March 3, 2008

Abstract

The optimization of expensive-to-evaluate functions generally relies on metamodel-based exploration strategies. Many deterministic global optimization algorithms used in the field of computer experiments are based on Kriging (Gaussian process regression). Starting with a spatial predictor including a measure of uncertainty, they proceed by iteratively choosing the point maximizing a criterion which is a compromise between predicted performance and uncertainty. Distributing the evaluation of such numerically expensive objective functions on many processors is an appealing idea. Here we investigate a multi-points optimization criterion, the *multipoints expected improvement* (q-$\mathbb{E}I$), aimed at choosing several points at the same time. An analytical expression of the q-$\mathbb{E}I$ is given when $q = 2$, and a consistent statistical estimate is given for the general case. We then propose two classes of heuristic strategies meant to approximately optimize the q-$\mathbb{E}I$, and apply them to Gaussian Processes and to the classical Branin-Hoo test-case function. It is finally demonstrated within the covered example that the latter strategies perform as good as the best Latin Hypercubes and Uniform Designs ever found by simulation (2000 designs drawn at random for every $q \in [1, 10]$).

Key words: Kriging, Expected Improvement, EGO, active learning, Monte-Carlo

*C.N.R.S. UMR 5146

1 Introduction

In many engineering applications, such as car crash tests, nuclear criticality safety, reservoir forecasting, the time needed to simulate the physical phenomena is so long that the experimenter can only afford a few simulation runs. It is common to see a deterministic simulator as a numerical black-box function

$$y : \mathbf{x} \in D \subset \mathbb{R}^d \to y(\mathbf{x}) \in \mathbb{R} \tag{1}$$

y is known at a Design of Experiments $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\} \in (\mathbb{R}^d)^d$, where $n \in \mathbb{N}$ is the number of initial runs or experiments. 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 that help understanding the function y with an accuracy that depends on n, the geometry of $\mathbf{X}$, and the regularity of y. This partial knowledge of y is needed to build simplified representations of the simulator, also called *surrogate models* or *metamodels*. A metamodel can be used for predicting values of y outside the initial design or visualizing the influence of each variable on y ([9],[13],[18]). It 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 ([9]).This paper proposes metamodel-based optimization 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, named the multi-points expected improvement. Calculations are performed in the framework of Gaussian processes. In particular, the metamodel considered is Ordinary Kriging (see eqs. 3, 4, and 46).

2 Gaussian processes and sequential optimization

2.1 Ordinary Kriging

Probabilistic metamodeling seems to be particularly adapted for the optimization of black-box functions, as analyzed and illustrated in ([7]). Our work follows ([9]), where Ordinary Kriging (OK) is used to derive a sequential optimization strategy (EGO). Kriging is an interpolation method originally developped in geostatistics ([1],[16]). It provides a predictor of spatial phenomena, with a measure of uncertainty quantifying the accuracy of the prediction at each site (A full derivation is proposed in the appendix). Ordinary Kriging is based on the assumption that y is a realization of a stationary Gaussian process Y with unknown constant mean and known covariance structure ([4]). In Kriging-based optimization, one often abusively plugs in maximum likelihood covariance hyperparameters without taking the estimation variance into account ([9]). Here we commit this abuse, and work with the classical Ordinary Kriging equations. This has the advantage of delivering a Gaussian posterior distribution, even if the uncertainty is slightly underestimated :

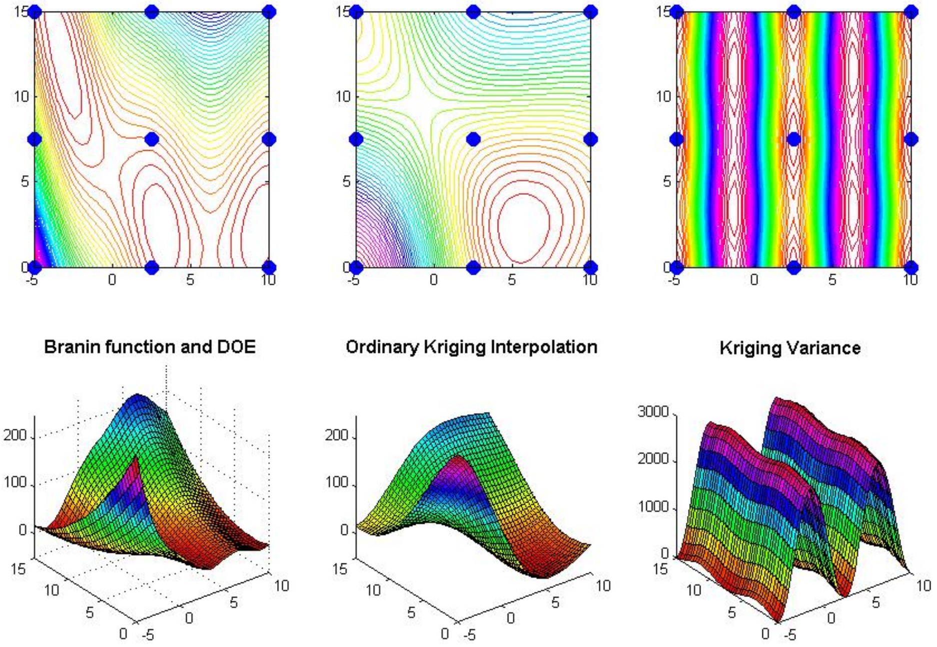
$$\forall \mathbf{x} \in D,\ [Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}(m_{OK}(\mathbf{x}), s^2_{OK}(\mathbf{x})) \tag{2}$$

where the kriging mean and variance functions are given by the following formulae ([16]):

$$m_{OK}(\mathbf{x}) = \mathbb{E}[Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}] = \left[c(\mathbf{x}) + \left(\frac{1 - c(\mathbf{x})^T \Sigma^{-1} \mathbb{1}_n}{\mathbb{1}_n^T \Sigma^{-1} \mathbb{1}_n} \right) \mathbb{1}_n \right]^T \Sigma^{-1} \mathbf{Y} \tag{3}$$

$$s^2_{OK}(\mathbf{x}) = Var[Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}] = \left[\sigma^2 - c(\mathbf{x})^T \Sigma^{-1} c(\mathbf{x}) + \frac{(1 - \mathbb{1}_n^T \Sigma^{-1} c(\mathbf{x}))^2}{\mathbb{1}_n^T \Sigma^{-1} \mathbb{1}_n} \right] \tag{4}$$

with $c(x) = \left(cov(Y(\mathbf{x}), Y(\mathbf{x}^1)), ..., cov(Y(\mathbf{x}), Y(\mathbf{x}^n)) \right)^T,\ \Sigma = \left(cov(Y(\mathbf{x}^i), Y(\mathbf{x}^j)) \right)_{i,j \in [1,n]}$, and $\sigma^2 = Var[Y(\mathbf{x})]$ (which is not depending on $\mathbf{x}$ since Y is stationary).


Branin function and DOE

Ordinary Kriging Interpolation

Kriging Variance

Figure 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 ([16]).

In other terms, under the Gaussian process assumptions that have been made, the random variable $Y(\mathbf{x})$ knowing previous observations $\{Y(\mathbf{x^1}), ..., Y(\mathbf{x^n})\}$ follows a normal distribution which mean and variance are $m_{OK}(\mathbf{x})$ and $s^2_{OK}(\mathbf{x})$, respectively.
A full bayesian interpretation can be found in ([18]), or more recently in ([10]). Classical properties of Ordinary Kriging 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 correlated random variables, where $\mathbf{x}^a$ and $\mathbf{x}^b$ are arbitrary points of D (see Appendix C.2).

The OK metamodel of the Branin-Hoo function (see eq. (25)) is plotted on fig. (2.1). The OK interpolation (upper middle) is made only on the basis of the 9 observations (as can be seen in eq. 3). Even if the shape is reasonably respected (lower middle), the contour of the interpolator shows an artificial optimal zone (upper middle, around the point $(6, 2)$). In other respects, the variance is not depending on the observations[1] (see eq. (4)). Note the particular shape of the variance, due to the strong anisotropy of the covariance function estimated by likelihood maximization.

2.2 Kriging-based optimization criteria

Such a Gaussian process regression has been used for optimization (minimization, by default). There is a detailed review of existing optimization methods relying on a metamodel in [7]. It analyzes and illustrates why directly optimizing a deterministic metamodel (like a spline, a polynomial, or the kriging mean) may be dangerous, and does not even necessarily lead to a local optimum. Kriging-based sequential optimization strategies (as developped in [9], and commented in [7]) 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 optimization algorithms produce one point at each iteration that maximizes a figure of merit (or criterion) 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 highest uncertainty: maximizing s_{OK}

The fundamental mistake of minimizing the Kriging mean (m_{OK}) when globally minimizing a function is that no account is done of the uncertainty associated with m_{OK}. At the extreme inverse, it is possible to define the next optimization iterate as the least known point in D,

$$\mathbf{x}' = argmax_{\mathbf{x} \in D} s_{OK}(\mathbf{x}) \tag{5}$$

This procedure defines a series of $\mathbf{x}'$s which will fill the space D (it is dense in D) and, in this sense, it will ultimately locate $\mathbf{x}^*$, a global optimum. Yet, since no use is made of

[1] phenomenon known as homoskedasticity of the Kriging variance with respect to the observations ([16])

previously obtained $\mathbf{Y}$ information (look at formula (4) for s_{OK}^2), there is no bias in favor of high performance regions. Maximizing the uncertainty is inefficient in practice.

2.2.2 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 two criteria problem

$$\begin{cases} \min_{\mathbf{x} \in D} m_{OK}(\mathbf{x}) \\ \text{and} \max_{\mathbf{x} \in D} s_{OK}(\mathbf{x}) \end{cases} \tag{6}$$

Let $\mathcal{P}$ denote the Pareto set of solutions [2]. Finding one (or many) elements in $\mathcal{P}$ remains a difficult problem since $\mathcal{P}$ typically contains an infinite number of points. A comparable approach called *direct* ([8])), although not based on Kriging, is described in ([8]) : the metamodel is piecewise constant and the uncertainty measure is an Euclidean 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 a parallelized version of *direct*.

2.2.3 Maximizing the probability of improvement

Among the numerous criteria presented in [7] and [12], the probability of improving the function beyond the currently known minimum $min(\mathbf{Y}) = \min\{y(\mathbf{x}^1), ..., y(\mathbf{x}^n)\}$ seems to be one of the most fundamental:

$$\begin{aligned} PI(\mathbf{x}) &= P(Y(\mathbf{x}) \leq \min(\mathbf{Y}) / Y(\mathbf{X}) = \mathbf{Y}) && (7) \\ &= \mathbb{E}[\mathbb{1}_{Y(\mathbf{x}) \leq min(\mathbf{Y})} / Y(\mathbf{X}) = \mathbf{Y}] = \Phi\left(\frac{\min(\mathbf{Y}) - m_{OK}(\mathbf{x})}{s_{OK}(\mathbf{x})}\right) && (8) \end{aligned}$$

$\min(\mathbf{Y})$ is sometimes replaced by some arbitrary target $T \in \mathbb{R}$. The PI criterion is known to provide a very local search whenever the value of T is close to $min(\mathbf{Y})$. Taking several T's is a remedy proposed by [7] to force global exploration.

2.2.4 Maximizing the expected improvement

An alternative solution is to maximize the *expected improvement*

$$EI(\mathbf{x}) = \mathbb{E}[\max\{0, \min(\mathbf{Y}) - Y(\mathbf{x})\} / Y(\mathbf{X}) = \mathbf{Y}] \tag{9}$$

that additionally takes into account the magnitude of the potential improvement. EI measures how much improvement is expected when sampling at x. *In fine*, the improvement

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

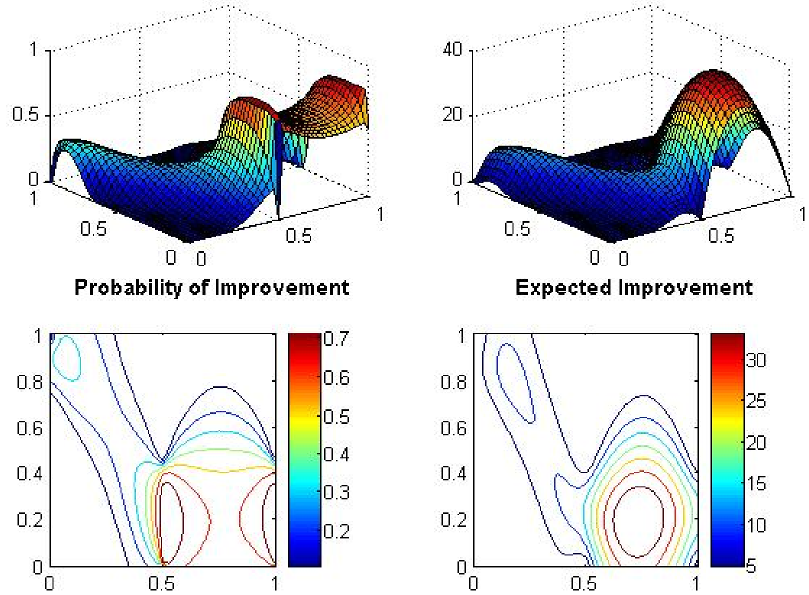


Figure 2: PI and EI surfaces of the Branin-Hoo function (same design of experiments, Kriging model, and covariance parameters as in fig. (2.1)). Maximizing PI leads to sample near the good points (associated with low observations) whereas maximizing EI leads here to sample between the good points. By 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 lower observed value.

will be 0 if the actual $y(\mathbf{x})$ is above $min(\mathbf{Y})$ and $min(\mathbf{Y})-y(\mathbf{x})$ in the opposite case. Since we know the conditional distribution of $Y(\mathbf{x})$, it is straightforward to calculate EI in closed form (see [9]):

$$\begin{aligned}EI(\mathbf{x}) &= \mathbb{E}[(min(\mathbf{Y}) - Y(\mathbf{x}))\mathbb{1}_{Y(\mathbf{x})\leq min(\mathbf{Y})}/Y(\mathbf{X}) = \mathbf{Y}] \\ &= (min(\mathbf{Y}) - m_{OK}(\mathbf{x}))\Phi\left(\frac{min(\mathbf{Y}) - m_{OK}(\mathbf{x})}{s_{OK}(\mathbf{x})}\right) + s_{OK}(\mathbf{x})\phi\left(\frac{min(\mathbf{Y}) - m_{OK}(\mathbf{x})}{s_{OK}(\mathbf{x})}\right)\end{aligned} \tag{10}$$

where ϕ and Φ stand for the probability density function and cumulative distribution function of the standard normal law $\mathcal{N}(0,1)$. EI represents a trade-off between promising and uncertain zones. EI 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 [8] for instance). The expected improvement and the probability of improvement are compared in fig. (2).

2.2.5 The *Stepwise Uncertainty Reduction* strategy

The stepwise uncertainty reduction (SUR) strategy has been introduced in ([5]) and extended to global optimization in ([12]). By looking at possible objective functions as conditional processes, $Y(\mathbf{x})/\mathbf{Y}$, it is possible to define $\mathbf{x}^*/\mathbf{Y}$, the random vector of the location of the minimizer of $Y(\mathbf{x})/\mathbf{Y}$, of density $p_{\mathbf{x}^*/\mathbf{Y}}(\mathbf{x})$. The uncertainty about the location of the optimum of $Y(x)$ 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,

$$\mathbf{x}' = argmin_{\mathbf{x}\in D} H(\mathbf{x}^*/\mathbf{Y}, Y(\mathbf{x})) \tag{11}$$

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 the number of dimensions to properly fill the space. The SUR criterion is different in nature from the other criteria presented so far in that it does not maximize an immediate (i.e. at the next iteration) payoff defined in terms of Y but rather lays the foundation of a more delayed payoff by gaining a more global knowledge on Y (reduce the entropy of its optima). The multi-points expected improvement criterion introduced in the present article also uses a delayed payoff measure.

2.2.6 The *Efficient Global Optimization* (EGO) algorithm

The EGO algorithm ([9]) 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 Kriging 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 and applied in [15], EGO has been considered as a reference and has inspired contemporary works in optimization of expensive-to-evaluate functions. For instance, ([11]) exposes some EGO-based methods for the optimization of noisy black-box functions. ([14]) proposes an adaptation of EGO to multi-objective optimization. EGO does not allow parallel evaluations of y, which is desirable for costly simulators (for instance, a crash-test simulation run typically lasts 24 hours). Here we present a criterion meant to choose an arbitrary number of points without intermediate evaluations of y.

3 The multi-points expected improvement

The main objective of this article is to propose and analyze a global optimization criterion, the multi-points expected improvement or q-points EI, that yields many (say q) points. Since the q-points EI is an extension of the expected improvement, all derivations are performed within the framework of Ordinary Kriging. Such criterion is the first step towards a parallelized version of the EGO algorithm [9]. It also departs, like the SUR criterion, from other criteria that look for an immediate payoff.

The q-points EI criterion (as already defined but not developed in ([15]) under the name "q-step EI") is the expectation of the improvement brought by the q considered points:

$$\begin{aligned} EI(\mathbf{x}^{n+1},...,\mathbf{x}^{n+q}) &= \mathbb{E}\left[\max\left\{(min(\mathbf{Y})-Y(\mathbf{x}^{n+1}))^{+},...,(min(\mathbf{Y})-Y(\mathbf{x}^{n+q}))^{+}\right\}/Y(\mathbf{X})=\mathbf{Y}\right] \\ &= \mathbb{E}\left[\left(\min\left(\mathbf{Y}\right)-\min\left(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+q})\right)\right)^{+}/Y(\mathbf{X})=\mathbf{Y}\right] \end{aligned} \tag{12}$$

Hence, the q-points 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 have to deal with a minimum of dependent random variables. Fortunately, classical results of multivariate statistics[3] provide us with the exact joint distribution of the q unknown responses conditionally on the observations:

$$[(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+q}))/Y(\mathbf{X})=\mathbf{Y}] \sim \mathcal{N}((m_{OK}(\mathbf{x}^{n+1}),...,m_{OK}(\mathbf{x}^{n+q})),S_q) \tag{13}$$

where the elements of the conditional covariance matrix S_q are:

$$\begin{aligned} (S_q)_{i,j} =& c(\mathbf{x}^{n+i}-\mathbf{x}^{n+j})-\mathbf{c}(\mathbf{x}^{n+i})^{T}\Sigma^{-1}\mathbf{c}(\mathbf{x}^{n+j}) \\ &+\sigma^{2}\left[\frac{(1-\mathbb{1}_n^{T}\Sigma^{-1}\mathbf{c}(\mathbf{x}^{n+i}))(1-\mathbb{1}_n^{T}\Sigma^{-1}\mathbf{c}(\mathbf{x}^{n+j}))}{\mathbb{1}_n^{T}\Sigma^{-1}\mathbb{1}_n}\right] \end{aligned} \tag{14}$$

A full derivation of the joint Simple and Ordinary Kriging predictors and some overall considerations about the minimum of dependent random variables are presented respectively in the Appendix C.2.-C.4. and A.1.

[3] Cochran's theorem for the projection of Gaussian vectors

3.1 Analytical calculation of 2-EI

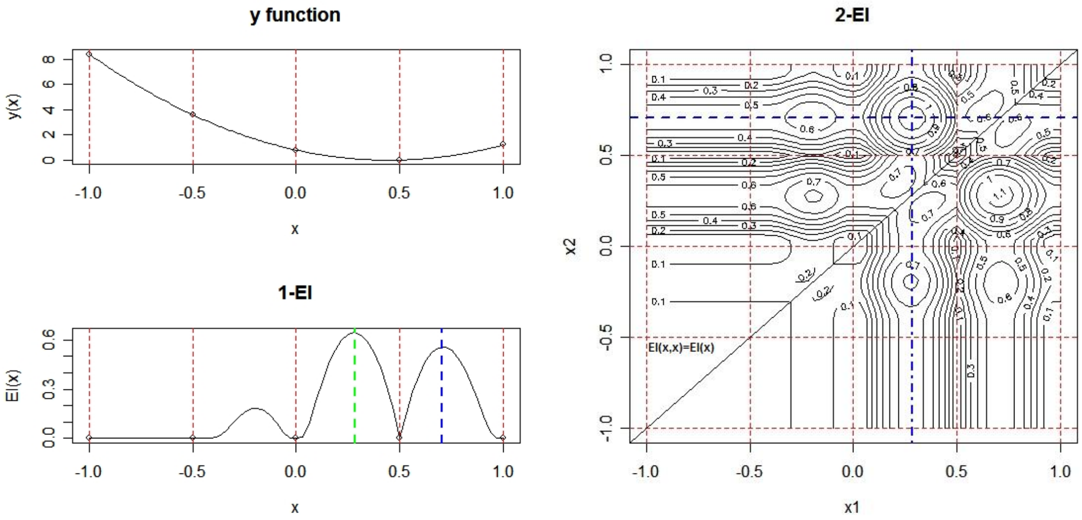


Figure 3: 1-point EI (lower left) and 2-points EI (right) functions associated with a monodimensional quadratic function ($y(x) = 4 \times (x - 0.45)^2$ known at $\mathbf{X} = \{-1, -0.5, 0, 0.5, 1\}$. The ordinary kriging has here a cubic covariance with parameters $\sigma^2 = 10, \text{ scale} = 0.9$).

2-EI can be derived as an expression depending on the mono- and bi-dimensional Gaussian cdf's. Using the following decomposition

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

one can analytically calculate $EI(\mathbf{x}^{n+1}, \mathbf{x}^{n+2})$. A complete derivation of the 2-points EI and some basic properties are proposed in the appendix A.2. and A.3.
fig. (3.1) represents the 1-EI and the 2-EI contour plots associated with a deterministic polynomial function known at 5 points. The 1-point EI advises here to sample between the "good points" of the initial design. The 2-points EI contour illustrates some general

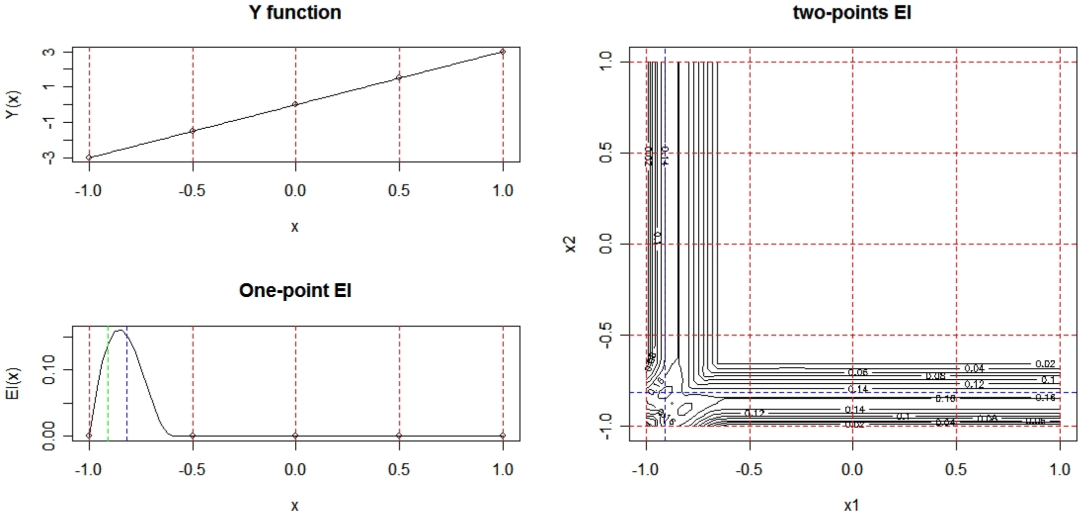


Figure 4: 1-point EI (lower left) and 2-points EI (right) functions associated with a monodimensional linear function ($y(x) = 3 \times x$) known at $\mathbf{X} = \{-1, -0.5, 0, 0.5, 1\}$. The ordinary kriging has here a cubic covariance with parameters $\sigma^2 = 10,\ \text{scale} = 1.4$).

properties: 2-EI is symmetric and its diagonal equals the 1-point EI, which 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 kriging). 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. Other examples illustrate 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 the 2-EI to the general case gives a complex expression depending on q-dimensional Gaussian cdf's. Hence, it seems that the direct computation of q-EI when q grows large would have to rely on numerical multivariate integral approximation techniques anyway. Therefore, directly evaluating q-EI by Monte-Carlo Simulation then makes sense. Thanks to Eqs. (13) and (14), the random vector $[(Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+q}))/\mathbf{Y}]$ can easily be simulated using the Mahalanobis decomposition of Gaussian vectors:

$$\forall k \in [1, n_{sim}],\ M_k = (m_{OK}(\mathbf{x}^{n+1}), ..., m_{OK}(\mathbf{x}^{n+q})) + [S_q^{\frac{1}{2}} N_k]^T, N_k \sim \mathcal{N}(\mathbf{0}_q, \mathbf{I}_q)\ i.i.d. \quad (15)$$

Computing the integral of any function (not necessarily linearly) depending on the vector $[(Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+q}))/\mathbf{Y}]$ can then be done in averaging the images of the simulated vectors by the considered function:

```
function Q-𝔼I($\mathbf{X}$, $\mathbf{Y}$, $\mathbf{X}^{new}$)
    $L = \mathrm{chol}(Var[Y(\mathbf{X}^{new})/Y(\mathbf{X}) = \mathbf{Y}])$          ▷ Cholesky decomposition of $S_q$
    for $i \leftarrow 1, n_{sim}$ do
        $N \sim \mathcal{N}(0, I_q)$                                  ▷ Drawing a vector $N$ at random
        $M_i = m_{OK}(\mathbf{X}^{new}) + LN$                            ▷ Simulating Y at $\mathbf{X}^{new}$
        $qI_{sim}(i)$=[min($\mathbf{Y}$) − min($M_i$)]$^+$             ▷ Simulating the improvement at $\mathbf{X}^{new}$
    end for
    $qEI_{sim} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} qI_{sim}(i)$                ▷ Empirical Expected Improvement
end function
```

A straightforward application of the Law of Large Numbers yields indeed

$$qEI_{sim} = \sum_{i=1}^{n_{sim}} \frac{[\min(\mathbf{Y}) - \min(M_i)]^+}{n_{sim}} \xrightarrow[n_{sim}\to+\infty]{} EI(\mathbf{x}^1, ..., \mathbf{x}^q) \text{ a.s.} \tag{16}$$

The Central Limit Theorem can finally be used to control the precision of the Monte Carlo approximation as a function of n_{sim} (see [2] for details concerning the variance estimation):

$$\sqrt{n_{sim}} \left(\frac{qEI_{sim} - EI(\mathbf{x}^1, ..., \mathbf{x}^q)}{\sqrt{Var[I(\mathbf{x}^1, ..., \mathbf{x}^q)]}} \right) \xrightarrow[n_{sim}\to+\infty]{} \mathcal{N}(0,1) \text{ in law} \tag{17}$$

4 Approximated q-EI maximization

In the last section, we presented a multi-points criterion meant to deliver a design of experiments in one step through the optimization problem

$$(\mathbf{x}'^{n+1}, \mathbf{x}'^{n+2}, ..., \mathbf{x}'^{n+q}) = argmax_{\mathbf{X}' \in D^q}[EI(\mathbf{X}')] \tag{18}$$

However, the computation of q-EI becomes intensive as q increases. Moreover, the optimization problem (18) is of dimension $d \times q$. Here we try to find pseudo-sequential strategies that approach the result of problem (18) while avoiding its numerical cost. Let us first come back to the notations. In the following, we will use the shortcut

$$EI[Y(\mathbf{Z}) = z](\mathbf{x}) = \mathbb{E}[(min(\mathbf{Y}, Y(\mathbf{Z})) - Y(\mathbf{x}))^+ / Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{Z}) = z] \tag{19}$$

where $\mathbf{Z}$ stands for a set of points in D and z is a vector of (true or assumed) images of $\mathbf{Z}$ by y. For instance, expressing q iterations of EGO (without hyperparameter updating) in this formalism yields

$$\left\{\begin{array}{r} \mathbf{x}^{n+1} = argmax_{\mathbf{x}\in D} EI(\mathbf{x}) = argmax_{\mathbf{x}\in D} EI[\ ](\mathbf{x}) \\ \forall j \in [1, q-1],\ \mathbf{x}^{n+j+1} = argmax_{\mathbf{x}\in D} EI[Y(\mathbf{x}^{n+j}) = y(\mathbf{x}^{n+j}), \\ ..., Y(\mathbf{x}^{n+1}) = y(\mathbf{x}^{n+1})](x) \end{array}\right. \tag{20}$$

Note that this formalism holds when the event "$Y(\mathbf{Z}) = z$" is replaced by an event of the form "$Y(\mathbf{Z})$". E.g. if $Y(\mathbf{Z})$ is random, $EI[Y(\mathbf{Z})](\mathbf{x}) = \mathbb{E}[(min(\mathbf{Y}, Y(\mathbf{Z})) - Y(\mathbf{x}))^{+}/Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{Z})]$ then becomes a random variable too, depending on the random variable $Y(\mathbf{Z})$. This is the basis of the following strategies.

4.1 A q-points design built with the 1-point expected improvement

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 best single point $\mathbf{x}^{n+1} = argmax_{\mathbf{x}\in D} EI(\mathbf{x})$, then feed the model and look for $\mathbf{x}^{n+2} = argmax_{\mathbf{x}\in D} EI[Y(\mathbf{x}^{n+1})](x)$, 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 $\mathbf{x}^{n+1}$ has already been visited, and $[Y(\mathbf{x}^{n+1})/\mathbf{Y} = Y(\mathbf{X})]$ is a random variable with known distribution. More precisely, the latter is $[Y(\mathbf{x}^{n+1})/\mathbf{Y} = Y(\mathbf{X})] \sim \mathcal{N}(m_{OK}(\mathbf{x}^{n+1}), s^2_{OK}(\mathbf{x}^{n+1}))$. Hence, the second site $\mathbf{x}^{n+2}$ can be computed as:

$$\mathbf{x}^{n+2} = argmax_{\mathbf{x}\in D} \mathbb{E}\left[EI[Y(\mathbf{x}^{n+1})](\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}\right] \tag{21}$$

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

$$\begin{aligned} &\mathbf{x}^{n+j+1} = argmax_{\mathbf{x}\in D} \mathbb{E}\left[EI[Y(\mathbf{x}^{n+j}), ..., Y(\mathbf{x}^{n+1})](x)/Y(\mathbf{X}) = \mathbf{Y}\right] \\ &= argmax_{\mathbf{x}\in D} \int_{\mathbf{u}\in\mathbb{R}^j} \left[EI[(Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+j-1})) = \mathbf{u}](\mathbf{x})\right] f_{(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+j}))/Y(\mathbf{X})=\mathbf{Y}}(\mathbf{u}) d\mathbf{u} \end{aligned} \tag{22}$$

where $f_{(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+j}))/Y(\mathbf{X})=\mathbf{Y}}(.)$ is the multi-Gaussian density of the joint kriging predictor at $(\mathbf{x}^{n+1}, ..., \mathbf{x}^{n+j})$. Although Eq. (22) is a sequentialized version of the q-points expected improvement maximization, it doesn't completely fulfill our objectives. There is still a multi-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 eq. (22).

4.2 Constant Liar and Kriging Believer strategies

Lying to escape intractable calculations

We propose to weaken the conditional knowledge taken into account at each iteration. This 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 The "kriging believer" heuristic

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:

$$\begin{cases} \mathbf{x}^{n+1} = argmax_{\mathbf{x}\in D} EI(\mathbf{x}),\ m_{OK}^{n}(\mathbf{x}^{n+1}) = \mathbb{E}[Y(\mathbf{x}^{n+1})/Y(\mathbf{X}) = \mathbf{Y}] \text{ and } \forall j \in [1, q-1] : \\ \mathbf{x}^{n+j+1} = argmax_{\mathbf{x}\in D} EI[Y(\mathbf{x}^{n+j}) = m_{OK}^{n+j-1}(\mathbf{x}^{n+j}), ..., Y(\mathbf{x}^{n+1}) = m_{OK}^{n}(\mathbf{x}^{n+1})](x) \\ m_{OK}^{n+j}(\mathbf{x}^{n+j+1}) = \mathbb{E}[Y(\mathbf{x}^{n+j+1})/Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{x}^{n+j}) = m_{OK}^{n+j-1}(\mathbf{x}^{n+j}), \\ ..., Y(\mathbf{x}^{n+1}) = m_{OK}^{n}(\mathbf{x}^{n+1})] \end{cases} \tag{23}$$

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

```
function KB($\mathbf{X}$, $\mathbf{Y}$, $q$)
    for $i \leftarrow 1, q$ do
        $\mathbf{x}^{n+i} = argmax_{\mathbf{x}\in D} EI(\mathbf{x})$
        $m_{OK}(\mathbf{x}^{n+i}) = \mathbb{E}[Y(\mathbf{x}^{n+i})/Y(\mathbf{X}) = \mathbf{Y}]$
        $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^{n+i}\}$
        $\mathbf{Y} = \mathbf{Y} \bigcup \{m_{OK}(\mathbf{x}^{n+i})\}$
    end for
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 a kriging surface 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 The "constant liar" heuristic:

Now consider a sequential strategy in which the model is actualized at each iteration with a value exogenously fixed by the user, and not necessarily connected with the Kriging predictor. The strategy referred to as the *constant liar* consists in lying with the same value L for every iteration: maximize the expected improvement (find x_{n+1}), actualize the model as if $y(x_{n+1}) = L$, and so on always with the same $L \in R$:

$$\left\{ \begin{array}{c} \mathbf{x}^{n+1} = argmax_{\mathbf{x}\in D} EI(\mathbf{x}) \text{ and } \forall j \in [1, q-1] : \\ \mathbf{x}^{n+j+1} = argmax_{\mathbf{x}\in D} EI[Y(\mathbf{x}^{n+j}) = L, ..., Y(\mathbf{x}^{n+1}) = L](\mathbf{x}) \end{array} \right. \tag{24}$$

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

```
function CL($\mathbf{X}$, $\mathbf{Y}$, $L$, $q$)
    for $i \leftarrow 1, q$ do
        $\mathbf{x}^{n+i} = argmax_{\mathbf{x}\in D} EI(\mathbf{x})$
        $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^{n+i}\}$
        $\mathbf{Y} = \mathbf{Y} \bigcup \{L\}$
    end for
end function
```

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 the initial design. 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.

5 Empirical comparisons

5.1 Application to 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 ([9],[15],[17]).

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

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 fig. (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 [9]). We build a 10-points optimization design with each strategy. We additionally estimated by Monte Carlo simulations ($n_{sim} = 10^4$) the probability of improvement and the expected improvement brought by the q first points of each strategy (here $q \in \{2, 6, 10\}$). The results are gathered in Table 1.

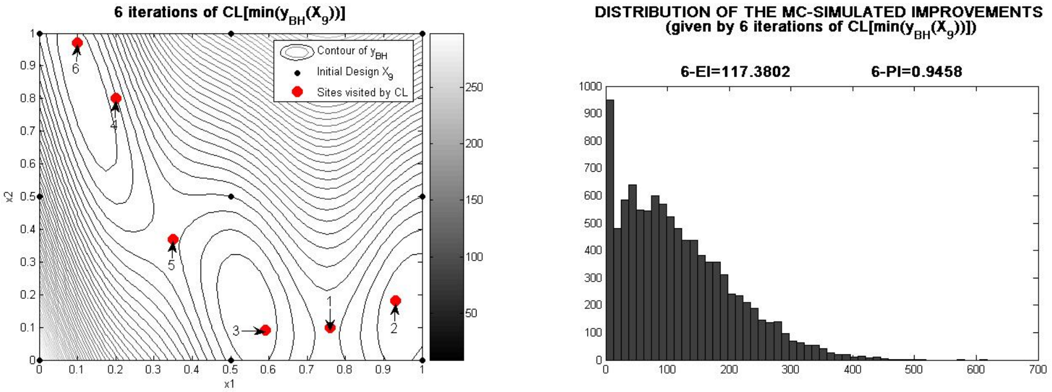


Figure 5: (Left) contour of the Branin-Hoo function with the initial design $\mathbf{X}_9$ (small black points) and the 6 first points given by the heuristic strategy CL[min($f_{BH}(\mathbf{X}_9)$)] (large bullets). (Right) Histogram of 10 000 Monte Carlo simulated values of the improvement brought by the 6-points CL[min($f_{BH}(\mathbf{X}_9)$)] strategy. The corresponding estimations of the 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 *Kriging Believer* (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

	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. (16), $n_{sim} = 10^4$).

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.

The results shown in Table 1 highlight a major drawback of the q-points 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-points EI is a more selective measure thanks to taking the magnitude of possible improvements into account. Nevertheless, the q-EI criterion overevaluates the improvement associated with all designs considered here. This effect (already pointed out in [15]) 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 now compare 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 empirical Expected Improvement distributions to the 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]

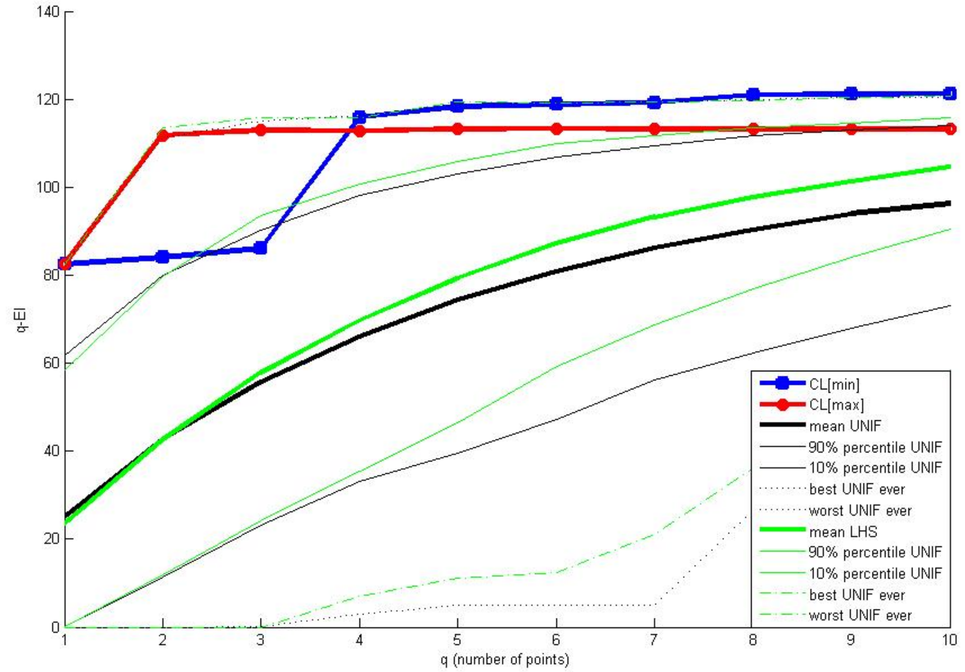


Figure 6: Comparaison of the q-EI associated with the q first points ($q \in [1, 10]$) given by the constant liar strategies (min and max), 2000 q-points designs taken uniformly at random for every q, and 2000 q-points LHS designs taken at random for every q.

for all $4 \leq q \leq 10$). Thr 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 heuristic strategies.

5.2 Kriging-based optimization of gaussian process realizations

With the intent to produce general results, we chose to study and compare the 3 heuristics KB, CL[$\min \mathbf{Y}$], and CL[$\max \mathbf{Y}$] presented in 2.3 in applying them to random functions.

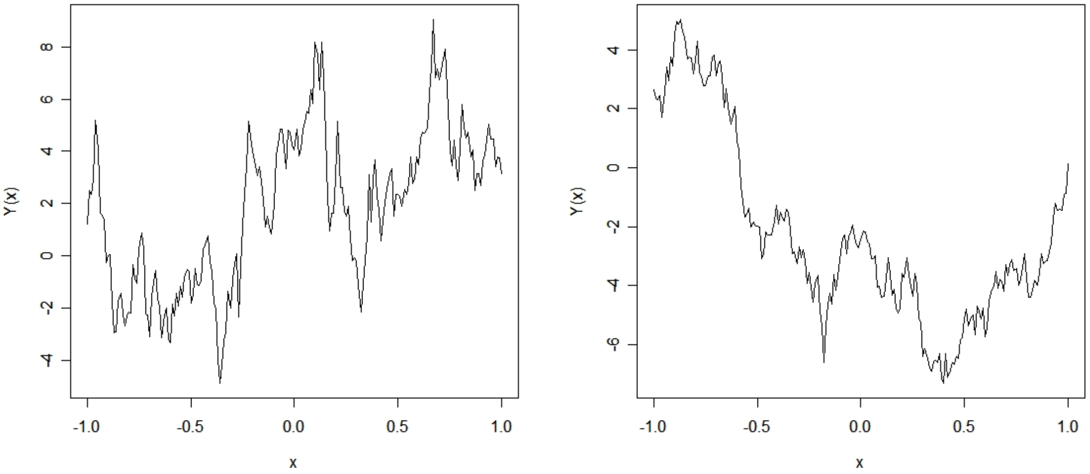


Figure 7: Two stationary Gaussian Process paths (both centered, with variance 10 and exponential covariance structure with respective correlation lengths 0.2 and 0.7). This family of Gaussian Process is often referred to as *Ornstein-Uhlenbeck* Process ([4])

Gaussian Process simulation is a handy way to work with such functions.[4]. We considered four experimental configurations (denoted by $k \in [1, 4]$) involving Gaussian Processes $Y^k(x)$ and 1000 realizations $\{y_i^k(x), i \in [1, 1000]\}$ of them for each configuration. For all configurations, the outputs varied between -1 and 1 ($D = [-1, 1]$), and the initial design of experiments was fixed to the 3-elements set $\mathbf{X} = \{-1, 0, 1\}$ (see fig. (8)). The other experimental parameters varied accordingly to the values specified in Table (5.2).

k	covariance	correlation length	variance	N_k
1	Exponential	0.3	40	2
2	Exponential	1	40	2
3	Exponential	0.3	40	10
4	Exponential	1	40	10

Table 2: Design of experiments for a comparison between the 3 heuristics

Formally, each heuristic strategy $\mathcal{S}$ (here $\mathcal{S} \in \{KB, CL[min], CL[max]\}$) provides a sequence of points $X^{k,1}(\mathcal{S}), ..., X^{k,N_k}(\mathcal{S})$. These points are random variables since they closely

[4] simulating mono- or multi-dimensional Gaussian processes on a grid (having m elements) is theorically (but not always numerically) straightforward, the cost being the inversion of an $m \times m$ covariance matrix.

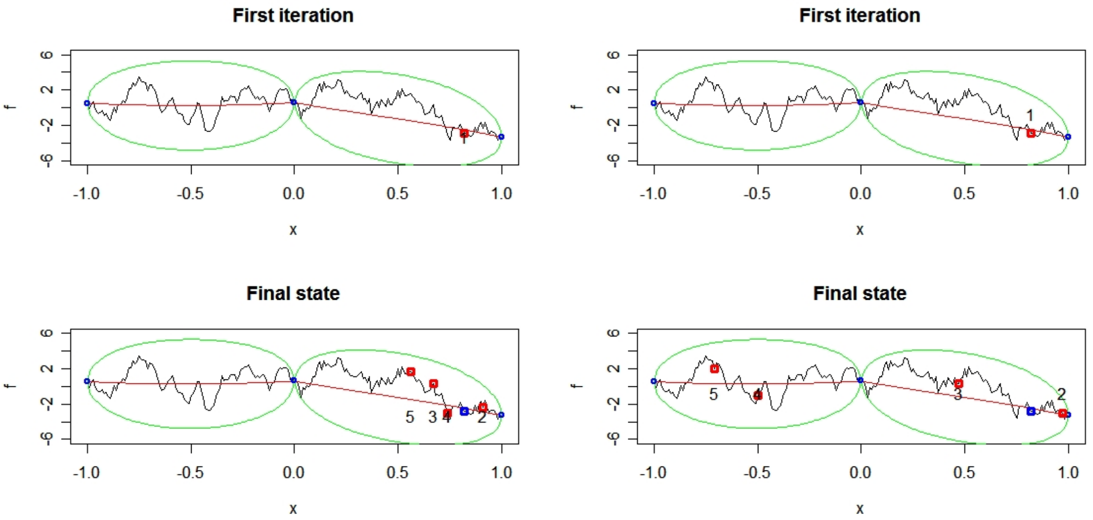


Figure 8: 5 iterations of CL[min($\mathbf{Y}$)] (left) and CL[max($\mathbf{Y}$)] (right) to one Gaussian Process path (scale 0.7, variance 10, exponential covariance). This example clearly illustrates how the CL strategy priveleges local search whenever $L = \min(\mathbf{Y})$, and possesses a more space-filling behaviour when $L = \max(\mathbf{Y})$.

depend on the process Y^k which is itself random. Here we wish to study the performances of each strategy (given a configuration) by looking at the behavior of the random variable

$$\Delta_k(\mathcal{S}) = \min\{Y^k(\mathbf{X}), Y^k(X^{k,1}(\mathcal{S})), ..., Y^k(X^{k,N_k}(\mathcal{S}))\} - \min_{x \in D}[Y^k(x)] \geq 0 \tag{27}$$

which measures how far we are from having perfectly optimized the process $Y^k(x)$ after having ran N_k iterates of the strategy $\mathcal{S}$. Hence, the closer the realizations of $\Delta_k(\mathcal{S})$ are to 0, the better $\mathcal{S}$ fullfils its goals as optimizer.

We studied the experimental performances of the three algorithms applied to the 1000 realizations ran for every configuration k. We considered the realizations of $\Delta_k(\mathcal{S})$,

$$\delta_k^i(\mathcal{S}) = \min\{y_i^k(\mathbf{X}), y_i^k(x_i^{k,1}(\mathcal{S})), ..., y_i^k(x_i^{k,N_k}(\mathcal{S}))\} - \min_{x \in D}[y_i^k(x)] \geq 0 \tag{28}$$

where the $x_i^{k,j}(\mathcal{S})$'s stand for the 1000 realizations of the $X^{k,j}(\mathcal{S})$'s. The results are summarized in figures (10) and (11). The histograms offer concentrated representations of the δ_k^i's distributions, i.e. the statistical performances of each strategy in all studied configurations. Values near 0 (on the extreme left of the histograms) mean succesful optimizations, whereas right tails stand for the cases of failure (best y value observed far beyond the

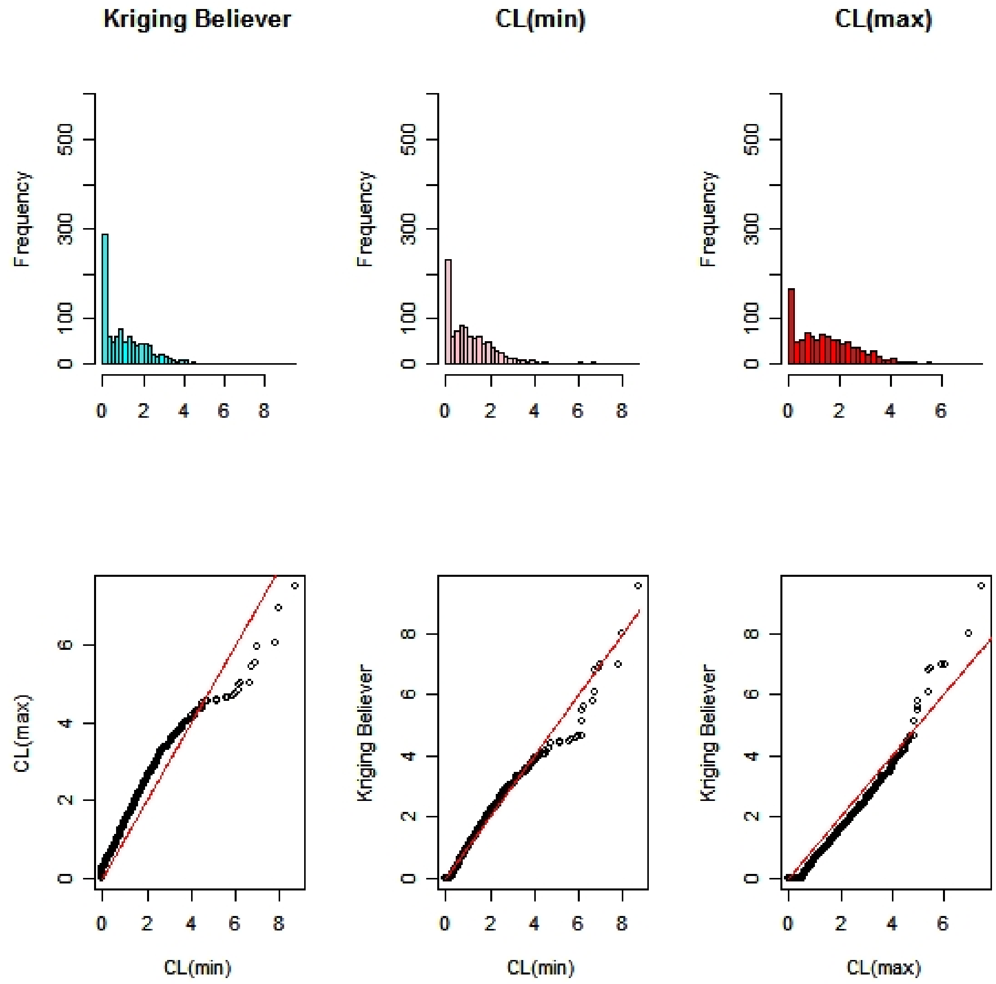


Figure 9: Comparison of the heuristic strategies $CL[min]$, $CL[max]$, KB applied to 1000 Gaussian process realizations with configurations 4. $CL[max]$ and KB keep their positions of best performers, respectively at the right and left extremes. Note the particular shape of the qq-plot between $CL[min]$ and $CL[max]$. The first one is statistically more likely to perform very well, but also more likely to fail dramatically. Conversely to configuration 1 (see appendix D), $CL[max]$ is here a good challenger for a risk averse user.

actual minimum). The *q-q plots* aim at comparing all couples of strategy in plotting the empirical quantiles (i.e. ranked values of the δ_k^i's) of the one against the empirical quantiles of the other. Such kind of graphic allows a far more subtile comparison between strategies than only scalar indexes like the mean or the median performance.

As shown on fig. (11), the Kriging Believer strategy does not behave pathologically anymore when using the exponential covariance: it seems in fact to give optimization results with a very good balance between high performances and risk covering. Even if the three strategies roughly give comparable results within this example, $CL[min]$ and KB appear indeed to provide more often extremely good results (small δ's) than $CL[max]$, which however has thiner tails than $CL[min]$. Note that if the comparaison between strategies is quite stable for small values of δ, this statement doesn't hold for high quantiles since the corresponding fluctuations are too large for samples of 1000 process realizations.

6 Conclusions

Gaussian Process regression is very convenient for metamodel-based optimization. Its probabilistic frame allows to build explicit criteria accounting for the exploration/exploitation trade-off, like the *expected improvement*, EI. The q-EI criterion developed here makes it possible to get an evaluation of the "optimization potential" given by a set of q experiments. It can be used to analytically derive EI-optimal singletons and couples. Monte-Carlo simulations offer the opportunity to evaluate the q-EI associated with any given design of experiment, whatever its size. 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 that they provide higher q-EI's than Latin Hypercubes and random uniform designs of experiments.

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. We wish to thank Xavier Bay, Raphael T. Haftka, Ben Smarslok, Yann Richet, Olivier Roustant, and Victor Picheny for their help and rich comments. Special thanks to the R project people ([6]) for developing and spreading such a useful freeware.

A More details on the 2-points Expected Improvement

A.1 Minimum of two random variables

Let us consider two real random variables U and V defined on the same probability space. In the case where U and V are independent, the cumulative distribution of the couple (U, V) is well known:

$$\forall x, y \in \mathbb{R},\ P(U \leq x, V \leq y) = P(U \leq x) \times P(V \leq y) = F_U(x) \times F_V(y)$$

This is the key to the distribution of $m = min(U, V)$ since $\forall x \in \mathbb{R},\ P(m \leq x) = 1 - P(m > x) = 1 - P(U > x, V > x) = 1 - P(U > x) \times P(V > x) = 1 - (1 - P(U \leq x)) \times (1 - P(V \leq x)) = P(U \leq x) + P(V \leq x) - P(U \leq x) \times P(V \leq x)$. For instance, in the case where $N_1, N_2 \sim \mathcal{N}(\mu, \sigma^2)$ independently, the distribution of their minimum is given by:

$$\forall x \in \mathbb{R},\ P(min(N_1, N_2) \leq x) = 2\left[\Phi(\frac{x-\mu}{\sigma})\right] - \left[\Phi(\frac{x-\mu}{\sigma})\right]^2 \qquad (29)$$

where Φ is the gaussian cumulative distribution function. Now we consider the case of two dependent random variables U and V. Since the last multiplicativity property doesn't hold anymore, all we can say is that $\forall x \in \mathbb{R},\ P(m \leq x) = P(U \leq x, V \leq x)$ which is the cumulative distribution function of the random vector (U, V) evaluated at (x, x). In the case where (U, V) is a dependent multigaussian vector $(N_1, N_2) \sim \mathcal{N}(\mu, \Sigma)$, we have the more general expression:

$$\forall x \in \mathbb{R},\ P(min(N_1, N_2) \leq x) = P(N_1 \leq x, N_2 \leq x) = CDF(\mu, \Sigma)(x, x) \qquad (30)$$

where CDF stands for the bi-gaussian cumulative distribution function. This is the integral of the bi-gaussian density function (corresponding to the distribution $\mathcal{N}(\mu, \Sigma)$) over the surface of the southwestern quadrant delimited by (x, x). This expression is so forth considered as analytically intractable and must be numerically approximated.

A.2 Analytical calculation of the 2-points Expected Improvement

Some classical results of conditional calculus allow us to precise this dependance and fix the notations. Let us first give shortened notations for the means, standard deviations, and covariance of the random variables $(Y_{OK}(\mathbf{x}^i) = [Y(\mathbf{x}^i)/Y(\mathbf{X}) = \mathbf{Y}], i \in \{1, 2\})$:

$$\begin{aligned}
&m_i = \mathbb{E}[Y(\mathbf{x}^i)/\mathbf{Y}] = m_{OK}(\mathbf{x}^i), \sigma_i = \sqrt{Var[Y(\mathbf{x}^i)/\mathbf{Y}]} = s_{OK}(\mathbf{x}^i),\\
&c_{1,2} = cov[Y_{OK}(\mathbf{x}^1), Y_{OK}(\mathbf{x}^2)/\mathbf{Y}] = C_{12} = \rho_{1,2}\sigma_1\sigma_2
\end{aligned}$$

Well-kwown results from linear regression (for instance) then give us conditional means and variances of one response knowing the other:

$$m_{2/1} = E[Y(\mathbf{x}^2)/\mathbf{Y}, Y_{OK}(\mathbf{x}^1)] = m_2 + \frac{c_{1,2}}{\sigma_1^2}(Y_{OK}(\mathbf{x}^1) - m_1),\ \sigma_{2/1}^2 = \sigma_2^2 - \frac{c_{1,2}^2}{\sigma_1^2} = \sigma_2^2(1 - \rho_{12}^2) \qquad (31)$$

$$m_{1/2} = E[Y(\mathbf{x}^1)/\mathbf{Y}, Y_{OK}(\mathbf{x}^2)]m_1 + \frac{c_{1,2}}{\sigma_2^2}(Y_{OK}(\mathbf{x}^2) - m_2),\ \sigma_{1/2}^2 = \sigma_1^2 - \frac{c_{1,2}^2}{\sigma_2^2} = \sigma_1^2(1 - \rho_{12}^2) \qquad (32)$$

At this stage we are in position to compute $EI(\mathbf{x}^1, \mathbf{x}^2)$. Starting here, we replace the complete notation $Y_{OK}(\mathbf{x}^i)$ by Y_i and forget the conditioning on $\mathbf{Y}$ for the sake of clarity.

Phase 1

$$\begin{aligned}
EI(\mathbf{x}^1, \mathbf{x}^2) &= E[I(\mathbf{x}^1, \mathbf{x}^2)] = E[max(0, min(\mathbf{Y}) - min(Y_1, Y_2))]\\
&= E[(min(\mathbf{Y}) - min(Y_1, Y_2))\mathbb{1}_{min(Y_1,Y_2) \leq min(\mathbf{Y})}]\\
&= E[(min(\mathbf{Y}) - min(Y_1, Y_2))\mathbb{1}_{min(Y_1,Y_2) \leq min(\mathbf{Y})}(\mathbb{1}_{Y_1 \leq Y_2} + \mathbb{1}_{Y_2 \leq Y_1})]\\
&= E[(min(\mathbf{Y}) - Y_1)\mathbb{1}_{Y_1 \leq min(\mathbf{Y})}\mathbb{1}_{Y_1 \leq Y_2}] + E[(min(\mathbf{Y}) - Y_2)\mathbb{1}_{Y_2 \leq min(\mathbf{Y})}\mathbb{1}_{Y_2 \leq Y_1}]
\end{aligned}$$

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

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

where $B(\mathbf{x}^1, \mathbf{x}^2) = E[(Y_1 - min(\mathbf{Y}))\mathbb{1}_{Y_1 \leq min(\mathbf{Y})}\mathbb{1}_{Y_2 \leq Y_1}]$. Informally, $B(\mathbf{x}^1, \mathbf{x}^2)$ 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-step expected improvement. Our aim in the next phases will be to give an explicit expression for $B(\mathbf{x}^1, \mathbf{x}^2)$.

Phase 2

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

At this point, it is worth noticing that $Y_1 = m_1 + \sigma_1 N_1$ with $N_1 \sim \mathcal{N}(0,1)$. Substituing this decomposition in the last expression of $B(\mathbf{x}^1, \mathbf{x}^2)$ leads to:

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

The two terms of this sum require some attention. We compute both of them in detail respectively in phase 3 and phase 4.

Phase 3

Using a classical property of conditional calculus [6], we have that:

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

Using the fact that $Y_2/Y_1 \sim \mathcal{N}(m_{2/1}(Y_1), s^2_{2/1}(Y_1))$, we obtain the following:

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

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

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

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

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

$$\begin{aligned} \int_{-\infty}^{\gamma_1} u\phi(u)\Phi(\alpha_1 u + \beta_1)du &= [-\phi(u)\Phi(\alpha_1 u + \beta_1)]_{-\infty}^{\gamma} + \int_{-\infty}^{\gamma_1} \alpha_1\phi(u)\phi(\alpha_1 u + \beta_1)du \\ &= -\phi(\gamma_1)\Phi(\alpha_1\gamma_1 + \beta_1) + \frac{\alpha_1}{2\pi}\int_{-\infty}^{\gamma_1} e^{\frac{-u^2 - (\alpha_1 u + \beta_1)^2}{2}} du \end{aligned}$$

[5] This expression should be rigorously noted $1 - \mathbb{1}_{Y_2 < Y_1}$. Since we work here with (continous) gaussian random variables, it sufficies however that their correlation is different from 1 for the expression to be exact ($\{Y_1 = Y_2\}$) is then neglectable). We implicitely do this assumption here and in the following.

[6] For all function ϕ in $L^2(\Omega)$, $E[X\phi(Y)] = E[E[X/Y]\phi(Y)]$

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:

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

Phase 4

We then finally compute the term:

$$E[1_{Y_1 \leq min(\mathbf{Y})} 1_{Y_2 \leq Y_1}] = E[1_{X \leq min(\mathbf{Y})} 1_{Z \leq 0}]$$

where $(X, Z) = (Y_1, Y_2 - Y_1)$ is following a two-dimensional normal distribution of mean $M = (m_1, m_2 - m_1)$, and variance 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:

$$E[1_{X \leq min(\mathbf{Y})} 1_{Z \leq 0}] = CDF(M, \Gamma)(min(\mathbf{Y}), 0)$$

where CDF stands for the bi-gaussian cumulative distribution function.

Proposition:

$$EI(\mathbf{x}^1, \mathbf{x}^2) = EI(\mathbf{x}^1) + EI(\mathbf{x}^2) + B(\mathbf{x}^1, \mathbf{x}^2) + B(\mathbf{x}^2, \mathbf{x}^1) \tag{33}$$

$$\text{with } B(\mathbf{x}^1, \mathbf{x}^2) = (m_{OK}(\mathbf{x}^1) - min(\mathbf{Y}))\delta(\mathbf{x}^1, \mathbf{x}^2) + \sigma_{OK}(\mathbf{x}^1)\epsilon(\mathbf{x}^1, \mathbf{x}^2)$$

$$\epsilon(\mathbf{x}^1, \mathbf{x}^2) = \alpha_1\phi(\frac{|\beta_1|}{\sqrt{(1+\alpha_1^2)}})\Phi(\frac{\gamma + \frac{\alpha_1\beta_1}{1+\alpha_1^2}}{(1+\alpha_1^2)^{-\frac{1}{2}}}) - \phi(\gamma)\Phi(\alpha_1\gamma + \beta_1),\ \delta(\mathbf{x}^1, \mathbf{x}^2) = CDF(\Gamma)\begin{pmatrix} min(\mathbf{Y}) - m_1 \\ m_1 - m_2 \end{pmatrix}$$

B Generalities about the q-points expected improvement

B.1 An alternative definition

After the definition of the bivariate expected improvement, it seems natural to define the multivariate expected improvement as:

$$EI(\mathbf{x}^1, ..., \mathbf{x}^q) = E[max(min(\mathbf{Y}) - min\{Y_{OK}(\mathbf{x}^1), ..., Y_{OK}(\mathbf{x}^q)\}, 0)]$$

Shortening again the notations, we have the equivalent definition:

$$EI(\mathbf{x}^1, ..., \mathbf{x}^q) = \sum_{i=1}^{q} E[(min(\mathbf{Y}) - Y_i)\mathbb{1}_{Y_i \leq min(\mathbf{Y})}(\Pi_{j \neq i}\mathbb{1}_{Y_i \leq Y_j})] \tag{34}$$

Proof of 34: like in phase 1, we use the property $1 = \sum_{i=1}^{q}(\Pi_{j \neq i}\mathbb{1}_{Y_i \leq Y_j})$ which only means that the smallest Y_i is among the Y_is!

B.2 First bounds on q-EI

$$EI(\mathbf{x}^1,...,\mathbf{x}^q) \leq \sum_{i=1}^{q} EI(\mathbf{x}^i) \tag{35}$$

Proof of 35: $\forall i \in [1,q]$, $(\Pi_{j\neq i}\mathbb{1}_{Y_i\leq Y_j}) \leq 1$ and hence $(min(\mathbf{Y})-Y_i)\mathbb{1}_{Y_i\leq min(\mathbf{Y})}(\Pi_{j\neq i}\mathbb{1}_{Y_i\leq Y_j}) \leq (min(\mathbf{Y})-Y_i)\mathbb{1}_{Y_i\leq min(\mathbf{Y})}$. The property follows from 34.

$$EI(\mathbf{x}^1,...,\mathbf{x}^q) \geq \max_{J\subsetneq[1,q]} EI(\{\mathbf{x}^i, i\in J\}) \geq \max_{1\leq i\leq q} EI(\mathbf{x}^i) \tag{36}$$

Proof of 36: Be $J \subsetneq [1,n]$. Both statements directly come from the inequality: $min(Y_i, i\in J) \geq min(Y_i, i \in [1,n])$.

$$\forall \sigma \in \Sigma_n,\ EI(\mathbf{x}^{\sigma(1)},...,\mathbf{x}^{\sigma(q)}) = EI(\mathbf{x}^1,...,\mathbf{x}^n) \tag{37}$$

Proof of 37: This follows the invariance of min by permutation.

C Joint predictions using Simple and Ordinary Kriging

Here we give some details about the calculation of the joint distribution obtained when simultaneously predicting at different points in the cases of Simple and Ordinary Kriging (SK and OK in the following). Let us first recall some basics about Kriging and Gaussian Processes.

C.1 Gaussian Processes for Machine Learning

A real (L^2) 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 $\mathbf{m_X}\in\mathbf{R}^n$ and a symmetric positive semi-definite matrix $\Sigma_\mathbf{X}\in\mathcal{M}_n(\mathbb{R})$ such that $(Y(\mathbf{x}^1),...,Y(\mathbf{x}^n))$ is a Gaussian Vector, following a multigaussian probability distribution $\mathcal{N}(\mathbf{m_X},\Sigma_\mathbf{X})$. More specifically, for all $i\in[1,n]$, $Y(\mathbf{x}^i)\sim\mathcal{N}(\mathbb{E}[Y(\mathbf{x}^i)], Var[Y(\mathbf{x}^i)])$ where $\mathbb{E}[Y(\mathbf{x}^i)]$ is the ith coordinate of $\mathbf{m_X}$ and $Var[Y(\mathbf{x}^i)]$ is the ith diagonal term of $\Sigma_\mathbf{X}$. 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 $\Sigma_\mathbf{X}$ 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 $\forall\mathbf{x}\in D$, $\mathbb{E}[Y(\mathbf{x})]=\mu$ where $\mu\in\mathbb{R}$. Y is said to be *second order stationary* if there exists a positive semidefinite function $c: D-D\longrightarrow\mathbb{R}$ 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 $\mathbf{X}$:

$$\Sigma_\mathbf{X} = (Cov[Y(\mathbf{x}_i),Y(\mathbf{x}_j)])_{i,j\in[1,n]} = (c(\mathbf{x}_i-\mathbf{x}_j))_{i,j\in[1,n]} = \begin{pmatrix} \sigma^2 & c(\mathbf{x}_1-\mathbf{x}_2) & ... & c(\mathbf{x}_1-\mathbf{x}_n) \\ c(\mathbf{x}_2-\mathbf{x}_1) & \sigma^2 & ... & c(\mathbf{x}_2-\mathbf{x}_n) \\ ... & ... & ... & ... \\ c(\mathbf{x}_n-\mathbf{x}_1) & c(\mathbf{x}_n-\mathbf{x}_2) & ... & \sigma^2 \end{pmatrix} \tag{38}$$

where $\sigma^2 := c(0)$. If Y is first and second order stationary, it is said *weakly stationary*. A major feature of Gaussian Processes 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 (in theory) Simple Kriging (SK) assumes a known mean, $\mu \in \mathbb{R}$. In Ordinary Kriging (OK), μ is estimated.

C.2 Conditioning Gaussian Vectors

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

$$\Sigma_V = \mathbb{E}[VV^T] = \begin{pmatrix} \Sigma_{V_1} & \Sigma_{cross}^T \\ \Sigma_{cross} & \Sigma_{V_2} \end{pmatrix} \tag{39}$$

Key properties of Gaussian vectors include that the orthogonal projection of a Gaussian vector 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 independance. 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 V_2, solving the equation:

$$\mathbb{E}[(V_1/ - \mathbb{E}[V_1/V_2])V_2^T] = 0 \tag{40}$$

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.40) 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[7]. We conclude that

$$\mathbb{E}[V_1/V_2] = \Sigma_{cross}^T \Sigma_{V_2}^{-1} V_2 \tag{41}$$

by uniqueness of the orthogonal projection in a Hilbert space. Using the independence between $(V_1 - \mathbb{E}[V_1/V_2])$ and V_2, we calculate the conditional covariance matrix Σ_{V_1/V_2}:

$$\begin{aligned} \Sigma_{V_1/V_2} &= \mathbb{E}[(V_1 - \mathbb{E}[V_1/V_2])(V_1 - \mathbb{E}[V_1/V_2])^T/V_2] \\ &= \mathbb{E}[(V_1 - AV_2)(V_1 - AV_2)^T] \\ &= \Sigma_{V_1} - A\Sigma_{cross} - \Sigma_{cross}^T A^T + A\Sigma_{V_2} A^T \\ &= \Sigma_{V_1} - \Sigma_{cross}^T \Sigma_{V_2}^{-1} \Sigma_{cross} \end{aligned} \tag{42}$$

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 $\mathbb{E}[V_1/V_2] = m_1 + \Sigma_{cross}^T \Sigma_{V_2}^{-1}(V_2 - m_2)$.

C.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:

$$\begin{cases} Y(\mathbf{x}) = \mu + \varepsilon(\mathbf{x}) \\ \varepsilon(\mathbf{x}) \text{ centered stationary GP with covariance function } c(.) \end{cases} \tag{43}$$

where $\mu \in \mathbb{R}$ is a known scalar. 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 $\mu \mathbb{1}_{n+q}$ and covariance matrix

$$\Sigma_{tot} = \begin{pmatrix} \Sigma & \Sigma_{cross}^T \\ \Sigma_{cross} & \Sigma_{new} \end{pmatrix} = \begin{pmatrix} \sigma^2 & c(\mathbf{x}_1 - \mathbf{x}_2) & ... & c(\mathbf{x}_1 - \mathbf{x}_{n+q}) \\ c(\mathbf{x}_2 - \mathbf{x}_1) & \sigma^2 & ... & c(\mathbf{x}_2 - \mathbf{x}_{n+q}) \\ ... & ... & ... & ... \\ c(\mathbf{x}_{n+q} - \mathbf{x}_1) & c(\mathbf{x}_{n+q} - \mathbf{x}_2) & ... & \sigma^2 \end{pmatrix} \tag{44}$$

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

We can directly apply eq. (41) and eq. (42) to derive the Simple Kriging Equations:

$$[Y(\mathbf{X}_{new})/Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}(m_{SK}(\mathbf{X}_{new}), \Sigma_{SK}(\mathbf{X}_{new})) \tag{45}$$

with $m_{SK}(\mathbf{X}_{new}) = \mathbb{E}[Y(\mathbf{X}_{new})/Y(\mathbf{X}) = \mathbf{Y}] = \mu\mathbb{1}_q + \Sigma_{cross}^T\Sigma^{-1}(\mathbf{Y} - \mu\mathbb{1}_q)$ 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.

C.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 $\mu \sim \mathcal{U}(\mathbb{R})$. y is thus seen as a realization of a random process Y, defined as the sum of μ and a centered GP [8]:

$$\begin{cases} Y(\mathbf{x}) = \mu + \varepsilon(\mathbf{x}) \\ \varepsilon(\mathbf{x}) \text{ centered stationary GP with covariance function } c(.) \\ \mu \sim \mathcal{U}(\mathbb{R}) \text{ (prior)} \end{cases} \tag{46}$$

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 $[Y(\mathbf{X}) = \mathbf{Y}/\mu] \sim \mathcal{N}(\mu\mathbb{1}_n, \Sigma)$, we get μ's posterior distribution:

$$[\mu/Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}\left(\hat{\mu}, \sigma_\mu^2\right) = \mathcal{N}\left(\frac{\mathbb{1}^T\Sigma^{-1}\mathbf{Y}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}, \frac{1}{\mathbb{1}_q^T\Sigma^{-1}\mathbb{1}_q}\right) \text{ (posterior)} \tag{47}$$

We can re-write the SK equations $[Y(\mathbf{X}_{new})/Y(\mathbf{X}) = \mathbf{Y}, \mu] \sim \mathcal{N}(m_{SK}(\mathbf{X}_{new}), \Sigma_{SK}(\mathbf{X}_{new}))$. Now it is very useful to notice that the conditional random vector $[(Y(\mathbf{X}_{new}), \mu)/Y(\mathbf{X}) = \mathbf{Y}]$ is Gaussian [9]. 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 $m_{OK}(\mathbf{X}_{new}) = \mathbb{E}[Y(\mathbf{X}_{new})/Y(\mathbf{X}) = \mathbf{Y}] = \mathbb{E}_\mu\left[\mathbb{E}[Y(\mathbf{X}_{new})/Y(\mathbf{X}) = \mathbf{Y}, \mu]\right]$, we find that $m_{OK}(\mathbf{X}_{new}) = \hat{\mu} + \Sigma_{cross}^T\Sigma^{-1}(\mathbf{Y} - \hat{\mu}\mathbb{1}_n)$. 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 get for all couples of points $(\mathbf{x}^{n+i}, \mathbf{x}^{n+j})$ $(i, j \in [1, q])$:

$$\begin{aligned} & Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})/Y(\mathbf{X}) = \mathbf{Y}] \\ =& \mathbb{E}\left[Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})/Y(\mathbf{X}) = \mathbf{Y}, \mu]\right] + Cov\left[\mathbb{E}[Y(\mathbf{x}^{n+i})/Y(\mathbf{X}) = \mathbf{Y}, \mu], \mathbb{E}[Y(\mathbf{x}^{n+j})/, Y(\mathbf{X}) = \mathbf{Y}, \mu]\right] \end{aligned} \tag{48}$$

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 $\mu + \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}(\mathbf{Y} - \mu\mathbb{1}_q)$ and $\mu + \mathbf{c}(\mathbf{x}^{n+j})^T\Sigma^{-1}(\mathbf{Y} - \mu\mathbb{1}_q)$ conditionally to the observations $Y(\mathbf{X}) = \mathbf{Y}$. Using eq. (47), we finally obtain:

$$\begin{aligned} & Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})/Y(\mathbf{X}) = \mathbf{Y}] \\ =& \mathbb{E}\left[Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})/Y(\mathbf{X}) = \mathbf{Y}, \mu]\right] \\ &+ Cov[\mathbf{E}[Y(\mathbf{x}^{n+i})/Y(\mathbf{X}) = \mathbf{Y}, \mu], \mathbf{E}[Y(\mathbf{x}^{n+j})/, Y(\mathbf{X}) = \mathbf{Y}, \mu]] \\ =& Cov_{SK}[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})/Y(\mathbf{X}) = \mathbf{Y}] \\ &+ Cov[\mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}(\mathbf{Y}) + \mu(1 + \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}\mathbb{1}_q), \mathbf{c}(\mathbf{x}^{n+j})^T\Sigma^{-1}(\mathbf{Y}) + \mu(1 + \mathbf{c}(\mathbf{x}^{n+j})^T\Sigma^{-1}\mathbb{1}_q)] \\ =& c(\mathbf{x}^{n+i} - \mathbf{x}^{n+j}) - \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}\mathbf{c}(\mathbf{x}^{n+j}) + \frac{(1 + \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}\mathbb{1}_q)(1 + \mathbf{c}(\mathbf{x}^{n+j})^T\Sigma^{-1}\mathbb{1}_q)}{\mathbb{1}_q^T\Sigma^{-1}\mathbb{1}_q} \end{aligned} \tag{49}$$

[8] The resulting random process Y is not Gaussian

[9] which can be proved by considering its Fourier transform

And the Ordinary Kriging Variance now appears as a particular case. For all $\mathbf{x} \in D$, we have indeed:

$$s_{OK}^2(\mathbf{x}) = Var[Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}] = \sigma^2 - \mathbf{c}(\mathbf{x})^T \Sigma^{-1} \mathbf{c}(\mathbf{x}) + \frac{(1 - \mathbb{1}_n^T \Sigma^{-1} \mathbf{c}(\mathbf{x}))^2}{\mathbb{1}_n^T \Sigma^{-1} \mathbb{1}_n} \tag{50}$$

D More graphics to compare the KB and CL strategies

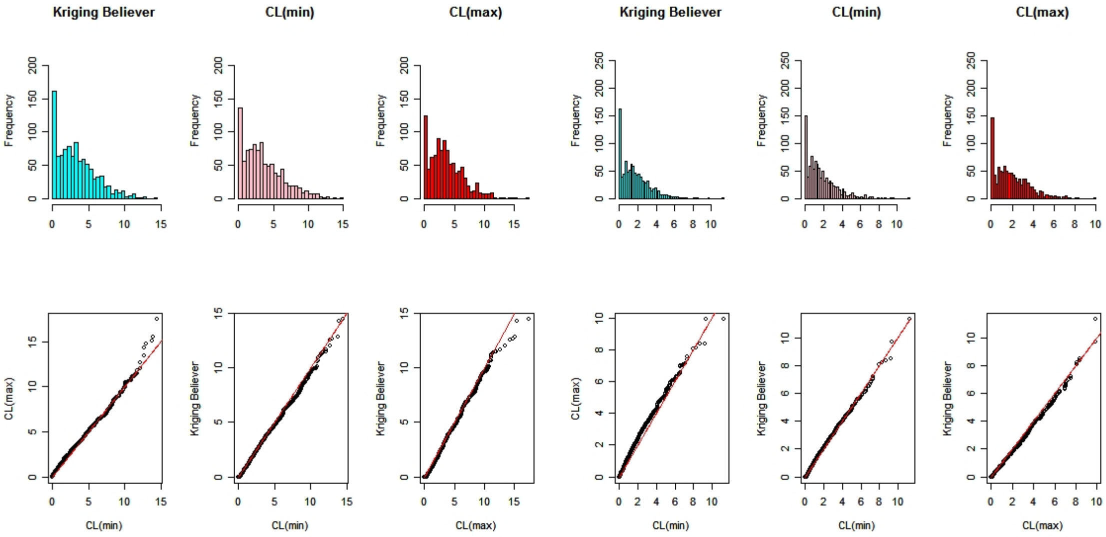


Figure 10: Comparison of the heuristic strategies $CL[min]$, $CL[max]$, KB applied to 1000 Gaussian process realizations with configurations 1 (left) and 3 (right).

This figure focuses on the δ_k^i's associated with 2 iterations of the three strategies applied to Gaussian processes with exponential covariance and respective scales 0.3 and 1 (see 5.2). In the first case ($k = 1$), the KB and $CL[min]$ show very similar results. $CL[max]$ has a slightly different right tail, and both first and third qq-plots illustrate how it is dominated in terms of extreme risk. This effect doesn't hold when the scale is 1 (right side of fig.(10)). All the strategies then behave almost equally. Note that the performances are uniformly better than with the previous configuration. This is because the realizations are more regular, and are as such easier to optimize starting with only 3 points.

We now look at (11), where 10 iterations are considered with the same sets of covariance parameters as previously (see 5.2). This time, clearer dissimilarities appear between the strategies. In configuration 2 (scale 0.3), $CL[max]$ shows better *right-tail* performances than both other strategies, which almost match. Note the dominance of KB in terms of extreme performance (near 0). These effect are amplified in configuration 4 where $CL[max]$ and KB keep their positions of best performers, respectively at the right and left extremes. Note the particular shape of the *qq*-plot between $CL[min]$ and $CL[max]$. The first one is statistically more likely to perform very well, but also more likely to fail dramatically. Conversely to configuration 1, $CL[max]$ is here a good challenger for a risk averse user.

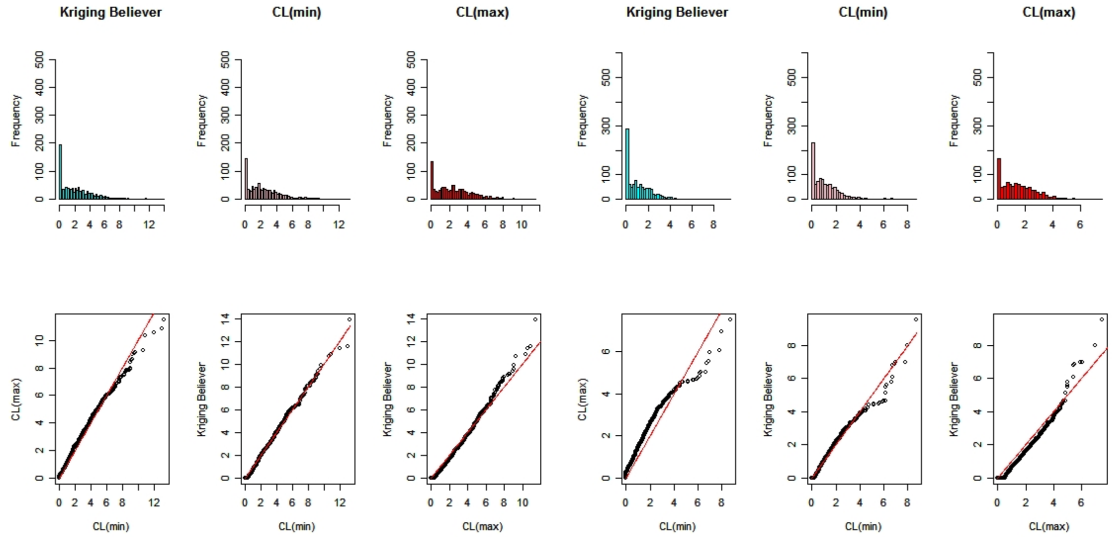


Figure 11: Comparison of the heuristic strategies $CL[min]$, $CL[max]$, KB applied to 1000 Gaussian process realizations with configurations 2 (left) and 4 (right).

References

[1] Journel A. Fundamentals of geostatistics in five lessons. Technical report, Stanford Center for Reservoir Forecasting, 1988.

[2] Ripley B.D. *Stochastic Simulation*. John Wiley and Sons, New York, 1987.

[3] Baker C.A., Watson L. T., Grossman B., Mason W. H., and Haftka R. T. Parallel global aircraft configuration design space exploration. *Practical parallel computing*, pages 79–96, 2001.

[4] Rasmussen C.E. and Williams K.I. *Gaussian Processes for Machine Learning*. M.I.T. Press, 2006.

[5] Geman D. and Jedynak B. An active testing model for tracking roads in satellite images. Technical report, Institut National de Recherches en Informatique et Automatique (INRIA), December 1995.

[6] R development Core Team. R: A language and environment for statistical computing, 2006.

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

[8] Jones D.R., Pertunen C.D., and Stuckman B.E. Lipshitzian optimization without the lipshitz constant. *Journal of Optimization Theory and Application*, (79), October 1993.

[9] Jones D.R., Schonlau M., and Welch W.J. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

[10] Sarah Goria. *Evaluation d'un projet minier: approche bayésienne et options réelles.* PhD thesis, Ecole des Mines de Paris, 2004.

[11] D. Huang, T.T. Allen, W. Notz, and N. Zheng. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization*, to appear.

[12] Villemonteix J., Vazquez E., and Walter E. An informational approach to the global optimization of expensive-to-evaluate functions. *Elsevier science direct*, 2006.

[13] Koehler J.R. and Owen A.B. Computer experiments. Technical report, Department of Statistics, Stanford University, 1996.

[14] Joshua Knowles. Parego: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE transactions on evolutionnary computation*, 2005.

[15] Schonlau M. *Computer Experiments and Global Optimization.* PhD thesis, University of Waterloo, 1997.

[16] Cressie N.A.C. *Statistics for spatial data.* Wiley series in probability and mathematical statistics, 1993.

[17] Queipo N.V., Verde A., Pintos S., and Haftka R.T. Assessing the value of another cycle in surrogate-based optimization. In *11th Multidisciplinary Analysis and Optimization Conference.* AIAA, 2006.

[18] Santner T.J., Williams B.J., and Notz W.J. *The Design and Analysis of Computer Experiments.* Springer, 2003.