Sequential dimension reduction for learning features of expensive black-box functions

Malek Ben Salem, François Bachoc, Olivier Roustant, Fabrice Gamboa, Lionel Tomaso

Sequential dimension reduction for learning features of expensive black-box functions

Malek Ben Salem [*] [†], François Bachoc[‡], Olivier Roustant[*], Fabrice Gamboa [‡], and Lionel Tomaso [†]

Abstract. Learning a feature of an expensive black-box function (optimum, contour line,...) is a difficult task when the dimension increases. A classical approach is two-stage. First, sensitivity analysis is performed to reduce the dimension of the input variables. Second, the feature is estimated by considering only the selected influential variables. This approach can be computationally expensive and may lack flexibility since dimension reduction is done once and for all. In this paper, we propose a so called *Split-and-Doubt* algorithm that performs sequentially both dimension reduction and feature oriented sampling. The 'split' step identifies influential variables. This selection relies on new theoretical results on Gaussian process regression. We prove that large correlation lengths of covariance functions correspond to inactive variables. Then, in the 'doubt' step, a doubt function is used to update the subset of influential variables. Numerical tests show the efficiency of the *Split-and-Doubt* algorithm.

Key words. Variable selection, Surrogate modeling, Design of experiments, Bayesian optimization

AMS subject classifications. 62L05, 62K99, 62F15

1. Introduction.

In design problems, the goal may be the estimation of a feature of an expensive black-box function (optimum, probability of failure, level set, ...). Several methods have been proposed to achieve this goal. Nevertheless, they generally suffer from the curse of dimensionality. Thus, their usage is limited to functions depending on a moderate number of variables. Meanwhile, most of real life problems are complex and may involve a large number of variables.

Let us focus first on high-dimensional optimization problems. In this context, we look for a good approximation of a global minimum of an expensive-to-evaluate black-box function $f : \Omega = [0,1]^D \to \mathbb{R}$ using a limited number of evaluations of f. That is, we aim at approximating $x^\star \in \Omega$ such that:

$$x^\star \in \arg\min_{x\in\Omega} f(x). \tag{1.1}$$

Bayesian optimization (BO) techniques have been successfully used in various problems [19, 18, 17, 27, 31]. These methods give interesting results when the number of evaluations of the function f is relatively low [15]. They are generally limited to problems of moderate dimension, typically up to about 10 [32]. Here, we are particularly interested in the case where the dimension D is large and the number of influential variables d, also called *effective*

[*]Mines de Saint-Étienne, UMR CNRS 6158, LIMOS, F-42023: 158 Cours Fauriel, Saint-Étienne.
[†]ANSYS, Inc: 11 Avenue Albert Einstein F-69100 Villeurbanne.
[‡]IMT Institut de Mathématiques de Toulouse 118 route de Narbonne - F-31062 TOULOUSE Cedex 9.
malek.ben-salem@emse.fr

dimension, is much smaller: $d << D$. In this case, there are different approaches to tackle the dimensionality problem.

A direct approach consists in first performing global sensitivity analysis. Then, the most influential variables are selected and used in the parametric study. Chen et al. [11] stated that "*Variables selection and optimization have both been extensively studied separately from each other*". Most of these methods are two-stage: First, the influential variables are selected and then optimization is performed on these influential variables. These strategies are generally computationally expensive. Furthermore, the set of selected variables does not take into account the new data. However, this new information may modify the results of the sensitivity analysis study. For an overview of global sensitivity analysis methods, one may refer to [16].

Some Bayesian optimization techniques are designed to handle the dimensionality problem. For instance, the method called Random EMbedding Bayesian Optimization (REMBO) selects randomly the subspace of influential variables [32, 9]. The main strengths of REMBO are that the selected variables are linear combinations of the input variables and that it works for huge values of D. However, the effective dimension d must be specified.

In this paper, we propose a versatile sequential dimension reduction method called *Split-and-Doubt*. The design is sequentially generated in order to achieve jointly two goals. The first goal is the estimation of the optimum (in the optimization case). The second one is the learning of the influential variables. In the "split" step, the algorithm selects the set of influential variables based on the values of the correlation lengths of Automatic Relevance Determination (ARD) covariances. We show theoretical results that support the intuition that large correlation lengths correspond to inactive variables. The "doubt" step questions the "split" step and helps correcting the estimation of the correlation lengths.

The paper is organized as follows. Section 2 presents the background and the notations. Section 3 introduces the *Split-and-Doubt*. The algorithm is based on theoretical results stated in Section 4. Finally, Section 5 illustrates the performance of the algorithm on various test functions. For readability, proofs are postponed to Section 6. Concluding remarks are given in Section 7.

2. General notations and background.

2.1. Gaussian Process Regression (GPR).

Kriging or Gaussian process regression (GPR) models predict the outputs of a function $f\colon \Omega = [0,1]^D \to \mathbb{R}$, based on a set of n observations [30, 23]. It is a widely used surrogate modeling technique. Its popularity is mainly due to its statistical nature and properties. Indeed, it is a Bayesian inference technique that provides an estimate of the prediction error distribution. This uncertainty is an efficient tool to construct strategies for various problems such as prediction refinement, optimization or inversion.

The GPR framework uses a centered real-valued Gaussian Process (GP) Y over Ω as a prior distribution for f. The predictions are given by the conditional distribution of Y given the observations $y = (y_1, \ldots, y_n)^\top$ where $y_i = f(x^{(i)})$ for $1 \le i \le n$. We denote by $k_\theta : \Omega \times \Omega \to \mathbb{R}$ the covariance function (or kernel) of Y: $k_\theta(x, x') = \mathrm{Cov}[Y(x), Y(x')]$ ($(x, x') \in \Omega^2$), by $X = \begin{pmatrix} x^{(1)}, & \ldots, & x^{(n)} \end{pmatrix}^\top \in \Omega^n$ the matrix of observation locations and by $Z = \begin{pmatrix} X & y \end{pmatrix}$ the matrix of observation locations and values where $x^{(i)} = (x_1^{(i)}, \ldots, x_D^{(i)})$ for $1 \le i \le n$. θ is a parameter that will be discussed later. Without loss of generality, we consider the simple kriging framework. The *a posteriori* conditional mean $m_{\theta,Z}$ and the *a posteriori* conditional variance $\widehat{\sigma}^2_{\theta,Z}$ are given by:

$$m_{\theta,Z}(x) = k_\theta(x,X)^\top K_\theta^{-1} y \tag{2.1}$$

$$\widehat{\sigma}^2_{\theta,Z}(x) = k_\theta(x,x) - k_\theta(x,X)^\top K_\theta^{-1} k_\theta(x,X). \tag{2.2}$$

Here, $k_\theta(x,X)$ is the vector $(k_\theta(x,x^{(1)}),\dots,k_\theta(x,x^{(n)}))^\top$ and $K_\theta = k_\theta(X,X)$ is the invertible matrix with entries $\Big(k_\theta(X,X)\Big)_{ij} = k_\theta(x^{(i)},x^{(j)})$, for $1 \le i,j \le n$.

Several methods are useful to select the covariance function. A common approach consists in assuming that the covariance function belongs to a parametric family. In this paper, we consider the Automatic Relevance Determination (ARD) kernels defined in (2.3). A review of classical covariance functions is given in [1].

$$k_\theta(x,y) = \sigma^2 \prod_{p=1}^{D} k\Big(\frac{d(x_p,y_p)}{\theta_p}\Big), \text{for } x,y \in \Omega. \tag{2.3}$$

Here, $d(,)$ is a distance on $\Omega \times \Omega$ and $k : \mathbb{R} \to \mathbb{R}$ is a fixed stationary covariance function. Without loss of generality, we suppose that the hyper-parameter σ is fixed while $\theta_1,\dots,\theta_D$ have to be estimated. The ARD kernels include most popular kernels such as the exponential kernel, the Matérn 5/2 kernel and the squared exponential kernel.

The hyper-parameters of these parametric families can be estimated by maximum Likelihood (ML) or cross validation (CV). Both methods have interesting asymptotic properties [2, 4, 5]. Nevertheless, when the number of observations is relatively low, the estimation can be misleading. These methods are also computationally demanding when the number of observations is large.

On one hand, estimating the correlation lengths by the maximum likelihood estimator gives the estimator $\widehat{\theta}^\star_{MLE} \in \arg\max_\theta l_Z(\theta)$ where the likelihood $l_Z(\theta)$ is given in (2.4).

$$l_Z(\theta) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{\det\Big(k_\theta(X,X)\Big)}} \exp\Big(-y^\top k_\theta(X,X)^{-1} y\Big). \tag{2.4}$$

On the other hand, the idea behind Cross-validation (CV) is to estimate the prediction errors by splitting the observations once or several times. One part is used as a test set while the remaining parts are used to construct the model. The Leave-One-Out Cross-Validation (LOO-CV) consists in dividing the n points into n subsets of one point each. Then, each subset plays the role of test set while the remaining points are used together as the training set. Using Dubrule's formula [13], the LOO-CV estimator is given in (2.5).

$$\widehat{\theta}^\star_{CV} \in \arg\min_\theta \frac{1}{n} y^\top K_\theta^{-1} \operatorname{diag}(K_\theta^{-1})^{-1} K_\theta^{-1} y. \tag{2.5}$$

For more insight on these estimators, one can refer to [3].

2.2. Derivative based global sensitivity measures: DGSM.

. Sobol' and Kucherenko [28, 29] proposed the so-called Derivative-based Global Sensitivity Measures (DGSM) to estimate the influence of an input variable on a function $f : \Omega = [0,1]^D \to \mathbb{R}$. For each variable x_i, the index ϑ_i is the global energy of the corresponding partial derivatives.

$$\vartheta_i(f) = \int_\Omega \left(\frac{\partial f(x)}{\partial \mathrm{x}_i}\right)^2 dx, i = 1, \dots, D. \tag{2.6}$$

DGSM provides a quantification of the influence of a single input on f. Indeed, assuming that f is of class C^1, then x_i is not influential iff $\frac{\partial f}{\partial \mathrm{x}_i}(x) = 0$, $\forall\ x \in \Omega$ iff $\vartheta_i = 0$. DGSM has recently shown its efficiency for the identification of non-influential inputs [24]. We further define the normalized DGSM $\widetilde{\vartheta}_i$ in (2.7). $\widetilde{\vartheta}_i$ measures the influence of x_i with regard to the total energy.

$$\widetilde{\vartheta}_i(f) = \frac{\vartheta_i(f)}{\sum\limits_{p=1}^{D} \vartheta_p(f)}, i = 1, \dots, D. \tag{2.7}$$

3. The Split-and-Doubt Design Algorithm (*Split-and-Doubt*).

3.1. Definitions.

Variable splitting. Let us consider the framework of a GPR using a stationary ARD kernel. Intuitively, large correlation length values correspond to inactive variables in the function. We prove this intuition in Proposition 4.1. The influential variables are selected in our algorithm according to the estimated values of their corresponding correlation lengths. We show also that the ML (and CV) estimator is able to assign asymptotically large correlation length values to the inactive variables (Propositions 4.3 and 4.4).

Let $\widehat{\theta^\star} = (\widehat{\theta^\star}_1, .., \widehat{\theta^\star}_D)$ be the ML estimation of the correlation lengths:

$$\widehat{\theta^\star} \in \arg\max_\theta l_Z(\theta).$$

The influential variables are then selected according to the estimated values of their corresponding correlation lengths. We split the indices into a set of influential variables I_M and a set of minor variables I_m as follows:

- $I_M = \{i; \widehat{\theta^\star}_i < T\}$
- $I_m = \{i; \widehat{\theta^\star}_i \geq T\}$

where $T \in \mathbb{R}$ is a suitable threshold. Let d_M (resp. d_m) be the size of I_M (resp. I_m). We further call $\Omega_m := [0,1]^{d_m}$ the minor subspace, that is the space of minor variables and $\Omega_M := [0,1]^{d_M}$ the major subspace, that is the subspace of major variables. We will use the set notation: for a set I of $\{1, \dots, D\}$, x_I will denote the vector extracted from x with coordinates x_i, $i \in I$. Hence, x_{I_M} (resp. x_{I_m}) denotes the sub-vector of x whose coordinates are in the major (resp. minor) subspace. For simplicity, we will also write $x = (x_{I_M}, x_{I_m})$, without specifying the re-ordering used to obtain x by gathering x_{I_M} and x_{I_m}.

Doubt. The doubt of a correlation length θ measures the influence on $m_{\theta,Z}$ of the variables from the minor subspace Ω_m. It is a decreasing function of the correlation lengths. We will use it to question the variable splitting.

Definition 3.1 (Doubt). *Let δ be the following function associated with a variable splitting (I_m, I_M). For all vector $\theta = (\theta_1, \dots, \theta_D) \in \mathbb{R}^D$:*

$$\delta(\theta) = \sum_{i \in I_m} \max(\theta_i^{-1} - T^{-1}, 0).$$

Contrast. Given two different correlation lengths $\theta^{(1)}$ and $\theta^{(2)}$ and a location x, the contrast measures the discrepancy between the predictions using $\theta^{(1)}$ and $\theta^{(2)}$ at x. It will be used to build a sequential design in the minor subspace.

Definition 3.2 (Prediction contrast). *For a point x and two correlation lengths $\theta^{(1)}$ and $\theta^{(2)}$, the prediction contrast $PC(x, \theta^{(1)}, \theta^{(2)})$ is*

$$PC(x, \theta^{(1)}, \theta^{(2)}) = \left| m_{\theta^{(1)},Z}(x) - m_{\theta^{(2)},Z}(x) \right|$$

.

3.2. The algorithm. The *Split-and-Doubt* algorithm performs a new variable selection at each iteration. It samples a point in two steps: a goal-oriented sampling in the major subspace and a sampling of the minor variables to question the variable selection done at the previous step. The *Split-and-Doubt* algorithm for optimization using the expected improvement (EI) criterion [18] is described below:

Algorithm 3.3. *Split-and-Doubt-EGO (f)*

- ***Algorithm parameters:*** *ℓ, kernel k, threshold T.*
- ***Start:*** *Inputs: $Z = (X, y)$.*
- ***Loop***
  1. *Estimate the correlation lengths:* $\widehat{\theta^\star} \in \arg\max_{\theta \in (\mathbb{R}_+^\star)^D} l_Z(\theta)$ *Eq.* (2.4).
  2. *Split the variables: Define the major set $I_M = \{i; \widehat{\theta}_i^\star < T\}$ and the minor set $I_m = \{i; \widehat{\theta}_i^\star \geq T\}$, $d_m = |I_m|$.*
  3. *Design in the major subspace: Compute $x_M^\star$ according to the objective function in the major subspace (by EI for instance): We compute a new GPR considering only the major variables to compute the EI. Let $Z_M = (X_{I_M}, y)$*

$$x_M^\star \in \arg\max_{x_M \in \Omega_M} EI_{Z_M}(x_M).$$

  4. *Doubt the variable splitting: Compute a challenger θ' for correlation lengths.*

$$\theta' \in \arg\max_{\theta \in (\mathbb{R}_+^\star)^D} \delta(\theta) \quad \textit{subject to } 2\left| \ln\left(\frac{l_Z(\theta)}{l_Z(\widehat{\theta^\star})} \right) \right| < \chi^2(\ell, d_m).$$

5. *Design in the minor subspace: Compute $x^\star_m$ by maximum contrast with the challenger θ'*

$$x^\star_m \in \underset{x_m \in \Omega_m}{\arg\max}\, PC\Big(x = (x^\star_M, x_m), \widehat{\theta^\star}, \theta'\Big).$$

6. *Update: Evaluate the new point output $y_{n+1} = f(x^{(n+1)})$ with $x^{(n+1)}_{I_m} = x^\star_M$ and $x^{(n+1)}_{I_m} = x^\star_m$ and add the new point to the design:*

$$X^\top \leftarrow \big(X^\top \quad x^{(n+1)}\big), \quad y^\top \leftarrow (y^\top, y_{n+1}).$$

- ***End loop***
- ***Outputs:*** $Z = (X, y)$.

Here, the algorithm is applied for optimization (Step 3). We used the Expected Improvement criterion (3.1).

$$EI_Z(x) = \mathbb{E}\Big[\max(\min_i y_i - Y(x), 0)|Z)\Big]. \tag{3.1}$$

It is possible to use any other classical optimization criterion to sample $x^\star_M$. We can use other criteria for other purposes such as contour estimation [21, 22, 8], probability of failure estimation [6] or surrogate model refinement [10].

The settings of the algorithm are mainly the kernel k, the limit ℓ and the threshold T. An other hidden setting is the search space for the ML estimator. We use a Matérn 5/2 kernel and we set $\ell = \mathrm{erf}(\frac{1}{\sqrt{2}})$ and an adaptive threshold $T = 20 \min_{i\in[1,D]} (\widehat{\theta^\star}_i)$.

3.3. Remarks on the steps of the *Split-and-Doubt* algorithm.

Remark on the doubt. When the observations do not carry enough information, it is hard to estimate accurately the correlation lengths. The use of such values can lead to unsatisfactory results [14, 7]. In our algorithm, the estimated correlation lengths are used to select the major variables. If this estimation is done once and for all, poor estimation can lead to considering a major variable inactive. So, it is important to always question the estimation. Therefore, we look for a “challenger kernel” at each iteration. Specifically, we are looking for correlation lengths that maximize the doubt and that are accepted by a likelihood ratio test. Indeed, this is why we limit the search space by a likelihood ratio deviation from the estimated correlation lengths $\widehat{\theta^\star}$: $\Theta_l = \{\theta; 2\Big|\ln\Big(\frac{l_Z(\theta)}{l_Z(\widehat{\theta^\star})}\Big)\Big| < l\}$. Notice that we used $l = \chi^2(\ell, d_m)$. Following [14, 11], the likelihood ratio test is compared to the χ^2 distribution to decide whether the correlation lengths are allowable or not.

Remark on the contrast. Sampling the coordinates in the non-influential variable subspace $\{x^\star_M\} \times \Omega_m = \{(x^\star_M, x_m), x_m \in \Omega_m\}$ aims at revealing the contrast between the maximum likelihood correlation lengths $\widehat{\theta^\star}$ and a challenging correlation length parameter is θ'. The main idea is to sample the point that helps either correcting the first estimation or reducing the allowable doubt space Θ in order to strengthen the belief in the estimated kernel.

We could have used an alternative direct approach. It consists in maximizing the likelihood ratio between two estimations of the correlation lengths in the future iterations.

Definition 3.4 (likelihood contrast). *For a point x and two correlation lengths $\theta^{(1)}$ and $\theta^{(2)}$, the likelihood contrast LC is:*

$$LC(x, \theta^{(1)}, \theta^{(2)}) = \mathbb{E}\left[\left| \ln\left(\frac{L(\theta^{(1)}, Z \cup (x, \widehat{Y}(x)))}{L(\theta^{(2)}, Z \cup (x, \widehat{Y}(x)))} \right) \right| \right]$$

where $\widehat{Y}(x) \sim \mathcal{N}\Big(m_{\theta_2, Z}(x), (\widehat{\sigma}_{\theta_2, Z}(x))^2 \Big)$.

However, this approach is computationally more expensive. Therefore, we prefer to use the prediction contrast (Definition 3.2).

3.4. Example: Illustration of the contrast effect. We illustrate here how the Doubt/Contrast strategy can help correcting the initial variable splitting. To do so, let us consider the following particular case. Let $f(x_1, x_2) = \cos(2\pi x_2)$. We assume that we have at hands four design points $x^{(1)} = (0, \frac{2}{3})$, $x^{(2)} = (\frac{1}{3}, 0)$, $x^{(3)} = (\frac{2}{3}, 1)$ and $x^{(4)} = (1, \frac{1}{3})$ and their corresponding responses $y_1 = y_4 = f(x^{(1)}) = f(x^{(4)}) = -0.5$ and $y_2 = y_3 = f(x^{(2)}) = f(x^{(3)}) = 1$. Here, the search space for the correlation lengths is $[0.5, 10]^2$. Let us assume that the first estimation will give the inaccurate estimation $\widehat{\theta^\star} = (0.5, 10)$. We display in Figure 1 the function, the design points, the log-likelihood of the correlation lengths in the search space $[0.5, 10]^2$ for the Matérn 5/2 kernel and the predictions using $k_{\widehat{\theta^\star}}$.

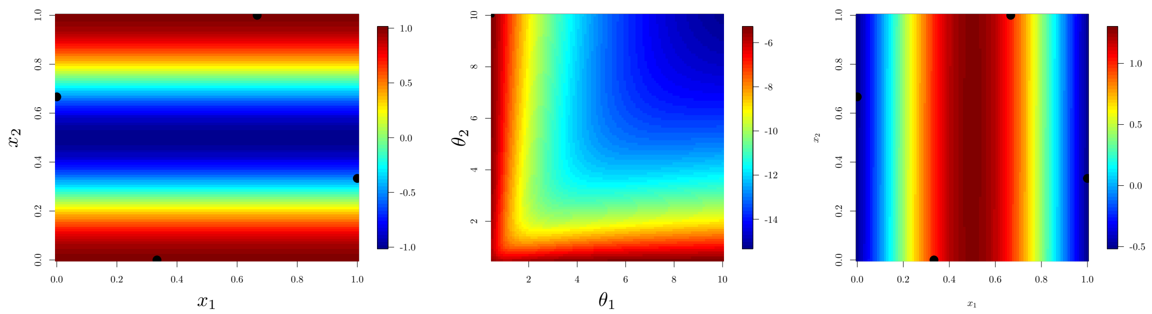


Figure 1: Left: $f(x_1, x_2) = \cos(2\pi x_2)$, the color code indicates the values of f and solid black circles indicate the design points. Middle: log-likelihood of the correlation lengths, solid black circle: $\widehat{\theta^\star}$. Right: The predictions given by the GPR using $k_{\widehat{\theta^\star}}$.

We set $T = 10$. So, $I_M = \{1\}$ and $I_m = \{2\}$. Notice that the challenger correlation lengths is $\theta' = (0.5, 0.5)$. It gives the maximum doubt $\delta(\theta') = \frac{1}{0.5} - \frac{1}{10} = 1.9$. In this example, we are sampling the fifth point $x^{(5)}$. The value sampled by the EI in the major space is $x_M^\star = 0.64$. We display in Figure 2 the prediction contrast $PC((x_M^\star, x_2), \widehat{\theta^\star}, \theta')$ in function of x_2.

To show the interests of the contrast strategy, we consider two cases: a) we add the point sampled by maximum contrast $(x_M^\star, x_m^\star)$ and b) we add the point with the minimum contrast $(x_M^\star, 1)$. Then, we evaluate these points. For both cases, we display the updated likelihood in Figure 3.

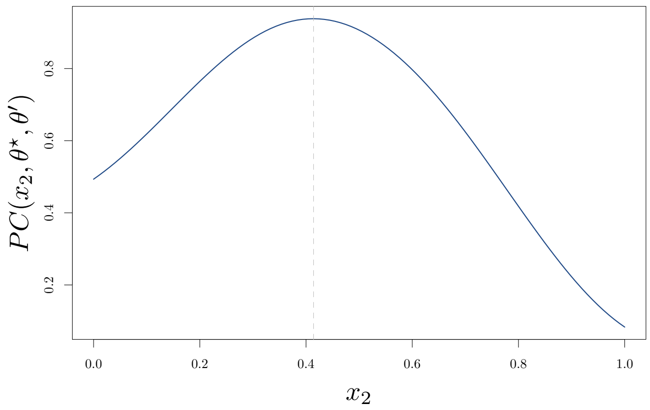


Figure 2: The prediction contrast in function of x_2.

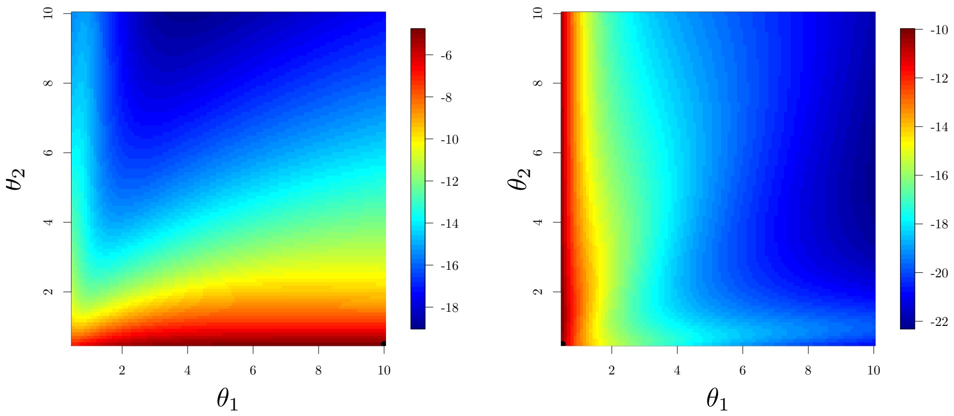


Figure 3: Left: The log likelihood of the correlation lengths if we add $\big((x_M^\star, x_m^\star), f(x_M^\star, x_m^\star)\big)$. Right: The log likelihood of the correlation lengths if we add $\big((x_M^\star, 1), f(x_M^\star, 1)\big)$.

On one hand, the log-likelihood has larger value for small values of θ_2 when we add the point $(x_M^\star, x_m^\star)$. Thus, the same inaccurate variables splitting is prevented thanks to the contrast strategy. On the other hand, if we choose $x_m = 1$ (small contrast value), we may still use the same variable splitting. Even if this 2-dimensional toy example is pathological, in high dimension it is hard to estimate accurately the correlation lengths when the number of design points is relatively small. In these cases, the Doubt/Contrast strategy can be helpful.

4. Links between correlation lengths and variable importance. In this section, we consider a deterministic function $f : \mathbb{R}^D \to \mathbb{R}$ to be modeled as a GP path. We consider the centered stationary GP with a covariance function k_θ defined by

$$k_\theta(h) = \prod_{i=1}^{D} k(h_i/\theta_i).$$

Here $k : \mathbb{R} \to \mathbb{R}$ is a fixed covariance function satisfying $k(0) = 1$ and $\theta \in (0, \infty)^D$ is the vector of correlation lengths. As an example, k may be the function $k(h) = e^{-h^2}$.

Intuitively, a small correlation length θ_i for the GP should correspond to an input variable x_i that has an important impact on the function value $f(x)$. Conversely, if the function f does not depend on x_i, then the length θ_i should ideally be infinite. This intuition is precisely the motivation for the *Split-and-Doubt* algorithm suggested in Section 3.

In this section, we show several theoretical results that confirm this intuition. First, we show that if the correlation length θ_i goes to zero (respectively infinity) then the derivative-based global sensitivity measure, obtained from the GP predictor for the input x_i, tends to its maximum value 1 (resp. its minimum value 0). Then, we show that an infinite correlation length θ_i can provide an infinite likelihood or a zero LOO mean square error, for the GP model, when the function f does not depend on x_i.

We use the additional following notations throughout the section. For $D, p, q \in \mathbb{N}^\star$, for a covariance function g on $\mathbb{R}^D$, for two $p \times D$ and $q \times D$ matrices X and Z, we denote by $g(X, Z)$ the $p \times q$ matrix defined by $[g(X, Z)]_{i,j} = g(X_i, Z_j)$ where M_l is the line l of a matrix M. When $d = 1$, $p = 1$ or $q = 1$, we identify the corresponding matrices with vectors. We further assume:

Assumption 1 (Invertibility assumption).*for any $p, d \in \mathbb{N}$, for any $\theta \in (0, \infty)^D$, for any $p \times d$ matrix X with two-by-two distinct lines, the matrix $k_\theta(X, X)$ is invertible.*

This assumption holds for instance when the spectral density of a stationary kernel is absolutely continuous. Further, for any vector u, u_{-i} is obtained from u by removing the i^{th} component of u.

4.1. Correlation lengths and derivative-based global sensitivity measures. Consider a function f to be observed at the locations $x^{(1)}, ..., x^{(n)} \in \Omega$, with $n \in \mathbb{N}$ and for a bounded domain $\Omega \subset \mathbb{R}^D$. Let X be the $n \times D$ matrix with lines given by $x^{(1)}, ..., x^{(n)}$, y be the vector of responses $y = (f(x^{(1)}), ..., f(x^{(n)}))^\top$ and Z the $(n+1) \times D$ matrix $Z = \begin{pmatrix} X & y \end{pmatrix}$.

Recall that the prediction of f at any line vector $x \in \Omega$, from the GP model, is given by $m_{\theta,Z}(x) = r_\theta(x)^\top K_\theta^{-1} y$, with $r_\theta(x) = k(x, X)$, $K_\theta = k_\theta(X, X)$. Then, we use the notation $\vartheta_i(\theta)$ for the DGSM index of the variable x_i on the predictor function $m_{\theta,Z}(x)$:

$$\vartheta_i(\theta) = \vartheta_i(m_{\theta,Z}) = \int_\Omega \left(\frac{\partial m_{\theta,Z}(x)}{\partial \mathrm{x}_i} \right)^2 dx.$$

We also use the following notation for the normalized DGSM index of the variable x_i:

$$\widetilde{\vartheta}_i(\theta) = \widetilde{\vartheta}_i(m_{\theta,Z}) = \frac{\vartheta_i(\theta)}{\sum\limits_{r=1}^{D} \vartheta_r(\theta)}.$$

The normalized DGSM index $\widetilde{\vartheta}_i(\theta)$ satisfies $0 \leq \widetilde{\vartheta}_i(\theta) \leq 1$. The larger this indice is, the more important the variable x_i is for $m_{\theta,Z}(x)$. In the two next propositions, we show that, under mild conditions, we have $\widetilde{\vartheta}_i(\theta) \to 1$ as $\theta_i \to 0$ and $\widetilde{\vartheta}_i(\theta) \to 0$ as $\theta_i \to \infty$. Hence, we give a theoretical support to the intuition that small correlation lengths correspond to important input variables.

Proposition 4.1. *Assume that the components of y are not all equal. Assume that k is continuously differentiable on $\mathbb{R}$. Let $i \in \{1, ..., D\}$ be fixed. For $j = 1, ..., n$ let $v^{(j)} = x^{(j)}_{-i}$. Assume that $v^{(1)}, ..., v^{(n)}$ are two by two distinct. Then, for fixed $\theta_{-i} \in (0, \infty)^{D-1}$*

$$\widetilde{\vartheta}_i(\theta) \underset{\theta_i \to \infty}{\longrightarrow} 0.$$

Proposition 4.2. *Assume that the components of y are not all equal. Consider the same notation as in Proposition 4.1. Assume that k is continuously differentiable on $\mathbb{R}$, that $k(t) \to 0$ as $|t| \to \infty$ and that Ω is an open set. Assume also that $x^{(1)}, ..., x^{(n)}$ are two-by-two distinct. Let $i \in \{1, ..., D\}$ be fixed. Then for fixed $\theta_{-i} \in (0, \infty)^{D-1}$*

$$\widetilde{\vartheta}_i(\theta) \underset{\theta_i \to 0}{\longrightarrow} 1.$$

In Propositions 4.1 and 4.2, the regularity conditions on k are mild, and the conditions on $x^{(1)}, ..., x^{(n)}$ hold in many cases, for instance when $x^{(1)}, ..., x^{(n)}$ are selected randomly and independently or from a latin hypercube procedure (see e.g. [26]).

4.2. Estimated correlation lengths and inactive variables. We first recall the likelihood function:

$$l_Z(\theta) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{\det\left(k_\theta(X, X)\right)}} \exp\left(-y^\top k_\theta(X, X)^{-1} y\right).$$

In the next proposition, we show that, if the function f does not depend on the variable x_i, then the likelihood $l_Z(\theta)$ goes to infinity when θ_i goes to infinity. This is a theoretical confirmation that maximum likelihood can detect inactive input variables and assign them large correlation lengths.

Proposition 4.3. *Assume that k is continuous. Assume that for any $\theta \in (0, \infty)^D$, the reproducing kernel Hilbert space (RKHS) of the covariance function k_θ contains all infinitely differentiable functions with compact supports on $\mathbb{R}^D$.*

Let $i \in \{1, ..., D\}$ be fixed. For $j = 1, ..., n$ let $v^{(j)} = x^{(j)}_{-i}$. Assume that

i) $x^{(1)}, ..., x^{(n)}$ are two-by-two distinct;

ii) $y_r = y_s$ if $v^{(r)} = v^{(s)}$;

iii) there exist $a, b \in \{1, ..., n\}$ with $a \neq b$ such that $v^{(a)} = v^{(b)}$.

Then, for fixed $\theta_{-i} \in (0, \infty)^{D-1}$

$$l_Z(\theta) \underset{\theta_i \to \infty}{\longrightarrow} \infty.$$

In Proposition 4.3, Conditions i), ii) and iii) are quite minimal. Condition i) ensures that the likelihood is well-defined, as the covariance matrix is invertible for all $\theta \in (0, \infty)^D$ (due to the invertibility assumption 1). Condition ii) holds when $f(x)$ does not depend on x_i. Condition iii) is necessary to have $l(\theta)$ going to infinity, since if $v^{(1)}, ..., v^{(n)}$ are two by two distinct, the determinant of $k_\theta(X, X)$ remains bounded from below as $\theta_i \to \infty$ (see also the proof of Proposition 4.1). Notice that Conditions ii) and iii) together imply that there is a pair of input points $x^{(a)}, x^{(b)}$ for which only the value of the i-th component changes and the value of f does not change, which means that the data set presents an indication that the input variable x_i is inactive.

We refer to, e.g., [33] for a reference to the RKHS notions that are used in this section. There are many examples of stationary covariance functions k satisfying the RKHS condition in Proposition 4.3. In particular, let $\widehat{k}_\theta$ be the Fourier transform of k_θ defined by $\widehat{k}_\theta(w) = \int_{\mathbb{R}^D} k_\theta(x) e^{-\mathrm{i} w^\top x} dx$ with $\mathrm{i}^2 = -1$. Then, if there exists $\tau < \infty$ such that $\widehat{k}_\theta(w)||w||^\tau \to \infty$ as $||w|| \to \infty$, then the RKHS condition of Proposition 4.3 holds. This follows from Theorem 10.12 in [33] and from the fact that an infinitely differentiable function with compact support ϕ has a Fourier transform $\widehat{\phi}$ satisfying $\widehat{\phi}(w)||w||^\gamma \to 0$ as $||w|| \to \infty$ for any $\gamma < \infty$. Hence, Lemma 4.3 holds in particular when k is the exponential covariance function with $k(t) = e^{-|t|}$. Lemma 4.3 also holds when k is the Matérn covariance function with

$$k\left(t\right) = \frac{1}{\Gamma\left(\nu\right) 2^{\nu-1}} \left(2\sqrt{\nu}|t|\right)^\nu K_\nu\left(2\sqrt{\nu}|t|\right),$$

where $0 < \nu < \infty$ is the smoothness parameter (see e.g. [30]). It should however be noted that the squared exponential covariance function k (defined by $k(t) = \exp(-t^2)$ with $t \in \mathbb{R}$) does not satisfy the condition of Lemma 4.3. [Notice that [34] study specifically the asymptotic behavior of the maximum likelihood estimation of a variance parameter for this covariance function, when the number of observations of a smooth function goes to infinity.]

In the next proposition, we study the LOO mean square prediction error

$$CV_Z(\theta) = \sum_{j=1}^n (y_j - \widehat{y}_{\theta,j})^2,$$

with $\widehat{y}_{\theta,j} = k_\theta(x^{(j)}, X_{-j}) k_\theta(X_{-j}, X_{-j})^{-1} y_{-j}$, where X_{-j} and y_{-j} are obtained, respectively, by removing the line j of X and the component j of y. We show that, as for the likelihood, inactive variables can be detected by this LOO criterion, since we can have $CV_Z(\theta) \to 0$ as $\theta_i \to \infty$ if the function f does not depend on x_i.

Proposition 4.4. *Let k satisfy the same conditions as in Proposition 4.3. Let $i \in \{1, ..., D\}$ be fixed.*

For $j = 1, ..., n$ let $v^{(j)} = x^{(j)}_{-i}$. Assume that

i) $x^{(1)}, ..., x^{(n)}$ are two-by-two distinct;

ii) $y_r = y_s$ if $v^{(r)} = v^{(s)}$;

iii) for all $r \in \{1, ..., n\}$ there exists $s \in \{1, ..., n\}$, $r \neq s$, such that $v^{(r)} = v^{(s)}$.

Then, for any fixed $\theta_{-i} \in (0, \infty)^{D-1}$, we have

$$CV_Z(\theta) \underset{\theta_i \to \infty}{\longrightarrow} 0.$$

In Proposition 4.4, Conditions i) and ii) are interpreted similarly as in Proposition 4.3. Condition iii), however, provides more restrictions than for the likelihood in Proposition 4.3. This condition states that for any observation point in the data set, there exists another observation point for which only the inactive input i is changed. This condition is arguably necessary to have $CV_Z(\theta) \to 0$.

5. Numerical examples.

5.1. Test sets. We illustrate the *Split-and-Doubt* algorithm on five benchmark optimization problems. The first four are classical synthetic functions: the two-dimensional Branin function, the general Ackley function in six dimensions, the six-dimensional Hartmann function and the general Rosenbrock function in five dimensions. The fifth one is the Borehole function [20]. It models the water-flow in a borehole. For each function, we added inactive input variables in order to embed them in a higher dimension $D^{(i)}$. The settings are summarized in Table 1.

Table 1: Optimization test functions.

$f^{(i)}$	$d^{(i)}$	$D^{(i)}$	Number of design points $n_0^{(i)}$	Number of iterations $N_{max}^{(i)}$
Hartmann 6-dim	6	15	30	30
Rosenbrock	5	20	40	60
Ackley	6	20	45	40
Borehole	6	25	30	25
Branin	2	25	30	50

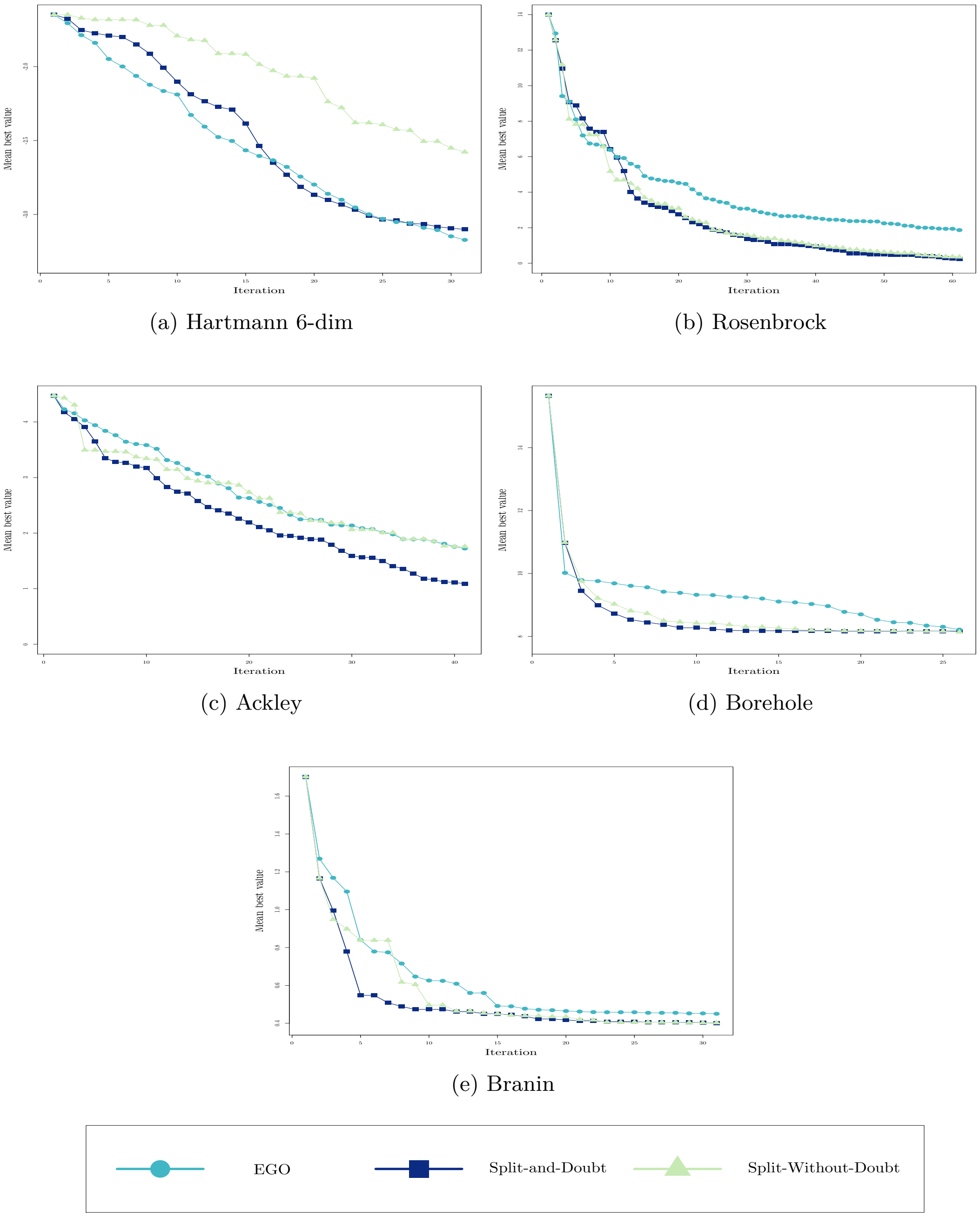
We launched the optimization process for these functions with three different optimization algorithms:

- EGO [18]: Implementation of the R package DiceOptim [25] using the default parameters.
- *Split-and-Doubt* algorithm with Matérn 5/2 covariance function.
- *Split-without-Doubt* algorithm: It uses the same variable splitting as *Split-and-Doubt* and generates the minor variables by uniform random sampling.

For each function $f^{(i)}$, we launched each optimization process for $N_{max}^{(i)}$ iterations starting with $N_{seed} = 20$ different initial designs of experiments of size $n_0^{(i)}$ generated by a maximin space-filling sampling.

5.2. Results. The results are represented by boxplots in Appendix A. We also display the mean best value evolution in Figure 4.

We can see that *Split-and-Doubt* gives better results than EGO for Rosenbrock, Ackley and Borehole function. EGO does not always converge for the first one and used more iterations for the second. These cases illustrate the efficiency of the dimension reduction for limited budget optimization. For Branin function the convergence is relatively fast for all the three



(a) Hartmann 6-dim

(b) Rosenbrock

(c) Ackley

(d) Borehole

(e) Branin

Figure 4: Comparison of 3 optimization strategies. Mean over N_{seed} of the best current values as a function of the number of iterations.

algorithms. This is due to the fact that the effective dimension is 2 and that the first design of experiments covers well these dimensions.

On one hand, sampling the minor variables at random or using the Doubt/Contrast strategy gives close results when the influential variables are easily determined. On the other hand, the efficiency of the Doubt/Contrast approach is visible on Hartmann and Ackley functions, by comparing the results of *Split-and-Doubt* and *Split-without-Doubt*. Notice that we start in general with a relatively small amount of design points. Thus, the initial estimation of the correlation lengths can be inaccurate. In these cases, the Doubt/Contrast approach is valuable to improve the estimation.

Finally, as we can see in Figure 5, *Split-and-Doubt* is faster than EGO in terms of computing time. The fact that we perform two optimization procedures in smaller spaces makes the algorithm faster than optimizing the EI in dimension D.

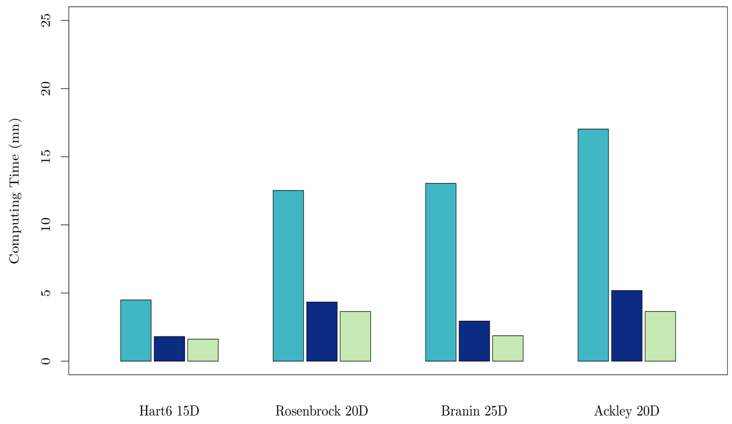


Figure 5: Mean computing time: Left: EGO, Middle: *Split-and-Doubt*, Right: *Split-Without Doubt* in minutes.

6. Proofs.

For the proofs of Propositions 4.1 and 4.2, we let $k'(t) = \partial k(t)/\partial t$.

Proof. [**Proof of Proposition 4.1**]

Without loss of generality, we consider $i = 1$ in the proof. Let $\theta_{-1} \in (0, \infty)^{D-1}$ be fixed. We have

$$\frac{\partial}{\partial \mathrm{x}_j} m_{\theta,Z}(x) = \left(\frac{\partial r_\theta(x)}{\partial \mathrm{x}_j} \right)^T K_\theta^{-1} y.$$

When $\theta_1 \to \infty$, K_θ converges to the $n \times n$ matrix $L_{\theta_{-1}}$ with $(L_{\theta_{-1}})_{pq} = \prod_{r=2}^{D} k([x_r^{(p)} - x_r^{(q)}]/\theta_r)$, by continuity of k. This matrix is invertible by assumption on k and $v^{(1)}, ..., v^{(n)}$. Hence

$||K_\theta^{-1}y||$ is bounded as $\theta_1 \to \infty$. We have for $j = 1, ..., n$

$$\left(\frac{\partial r_\theta(x)}{\partial \mathrm{x}_1}\right)_j = \frac{1}{\theta_1} k'([x_1 - x_1^{(j)}]/\theta_1) \prod_{p=2}^{D} k([x_p - x_p^{(j)}]/\theta_p).$$

Observe that k is continuously differentiable and that Ω is bounded. Hence by uniform continuity as $\theta_1 \to \infty$

$$\sup_{x\in\Omega} \left|\left|\frac{\partial r_\theta(x)}{\partial \mathrm{x}_1}\right|\right| \to 0.$$

Hence, $\vartheta_1(\theta) \to 0$ as $\theta_1 \to \infty$. Let now for $x = (u, v)$ with $u \in \mathbb{R}$, $l_{\theta_{-1}}(x)$ be the $n \times 1$ vector defined by $[l_{\theta_{-1}}(x)]_j = \prod_{r=1}^{d-1} k([v_r - v_r^{(j)}]/\theta_{r+1})$ (we recall that for $j = 1, ..., n$, $v^{(j)} = x_{-1}^{(j)}$). Let $\widehat{g}_{\theta_{-1}}(x) = l_{\theta_{-1}}(v) L_{\theta_{-1}}^{-1} y$. Then for $m = 2, ..., D$, by the triangle and Cauchy-Schwarz inequalities

$$\begin{aligned} &\left|\frac{\partial m_{\theta,Z}(x)}{\partial \mathrm{x}_m} - \frac{\partial \widehat{g}_{\theta_{-1}}(x)}{\partial \mathrm{x}_m}\right| \\ &\leq \left|\left|\frac{\partial r_\theta(x)}{\partial \mathrm{x}_m} - \frac{\partial l_{\theta_{-1}}(x)}{\partial \mathrm{x}_m}\right|\right| . \left|\left|K_\theta^{-1} y\right|\right| + \left|\left|\frac{\partial l_{\theta_{-1}}(x)}{\partial \mathrm{x}_m}\right|\right| . \left|\left|K_\theta^{-1} y - L_{\theta_{-1}}^{-1} y\right|\right| . \end{aligned} \tag{6.1}$$

In (6.1), the vector in the first norm has component $r \in \{1, ..., n\}$ equal to

$$\left(k((u - u_r)/\theta_1)) - 1\right) \frac{1}{\theta_m} k'([v_{m-1} - v_{m-1}^{(r)}]/\theta_m) \prod_{\substack{p=2,...,D \\ p\neq m}} k([v_{p-1} - v_{p-1}^{(r)}]/\theta_p)$$

which goes to 0 as $\theta_1 \to \infty$, uniformly over $x \in \Omega$, by uniform continuity. The second norm in (6.1) is bounded as discussed above. The third norm in (6.1) does not depend on θ_1 and is thus bounded uniformly over $x \in \Omega$ as $\theta_1 \to \infty$. The fourth norm in (6.1) goes to 0 as $\theta_1 \to \infty$ as discussed above.

Hence, uniformly over $x \in \Omega$,

$$\left|\frac{\partial m_{\theta,Z}(x)}{\partial \mathrm{x}_m} - \frac{\partial \widehat{g}_{\theta_{-1}}(x)}{\partial \mathrm{x}_m}\right| \underset{\theta_1\to\infty}{\longrightarrow} 0.$$

Furthermore, the function $\widehat{g}_{\theta_{-1}}$ is continuously differentiable and non-constant on Ω because $\widehat{g}_{\theta_{-1}}(x^{(r)}) = y_r$ for $r = 1, ..., n$ and because the components of y are not all equal. This implies that

$$\liminf_{\theta_1\to\infty} \sum_{m=2}^{D} \vartheta_m(\theta) > 0,$$

which concludes the proof. ■

Proof. [**Proof of Proposition 4.2**] As before, we consider $i = 1$ in the proof. We have for $m = 2, ..., D$ and $r = 1, ..., n$

$$\left(\frac{\partial r_\theta(x)}{\partial \mathrm{x}_m}\right)_r = k([x_1 - x_1^{(r)}]/\theta_1) \frac{1}{\theta_m} k'([x_m - x_m^{(r)}]/\theta_m) \prod_{\substack{j=2,...,D \\ j\neq m}} k([x_j - x_j^{(r)}]/\theta_j).$$

Hence, $||\partial r_\theta(x)/\partial \mathrm{x}_m||$ is bounded as $\theta_1 \to 0^+$ uniformly in $x \in \Omega$ from the assumptions on k.

For $j = 1, \ldots, n$, let u_j be the first component of $x^{(j)}$ and let $v^{(j)} = x^{(j)}_{-1}$. As $\theta_1 \to 0^+$, the matrix K_θ converges to the $n \times n$ matrix $N_{\theta_{-1}} = \left[\mathbf{1}_{u_p=u_q}(L_{\theta_{-1}})_{pq}\right]_{p,q=1,\ldots,n}$ with the notation of the proof of Proposition 4.1. The matrix $N_{\theta_{-1}}$ is invertible because its submatrices are invertible. This is so because for any $p = 1, ..., n$ the subset $\{v^{(q)}; q = 1, ..., n, u_q = u_p\}$ is composed of two-by-two distinct elements since $x^{(1)}, ..., x^{(n)}$ are two-by-two distinct.

Hence, $||K_\theta^{-1}y||$ is bounded as $\theta_1 \to 0^+$ and so $\sum_{m=2}^{D} \vartheta_m(\theta)$ is bounded as $\theta_1 \to 0^+$.

Let now $j \in \{1, ..., n\}$ for which $y_j \neq 0$. Let $\delta > 0$, not depending on θ_1, be small enough so that $\prod_{r=1}^{D}[x_r^{(j)} - \delta, x_r^{(j)} + \delta] \in \Omega$. Then we have

$$\sup_{s\in[-\delta,\delta]^D;|s_1|=\sqrt{\theta_1}} \left|m_{\theta,Z}(x^{(j)} + s)\right| \underset{\theta_1\to 0^+}{\longrightarrow} 0. \tag{6.2}$$

Indeed, we have

$$\left(r_\theta(x^{(j)} + s)\right)_p = k\left(\frac{u_p - u_j - s_1}{\theta_1}\right)\prod_{r=2}^{D} k\left(\frac{x_r^{(p)} - x_r^{(j)} - s_r}{\theta_r}\right).$$

The product above is bounded uniformly over $s \in [-\delta, \delta]^D$ by uniform continuity of k. Also, whether $u_p - u_j = 0$ or $u_p - u_j \neq 0$, we have

$$\sup_{|s_1|=\sqrt{\theta_1}} k\left(\frac{u_p - u_j - s_1}{\theta_1}\right) \underset{\theta_1\to 0^+}{\longrightarrow} 0.$$

Finally, $||K_\theta^{-1}y||$ is bounded as $\theta_1 \to 0^+$ as discussed above. Hence (6.2) is proved. Also, let $E = \{u_j\} \times \prod_{r=2}^{D}[x_r^{(j)} - \delta, x_r^{(j)} + \delta]$. Then as $\theta_1 \to 0^+$, uniformly over $x \in E$, for $p = 1, ..., n$, we have

$$(r_\theta(x))_p \underset{\theta_1\to 0^+}{\longrightarrow} \mathbf{1}_{\{u_p=u_j\}}\prod_{r=2}^{D} k\left(\frac{x_r - (x_p)_r}{\theta_r}\right).$$

Also $K_\theta^{-1}y \underset{\theta_1\to 0^+}{\longrightarrow} N_{\theta_{-1}}y$ as discussed above. Hence, as $\theta_1 \to 0^+$, $m_{\theta,Z}(x)$ converges uniformly over $x \in E$ to a function value $\widehat{g}_{\theta_{-1}}(x)$, with $\widehat{g}_{\theta_{-1}}(x)$ continuous with respect to $x \in E$. Since $m_{\theta,Z}(x^{(j)}) = y_j$, we can choose the $\delta > 0$ (still independently of θ_1) so that it also satisfies

$$\liminf_{\theta_1\to 0^+} \inf_{x\in E} |m_{\theta,Z}(x)| \geq \frac{|y_j|}{2}. \tag{6.3}$$

We have

$$
\begin{aligned}
\int_{\Omega} \left(\frac{\partial m_{\theta,Z}(x)}{\partial \mathrm{x}_1} \right)^2 dx &\geq \int_{\prod_{r=1}^{D}[x_r^{(j)}-\delta, x_r^{(j)}+\delta]} \left(\frac{\partial m_{\theta,Z}(x)}{\partial \mathrm{x}_1} \right)^2 dx \\
&= \int_{\prod_{r=2}^{D}[x_r^{(j)}-\delta, x_r^{(j)}+\delta]} dx_{-1} \int_{x_1^{(j)}-\delta}^{x_1^{(j)}+\delta} dx_1 \left(\frac{\partial m_{\theta,Z}(x)}{\partial \mathrm{x}_1} \right)^2 \\
&\geq \int_{\prod_{r=2}^{D}[x_r^{(j)}-\delta, x_r^{(j)}+\delta]} dx_{-1} \int_{x_1^{(j)}-\sqrt{\theta_1}}^{x_1^{(j)}} dx_1 \left(\frac{\partial m_{\theta,Z}(x)}{\partial \mathrm{x}_1} \right)^2 \\
\text{(Jensen:)} &\geq \int_{\prod_{r=2}^{D}[x_r^{(j)}-\delta, x_r^{(j)}+\delta]} dx_{-1} \sqrt{\theta_1} \left(\frac{1}{\sqrt{\theta_1}} \int_{x_1^{(j)}-\sqrt{\theta_1}}^{x_1^{(j)}} dx_1 \frac{\partial m_{\theta,Z}(x)}{\partial \mathrm{x}_1} \right)^2 \\
&\geq (2\delta)^{D-1} \frac{1}{\sqrt{\theta_1}} \left(\inf_{x \in E} |m_{\theta,Z}(x)| - \sup_{s \in [-\delta,\delta]^D; |s_1| = \sqrt{\theta_1}} \left| m_{\theta,Z}(x^{(j)} + s) \right| \right)^2 \\
&\underset{\theta_1 \to 0^+}{\longrightarrow} \infty,
\end{aligned}
$$

from (6.2) and (6.3). This concludes the proof. ■

Proof. [**Proof of Proposition 4.3**] Without loss of generality, we consider $i = 1$ in the proof. Let us consider the 2×2 submatrix of $k_\theta(X, X)$ obtained by extracting the lines and columns a, b, with a, b as in Condition iii) of the lemma. Then as $\theta_1 \to \infty$ this submatrix converges to the singular matrix $((1,1)^\top, (1,1)^\top)$. Hence, we have, as $\theta_1 \to \infty$, $|k_\theta(X, X)| \to 0$ (since $k_\theta(X, X)$ has components bounded in absolute value by 1). Hence, it is sufficient to show that $y^\top k_\theta(X, X)^{-1} y$ is bounded in order to conclude the proof.

Let X_{θ_1} be obtained from X by dividing its first column by θ_1 and by leaving the other columns unchanged. Let $x^{(\theta_1,j)}$ be the transpose of the line j of X_{θ_1}, for $j = 1, ..., n$. Let $\bar{\theta} = (1, \theta_{-1})$. Then, $y^t k_\theta(X, X)^{-1} y = y^t k_{\bar{\theta}}(X_{\theta_1}, X_{\theta_1})^{-1} y$.

We now use tools from the theory of RKHSs and refer to, e.g., [33] for the definitions and properties of RKHSs used in the rest of the proof. Let $\mathcal{H}$ be the RKHS of $k_{\bar{\theta}}$. Let $\alpha^{(\theta_1)} = k_{\bar{\theta}}(X_{\theta_1}, X_{\theta_1})^{-1} y$. Then, $f_{\theta_1} : \mathbb{R}^D \to \mathbb{R}$ defined by $f_{\theta_1}(x) = \sum_{j=1}^{n} \alpha_j^{(\theta_1)} k_{\bar{\theta}}(x - x^{(\theta_1,j)})$ is the function of $\mathcal{H}$ with minimal RKHS norm $||.||_{\mathcal{H}}$ satisfying $f_{\bar{\theta}_1}(x^{(\theta_1,j)}) = y_j$ for $j = 1, ..., n$.

As $\theta_1 \to \infty$, the points $x^{(\theta_1,1)}, ..., x^{(\theta_1,n)}$ converge to the points $w^{(1)}, ..., w^{(n)}$ with $w^{(i)} = (0, [v^{(i)}]^\top)^\top$. We observe that, by assumption, $y_r = y_s$ for $w^{(r)} = w^{(s)}$. Hence, there exists $\epsilon > 0$ small enough and p column vectors $c^{(1)}, ..., c^{(p)}$ in $\mathbb{R}^D$ with the following properties: (i) each Euclidean ball with center $c^{(m)}$, $m = 1, ..., p$, and radius 2ϵ does not contain two $w^{(r)}, w^{(s)}$ with $y_r \neq y_s$, $r, s \in \{1, ..., n\}$; (ii) each w_j, $j = 1, ..., n$, is contained in a ball with center $c^{(m)}$ with $m \in \{1, ..., p\}$ and radius ϵ; (iii) the p balls with centers $c^{(1)}, ..., c^{(p)}$ and radii 2ϵ are two-by-two non-intersecting. We can also assume that each ball with center $c^{(m)}$, $m = 1, ..., p$ and radius ϵ contains at least one $w^{(j(m))}$ with $j(m) \in \{1, ..., n\}$ and we write $z_m = y_{j(m)}$.

Then, from Lemma 6.2, there exists an infinitely differentiable function g with compact support on $\mathbb{R}^d$ so that for $m = 1, ..., p$, $g(x) = z_m$ for $||x - c^{(m)}|| \leq 2\epsilon$. Hence, for θ_1 large enough, the function g satisfies $g(x^{(\theta_1,j)}) = y_j$ for $j = 1, ..., n$.

Hence, $||f_{\theta_1}||_{\mathcal{H}} \leq ||g||_{\mathcal{H}}$ for θ_1 large enough, where $||g||_{\mathcal{H}}$ does not depend on θ_1. Finally, a simple manipulation of $||.||_{\mathcal{H}}$ (see again [33] for the definitions), provides

$$\begin{aligned}||f_{\theta_1}||_{\mathcal{H}} &= \sum_{r,s=1}^{n} \alpha_r^{(\theta_1)}\alpha_s^{(\theta_1)} k_{\bar{\theta}}(x_r^{(\theta_1)} - x_s^{(\theta_1)}) \\ &= y^\top k_\theta(X,X)^{-1} k_\theta(X,X) k_\theta(X,X)^{-1} y \\ &= y^\top k_\theta(X,X)^{-1} y.\end{aligned}$$

This concludes the proof. ■

Proof. [**Proof of Proposition 4.4**] Without loss of generality, we consider $i = 1$ in the proof. Also, up to renumbering the lines of X and components of y, it is sufficient to show that, for fixed $\theta_{-1} \in (0,\infty)^D$, as $\theta_1 \to \infty$, $\widehat{y}_{\theta,n} \to y_n$. We use the same notation $\bar{\theta}$, $\mathcal{H}$ and $x^{(\theta_1,j)}$ as in the proof of Proposition 4.3. Then, we have $\widehat{y}_{\theta,n} = f_{\theta_1}(x^{(\theta_1,n)})$, where $f_{\theta_1} \in \mathcal{H}$ is the function with minimal norm $||.||_{\mathcal{H}}$ satisfying $f_{\theta_1}(x^{(\theta_1,j)}) = y_j$ for $j = 1, ..., n-1$.

Furthermore, from the proof of Proposition 4.3, there exists a function $g \in \mathcal{H}$, not depending on θ_1 satisfying, for θ_1 large enough, $g(x^{(\theta_1,j)}) = y_j$ for $j = 1, .., n$. This shows that $||f_{\theta_1}||_{\mathcal{H}}$ is bounded as $\theta_1 \to \infty$. Let $m \in \{1, ..., n-1\}$ be so that $v^{(m)} = v^{(n)}$ (the existence is assumed in Condition iii)). Let also, for $x \in \mathbb{R}^D$, $k_{\bar{\theta},x} \in \mathcal{H}$ be the function $k_{\bar{\theta}}(x - .)$. Then we have (see again [33]), with $(.,.)_{\mathcal{H}}$ the inner product in $\mathcal{H}$

$$\begin{aligned}|\widehat{y}_n - y_n| &= \left| f_{\theta_1}(x^{(\theta_1,n)}) - f_{\theta_1}(x^{(\theta_1,m)}) \right| \\ &= \left| (f_{\theta_1}|k_{\bar{\theta},x^{(\theta_1,n)}})_{\mathcal{H}} - (f_{\theta_1}|k_{\bar{\theta},x^{(\theta_1,m)}})_{\mathcal{H}} \right| \\ &\leq ||f_{\theta_1}||_{\mathcal{H}} ||k_{\bar{\theta},x^{(\theta_1,n)}} - k_{\bar{\theta},x^{(\theta_1,m)}}||_{\mathcal{H}} \\ &= ||f_{\theta_1}||_{\mathcal{H}} \sqrt{k_{\bar{\theta}}(x^{(\theta_1,n)} - x^{(\theta_1,n)}) + k_{\bar{\theta}}(x^{(\theta_1,m)} - x^{(\theta_1,m)}) - 2k_{\bar{\theta}}(x^{(\theta_1,n)} - x^{(\theta_1,m)})}.\end{aligned}$$

In the above display, the square root goes to zero as $\theta_1 \to \infty$ because $x^{(\theta_1,n)} - x^{(\theta_1,m)}$ goes to zero and $k_{\bar{\theta}}$ is continuous. This concludes the proof. ■

Lemma 6.1. *For any $0 < \epsilon_1 < \epsilon_2 < \infty$, there exists an infinitely differentiable function $g : \mathbb{R} \to \mathbb{R}$ satisfying $g(u) = 1$ for $|u| \leq \epsilon_1$ and $g(u) = 0$ for $|u| \geq \epsilon_2$.*

Proof. Let $h : \mathbb{R} \to \mathbb{R}$ be defined by $h(t) = \exp(-1/(1-t^2))\mathbf{1}\{t \in [-1,1]\}$. Then h is infinitely differentiable. Hence, g can be chosen of the form

$$g(t) = \begin{cases} A\int_{-\infty}^{t} h\left(B\left[u + \frac{\epsilon_1+\epsilon_2}{2}\right]\right) du & \text{if } \ t \leq 0 \\ A\int_{t}^{\infty} h\left(B\left[u - \frac{\epsilon_1+\epsilon_2}{2}\right]\right) du & \text{if } \ t \geq 0 \end{cases},$$

with $2/(\epsilon_2 - \epsilon_1) < B < \infty$ and $A = B/(\int_{-\infty}^{\infty} h(u)du)$. It can be checked that g is infinitely differentiable and satisfies the conditions of the lemma. ■

Lemma 6.2. *Let $d, p \in \mathbb{N}$. Let $x^{(1)}, ..., x^{(p)}$ be two-by-two distinct points in $\mathbb{R}^D$ and $\epsilon > 0$ be so that the p closed Euclidean balls with centers $x^{(i)}$ and radii ϵ are disjoint. Let $y_1, ..., y_p \in \mathbb{R}$ be arbitrary. Then there exists an infinitely differentiable function $r : \mathbb{R}^D \to \mathbb{R}$, with compact support, satisfying for $i = 1, ..., p$, $g(u) = y_i$ when $||u - x^{(i)}|| \leq \epsilon$.*

Proof. Let $l = \min_{i \neq j} ||x^{(i)} - x^{(j)}||$ and observe that $\epsilon < 2l$. Let g satisfies Lemma 6.1 with $\epsilon_1 = \epsilon^2$ and $\epsilon_2 = l^2/4$. Then the function r defined by $r(u) = \sum_{i=1}^{p} y_i g(||u - x^{(i)}||^2)$ satisfies the conditions of the lemma. ■

7. Conclusion. Performing Bayesian optimization in high dimension is a difficult task. In many real-life problems, some variables are not influential. Therefore, we propose the so called *Split-and-Doubt* algorithm that performs sequentially both dimension reduction and feature oriented sampling. The "split" step (model reduction) is based on a property of stationary ARD kernel of Gaussian process regression. We proved that large correlation lengths correspond to inactive variables. We also showed that classical estimators such ML and CV may assign large correlation lengths to inactive variables.

The "doubt" step questions the "split" step and helps correcting the estimation of the correlation lengths. It is possible to use this strategy for different feature learning purposes such as refinement, optimization and inversion. The optimization *Split-and-Doubt* algorithm has been evaluated on classical benchmark functions embedded in larger dimensional spaces by adding useless input variables. The results show that *Split-and-Doubt* is faster than classical EGO in the whole design space and outperforms it for most of the discussed tests.

The main limitation of *Split-and-Doubt* is that we perform correlation length estimation in the whole design space. This computation is expensive. To overcome this problem, one can use fast maximum likelihood approximation techniques [12]. Future research may investigate such methods and extend *Split-and-Doubt* to constrained optimization.

8. Acknowledgments. Malek Ben Salem is funded by a CIFRE grant from the ANSYS company, subsidized by the French National Association for Research and Technology (ANRT, CIFRE grant number 2014/1349). Part of this research was presented at the Chair in Applied Mathematics OQUAIDO, gathering partners in technological research (BRGM, CEA, IFPEN, IRSN, Safran, Storengy) and academia (CNRS, Ecole Centrale de Lyon, Mines Saint-Etienne, University of Grenoble, University of Nice, University of Toulouse) around advanced methods for Computer Experiments. We thank the participants for their feedback.

Appendix A. In this section, we use boxplots to display the evolution of the best value of the optimization test bench. For each iteration, we display: Left: EGO in light blue, Middle: *Split-and-Doubt* in dark blue, Right: *Split-without-Doubt* in light green.

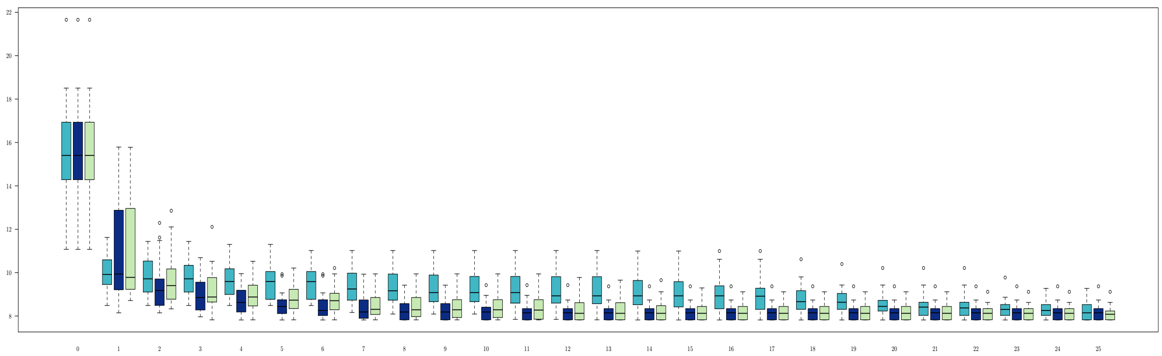


Figure 6: Borehole: Box plots convergence.

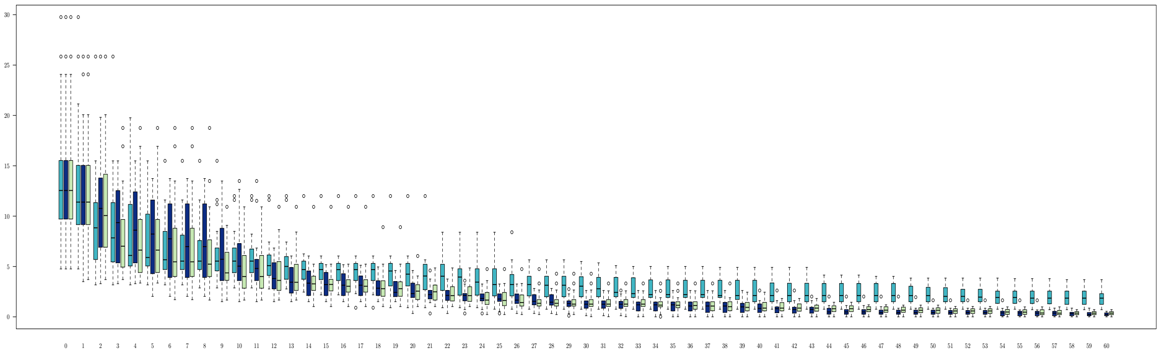


Figure 7: Rosenbrock: Box plots convergence.

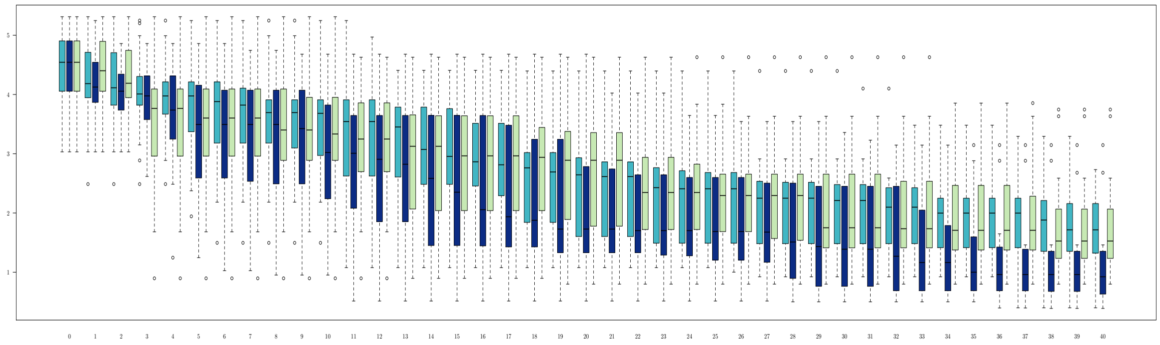


Figure 8: Ackley: Box plots convergence.

REFERENCES

[1] P. Abrahamsen, *A review of Gaussian random fields and correlation functions*, Norsk Regnesentral/Norwegian Computing Center, 1997.

[2] F. Bachoc, *Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification*, Computational Statistics and Data Analysis, 66 (2013), pp. 55 – 69.

[3] F. Bachoc, *Estimation paramétrique de la fonction de covariance dans le modèle de Krigeage par processus Gaussiens: application à la quantification des incertitues en simulation numérique*, PhD thesis, Paris 7, 2013.

[4] F. Bachoc, *Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes*, Journal of Multivariate Analysis, 125 (2014), pp. 1 – 35.

[5] F. Bachoc, A. Lagnoux, and T. M. Nguyen, *Cross-validation estimation of covariance parameters under fixed-domain asymptotics*, Journal of Multivariate Analysis, (2017).

[6] J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez, *Sequential design of computer experiments for the estimation of a probability of failure*, Statistics and Computing, 22 (2012), pp. 773–793.

[7] R. Benassi, J. Bect, and E. Vazquez, *Robust Gaussian process-based global optimization using a fully*

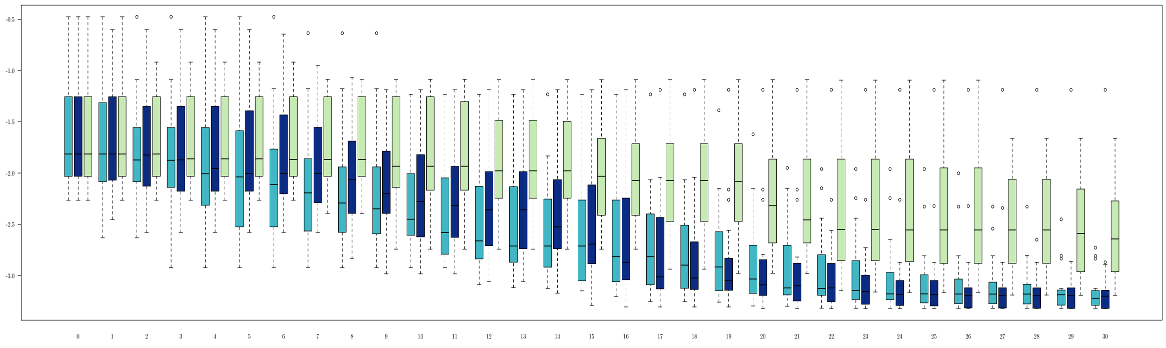


Figure 9: Hartmann 6-dim: Box plots convergence.

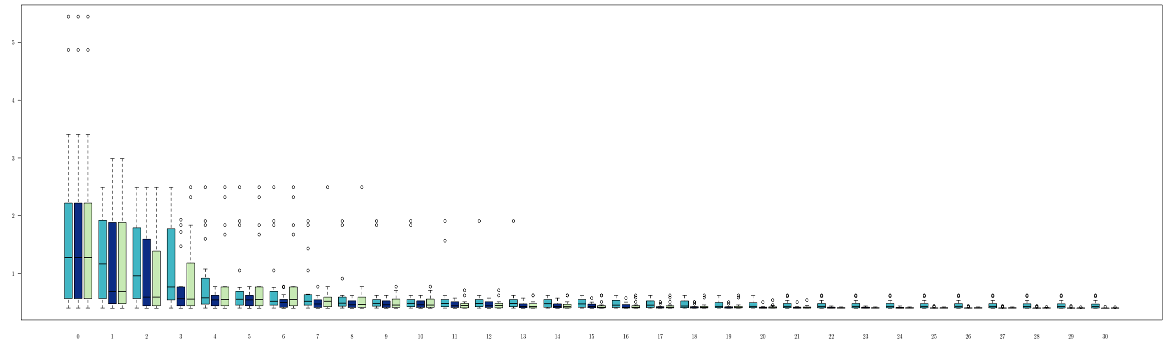


Figure 10: Branin: Box plots convergence.

Bayesian expected improvement criterion., LION, 5 (2011), pp. 176–190.

[8] B. J. Bichon, M. S. Eldred, L.P. Swiler, S. Mahadevan, and J. M. McFarland, *Efficient global reliability analysis for nonlinear implicit performance functions*, AIAA journal, 46 (2008), pp. 2459–2468.

[9] M.Binois, D. Ginsbourger, and O. Roustant, *A Warped Kernel Improving Robustness in Bayesian Optimization Via Random Embeddings*, Springer International Publishing, Cham, 2015, pp. 281–286.

[10] D. Busby, C. L. Farmer, and Armin Iske, *Hierarchical nonlinear approximation for experimental design and statistical data fitting*, SIAM Journal on Scientific Computing, 29 (2007), pp. 49–69.

[11] B. Chen, A. Krause, and R. M. Castro, *Joint optimization and variable selection of high-dimensional Gaussian processes*, in Proceedings of the 29th International Conference on Machine Learning (ICML-12), John Langford and Joelle Pineau, eds., New York, NY, USA, 2012, ACM, pp. 1423–1430.

[12] J .H.S. De Baar, R. P. Dwight, and H. Bijl, *Speeding up kriging through fast estimation of the hyperparameters in the frequency-domain*, Computers & geosciences, 54 (2013), pp. 99–106.

[13] O. Dubrule, *Cross validation of kriging in a unique neighborhood*, Mathematical Geology, 15 (1983), pp. 687–699.

[14] A. I.J. Forrester and D. R. Jones, *Global optimization of deceptive functions with sparse sampling*, in 12th AIAA/ISSMO multidisciplinary analysis and optimization conference, vol. 1012, 2008.

[15] F. Hutter, H. H. Hoos, and K. Leyton-Brown, *Sequential model-based optimization for general*

algorithm configuration., LION, 5 (2011), pp. 507–523.

[16] B. Iooss and Paul Lemaître, *A review on global sensitivity analysis methods*, in Uncertainty Management in Simulation-Optimization of Complex Systems, Springer, 2015, pp. 101–122.

[17] D. J. Jones, *A taxonomy of global optimization methods based on response surfaces*, Journal of global optimization, 21 (2001), pp. 345–383.

[18] D. J. Jones, M. Schonlau, and W. J. Welch, *Efficient global optimization of expensive black-box functions*, Journal of Global optimization, 13 (1998), pp. 455–492.

[19] J. Mockus, *The Bayesian approach to global optimization*, System Modeling and Optimization, (1982), pp. 473–481.

[20] M. D. Morris, T. J. Mitchell, and D. Ylvisaker, *Bayesian design and analysis of computer experiments: use of derivatives in surface prediction*, Technometrics, 35 (1993), pp. 243–255.

[21] V. Picheny, D. Ginsbourger, O. Roustant, R.T. Haftka, and N.H. Kim, *Adaptive designs of experiments for accurate approximation of a target region*, AMSE. J. Mech. Des., 132 (2010), pp. 071008–071008–9.

[22] P. Ranjan, D. Bingham, and G. Michailidis, *Sequential experiment design for contour estimation from complex computer codes*, Technometrics, 50 (2008).

[23] C. E. Rasmussen and C. K.I. Williams, *Gaussian processes for machine learning*, vol. 1, MIT press Cambridge, 2006.

[24] O. Roustant, J. Fruth, B. Iooss, and Sonja Kuhnt, *Crossed-derivative based sensitivity measures for interaction screening*, Mathematics and Computers in Simulation, 105 (2014), pp. 105–118.

[25] O. Roustant, D. Ginsbourger, and Y. Deville, *Dicekriging, diceoptim: Two r packages for the analysis of computer experiments by kriging-based metamodelling and optimization*, Journal of Statistical Software, 51 (2012), p. 54p.

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

[27] M. J. Sasena, *Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations*, PhD thesis, University of Michigan Ann Arbor, MI, 2002.

[28] I.M. Sobol' and A.L. Gershman, *On an alternative global sensitivity estimator*, in Proceedings of SAMO 1995, Belgirate, Italy, SAMO, 1995, pp. 40–42.

[29] I.M. Sobol' and S. Kucherenko, *Derivative based global sensitivity measures and their link with global sensitivity indices*, Mathematics and Computers in Simulation, 79 (2009), pp. 3009 – 3017.

[30] M.L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York, 1999.

[31] S. Streltsov and P. Vakili, *A non-myopic utility function for statistical global optimization algorithms*, Journal of Global Optimization, 14 (1999), pp. 283–298.

[32] Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and Nando de Feitas, *Bayesian optimization in a billion dimensions via random embeddings*, Journal of Artificial Intelligence Research, 55 (2016), pp. 361–387.

[33] H. Wendland, *Scattered data approximation*, vol. 17, Cambridge university press, 2004.

[34] W. Xu and M. L. Stein, *Maximum likelihood estimation for a smooth Gaussian random field model*, SIAM/ASA Journal on Uncertainty Quantification, 5 (2017), pp. 138–175.