SIAM/ASA J. UNCERTAINTY QUANTIFICATION
Vol. 7, No. 4, pp. 1369–1397

Gaussian Process-Based Dimension Reduction for Goal-Oriented Sequential Design*

Malek Ben Salem†, François Bachoc‡, Olivier Roustant§, Fabrice Gamboa‡, and Lionel Tomaso¶

Abstract. Several methods are available for goal-oriented sequential design of expensive black-box functions. Yet, it 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 goal-oriented sampling is achieved 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 the goal-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

DOI. 10.1137/18M1167930

1. Introduction.

In design problems, each study may have a specific goal (e.g., finding an optimum or a level set). Several methods have been proposed to achieve such goals. 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 real-life problems are complex and may involve a large number of variables.

Let us focus first on optimization problems in relatively high dimension, typically several tens. 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

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

*Received by the editors January 30, 2018; accepted for publication (in revised form) August 30, 2019; published electronically December 12, 2019.

https://doi.org/10.1137/18M1167930

Funding: The work of the first author was supported by a CIFRE grant from the ANSYS company, subsidized by the French National Association for Research and Technology (ANRT, CIFRE grant 2014/1349).

†Mines de Saint-Étienne, UMR CNRS 6158, Limos, F-42023, Saint-Étienne and Ansys, Inc F-69100 Villeurbanne, France (malek.ben-salem@emse.fr).

‡IMT Institut de Mathématiques de Toulouse, F-31062 Toulouse, Cedex 9, France (Francois.Bachoc@math.univ-toulouse.fr, fabrice.gamboa@math.univ-toulouse.fr).

§Mines de Saint-Étienne, UMR CNRS 6158, Limos, F-42023, Saint-Étienne, France (roustant@emse.fr).

¶ANSYS, Inc, F-69100 Villeurbanne, France (Lionel.Tomaso@ansys.com).

Bayesian optimization techniques have been successfully used in various problems [24, 21, 20, 33, 39]. These methods give interesting results when the number of evaluations of the function f is relatively low [18]. They are generally limited to problems of moderate dimension, typically up to about 10 [41]. Here, we are particularly interested in the case where the dimension D is large and the number of influential variables d, also called *effective dimension*, is much smaller: $d \ll D$. 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. One may use a goal-oriented sensitivity index to quantify the importance of a variable with respect to the goal under study [16]. For an overview, we refer the reader to [27]. Chen, Krause, and Castro [12] 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 [19]. Other dimension reduction techniques do not necessarily identify the influential variables, though they may identify an influential subspace [23, 11, 41, 9].

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 [41, 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 correct 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. Concluding remarks are given in section 6. For readability, proofs are postponed to section A.

2. General notations and background.

2.1. Gaussian process regression.

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 [38, 29]. 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 Y over Ω as a prior distribution for f. The predictions are given by the conditional distribution of Y given the observations $y = (y_1, \dots, y_n)^\top$, 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 $X = (x^{(1)}, \dots, x^{(n)})^\top \in \Omega^n$ the matrix of observation locations, and by $Z = (X \quad y)$ the matrix of observation locations and values, where $x^{(i)} = (x_1^{(i)}, \dots, x_D^{(i)})$ 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 $(k_\theta(X, X))_{ij} = k_\theta(x^{(i)}, x^{(j)})$ 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 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 $\mathbb{R} \times \mathbb{R}$, and $k : \mathbb{R} \to \mathbb{R}$ is a fixed stationary covariance function. Without loss of generality, we suppose that the hyperparameter 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 hyperparameters of these parametric families can be estimated by different methods, such as, among others, 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 the one hand, estimating the correlation lengths by the ML estimator gives the estimator $\widehat{\theta}^\star_{MLE} \in \arg\max_\theta l_Z(\theta)$, where the likelihood $l_Z(\theta)$ is given as

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

On the other hand, the idea behind 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 [14], the LOO-CV estimator is given as

$$\widehat{\theta}^{\star}_{CV} \in \arg\min_{\theta} \frac{1}{n} y^{\top} K_{\theta}^{-1} \operatorname{diag}(K_{\theta}^{-1})^{-2} K_{\theta}^{-1} y. \tag{2.5}$$

For more insight on these estimators, one can refer to [3]. In high dimension, an accurate estimation of the correlation lengths is challenging. Several methods have been proposed to overcome this problem, such as [42, 22].

2.2. Derivative-based global sensitivity measures. Sobol and Kucherenko [35, 37] 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:

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

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 noninfluential inputs [30]. 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. It is close to the normalized criterion defined in [40]. Here, we do not take into account the parameters defining the supports of the input variable, contrary to [40]:

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

3. The Split-and-Doubt DesignAlgorithm (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 $\widehat{\theta^{\star}} = (\widehat{\theta^{\star}}_1, \ldots, \widehat{\theta^{\star}}_D)$ 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) $I_M = \{i; \widehat{\theta^\star}_i < T\}$,
(ii) $I_m = \{i; \widehat{\theta^\star}_i \geq T\}$,

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 subvector 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 reordering used to obtain x by gathering x_{I_M} and x_{I_m}.

Doubt. Notice that the correlation lengths estimation $\widehat{\theta^\star}$ defines the variable splitting. Naturally, we are concerned if the estimation is misleading, specifically, if an influential variable is misclassified as a minor variable. Therefore, if a rival correlation length vector θ is available, we define the so-called doubt function. It reflects how much doubt the rival estimation would cast on the definition of I_m. It is a decreasing function of the correlation lengths of the minor variables. 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 vectors $\theta = (\theta_1,\dots,\theta_D) \in [0,\infty]^D$,*

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

Contrast. Given two different correlation length vectors $\theta^{(1)}$ 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 length vectors $\theta^{(1)}$ 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 [21] is described in Algorithm 3.1.

Here, Algorithm 3.1 is applied for optimization. We used the EI criterion:

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

It is possible to use any other classical optimization criterion to sample $x_M^\star$ (step 6 of Algorithm 3.1). We can use other criteria for other purposes, such as contour estimation [26, 28, 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 $\ell = \operatorname{erf}(\frac{1}{\sqrt{2}})$ and an adaptive threshold $T = 20 \min_{i \in [1,D]}(\widehat{\theta^\star}_i)$.

Algorithm 3.1 Split-and-Doubt-EGO (f)

1: **Algorithm parameters:** ℓ, kernel k, threshold T.
2: **Start:** Inputs: $Z = (X, y)$.
3: **while** Stop conditions are not satisfied **do**
4: Estimate the correlation lengths (for instance by ML): $\widehat{\theta^\star} \in \underset{\theta \in (\mathbb{R}_+^\star)^D}{\arg\max}\, l_Z(\theta)$ Eq. (2.4)
5: Split the variables into major and minor variables using the correlation lengths estimation: Define $I_M = \{i; \widehat{\theta_i^\star} < T\}$ and $I_m = \{i; \widehat{\theta_i^\star} \geq T\}$, $d_m = |I_m|$.
6: Perform goal-oriented 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 $Z_M = (X_{I_M}, y)$

$$x_M^\star \in \underset{x_M \in \Omega_M}{\arg\max}\, \mathrm{EI}_{Z_M}(x_M).$$

7: Doubt the variable splitting by challenging the correlation lengths estimation: Compute a challenger θ' for correlation lengths.

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

8: Design in the minor subspace to reveal whether $\widehat{\theta^\star}$ or θ' is more accurate: Compute $x_m^\star$ by maximum contrast with the challenger θ'

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

9: Update: Evaluate the new point output $y_{n+1} = f(x^{(n+1)})$ with $x_{I_M}^{(n+1)} = x_M^\star$ and $x_{I_m}^{(n+1)} = x_m^\star$ and add the new point to the design:

$$X^\top \leftarrow \begin{pmatrix} X^\top & x^{(n+1)} \end{pmatrix}, \quad y^\top \leftarrow (y^\top, y_{n+1}).$$

10: **end while**
11: **return Outputs:** $Z = (X, y)$.

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 [15, 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 $\widehat{\theta^\star}$: $\Theta_l = \{\theta; 2|\ln(\frac{l_Z(\theta)}{l_Z(\widehat{\theta^\star})})| < l\}$. Notice that we used $l = \chi^2(\ell, d_m)$. Following [15, 12],

the likelihood ratio test is compared to the χ^2 distribution to decide whether the correlation lengths are allowable.

Remark on the contrast. Sampling the coordinates in the noninfluential variable subspace $\{x_M^\star\}\times\Omega_m = \{(x_M^\star, x_m), x_m \in \Omega_m\}$ aims at revealing the contrast between the ML correlation lengths $\widehat{\theta^\star}$ and a challenging correlation length parameter 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 3.3 (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 $\widehat{Y}(x) \sim \mathcal{N}(m_{\theta_2,Z}(x), (\widehat{\sigma}_{\theta_2,Z}(x))^2)$.

However, this approach is computationally more expensive. Indeed, it requires the computation of the expected log-likelihood in case we add the new point, which is not available in closed form and thus more expensive than computing the predictions. Therefore, we 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 correct an inaccurate variable splitting. To do so, let us consider the following example. Let $f(x_1, x_2) = \cos(2\pi x_2)$. We assume that we have at hand 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 $[0.5, 10]^2$.

Misleading estimation. The log-likelihood of the correlation lengths in the search space $[0.5, 10]^2$ for the Matérn 5/2 kernel is displayed in Figure 1. Notice that the likelihood is maximized for different values of θ and that $\widehat{\theta^\star} = (0.5, 10)$ is among these values:

$$(0.5, 10) \in \underset{\theta\in(\mathbb{R}_+^\star)^D}{\arg\max}\, l_Z(\theta).$$

We also display in Figure 1 the function f, the design points, and the predictions using $k_{\widehat{\theta^\star}}$. This example shows that a limited number of observations may lead to inaccurate correlation lengths and consequently inaccurate predictions.

Doubt/Contrast strategy. Adding more points will arguably improve the quality of the correlation lengths estimation. Here, we want to highlight that the improvement due to the Doubt/Contrast strategy is not only related to the fact that more points are sampled. To do so, we set $T = 10$. So, $I_M = \{1\}$ and $I_m = \{2\}$. Notice that the challenger 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 $PC((x_M^\star, x_2), \widehat{\theta^\star}, \theta')$ as a function of x_2.

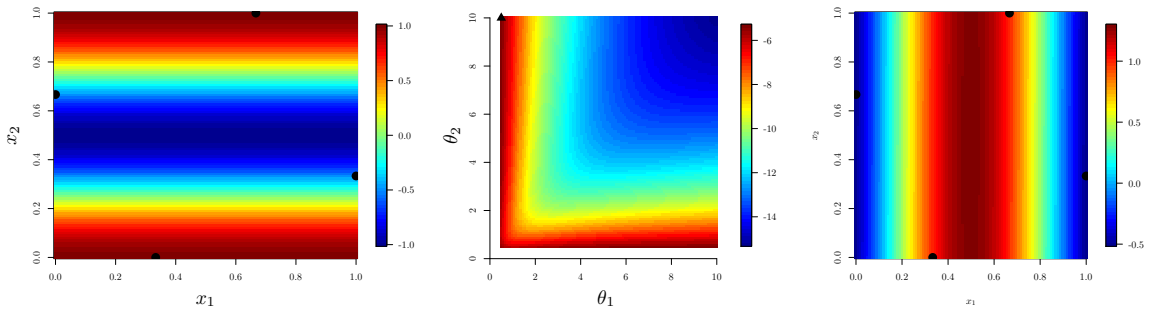


Figure 1. *Left:* $f(x_1, x_2) = \cos(2\pi x_2)$*; the color code indicates the values of* f*, and the solid black circles indicate the design points. Middle: log-likelihood of the correlation lengths, solid black triangle:* $\widehat{\theta^\star}$*. Right: predictions given by the GPR using* $k_{\widehat{\theta^\star}}$*.*

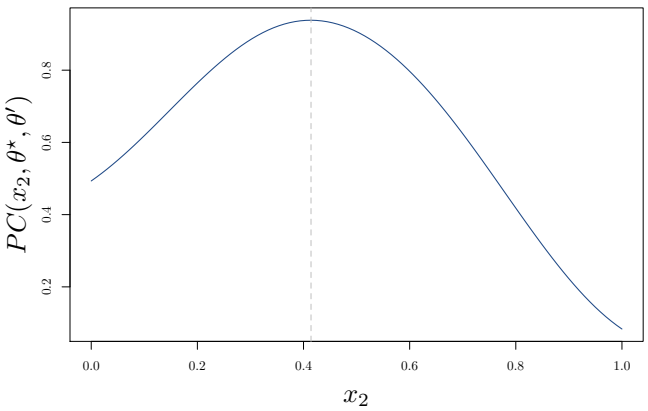


Figure 2. *The prediction contrast in function of* x_2*.*

Let us now consider two cases: (a) We add the point sampled by the maximum contrast $(x_M^\star, x_m^\star)$, and (b) we add the point with the minimum contrast $(x_M^\star, 1)$. For both cases, we display the updated likelihood function in Figure 3.

Notice the following:

(a) For $x_m = x_m^\star$ (maximum contrast), the log-likelihood has larger value for small values of θ_2. Thus, the same inaccurate variable splitting is prevented.

(b) For $x_m = 1$ (small contrast value), we may still use the same misleading variable splitting.

Even if this two-dimensional toy example is pathological, it illustrates the interests of the Doubt/Contrast strategy. This strategy can be valuable in relatively high dimension when the number of observations is relatively small to estimate accurately the correlation lengths.

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 a

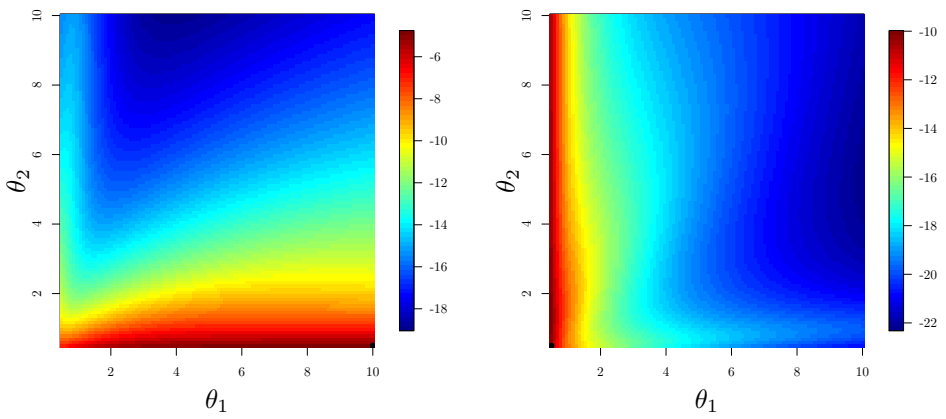


Figure 3. *Left: log-likelihood of the correlation lengths if we add* $((x_M^\star, x_m^\star), f(x_M^\star, x_m^\star))$. *Right: log-likelihood of the correlation lengths if we add* $((x_M^\star, 1), f(x_M^\star, 1))$.

centered stationary GP with 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 (resp., 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 V and W, we denote by $g(V, W)$ the $p \times q$ matrix defined by $[g(V, W)]_{i,j} = g(V_i, W_j)$, 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 the following.

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

This assumption holds for instance when the spectral density of the stationary kernel k is strictly positive almost everywhere. Further, for any vector u, we let u_{-i} be obtained from u by removing the ith component of u.

4.1. Correlation lengths and derivative-based global sensitivity measures. Consider a function f to be observed at the locations $x^{(1)}, \dots, x^{(n)} \in \Omega$ 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 $x^{(1)}, \dots, x^{(n)}$, y be the vector of responses $y = (f(x^{(1)}), \dots, f(x^{(n)}))^\top$, and Z be the $n \times (D+1)$ 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 $r_\theta(x) = k(x, X)$ and $K_\theta = k_\theta(X, X)$. 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:

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

The normalized DGSM index $\widetilde{\vartheta}_i(\theta)$ satisfies $0 \le \widetilde{\vartheta}_i(\theta) \le 1$. 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 $i \in \{1, \dots, D\}$ be fixed. For $j = 1, \dots, n$, let $v^{(j)} = x_{-i}^{(j)}$. Assume that $v^{(1)}, \dots, v^{(n)}$ 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 $x^{(1)}, \dots, x^{(n)}$ are two-by-two distinct. Let $i \in \{1, \dots, D\}$ 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 $x^{(1)}, \dots, x^{(n)}$ hold in many cases, for instance, when $x^{(1)}, \dots, x^{(n)}$ are selected randomly and independently or from a Latin hypercube procedure (see, e.g., [32]).

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

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

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

(iii) *there exist $a, b \in \{1,\ldots,n\}$ 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 $l_Z(\theta)$ going to infinity since if $v^{(1)},\ldots,v^{(n)}$ 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 ith 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 the reader to, e.g., [43] 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 [43, Theorem 10.12] 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

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

where $0 < \nu < \infty$ is the smoothness parameter (see, e.g., [38]). 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 [44] study specifically the asymptotic behavior of the ML 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 $i \in \{1, \dots, D\}$ be fixed.*

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

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

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

(iii) *for all $r \in \{1, \dots, n\}$, there exists $s \in \{1, \dots, n\}$, $r \neq s$, 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. Optimization test problems.

5.1. Numerical tests.

5.1.1. Test procedure.

Test functions. We consider five different analytical functions defined in Appendix C. The first four functions are classical synthetic functions: the general Ackley function (Appendix C.1), the Branin function (Appendix C.3), the six-dimensional Hartmann function (Appendix C.4), and the general Rosenbrock function (Appendix C.5).

The fifth function (Appendix C.2) is the Borehole function [25]. It models the water-flow in a borehole.

Considered algorithms. In this section, we will consider four different algorithms:

- EGO [21]: Implementation of the R package DiceOptim [31] using the default parameters.
- Split-and-Doubt algorithm for optimization (Algorithm 3.1) using the 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.
- HSIC-OPT: Implementation of the authors of [36]. The method is based on the Hilbert-Schmidt Independence Criterion with indicator thresholding. It includes the sensitivity of a variable conditional to its objective reaching a certain performance level. The HSIC-OPT method first selects variables according to the HSIC sensitivity indices. Then an optimization is carried out when the nonselected variables are kept at fixed values.

Table 1
Optimization test functions.

$f^{(i)}$	$d^{(i)}$	$D^{(i)}$	No. of Design points $n_0^{(i)}$	$N_{\max}^{(i)}$	Definition
Ackley (6d 20D)	6	20	45	40	Appendix C.1
Borehole (8d 25D)	8	25	15	25	Appendix C.2
Borehole (8d 10D)	8	10	15	25	Appendix C.2
Branin (2d 25D)	2	25	30	15	Appendix C.3
Hartmann6 (6d 15D)	6	15	30	30	Appendix C.4
Hartmann6 (6d 8D)	6	8	15	25	Appendix C.4
Rosenbrock (20d 100D)	20	100	50	50	Appendix C.5

General setup. We consider seven test cases, summarized in Table 1. For each test case $i = 1, \ldots, 7$, we consider a test function $f^{(i)}$, to which we add inactive input variables in order to embed it in a higher-dimension $D^{(i)}$.

For each function $f^{(i)}$, we launch each optimization process for $N_{\max}^{(i)}$ iterations starting with $N_{seed} = 20$ different initial designs of experiments (DOE) of size $n_0^{(i)}$ generated by a latin hypercube design. At each iteration, each algorithm runs an evaluation of a new point optimization point.

5.1.2. Role of the Doubt/Contrast. Before studying the benchmark results of Table 1, we will study the role of the Doubt step to accurately perform the variable splitting. Let us consider the extended Rosenbrock function in 5 dimension embedded in 20D (Rosenbrock 5d 20D; see Appendix C.5). For this function, we used $n_0 = 40$ and $N_{\max} = 60$. We run an optimization for EGO: Split-and-Doubt and Split-without-Doubt.

The figures displaying the results show the evolution of the best value in function of the number of iteration where the value at iteration 1 is the best value in the initial design space. Figure 4 displays the evolution the mean best value (over N_{seed}), while Figure 9 displays the evolution using boxplots.

The difference between Split-and-Doubt and Split-without-Doubt is due to the Doubt and Contrast steps. In fact, if there is no misclassification of the major variables, the two algorithms should have the same results. Notice that we started with a small amount of design points (40 in dimension 20). Thus, the initial estimation of the correlation lengths can be inaccurate. In these cases, the Doubt/Contrast approach is valuable to improve the estimation. To further highlight this idea, we display in Figure 5 the percentage of undetected influential variables and the misclassification rate of all the variables for both Split-and-Doubt and Split-without-Doubt for the Rosenbrock function.

Among the 20 DOE, the Split-and-Doubt detects all the major variables for 19 cases starting from iteration 15 and for all the DOE starting from iteration 37. However, the Split-without-Doubt struggles to select properly all the influential variables even in the last iterations. Considering all the variables, the misclassification rate decreases rapidly for the Split-and-Doubt. However, for one test a minor variable is considered influential until the end. This can be explained by the nature of the doubt function that aims at correcting only a misclassification of an influential variable.

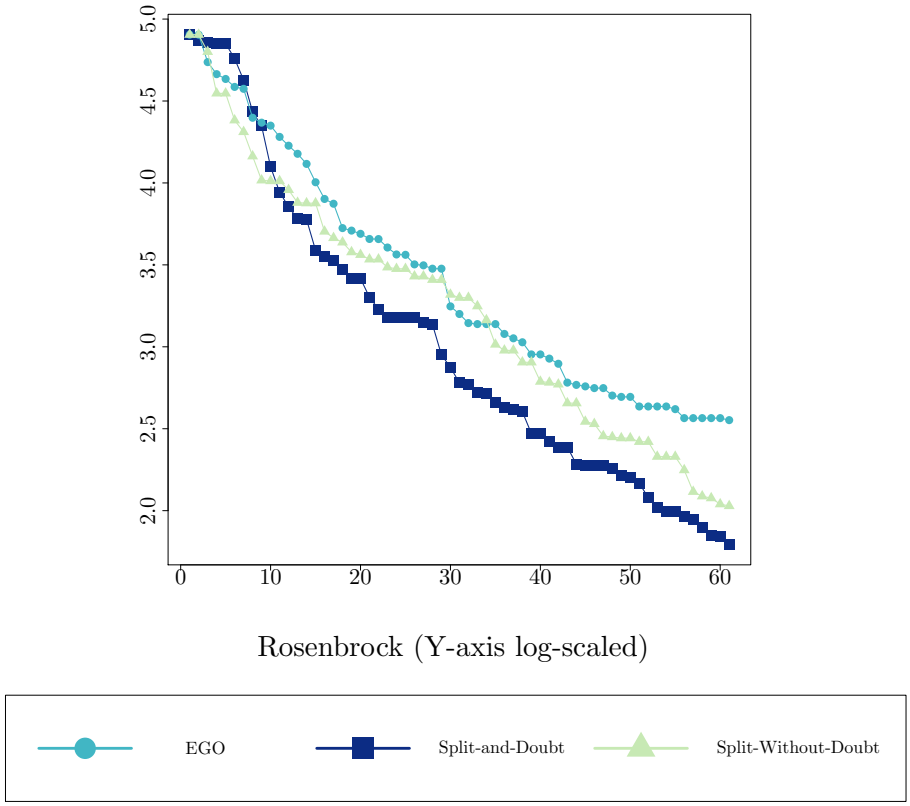
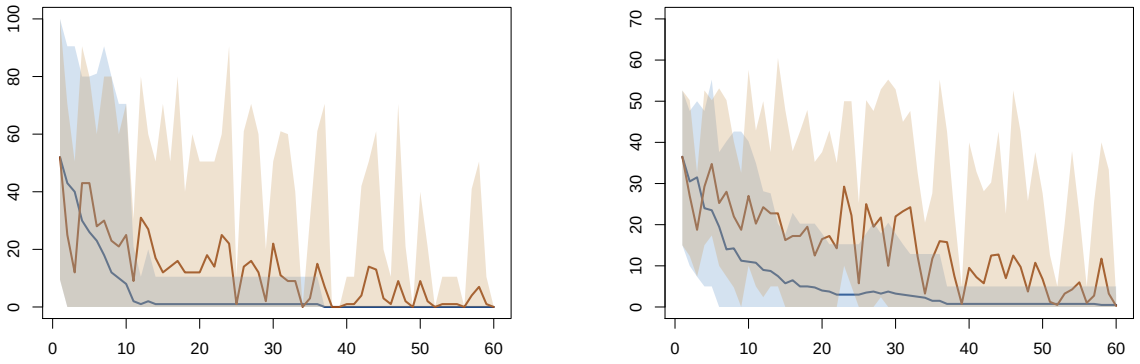


Figure 4. *Comparison of three optimization strategies. Mean over N_{seed} of the best current values as a function of the number of iterations (new added points).*



(a) Misclassification rate of major variables (b) Misclassification rate of all the variables

Figure 5. *Rosenbrock 20D function. Solid line: mean value over 20 repetitions; colored area: 95% confidence interval; blue: Split-and-Doubt; red: Split-without-Doubt; x-axis: iteration number.*

5.1.3. Benchmark.

Setup. The setup of the benchmark is summarized in Table 1.

Results. The results are represented by boxplots in Appendix B. We also display the mean best value evolution in Figure 6.

First, notice that either Split-and-Doubt outperforms Split-without-Doubt (Ackley, Hartmann, and Branin) or both algorithms have slightly similar results (Borehole). This is due to the accuracy of the estimation of the correlation length. When Split-and-Doubt outperforms Split-without-Doubt, this is due the Doubt/Contrast strategy, as explained in the previous subsection.

Further, we can see that Split-and-Doubt gives better results than EGO for all the functions except for the Borehole (8d 10D) function. These cases illustrate the efficiency of the dimension reduction for limited budget optimization. However, the case of the Borehole (8d 10D) function shows that there is no free lunch, so to speak. Indeed, in this case, variables that are only slightly influential may be wrongly discarded as noninfluential by Split-and-Doubt.

Finally, let us compare the results of HSIC-OPT and Split-and-Doubt. The performances look overall similar. In some cases, HSIC-OPT outperforms Split-and-Doubt and vice versa. However, HSIC-OPT outperforms Split-and-Doubt in one case out of six (Ackley 6d 20D) and (not until the final iteration) in a second case (Rosenbrock 20d 100D), while Split-and-Doubt outperforms HSIC-OPT significantly in three cases (Borehole and Branin), and the advantage of Split-and-Doubt remains until the final iteration. Furthermore, before the optimization iterations, the HSIC-OPT procedure actually requires additional function evaluations to compute the HSIC. Hence, at a given iteration number, HSIC-OPT has actually used more function evaluations than Split-and-Doubt. Thus, Split-and-Doubt fulfill both the objectives of optimization and dimension reduction, with limited evaluation budget, better than HSIC-OPT.

Hence, for the two above reasons, our conclusion is that Split-and-Doubt generally outperforms HSIC-OPT on the set of tests considered here.

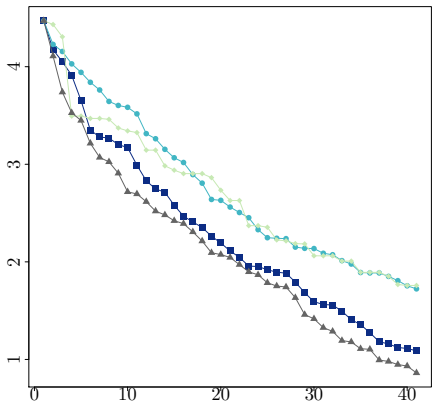
Advantages and drawbacks. The main advantages of Split-and-Doubt are the following:

+ Split-and-Doubt is relevant when some variables are not influential.
+ The Doubt/Contrast strategy is relevant to correct an inaccurate estimation.

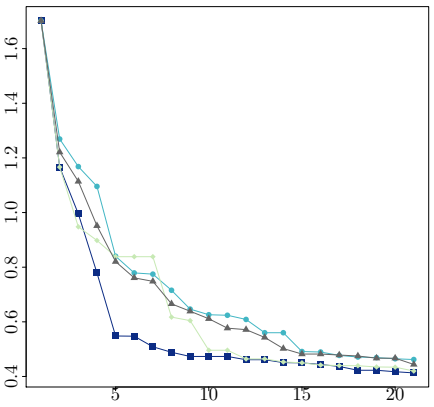
On the other hand, the main drawback of Split-and-Doubt is that input variables with mild global influences can be wrongly classified as noninfluential (this depends on the threshold parameter T).

Computing time. Figure 7 illustrates the computing time in seconds of Split-and-Doubt, Split-without-Doubt, and EGO for four benchmark functions (Hartmann, Branin, Rosenbrock, and Ackley). Clearly, Split-and-Doubt is faster than EGO. The fact that we perform two optimization procedures in smaller spaces makes the algorithm faster than optimizing the EI in dimension D.

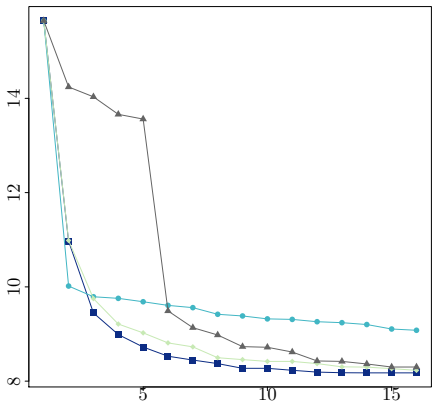
5.2. Application to a differential signaling model. We consider a model of a pair of microstrip lines that transition through the vertical interconnect access (via) to a pair of striplines on a lower layer; see Figure 8 for an illustration. The two microstrip lines are each assigned a terminal in the coupled microstrip port. This is similar for the two striplines at the opposite end. The conductors are copper, and a radiation boundary is applied to the air box.



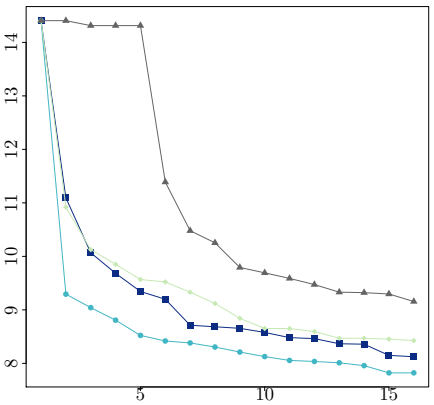
(a) Ackley (6d 20D)



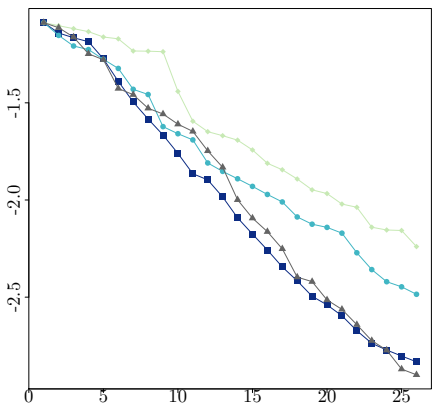
(b) Branin (2d 25D)



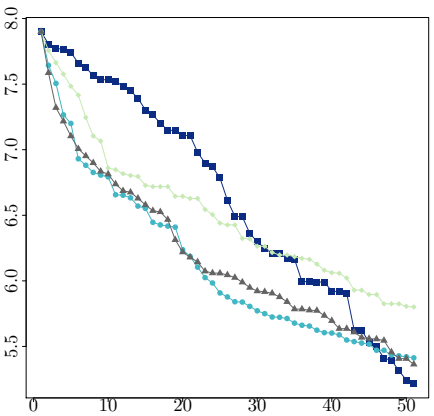
(c) Borehole (8d 25D)



(d) Borehole (8d 10D)



(e) Hartmann6 (6d 8D)



(f) Rosenbrock (20d 100D Y-axis log-scaled)

Figure 6. *Comparison of four optimization strategies. Mean over N_{seed} of the best current values as a function of the number of iterations (new added points).*

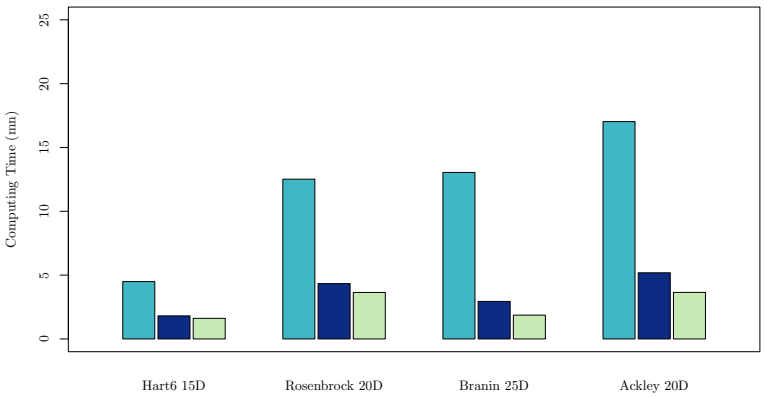


Figure 7. *Mean computing time. Left: EGO. Middle: Split-and-Doubt. Right: Split-without-Doubt in minutes. Functions from left to right: Hartmann* ($6d$ $15D$)*, Rosenbrock* ($5d$ $20D$)*, Branin* ($2d$ $25D$)*, Ackley* ($6d$ $20D$)*.*

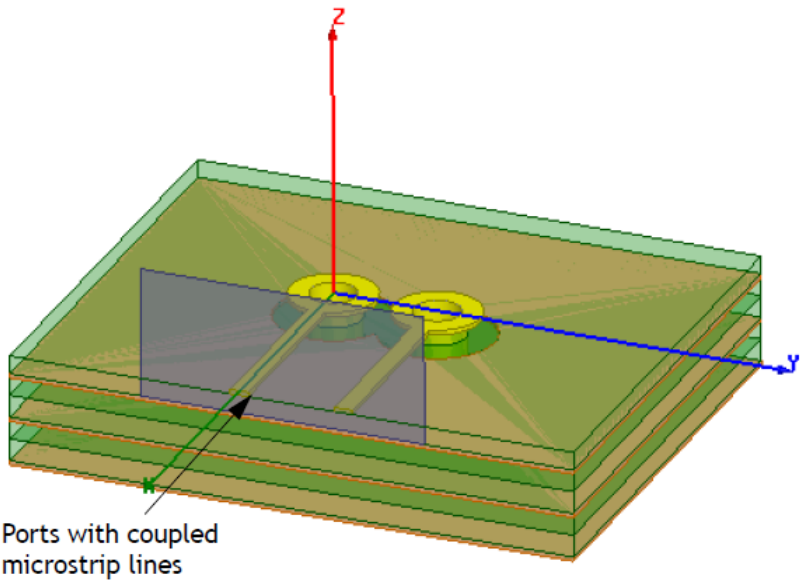


Figure 8. *Differential signaling model.*

The design problem has 21 geometrical variable. The cost function to minimize represents the objective of ensuring a threshold on the transfer function. The objective function is computed by a simulation using Ansys HFSS. To perform the optimization, we used Split-and-Doubt and a gradient-based algorithm (NLPQL [34]) as a reference method.

As we can see in Table 2, Split-and-Doubt outperforms the reference algorithm. This is in part due to the local nature of this gradient-based algorithm. We also remark that the experts in electromagnetism that provided us with the differential signaling model are satisfied with

Table 2
Results of the optimization methods.

Algorithm	Reference value	Best value	Discovered at iteration
Split-And-Doubt	154.6	7.6	67
NLPQL	154.6	68.2	83

the optimization results of Split-and-Doubt. In fact, the cost function should ideally be 0. But there is no theoretical guarantee that such design exists. The value design discovered by the Split-and-Doubt provides a close enough value compared to the other optimizer. Additionally, the list of noninfluential variables provided by Split-and-Doubt is in agreement with the intuition of the electromagnetism experts.

6. 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, which performs sequentially both dimension reduction and goal-oriented sampling. The split step (model reduction) is based on a property of stationary ARD kernels of the GPR. 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 correct the estimation of the correlation lengths. It is possible to use this strategy for study 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. In various test cases, Split-and-Doubt also outperforms the HSIC-OPT procedure of [36], which performs dimension reduction and optimization but more separately than for Split-and-Doubt and requires more function evaluation to perform dimension reduction.

The main benefit of Split-and-Doubt, for instance, for optimization, is that variable selection and optimization are performed at the same time. Furthermore, the Doubt step allows efficiently correcting misclassifications of input variables as influential or noninfluential. A main limitation of Split-and-Doubt is that variables that are mildly influential can be wrongly classified as noninfluential. Another 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 ML approximation techniques [13]. Split-and-Doubt is also based on the GPR, so, without projection, it requires an initial number of points larger than D. Future research may investigate fast likelihood methods and extend Split-and-Doubt to constrained optimization.

Appendix A. Proofs. We first give the sketches of the proofs of Propositions 4.1, 4.2, 4.3, and 4.4. Then we give the full proofs. The intermediary Lemma A.1 for the proof of Propositions 4.3 and 4.4 is given at the end of the proof section. For the proofs of Propositions 4.1 and 4.2, we let $k'(t) = \partial k(t)/\partial t$.

A.1. Sketches of the proofs.

Sketch of proof of Proposition 4.1. We let $i = 1$ without loss of generality. We first observe that, when computing $\partial m_{\theta,Z}(x)/\partial \mathrm{x}_1$, a $1/\theta_1$ appears as a factor. This and the regularity of k and k' enable us to show that $\vartheta_1(\theta) \to 0$ as $\theta_1 \to \infty$.

Then, we show that for $m = 2, \dots, D$, $\partial m_{\theta,Z}(x)/\partial \mathrm{x}_m$ converges uniformly to $\partial \widehat{g}_{\theta_{-1}}(x)/\partial \mathrm{x}_m$, where $\widehat{g}_{\theta_{-1}}(x)$ is a GP predictor of $f(x)$ that does not depend on x_1 and θ_1. From the interpolation property of $\widehat{g}_{\theta_{-1}}$, we show that this function is not constant on Ω so that $\sum_{m=2}^{D} \int_\Omega (\partial \widehat{g}_{\theta_{-1}}(x)/\partial \mathrm{x}_m)^2 dx > 0$, which implies that $\liminf_{\theta_1 \to \infty} \sum_{m=2}^{D} \vartheta_m(\theta) > 0$. ∎

Sketch of proof of Proposition 4.2. We let $i = 1$ without loss of generality. We first show that $\vartheta_2(\theta), \dots, \vartheta_D(\theta)$ are bounded as $\theta_1 \to 0^+$, using the regularity of k and k' and using that K_θ converges to an invertible matrix as $\theta_1 \to 0^+$. The limit matrix is shown to be invertible from Assumption 1.

Then we show that $\vartheta_1(\theta) \to \infty$ as $\theta_1 \to 0^+$. For this, we consider j so that $y_j \neq 0$, and we observe that $m_{\theta,Z}(x^{(j)}) = y_j$. As $\theta_1 \to 0^+$, we show that $m_{\theta,Z}(x)$ goes to 0 when $(x_2, \dots, x_D) = (x_2^{(j)}, \dots, x_D^{(j)})$ and $|x_1 - x_1^{(j)}| = \sqrt{\theta_1}$. This means that the derivative of $m_{\theta,Z}$ w.r.t. x_1 can be very large, from which we show $\int_\Omega (\partial m_{\theta,Z}(x)/\partial \mathrm{x}_1)^2 \to \infty$ by using in particular the Jensen inequality. ∎

Sketch of proof of Proposition 4.3. We first show that $\det(k_\theta(X, X)) \to 0$ as $\theta_1 \to \infty$ from condition (iii) of the proposition since $k_\theta(x^{(a)}, x^{(b)}) \to k(0)^D = 1 = k_\theta(x^{(a)}, x^{(a)}) = k_\theta(x^{(b)}, x^{(b)})$ as $\theta_1 \to \infty$ with a, b as in this condition. Hence, it remains to show that $y^\top k_\theta(X, X)^{-1} y$ is bounded as $\theta_1 \to \infty$.

We express $y^\top k_\theta(X, X)^{-1} y$ as the RKHS norm $\|f_{\theta_1}\|_{\mathcal{H}}$ of a function f_{θ_1}. This function f_{θ_1} has minimal RKHS norm among all functions with bounded RKHS norms that satisfy specific interpolation conditions. We show that there exists a function g, not depending on θ_1, that also satisfies these interpolation conditions. This is where we use the intermediary Lemma A.1 at the end of the proof section. Hence, by minimality, we have $\|f_{\theta_1}\|_{\mathcal{H}} \leq \|g\|_{\mathcal{H}}$, and so $\|f_{\theta_1}\|_{\mathcal{H}} = y^\top k_\theta(X, X)^{-1} y$ is bounded. ∎

Sketch of proof of Proposition 4.4. We use the same function f_{θ_1} as in the proof of Proposition 4.3. Using again RKHS concepts, we show that, for $m \in \{1, \dots, n-1\}$ so that $v^{(m)} = v^{(n)}$ (see condition (iii)), we have $|\hat{y}_{\theta,n} - y_n| \leq \|f_{\theta_1}\|_{\mathcal{H}} \psi((x_1^{(m)}/\theta_1, v^{(m)}) - (x_1^{(n)}/\theta_1, v^{(n)}))$, where ψ is a continuous function. Since $\|f_{\theta_1}\|_{\mathcal{H}}$ is bounded as seen in the proof of Proposition 4.3, we obtain $|\hat{y}_{\theta,n} - y_n| \to 0$ as $\theta_1 \to \infty$. This concludes the proof by symmetry. ∎

A.2. Complete proofs.

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 and because $k(0) = 1$. This matrix is invertible by Assumption 1 since

$v^{(1)}, \dots, v^{(n)}$ are two-by-two distinct. Hence, $\|K_\theta^{-1} y\|$ is bounded as $\theta_1 \to \infty$. We have for $j = 1, \dots, n$

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

Recall that k is continuously differentiable and that Ω is bounded. Hence, by uniform continuity of k and k' and by the triangle inequality, as $\theta_1 \to \infty$,

$$\sup_{x \in \Omega} \left\| \frac{\partial r_\theta(x)}{\partial \mathrm{x}_1} \right\|^2 \le \frac{1}{\theta_1} n \sup_{x,w \in \Omega} \left(|k'([x_1 - w_1]/\theta_1)| \prod_{p=2}^{D} |k([x_p - w_p]/\theta_p)| \right)^2 \to 0.$$

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 $j = 1, \dots, n$, $v^{(j)} = x_{-1}^{(j)}$). Let $\widehat{g}_{\theta_{-1}}(x) = l_{\theta_{-1}}(x) L_{\theta_{-1}}^{-1} y$. Then for $m = 2, \dots, D$, by the triangle and Cauchy–Schwarz inequalities,

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

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

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

which goes to 0 as $\theta_1 \to \infty$, uniformly over $x \in \Omega$, by uniform continuity of k and since $k(0) = 1$. The second norm in (A.1) is bounded as discussed above. The third norm in (A.1) does not depend on θ_1 and is thus bounded uniformly over $x \in \Omega$ as $\theta_1 \to \infty$ by uniform continuity of k and k'. The fourth norm in (A.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 nonconstant on Ω because $\widehat{g}_{\theta_{-1}}(x^{(r)}) = y_r$ for $r = 1, \dots, n$ and because the components of y are not all equal. This implies that $\sum_{m=2}^{D} \int_\Omega (\partial \widehat{g}_{\theta_{-1}}(x)/\partial \mathrm{x}_m)^2 dx > 0$ and so that

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

which concludes the proof. ∎

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

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

Hence, $\|\partial r_\theta(x)/\partial \mathrm{x}_m\|$ is bounded as $\theta_1 \to 0^+$ uniformly in $x \in \Omega$ because k is a bounded function on $\mathbb{R}$ (since it has limit zero at $\pm\infty$) and because k' is a bounded function on any compact set.

For $j = 1, \dots, n$, let u_j be the first component of $x^{(j)}$, and let $v^{(j)} = x_{-1}^{(j)}$. As $\theta_1 \to 0^+$, the matrix K_θ converges to the $n \times n$ matrix $N_{\theta_{-1}} = [\mathbf{1}_{u_p=u_q}(L_{\theta_{-1}})_{pq}]_{p,q=1,\dots,n}$ with the notation of the proof of Proposition 4.1. Indeed, $(K_\theta)_{ab} = k((u_a - u_b)/\theta_1)\prod_{p=2}^D k((x_p^{(a)} - x_p^{(b)})/\theta_p)$, $k(0) = 1$ and $k(t) \to 0$ as $t \to \pm\infty$.

Let us show that the matrix $N_{\theta_{-1}}$ is invertible. Let $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ be a permutation for which $u_{\sigma(1)} \le \cdots \le u_{\sigma(n)}$. Let $\bar{N}_{\theta_{-1}}$ be defined by $(\bar{N}_{\theta_{-1}})ab = (N_{\theta_{-1}})\sigma(a)\sigma(b)$. Then $\bar{N}_{\theta_{-1}}$ is block diagonal, and each of its blocks is of the form

$$B = \left(\prod_{p=1}^{D-1}((z_{m+1}^{(a)} - z_{m+1}^{(b)})/\theta_{m+1})\right)_{a,b=1,\dots,s}$$

with $s \in \mathbb{N}$, with $\{z^{(1)}, \dots, z^{(s)}\} \subset \{x^{(1)}, \dots, x^{(n)}\}$, and with $z_1^{(1)} = \cdots = z_1^{(s)}$. Since $x^{(1)}, \dots, x^{(n)}$ are two-by-two distinct, then also $\bar{z}^{(1)}, \dots, \bar{z}^{(s)}$ are two-by-two distinct with $\bar{z}^{(r)} = (z_2^{(r)}, \dots, z_D^{(r)})$. As a consequence the block matrix B is invertible by Assumption 1. Hence, $\bar{N}_{\theta_{-1}}$ is invertible, and so also $N_{\theta_{-1}}$ is invertible.

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

Let now $j \in \{1, \dots, n\}$ for which $y_j \neq 0$ (the existence is guaranteed since the components of y are not all equal by assumption). Let $\delta > 0$, not depending on θ_1, be small enough so that $\prod_{r=1}^D [x_r^{(j)} - \delta, x_r^{(j)} + \delta] \in \Omega$ (Ω is assumed to be open). Then we have

$$\sup_{s\in[-\delta,\delta]^D;|s_1|=\sqrt{\theta_1}} \left|m_{\theta,Z}(x^{(j)} + s)\right| \underset{\theta_1\to 0^+}{\longrightarrow} 0. \tag{A.2}$$

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, for $|s_1| = \sqrt{\theta_1}$, if $u_p - u_j = 0$, we have $|u_p - u_j - s_1|/\theta_1 = 1/\sqrt{\theta_1}$, and if $u_p - u_j \neq 0$, we have, for θ_1 small enough, $|u_p - u_j - s_1|/\theta_1 \ge |u_p - u_j|/(2\theta_1)$. Hence, 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, (A.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 $p = 1, \ldots, n$, we have, since $k(0) = 1$ and $k(t) \to 0$ as $|t| \to \infty$,

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

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 have also $\widehat{g}_{\theta_{-1}}(x^{(j)}) = y_j$, and so, from the uniform convergence, we can choose the $\delta > 0$ (still independently of θ_1) so that it also satisfies

$$\liminf_{\theta_1 \to 0^+} \inf_{x \in E} |m_{\theta,Z}(x)| = \inf_{x \in E} \left|\widehat{g}_{\theta_{-1}}(x)\right| \geq \frac{|y_j|}{2}. \tag{A.3}$$

We have

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

from (A.2) and (A.3). This concludes the 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, as $\theta_1 \to \infty$, the smallest eigenvalue of $k_\theta(X, X)$ goes to 0. The largest eigenvalue of $k_\theta(X, X)$ is bounded as $\theta_1 \to \infty$ since $k_\theta(X, X)$ has components bounded by 1 in absolute value and has fixed dimension from the Gershgorin circle theorem. Hence, again because $k_\theta(X, X)$ has bounded dimension, we have $\det(k_\theta(X, X)) \to 0$ as $\theta_1 \to \infty$.

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 $j = 1, \dots, n$. 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., [43] 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 $j = 1, \dots, n$.

As $\theta_1 \to \infty$, the points $x^{(\theta_1, 1)}, \dots, x^{(\theta_1, n)}$ converge to the points $w^{(1)}, \dots, w^{(n)}$ 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 $c^{(1)}, \dots, c^{(p)}$ in $\mathbb{R}^D$ with the following properties: (i) each Euclidean ball with center $c^{(m)}$, $m = 1, \dots, p$, and radius 2ϵ does not contain two $w^{(r)}, w^{(s)}$ with $y_r \neq y_s$, $r, s \in \{1, \dots, n\}$; (ii) each w_j, $j = 1, \dots, n$, is contained in a ball with center $c^{(m)}$ with $m \in \{1, \dots, p\}$ and radius ϵ; and (iii) the p balls with centers $c^{(1)}, \dots, c^{(p)}$ and radii 2ϵ are two-by-two nonintersecting. We can also assume that each ball with center $c^{(m)}$, $m = 1, \dots, p$ and radius ϵ contains at least one $w^{(j(m))}$ with $j(m) \in \{1, \dots, n\}$ and we write $z_m = y_{j(m)}$.

Then, from Lemma A.1, there exists an infinitely differentiable function g with compact support on $\mathbb{R}^d$ so that for $m = 1, \dots, p$, $g(x) = z_m$ 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 $j = 1, \dots, n$.

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 [43] 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 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 $j = 1, \dots, n-1$.

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 $j = 1, \dots, n$. This shows that $\|f_{\theta_1}\|_{\mathcal{H}}$ is bounded as $\theta_1 \to \infty$. Let $m \in \{1, \dots, n-1\}$ 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 [43]), with $(.,.)_{\mathcal{H}}$ the inner product in $\mathcal{H}$,

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

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. ∎

A.3. Intermediary lemma for the proofs of Propositions 4.3 and 4.4.

Lemma A.1. *Let $d, p \in \mathbb{N}$. Let $x^{(1)}, \ldots, x^{(p)}$ 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 $y_1, \ldots, y_p \in \mathbb{R}$ be arbitrary. Then there exists an infinitely differentiable function $g : \mathbb{R}^d \to \mathbb{R}$, with compact support, satisfying for $i = 1, \ldots, p$, $g(u) = y_i$ when $\|u - x^{(i)}\| \leq \epsilon$.*

Proof of Lemma A.1. We first show the following technical lemma.

Lemma A.2. *For any $0 < \epsilon_1 < \epsilon_2 < \infty$, there exists an infinitely differentiable function $\tilde{g} : \mathbb{R} \to \mathbb{R}$ satisfying $\tilde{g}(u) = 1$ for $|u| \leq \epsilon_1$ and $\tilde{g}(u) = 0$ 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, $\tilde{g}$ can be chosen of the form

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

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

Coming back to the proof of Lemma A.1, let $l = \min_{i \neq j} \|x^{(i)} - x^{(j)}\|$ and observe that $\epsilon < 2l$. Let $\tilde{g}$ satisfies Lemma A.2 with $\epsilon_1 = \epsilon^2$ and $\epsilon_2 = l^2/4$. Then the function g defined by $g(u) = \sum_{i=1}^{p} y_i \tilde{g}(\|u - x^{(i)}\|^2)$ satisfies the conditions of the lemma. ∎

Appendix B. Optimization test results. In this section, we use boxplots to display the evolution of the best value of the optimization test bench. For the illustration of subsection 5.1.2, we diplay the result in Figure 9. For each iteration, the algorithm results are displayed, respectively, from left to right: EGO in light blue, Split-and-Doubt in dark blue, and Split-without-Doubt in light green.

The results of the optimization benchmark (Table 1) are displayed in Figures 10–15. For each iteration, the algorithm results are diplayed, respectively, from left to right: EGO in light blue, Split-and-Doubt in dark blue, Split-without-Doubt in light green, and HSIC-OPT in gray.

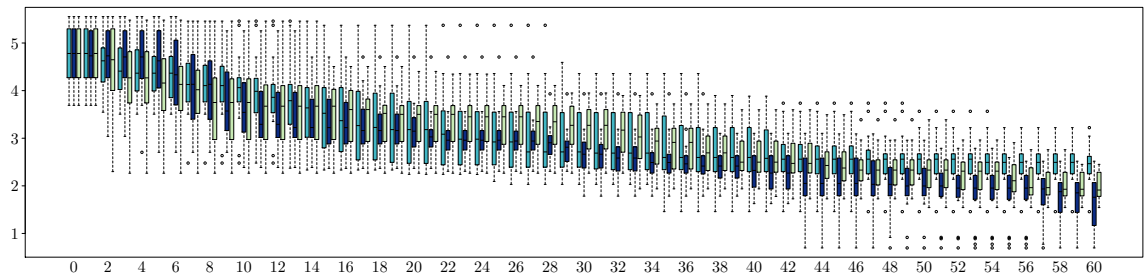


Figure 9. *Rosenbrock (5d 20D). Boxplot convergence. X-axis: number of iterations; Y-axis (log-scaled), the best discovered value.*

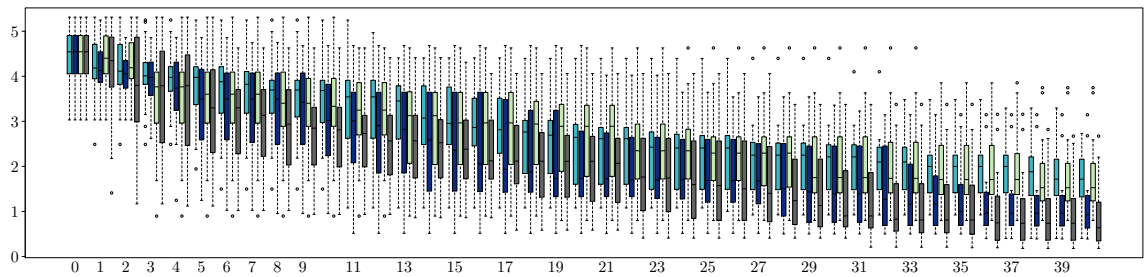


Figure 10. *Ackley (6d 20D). Boxplot convergence. X-axis: number of iterations; Y-axis, the best discovered value.*

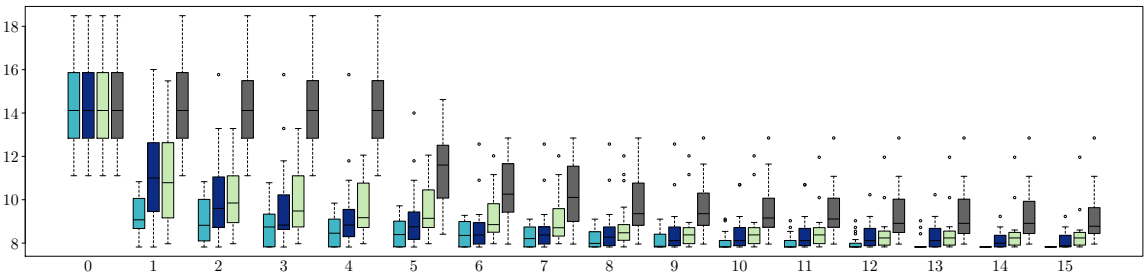


Figure 11. *Borehole (8d 10D). Boxplot convergence. X-axis: number of iterations; Y-axis, the best discovered value.*

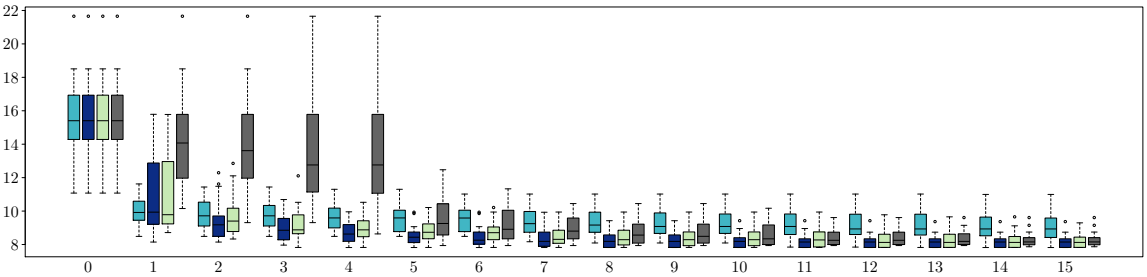


Figure 12. *Borehole (8d 25D). Boxplot convergence. X-axis: number of iterations; Y-axis, the best discovered value.*

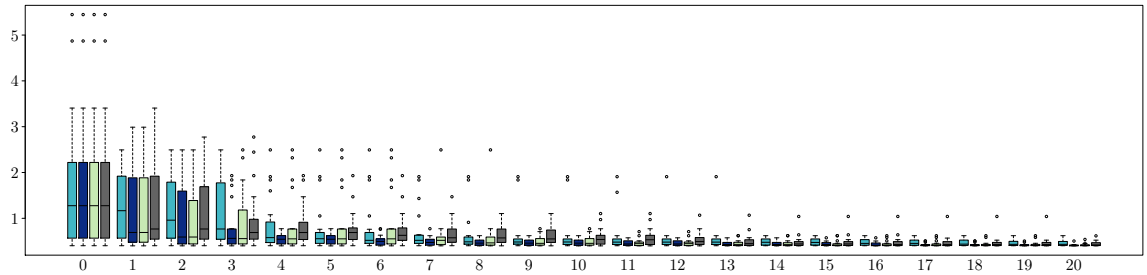


Figure 13. *Branin (2d 25D). Boxplot convergence. X-axis: number of iterations; Y-axis, the best discovered value.*

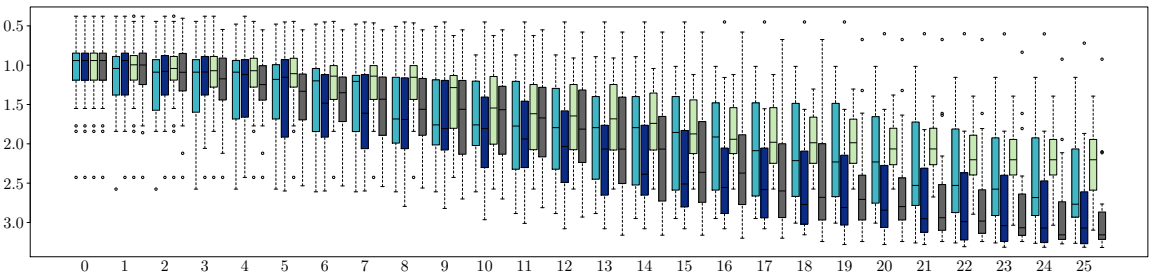


Figure 14. *Hartmann (6d 8D). Boxplot convergence. X-axis: number of iterations; Y-axis, the best discovered value.*

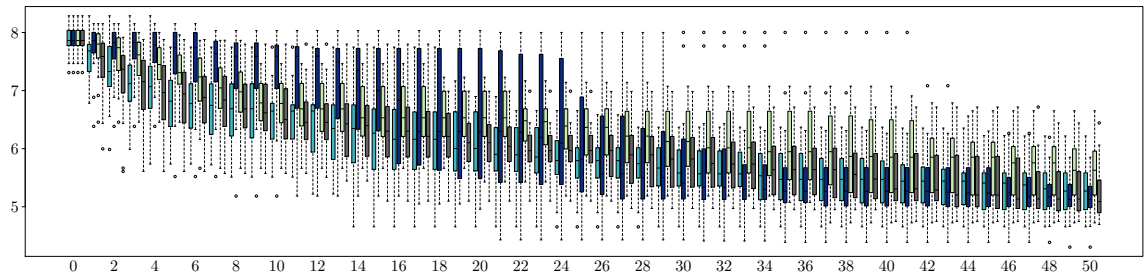


Figure 15. *Rosenbrock (20d 100D). Boxplot convergence. X-axis: number of iterations; Y-axis, the best discovered value.*

Appendix C. Test functions.

In this section, the analytical test functions are defined.

C.1. Ackley function.

For $(x_1, \ldots, x_d) \in [-3, 3]^d$,

$$f(x_1, \ldots, x_d) = -a. \exp\left(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{n}\cos(cx_i)\right) + a + \exp(1), \tag{C.1}$$

where $a = 20$, $b = 0.2$, and $c = 2\pi$.

C.2. Borehole function. For $r_w \in [0.05, 0.15]$, $r \in [100, 50000]$, $T_u \in [63070, 115600]$, $H_u \in [990, 1110]$, $T_l \in [63.1, 116]$, $H_l \in [700, 820]$, $L \in [1120, 1680]$, $K_w \in [9855, 12045]$,

$$f(r_w, r, T_u, H_u, T_l, H_l, L, K_w) = \frac{2\pi T_u (H_u - H_l)}{\ln\left(\frac{r}{r_w}\right)\left(1 + \frac{2LT_u}{\ln\left(\frac{r}{r_w}\right) r_w^2 K_w} + \frac{T_u}{T_l}\right)}. \tag{C.2}$$

C.3. Branin function. For $(x_1, x_2) \in [-5, 10] \times [0, 15]$,

$$f(x_1, x_2) = \left(x_2 - \left(\frac{5.1}{4\pi^2}\right) x_1^2 + \left(\frac{5}{\pi}\right) x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10. \tag{C.3}$$

C.4. Hartmann6 function. For $(x_1, \ldots, x_6) \in [0, 1]^6$,

$$f(x_1, \ldots, x_6) = -\sum_{i=1}^{4} \alpha_i \exp\left(-\sum_{j=0}^{6} A_{ij}(x_j - P_{ij})^2\right), \tag{C.4}$$

where

$$\alpha = (1, 1.2, 3, 3.2)^\top, \qquad A = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix},$$

and P is given by (C.5) .

$$P = 10^{-4} \begin{pmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991 \\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650 \\ 4047 & 8828 & 8732 & 5743 & 1091 & 381 \end{pmatrix}. \tag{C.5}$$

.

C.5. Rosenbrock function. For $(x_1, \ldots, x_d) \in [-2, 2]^d$,

$$f(x_1, \ldots, x_d) = \sum_{i=1}^{d-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]. \tag{C.6}$$

Acknowledgments. Part of the research of the first author 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. We warmly thank Adrien Spagnol and David Prestaux for their help.

REFERENCES

[1] P. Abrahamsen, *A Review of Gaussian Random Fields and Correlation Functions*, Norsk Regnesentral/Norwegian Computing Center, Oslo, Norway, 1997.

[2] F. Bachoc, *Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification*, Comput. Statist. Data Anal., 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*, Ph.D. thesis, Paris 7, 2013.

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

[5] F. Bachoc, A. Lagnoux, and T. M. Nguyen, *Cross-validation estimation of covariance parameters under fixed-domain asymptotics*, J. Multivariate Anal., 160 (2017), pp. 42–67.

[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*, Statist. Comput., 22 (2012), pp. 773–793.

[7] R. Benassi, J. Bect, and E. Vazquez, *Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion*, Learn. Intell. Optim., (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 J., 46 (2008), pp. 2459–2468.

[9] M. Binois, D. Ginsbourger, and O. Roustant, *A warped kernel improving robustness in Bayesian optimization via random embeddings*, Learn. Intell. Optim., (2015), pp. 281–286.

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

[11] P. G. Constantine, E. Dow, and Q. Wang, *Active subspace methods in theory and practice: Applications to kriging surfaces*, SIAM J. Sci. Comput., 36, pp. A1500–A1524.

[12] 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, Omnipress, (2012), pp. 1423–1430.

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

[14] O. Dubrule, *Cross validation of kriging in a unique neighborhood*, Math. Geol., 15 (1983), pp. 687–699.

[15] 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).

[16] J. C Fort, T. Klein, and N. Rachdi, *New sensitivity analysis subordinated to a contrast*, Comm. Statist. Theory Methods, 45 (2016), pp. 4349–4364.

[17] K. Fukumizu and C. Leng, *Gradient-based kernel dimension reduction for regression*, J. Amer. Statist. Assoc.. 4 (2014) pp. 359–370.

[18] F. Hutter, H. H. Hoos, and K. Leyton-Brown, *Sequential model-based optimization for general algorithm configuration*, LION, 5 (2011), pp. 507–523.

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

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

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

[22] C. Linkletter, D. Bingham, N. Hengartner, D. Higdon, and K. Q. Ye, *Variable selection for Gaussian process models in computer experiments*, Technometrics, 48 (2006), pp. 478–490.

[23] X. Liu and S. Guillas, *Dimension reduction for Gaussian process emulation: An application to the influence of bathymetry on tsunami heights*, SIAM/ASA J. Uncertain. Quantif., 5 (2017), pp. 787–812.

[24] J. Mockus, *The Bayesian approach to global optimization*, Syst. Model. Optim., (1982), pp. 473–481.

[25] 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.

[26] 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.
[27] H. Raguet and A. Marrel, *Target and Conditional Sensitivity Analysis with Emphasis on Dependence Measures*, preprint, arXiv:1801.10047, 2018.
[28] P. Ranjan, D. Bingham, and G. Michailidis, *Sequential experiment design for contour estimation from complex computer codes*, Technometrics, 50 (2008).
[29] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Vol. 1, MIT Press, Cambridge, MA, 2006.
[30] O. Roustant, J. Fruth, B. Iooss, and S. Kuhnt, *Crossed-derivative based sensitivity measures for interaction screening*, Math. Comput. Simulation, 105 (2014), pp. 105–118.
[31] O. Roustant, D. Ginsbourger, and Y. Deville, *Dicekriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodelling and optimization*, J. Stat. Softw., 51 (2012), pp. 1–55.
[32] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*, Springer, New York, 2003.
[33] M. J. Sasena, *Flexibility and Efficiency Enhancements for Constrained Global Design Optimization with Kriging Approximations*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 2002.
[34] K. Schittkowski, *NLPQL: A FORTRAN subroutine solving constrained nonlinear programming problems*, Ann. Oper. Res., 5 (1986), pp. 485–500.
[35] 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.
[36] A. Spagnol, R. Le Riche, and S. Da Veiga , *Global sensitivity analysis for optimization with variable selection*, SIAM/ASA J. Uncertain. Quantif., 7 (2019), pp. 417–443.
[37] I. M. Sobol and S. Kucherenko, *Derivative based global sensitivity measures and their link with global sensitivity indices*, Math. Comput. Simul., 79 (2009), pp. 3009–3017.
[38] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York, 1999.
[39] S. Streltsov and P. Vakili, *A non-myopic utility function for statistical global optimization algorithms*, J. Global Optim., 14 (1999), pp. 283–298.
[40] S. Touzani and D. Busby, *Screening method using the derivative-based global sensitivity indices with application to reservoir simulator*, Oil Gas Sci. Technol. Rev. dÍFP Energies nouvelles, 4 (2014), pp. 619–632.
[41] Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. de Feitas, *Bayesian optimization in a billion dimensions via random embeddings*, J. Artificial Intelligence Res., 55 (2016), pp. 361–387.
[42] W. J. Welch, R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris, *Screening, predicting, and computer experiments*, Technometrics, 34 (1992), pp. 15–25.
[43] H. Wendland, *Scattered Data Approximation*, Vol. 17, Cambridge University Press, Cambridge, MA, 2004.
[44] W. Xu and M. L. Stein, *Maximum likelihood estimation for a smooth Gaussian random field model*, SIAM/ASA J. Uncertain. Quantif., 5 (2017), pp. 138–175.