Global Optimization for mixed categorical-continuous variables based on Gaussian process models with a randomized categorical space exploration step

M. Munoz Zuniga[a], D. Sinoquet[a]

[a]IFP Energies Nouvelles, 1-4 Avenue du bois préau 92852 Rueil-Malmaison Cedex - France

ARTICLE HISTORY
Compiled August 16, 2019

ABSTRACT
Real industrial studies often give rise to complex optimization problems involving mixed variables and time consuming simulators. To deal with these difficulties we propose the use of a Gaussian process regression surrogate with a suitable kernel able to capture simultaneously the output correlations with respect to continuous and categorical/discrete inputs without relaxation of the categorical variables. The surrogate is integrated into the Efficient Global Optimization method based on the maximization of the Expected Improvement criterion. This maximization is a Mixed Integer Non-Linear problem which is solved by means of an adequate optimizer: the Mesh Adaptive Direct Search, integrated into the NOMAD library. We introduce a random exploration of the categorical space with a data-based probability distribution and we illustrate the full strategy accuracy on a toy problem. Finally we compare our approach with other optimizers on a benchmark of functions.

KEYWORDS
Derivative Free Optimization; Surrogate Models; EGO; NOMAD; Categorical variables

1. Introduction

The field of research around mixed integer non-linear programming (MINLP) has recently focused on designing algorithms specifically dedicated to finding global solutions. Nevertheless, this latter task is all the more difficult than the model to optimize is expensive to evaluate and thus often relies on the construction of a cheap-to-evaluate surrogate. Hence, iterative surrogate-based approaches have been developed in the literature and can be decomposed, as presented in Muller et al. (2013), as

(1) Build an initial experimental design and evaluate the optimized function
(2) Compute the surrogate model based on the available evaluations
(3) Select the next sample point(s) with respect to some improvement surrogate-based criteria
(4) Update the surrogate model with the new evaluate point(s)
(5) Iterate through (3) and (4) until a predefined stopping criterion has been met.

In this context, the Efficient Global Optimization (EGO) method (Jones et al., 1998)

CONTACT M. Munoz Zuniga. Email: miguel.munoz-zuniga@ifpen.fr

has given noticeable results, in particular, when dealing with expensive blackbox simulators, as in Kanazaki et al. (2015), Hamza and Shalaby (2014), Comola et al. (2016), for instance. EGO is based on a Gaussian process (GP) surrogate (Rasmussen, 2006) and an adaptive strategy where one or more new points are iteratively selected, to be evaluated, with respect to the so called Expected Improvement (EI) criterion. This criterion offers a trade-off between exploitation and exploration by adding points around potential optima and unexplored areas. We propose to extend the EGO strategy to mixed discrete-continuous inputs and will specifically focus on the categorical case where no order is presumed on the discrete variables. To deal with discrete variables we make use of a dedicated kernel proposed in the literature of GP based surrogates (Zhou et al., 2011; Qian et al., 2008). We then integrate this surrogate into the EGO strategy. Once the surrogate is able to deal with categorical variables, the difficulty lies in the optimization of the EI criterion. This latter sub-optimization problem is also a mixed continuous-discrete one, but with an EI function that is relatively cheap to evaluate. Up to this point our strategy is very similar with the recent work of Pelamatti et al. (2018). But instead of relaxing the categorical variables into ordered integer or continuous variables, we use the Mesh Adaptive Direct Search algorithm (Audet and J. E. Dennis, 2006) available in the NOMAD library to solve the mixed categorical-continuous sub-problem. This algorithm initially requires a notion of proximity in order to explore the input space and can therefore straightforwardly deal with continuous and ordered integer variables. In the presence of categorical variables the discrete space exploration is left to the user in the NOMAD implementation. We take advantage of this latter opportunity and develop a random exploration strategy of the discrete space, given by a discrete probability distribution that can evolve as the optimization proceeds.

In comparison with the current literature that we are aware of, the contribution of this article relies on three points. Firstly, we propose to deal, from the beginning to the end, with both continuous and discrete variables in the surrogate based optimization without relaxing categorical variables to integer or continuous ones. Secondly we construct a discrete distribution used to explore the categorical variables randomly within the EI maximization with NOMAD. Lastly we propose a comprehensive presentation and comparison with some well established radial basis functions (RBF) surrogate-based approaches of the literature. Indeed most of the literature on RBF-based optimization does not discuss the importance of kernels and hyper-parameter choices and the key implication/simplification it can involve in the problem formulation.

In section 2, we will introduce the main context and notations, followed in section 3 by a review of the GP surrogate approach: model, kernels and discussion. In section 4, the EGO algorithm framework adapted to the discrete case is presented, with a focus on NOMAD algorithm for the Expected Improvement criterion optimization. In section 5, we introduce a novel discrete probability distribution for the random sampling of the categorical variables within NOMAD. In section **??**, we present a couple of surrogate-based methods for comparison and discuss the similarities and differences with our approach. Finally in section 6 we apply the proposed methodology to a benchmark of test functions and we compare our method with other surrogate optimization methods based on RBF.

2. Context and notations

Our approach can be placed in the framework of "engineering models" optimization as introduced in Swiler et al. (2014). Our purpose is to tackle optimization problems involving computationally expensive simulators with a moderate number of optimization variables. Indeed, in the high-dimensional input cases, the number of data points necessary to capture the function structure increases more rapidly when categorical variables are involved; therefore, dealing with high-dimensional inputs with expensive-to-evaluate simulators is rather difficult. We give a hint on the reason why this difficulty arises in section 3.3 in the context of Gaussian process based optimization.

The function to optimize will be denoted by f and the mixed parameters by $w = (x, z)$ where x represents the continuous variable vector of dimension p and z the integer/categorical one with dimension q. The discrete vector z is supposed to be defined on $I = \prod_{i=1}^{q} I_i$ where I_i is a finite discrete set. If z_i is categorical then no order is pre-supposed and $I_i = \{1, ..., m_i\}$ where m_i is the number of levels of the i-th categorical variable. The integers 1 to m_i are simply representation of the levels. In the integer case, I_i is defined as $I_i = \{a_1, ..., a_{m_i}\}$ where the a_j's are ordered integer numbers such that $j \leq k$ implies $a_j \leq a_k$.

Our aim is to solve the following optimization problem:

$$\begin{aligned} &\underset{w \in \mathbb{R}^p \times I}{\text{minimize}} && f(w) \\ &\text{subject to} && x \in B \subset \mathbb{R}^p, \end{aligned} \tag{1}$$

where B defines bound constraints. In the sequel we suppose that an initial design of experiment (DoE) is given: $w = \{w^1, ..., w^{n_0}\}$, with $w^i \in \mathbb{R}^p \times I$ and the corresponding responses $\mathbf{y} = (y_1, ..., y_{n_0})$ such that $y_i = f(w^i)$. The DoE will be iteratively enriched with respect to an optimization scope as explained in the next sections. The chosen initial DoE, of size n_0, is the concatenation of independent Latin hypercube samplings (LHS) (McKay et al., 1979; Santner et al., 2003), with respect to the continuous variables, each of size $n_{0,i}$. We will further specify the chosen $n_{0,i}$ in the numerical section. This kind of DoE displays good properties for GP surrogates in a sample of test cases presented in Swiler et al. (2014). The results are competitive with sliced-LHS (Qian, 2012) which are theoretically an adequate option for mixed designs but more difficult to obtain. For these reasons, we just settle for the independent LHS DoE.

We now describe the first step of the proposed methodology which consists of the construction of a Gaussian process surrogate of f.

3. Gaussian process surrogate with mixed inputs

3.1. GP Surrogate Model

Gaussian process models are flexible and efficient surrogates of complex computer codes. The popularity of GP stems among other things from the availability of the prediction distribution estimate. For the purpose of optimization, the availability of the prediction distribution estimation opens up the possibility to devise refinement strategies based on some measure of improvement in the regions of interest. We will

discuss this latter point in the next section.
We now present a specific Bayesian approach of the GP regression, as presented in Helbert et al. (2009), that will be used in the numerical tests. The GP regression is based on the hypothesis that the function f is a realization of a Gaussian process $f_{\mathcal{G}}$ defined by a linear regression trend $\mu : \mathbb{R}^p \times I \to \mathbb{R}$, a constant variance σ^2 and a correlation function $K_\theta : (\mathbb{R}^p \times I)^2 \to [-1, 1]$ with hyper-parameters θ. The constant variance and the hyper-parameters are assumed known at this stage. For the sake of simplicity, the trend regression term will only take into account the continuous variables such that

$$\mu(w) = \sum_{i=0}^{l} \beta_i h_i(x),$$

where, for $i = 0, 1, ..., l$, h_i are known functions (that have been chosen by the user from his prior information on the function trend), β_i are random coefficients modeled with some improper prior distribution and l is a non-negative integer. A mixed variable trend could be considered but not treated in this work. The conditional random process

$$f_{\mathcal{G}}^c := f_{\mathcal{G}} \text{ knowing that } f_{\mathcal{G}}(w^1) = y_1, ..., f_{\mathcal{G}}(w^n) = y_n$$

is then known to also be Gaussian. $f_{\mathcal{G}}^c$ has known mean μ_c and correlation function K_c such that

$$\mu_c(w) = h^T(x)\hat{\beta} + r^T(w)R^{-1}(y - H\hat{\beta}) \tag{2}$$

and, in particular, the prediction variance $\sigma_c^2(w) = \sigma^2 K_c(w, w)$ is given by

$$\sigma_c^2(w) = \sigma^2\Big[1 - r^T(w)R^{-1}r(w) + v^T(w)(H^T R^{-1}H)^{-1}v(w)\Big] \tag{3}$$

where the correlation matrix R of the DoE is defined by $R_{ij} = K_\theta(w^i, w^j)$, $i, j = 1, \ldots, n$. The cross-correlation vector between the prediction and the observations is denoted by $r(w) = \left[K_\theta(w, w^i)\right]_{i=1}^n$, while H is the matrix defined by $H_{ij} = h_j(x^i)$, $1 \leq i \leq n$, $1 \leq j \leq d$ and $v(w) = H^T R^{-1} r(w) - h(x)$. The vector $\hat{\beta}$ is explicitly given by

$$\hat{\beta} = (H^T R^{-1} H)^{-1} H^T R^{-1} y. \tag{4}$$

Hence, for known hyper-parameters θ and variance σ^2, the predictor is given by μ_c (2) as in Sacks et al. (1989).
We calibrate the hyper-parameters and the variance by maximizing the log-likelihood. This log-likelihood is the logarithm of the probability of observing the experimental data with our GP model parametrized by θ and σ^2, *i.e.*

$$\mathcal{L}(\theta, \sigma^2) = \ln\Big[\frac{1}{(2\pi\sigma^2)^{n/2}|R|^{1/2}} \exp\big(-\frac{1}{2\sigma^2}(y - H\hat{\beta})^T R^{-1}(y - H\hat{\beta})\big)\Big]$$

The first-order optimality conditions result in analytical formula for σ^2 as a function of θ, namely,

$$\sigma^2(\theta) = \frac{1}{n}(y - H\hat{\beta})^T R^{-1}(y - H\hat{\beta}). \tag{5}$$

This latter expression of $\sigma^2(\theta)$ is plugged in the log-likelihood. The "plugged-in" or "concentrated" log-likelihood then boils down to

$$\mathcal{L}(\theta, \sigma^2(\theta)) = -\frac{1}{2}\Big[n \ln\left(\sigma^2(\theta)\right) + \ln(|R|) + n + n\ln(2\pi)\Big] \tag{6}$$

and has to be maximized with respect to θ. The gradient of the "concentrated" log-likelihood is analytical so that this non-linear optimization problem is generally tackled with a multi-start BFGS algorithm (Roustant et al., 2012).

Finally, the surrogate of the objective function is the conditional mean μ_c given by (2) in which we have plugged-in the hyper-parameter solution of the log-likelihood (6) optimization. In the sequel, we designate this surrogate by $\hat{f}$. For the prediction variance we also use the version with the plugged-in optimal hyper-parameters in (3) denoted by $\hat{\sigma}_c$ in the following.

We will now give more insight on the importance of the correlation kernel choice and the nature of the θ hyper-parameters in the continuous-discrete mixed variables context.

3.2. Correlation kernel for mixed-inputs

In this section we do not intend to present in details the large amount of literature on the correlation kernel choice and its implications in GP. Our aim is to give a sufficient intuition of its importance and to present the kernel we selected.
We first notice in (2) that the prediction at any point w can be written as the sum of a trend term and a linear combination of $r(w) = K_\theta(w, w_i)$. Hence the GP predictor is deeply impacted by the kernel choice. The mixed kernel, defined in $(\mathbb{R}^p \times I)^2$, is typically constructed with the association of two separate kernels: one for the continuous part defined in $(\mathbb{R}^p)^2$ and another one for the categorical part defined in I^2 such that

$$K_\theta(w, w') = K_{\theta_{\text{Cont}}}(x, x') \times K_{\theta_{\text{Cat}}}(z, z')$$

with $w = (x, z) \in \mathbb{R}^p \times I$ and $w' = (x', z') \in \mathbb{R}^p \times I$. The vectors θ_{Cont} and θ_{Cat} are the hyper-parameters associated with their respective kernels. These latter will be defined in next paragraphs.

Correlation kernel for continuous variables. The continuous kernel part is the standard product of 1-D correlation kernels such that

$$K_{\theta_{\text{Cont}}}(x, x') = \prod_{i=1}^{p} K_{\theta_i}(x_i, x_i')$$

with $\theta_{\text{Cont}} = (\theta_1, ..., \theta_p)$. In our context these hyper-parameters are called correlation lengths.

For the continuous part, the degree of smoothness of the stationary GP surrogate is directly linked to the degree of smoothness of the associated positive definite correlation kernel (Rasmussen (2006), section 4.1.1). Moreover, for a 1-D kernel, the correlation length is associated with a notion of regularity, which is defined in Adler (1981) as the mean number of up-crossings of a given level u by the GP (a continuous random process achieves an up-crossing of a given level when its values crosses the level from below). The smaller (higher) the correlation length the higher (smaller) is the mean number of up-crossings. In the numerical applications we selected the anisotropic stationary Matern-5/2 correlation kernel which offers enough flexibility to adequately capture the variability of numerous objective function depending on the choice of the correlation lengths:

$$K_{\theta_{Cont}}(x^i, x^j) = \prod_{k=1}^{p} \left(1 + \frac{\sqrt{5}|x_k^i - x_k^j|}{\theta_k} + \frac{5(x_k^i - x_k^j)^2}{3\theta_k^2}\right) \exp\left(-\frac{\sqrt{5}|x_k^i - x_k^j|}{\theta_k}\right).$$

Note that the correlation lengths, θ_{Cont}, do not depend on the categorical levels which by construction implies that the correlation lengths are similar for all levels. In order to limit the number of hyper-parameters (and therefore the potential number of black-box simulations required to assess them) we assume this independence between the continuous hyper-parameters and the categorical levels. A version with dependence between continuous and categorical parameters is presented in Qian et al. (2008) and used in Han et al. (2009) in a Bayesian context. Thus the continuous/categorical independence hypothesis could be relaxed but at a computational cost which we will avoid for the benchmark tests presented in this paper.

Correlation kernel for categorical variables. In order to treat the categorical variables, different types of correlation kernel can be constructed and recent works in the literature focus on the implications and the relevance of the choice of these kernels. Again, here our goal is only to give an understanding of the nature and importance for the model of the kernel for categorical variables. For more details we invite the reader to see for instance: Pinheiro and Bates (1996), Qian et al. (2008), Pinheiro and Bates (2009), Zhou et al. (2011), Zhang and Notz (2015), Roustant et al. (2018) and Pelamatti et al. (2018).

Hereafter, the i^{th} categorical level is the value taken by z_i, and a global-level, denoted by c, is defined as a set of values assigned to the vector of categorical variables z. The total number of global-levels is given by

$$N_{\text{GL}} = \prod_{i=1}^{q} m_i, \tag{7}$$

with m_i the number of levels of the i^{th} categorical variable. The most flexible approach is then achieved by choosing the correlation kernel, defined on the finite set I^2, as the correlation matrix T whose N_{GL}^2 elements are the correlations between global-levels pairs. This correlation kernel is thus defined as

$$K_{\theta_{Cat}}(c_i, c_j) = T_{c_i, c_j}, \tag{8}$$

for two global-levels c_i and c_j. Since the hyper-parameters represent correlations between global-levels, the matrix T must be unit diagonal, symmetric and positive definite. In total generality, the elements of the matrix T are the $N_{\mathrm{GL}}(N_{\mathrm{GL}}-1)/2$ categorical hyper-parameters. Another approach considers independently the correlation between levels for each categorical variable. This involves q correlation matrices $T^{(k)}$ of size m_k^2 and the kernel is defined as

$$K_{\theta_{Cat}}(c_i, c_j) = \prod_{k=1}^{q} T^{(k)}_{c_i^k, c_j^k}, \tag{9}$$

where c_i and c_j are two global-levels with, respectively, k-component: c_i^k and c_j^k. In this case the number of hyper-parameters is reduced to $\sum_{i=1}^{q} m_i(m_i-1)/2$. In both cases, the kernels (8) and (9) take values in $[-1, 1]$. The latter approach can be justified by an hypothesis on the underlying structure of the GP model; i.e., it is supposed to be a weighted sum of independent GPs with the same correlation function K_{Cont} (one GP per level), see Qian et al. (2008) for further details.

The reduced number of hyper-parameters therefore comes with an underlying hypothesis on the structure of the GP which might not be adequate in some cases, in the sense that even the maximum log-likelihood solution θ could give a poor representation of the objective function if the underlying GP model structure is too far from the real function. In this case the approach (8) with N_{GL} hyper-parameters might be more appropriate if affordable. In the sequel we will use (9) as categorical kernel for computational cost reasons.

To simplify the log-likelihood optimization task which is a difficult positive definite constrained optimization problem with respect to the correlation matrix, we adopt the spherical parametrization of the Cholesky decomposition of each matrix $T^{(k)}$ as in Zhou et al. (2011). This latter trick transforms the previous constrained log-likelihood optimization problem into a box constrained one that can be solved with a BFGS algorithm (Byrd et al., 1995). In the sequel, θ_{Cat} is the categorical hyper-parameter vector of size

$$N_{\theta_{\mathrm{SC}}} = \sum_{i=1}^{q} \frac{m_i(m_i-1)}{2} \tag{10}$$

composed of the spherical coordinates associated with the correlation matrices.

3.3. Discussion on the dimension of the input variables and hyper-parameters

We could further try to reduce the number of correlation parameters to be estimated but this would imply further simplifications of the model. Since our aim is to use this model to approximate black-box functions, on which little information on regularity is known, we prefer to adopt a flexible kernel. In fact our objective is to find the best trade-off between the flexibility of the model (good approximation skills) and its estimation cost (number of experiments required to well determine the hyper-parameters). We will discuss these points in the following.

Kernels for categorical variables often boil down to the product of correlation factors between the different levels. To save computational expenses, one can assume the same correlation between all the levels or at least per group (group to be defined as proposed in Roustant et al. (2018)). In this paper, we decided to not introduce this kind of prior information. Nevertheless, the kernel flexibility comes with a price. Indeed, the selected categorical kernel (9) involves a number of correlations to estimate that can increase rapidly with the number of levels as shown in (10). The more hyper-parameters we introduce, the greater is the flexibility of the surrogate model but in return more data points are required to capture enough information to "feed the flexibility". In this context, we will limit ourselves to applications with a few categorical variables with moderate number of levels. A large range of industrial problems falls in this framework (Swiler et al., 2014). Indeed, prior knowledge of mechanical engineers is often used to limit the number of possible levels of the categorical variables. For instance, in optimal design of a mechanical system, a few types of materials, predefined shapes or structures are selected beforehand.

4. Efficient Global Optimization with mixed inputs

4.1. Global optimization based on the Expected Improvement criterion

Once the hyper-parameters are tuned on the current DoE, the GP model is completely defined and can be used in an adaptive optimization scheme. The Efficient Global optimization strategy (Jones et al., 1998) relies on the posterior distribution of the GP model which enables us to assess the distribution of the following random improvement of f minimization:

$$I(w) = f_{min} - f_{\mathcal{G}}^{c}(w) \tag{11}$$

and the related Expected Improvement (EI)

$$EI(w) = \mathbb{E}\big(\max\big(0, I(w)\big)\big) \tag{12}$$

where $f_{min} = \min(y_1, ..., y_{n_k})$ and n_k the size of the DoE at the k-th iteration of the method. This criterion gives a measure of the expected improvement achievable at a new point w (i.e. expectation to go below the current minimum) based on the known responses at the known available simulated points $w_1, ..., w_{n_k}$ and the GP surrogate estimated distribution. The Gaussian hypothesis enables the implementation of the following closed formula of the EI criterion (Schonlau, 1997)

$$EI(w) = (f_{\min} - \hat{f}(w))\Phi\Big(\frac{f_{\min} - \hat{f}(w)}{\hat{\sigma}_c(w)}\Big) + \hat{\sigma}_c(w)\phi\Big(\frac{f_{\min} - \hat{f}(w)}{\hat{\sigma}_c(w)}\Big) \tag{13}$$

where ϕ is the standard univariate Gaussian distribution and Φ its cumulative distribution function. This criterion offers a built-in exploration-exploitation measure for the optimization strategy.

We can now define the next point to simulate in our optimization scheme as the point that maximizes the EI criterion. After evaluating the new point we can update the GP model with the same hyper-parameters or update the θ's too by maximizing the updated log-likelihood. In the numerical results we will update the GP and the hyper-parameters at each step.

At this stage, we can mention the work of Taddy et al. (2009) and Gramacy and Taddy (2010) where a Tree GP (TGP) strategy is presented. On a tree structure, they propose to construct a GP surrogate for each global-level independently: corresponding to the leaves of the tree. The EI criterion (only depending on the continuous variables) is evaluated on a sampled grid and the locations ranked with respect to their EI values in an iterative manner. This seems to us rather costly and does not take into account any correlation between the levels. Indeed a dense grid, per global-level, has to be evaluated at each step. Nevertheless, this strategy is interesting as it enables to select, from the ranking, not only one point but a batch of points to be evaluated at each iteration. Adding only one point at the time would be equivalent to a simple grid search per global-levels. We believe that doing one optimization per global-level would give better results. This latter "per level" optimization strategy will be evaluated in the numerical section and as mentioned will be considered equivalent or better than a TGP with one point added at the time. Since our actual implementation of EGO does not add batches of points, we will not compare it to the corresponding batch-TGP in this paper.

4.2. NOMAD for the EI criterion optimization

The maximization of the EI criterion is also a mixed continuous-discrete problem but with an objective function relatively cheap to evaluate. Sampling strategies are sometimes preferred to optimize the improvement criterion (Muller et al., 2013) but this seems inefficient since the parameter space to explore can be very large especially with categorical variables.

To tackle this EI-maximization task we used the derivative free Mesh Adaptive Direct Search (MADS) algorithm (Audet and J. E. Dennis, 2006). MADS is a robust optimization method which can be used on a very wide range of non-linear optimization problems. Nevertheless, when the direct search algorithm is not coupled with a surrogate model, the number of simulations required to reach the optimum can sometimes be impractical when dealing with expensive-to-evaluate black-box models. For these reasons we decided not to use MADS on the main optimization problem (1) but it appears as a good tool for the EI maximization sub-problem.

We will now briefly describe the MADS algorithm and introduce our randomized approach for the categorical space exploration in the next section.

The MADS algorithm consists of iteratively evaluating new trial points on an adaptive mesh. Each iteration is divided into two steps: the search and the poll steps. In the search step, a given number of trial mesh points are evaluated around the current best point. If a better point (smaller value of the function) is found the mesh is coarsened, if no improvement is achieved, the poll step is invoked. In the poll step, new points to evaluate are chosen along random positive spanning directions within a limited distance of the current best point. This distance is controlled by a poll size parameter which is greater than or equal to the current mesh size. This latter choice enables a possible dense exploration within the area

centered on the current best point and ensures convergence of the algorithm under mild hypotheses on the objective function (Audet and J. E. Dennis, 2006). Then, (Abramson et al., 2009) added to the MADS algorithm the notion of discrete neighbourhood, introduced in (Audet and Dennis, 2000). Also an additional extended poll step is introduced and triggered when no improvement is found in the two previous steps: a poll step is then performed around each point associated with an objective function value close enough to the current best one. This additional step can, in practice, help the algorithm to escape from some local optima. The described method is implemented in the NOMAD software and offers the option of a user-defined neighbourhood notion for the categorical variables for the poll step. We will use this capacity to define a probability based notion of proximity in our approach.

5. Random sampling of categorical variables within NOMAD for EI optimization

Often in the literature, categorical variables are coded as integers and then treated as real numbers or ordered integers. In this way the notion of neighbourhood is straightforward: often based on some l^p norm. In other cases categorical variables are coded as binaries (Potdar et al., 2017). In the binary space a notion of proximity is based on the number of flips necessary to pass from one global-level to another one (Hamming distance). These approaches can be useful when the user is able to define an imposed order on the categories or if the "binary flipping" proximity model has a real physical meaning. If this kind of information is not available, the ordering or/and the notion of proximity is clearly arbitrary and might skew the exploration in the optimization.

In this context it seems natural to assume that, without any prior information available, no proximity assumption should be introduced. A better approach seems to model the categorical variables as random variables with a discrete probability on the global-levels. So that, at each stage k of the optimization process of (1) (n_k points have been evaluated), the probability for the randomized categorical vector Z_k to take the global-level c_i is $p_{k,i}$ for $i = 1, ..., N_{\mathrm{GL}}$.
Within NOMAD, the aim of the extended poll step is to propose a new categorical position from the current one. Since no proximity assumption between categorical variables is considered, we suppose that the proposed random position, Z_k^{Poll}, should be chosen independently from the current one: $Z_k^{Current}$. But Z_k^{Poll} has to be different from the current position. So that the probability of switching from one current global-level to another in the poll step is given by

$$\mathbb{P}(Z_k^{Poll} = c_i | Z_k^{Current} = c_j, Z_k^{Poll} \neq c_j) = \frac{p_{k,i}}{1 - p_{k,j}} \tag{14}$$

for $i \neq j$ and 0 otherwise. Hence, at each step in the NOMAD algorithm, the user-defined neighbor of the current categorical variable is determined by sampling a global-level poll point according to the probability (14). Multiple global-level-polls points could be sampled similarly.
A first and natural idea to model the discrete probabilities $(p_{k,i})_{i=1,...,N_{GL}}$ is to consider

a uniform distribution so that

$$\mathbb{P}(Z_k^{Poll} = c_i | Z_k^{Current} = c_j, Z_k^{Poll} \neq c_j) = \frac{1}{N_{\text{GL}} - 1}. \tag{15}$$

In this case all global-levels (different from the current one) have the same probability to be chosen in the extended poll step of iteration k. In order to take advantage of the information on the distribution of the objective function evaluations on each global-level, we propose another approximation of $p_{k,i}$ in (14) based on a non-uniform discrete probability on the global-levels. We know, in particular, which global-levels have been more or less explored and which global-level is associated with the smallest objective function values. This information is already integrated within the EI function definition and its maximization should provide us the best next points in unexplored areas or areas where minimal objective function values are expected. Nevertheless, the extended poll step based on (15) does not explicitly use this information. We thus propose to integrate, in the extended poll step, the information learned from simulations of previous optimization iterations. This leads us to explore the global-levels randomly with respect to an "informative" probability distribution defined hereafter.

To give a hint on the relevance of the proposed discrete distribution, we can study the improvement criterion

$$I_{Cat}(z) = \max(0, f_{\min} - M(z)),$$

with $f_{\min} = \min(y_1, ..., y_{n_k})$ and

$$M(z) = \min_{x \in B} f_{\mathcal{G}}^c(x, z)$$

and $z \in \{c_1, ..., c_{N_{GL}}\}$. For each z, $M(z)$ is a real random variable with cumulative distribution Ψ_z [1], mean $\bar{M}(z)$ and standard deviation $\sigma_M(z)$. The expectation of this criterion can be developed as

$$\mathbb{E}(I_{Cat}(z)) = (f_{min} - \bar{M}(z))\Psi_z(f_{min}) + \sigma_M(z)\mathbb{E}\Big[\frac{M(z) - \bar{M}(z)}{\sigma_M(z)} \mathbb{1}_{M(z) \leq f_{min}}\Big], \tag{16}$$

where $\mathbb{1}$ stands for the indicator function. Unfortunately, the quantity (16) can not be further computed since the distribution of $M(z)$ is unknown and too costly to approximate empirically. Nevertheless, we will see that the proposed discrete probability presents some similarities with the terms involved in (16). We also should keep in mind that the prior information we hope to integrate in the EI exploration (13) has to be available at a small computational cost at least smaller than an intensive sampling of EI function, which can become cumbersome when the number of global-levels increases.

We propose that the discrete distribution on the global-levels integrates the density of evaluated points at each global-level. To achieve this goal, we introduce $p_{k,i}^g$ as the probability that the global-level c_i has not been fully explored. The quantity $p_{k,i}^g$ should thus be close to one when the global-level c_i needs more exploration. This quantity is

[1] Ψ_z is not necessary Gaussian

similar to the term $\sigma_M(c_i)$ in (16) which is large when uncertainty for the global-level c_i is large. Furthermore, the discrete distribution should integrate the potential of each global-level to contain the global minimum. For this purpose we introduce $p^m_{k,i}$ as the probability that the global-level c_i has a high potential of containing the minimum. Therefore, $p^m_{k,i}$ will present some similarity with the term $\Psi_z(f_{min})$ involved in (16). Finally, $p_{k,i}$, the probability of global-level c_i, is given by a weighted sum of the two previously introduced probability distributions, that is

$$p_{k,i} = \mathbb{P}(Z_k = c_i) = \alpha_k p^g_{k,i} + (1 - \alpha_k) p^m_{k,i} \qquad \forall i = 1, ..., N_{\mathrm{GL}} \tag{17}$$

where k stands for the optimization iteration and $\alpha_k \in [0, 1]$ is a weight parameter. As k increases, the discrete distribution $p^g_{k,i}$ should converge to the zero discrete distribution since all global-levels will be fully explored. The probabilities $p^m_{k,i}$ should converge to a discrete distribution that is zero for all i except for the one associated with the global-level containing the global minimum.

We define hereafter approximations of $p^g_{k,i}$ and $p^m_{k,i}$. The probability that the global-level c_i is not fully explored is approximated by

$$\hat{p}^g_{k,i} = 1 - \left(\frac{n_{k,i}}{n_k}\right)^l, \tag{18}$$

with $n_{k,i}$ the number of evaluated points in the global-level c_i at iteration k, n_k the total number of evaluations of the objective function at iteration k and $l > 0$ ($l = 1/2$ in the numerical results). Then, the probability $\hat{p}^g_{k,i}$ depends on the proportion of points currently evaluated in the corresponding level. Selecting a decreasing sequence α_k in (17), with respect to k, will force $\alpha_k \hat{p}^g_{k,i}$ to go towards zero, and in a sense to mimic $\sigma_M(c_i)$ in (16).

The probability that the global-level c_i contains the minimum is approximated by

$$\hat{p}^m_{k,i} = \frac{S^R_{k,i}}{\sum_{j=1}^{N_{\mathrm{GL}}} S^R_{k,j}}, \tag{19}$$

with $S^R_{k,i}$ an approximation of $\Psi_z(f_{min})$ in (16) for $z = c_i$. The chosen model for $S^R_{k,i}$ is detailed in appendix A. It is defined as a function of the mean of the function values available for global-level c_i and the associated standard deviation.

Hence, NOMAD extended poll provides a global-level sampled according to (14) with $p_{k,i}$ given by (17) and $p^g_{k,i}$, $p^m_{k,i}$ respectively by (18) and (19).

To summarize, the randomized approach for the extended poll step is based on available evaluated objective function data (initial DoE and from previous iterations), and is a trade-off between focusing on under-explored global-levels and the ones with potential optimality. The selected discrete distribution can be seen as a prior information integrated in the extended poll step. In the uniform case (15), the prior is non-informative, and we expect the NOMAD optimization to converge asymptotically, since all levels will be fully explored with probability one. When the proposed poll step proposition is given by (14) combined with (17), as described, we expect an accelerated recovery of the optimal EI which is illustrated by our numerical tests. The proof of convergence of the NOMAD algorithm with this latter extended poll scheme

seems tricky and dependent on the tuning parameters α_k, $b_{k,i}$ and $\hat{\sigma}_{k,i}$. Nevertheless, we are confident that it will at least converge to a local minimum which seems sufficient for the iterative (with respect to k) EI optimization. Convergence analysis will be the subject of a further work.

6. Numerical results

6.1. Two competitive approaches

In the numerical part we will compare our approach, denoted by Cat-EGO, to two different methods based on a radial basis function (RBF) surrogate: RBFOpt and MISO-CSTV(f). We will refer to the latter one as MISO. For details on the two algorithms we refer the reader to the corresponding papers, Gutmann (2001); Costa and Nannicini (2018) for RBFOpt and Muller (2016) for MISO.

Hereafter, we list the main differences bewteen these two methods and our approach. First of all, MISO performs a local continuous optimization with categorical variables fixed at the values corresponding to the current best point. Concerning RBFOpt, an automatic selection of the kernel within the predefined set $\mathcal{K}$ is performed by using a cross-validation scheme. One last significant difference between the MISO/RBFOpt strategies and ours is the initial DoE size. It is of the order of magnitude of $2(p+q+1)$ for MISO and RBFOpt. This value corresponds to an approximation of the minimal number of data points required for the surrogate to be well fitted. Since our kernel has much more hyper-parameters, an adequate size of the initial DoE has to be much larger in order to sufficiently feed the surrogate model learning stage. Our DoE size is generally between $3 \times p \times N_{\theta_{SC}}$, with $N_{\theta_{SC}}$ given by (10), and $3 \times p \times N_{\mathrm{GL}}$ with N_{GL} given by (7). This latter size gives more robustness to the hyper-parameter optimization but becomes rapidly prohibitive if the problem involves a large number of global-levels N_{GL}.

The differences in DoE size appear as a consequence of the treatment of the categorical variables as continuous by both MISO and RBFOpt. This offers a much simpler function basis approach but also drastically reduces the potential to learn information within categorical global-levels. More precisely, no correlation is estimated between the global-levels. This correlation information helps for the global-level exploration: indeed, if two global-levels are detected as strongly correlated, then only one has to be explored. Another consequence of treating categorical variables as continuous is that a continuous interpolation between the global-levels is done: it assumes that the underlying regularity of this hypothetical continuous approximation of the categorical variables can be captured by the selected radial basis function. Moreover, the radial basis function model implicitly assumes stationarity of the approximated function with respect to all variables, which is a very strong assumption when imposed on the categorical variables (when treated as continuous).

6.2. *A baseline function: exploration improvement for categorical variables*

We consider a two-dimensional toy problem with one categorical variable with 10 levels defined as

$$f(x,z) = \begin{cases} \cos(3.6\pi(x-2)) + x - 1 & \text{if} \quad z = 1, \\ 2\cos(1.1\pi\exp(x)) - \frac{x}{2} + 2 & \text{if} \quad z = 2, \\ \cos(2\pi x) + \frac{1}{2}x & \text{if} \quad z = 3, \\ x(\cos(3.4\pi(x-1)) - \frac{x-1}{2}) & \text{if} \quad z = 4, \\ -\frac{x^2}{2} & \text{if} \quad z = 5, \\ 2\cos(\frac{\pi}{4}\exp(-x^4))^2 - \frac{x}{2} + 1 & \text{if} \quad z = 6, \\ x\cos(3.4\pi x) - \frac{x}{2} + 1 & \text{if} \quad z = 7, \\ x(-\cos(7\frac{\pi}{2}x) - \frac{x}{2}) + 2 & \text{if} \quad z = 8, \\ -\frac{x^5}{2} + 1 & \text{if} \quad z = 9, \\ -\cos(5\frac{\pi}{2}x)^2\sqrt{x} - \frac{\ln(x+0.5)}{2} - 1.3 & \text{if} \quad z = 10. \end{cases}$$

This problem has several local minima with close function values (see Figure 1). Some correlations between the individual one-dimensional functions associated with given levels can be observed: *e.g.* functions at levels 7 and 10 are strongly correlated whereas functions at levels 4 and 7 are anti-correlated.
NOMAD implementation for integer variables is compared on the minimization of this function with the two proposed randomized poll steps for categorical variables: uniform sampling and improved sampling which takes into account the current simulated data distribution. The methods are run 100 times with different initial points (the best of 5 randomly sampled points) in order to measure the robustness of the methods. Figure 2 and 3 illustrate the effect of randomized categorical poll steps on NOMAD results: the percentiles of runs reaching the global minimum area increases from 40% to 60% (Figure 3) and the associated objective functions are closer (see outliers in Figure 2 (right)). In comparison to the uniform strategy, the improved approach shows more robust results with an higher rate of global minima discovery.

In a second analysis, we compare the three NOMAD implementations on a "real" EI function optimization. For this aim, we ran 800 times the EGO algorithm up to the tenth iteration: EGO is run 80 times from random design of 5 experiments and each iteration is performed 10 times to take into account the randomness of the EI minimization methods. We present the results obtained for the next optimization (for each of the 800 runs) of the EI function (the eleventh iteration) with the three NOMAD implementations and two "per level" optimizations. The last two methods consist of continuous optimizations with NOMAD and a multi-start BFGS method with fixed categorical variable. For the mentioned example, ten 1D continuous optimizations are run and the best EI of the 10 runs is considered as the solution. For NOMAD methods applied on the mixed continuous-categorical space, 30 individual initial points are uniformly sampled in the full space and NOMAD is run from the best point (minimal EI). NOMAD-per-level is initialized by the best point of 3 uniformly sampled points for each level and BFGS is run 3 times from 3 uniformly sampled initial points per level.
Figure 4 displays the absolute EI errors of the solutions of the 5 methods compared to the best EI among the solutions of all methods. We observe that the EI errors

obtained by the 5 methods are very close. NOMAD-per-level is more robust with a very small errors for all runs, whereas the BFGS per level approach is less robust. The latter method might converge to local solutions, whereas, NOMAD is a more global method. On the other hand, in Figure 5 we observe that the number of iterations necessary to reach the global EI is much larger for NOMAD-per-level and NOMAD-integer, and slightly larger for the BFGS method compared to the 2 NOMAD methods with the randomized poll steps for categorical variables. The accuracy of the strategies that consider each level independently comes with an higher number of simulation cost. On this example, no significant differences are noticeable between the uniform and the improved strategies. The same analysis has been achieved with different EI-shaped functions associated with different iterations of EGO, and similar conclusions are obtained on these cases.

Since the EI maximization is only a sub-optimization problem of the global cat-EGO strategy, we now present the results of Cat-EGO on the toy example for the EI maximization with the two proposed randomized NOMAD implementations and the two "per-level" approaches with NOMAD and multi-start BFGS. The Cat-EGO results are compared for 2 simulation budgets of 40 and 50. First, in Figures 6 and 7, we observe that the random sampling based strategies (Uniform and Improved sampling) are more robust with regard to reaching the global optima with high accuracy (0.001) for all the stopping criteria. The lack of robustness of multi-start BFGS (for high accuracy) can be explained by the local optimization approach of BFGS on the EI function which makes difficult "exploitation" of Cat-EGO. To improve the exploration in this method, we should add more initial points for the multi-start approach but it will become cumbersome for higher dimensions. On the other hand, after a sufficient number of iterations (as for 40 and 50 simulations), the multi-start BFGS strategy is very efficient in finding the settings of the categorical variables at optimality (see percentages of success with accuracy 0.1: 85% and 90%). NOMAD per level is less efficient in reaching the "right" level (74% and 78%) but is more accurate in the optimization in continuous variable due certainly to its global optimization skills (68% and 75% for accuracy 0.001 comapred to 49% and 63% for multi-start BFGS approach). In comparison to the two random sampling NOMAD strategies, the 2 "per-level" EI maximization approaches are less efficient in reaching the global optimum with high accuracy for the 2 simulation budgets: from 72% to 86% for accuracy 0.001.
The uniform sampling approach gives better results than the improved sampling approach when the number of simulations is large enough: for a budget of 50 simulations, its percentage of success for the two accuracies are larger. The improved sampling results in better percentages of success during the first iterations (see results for the budget of 40 simulations) but seems to bias approximatively 5% to 7% of the 100 runs, leading to a smaller percentage of success in comparison with the uniform sampling after 50 simulations. We remind that the tuning parameter α_k (17) was set to zero in the numerical results and we expect a better behavior from the improved strategy with an adequate adaptive tuning of α_k, increasing the exploration of levels of the categorical variables compared to exploitation when necessary (the uniform sampling performs only exploration).

In the next section the Cat-EGO will always be run with the NOMAD improved random sampling poll step.

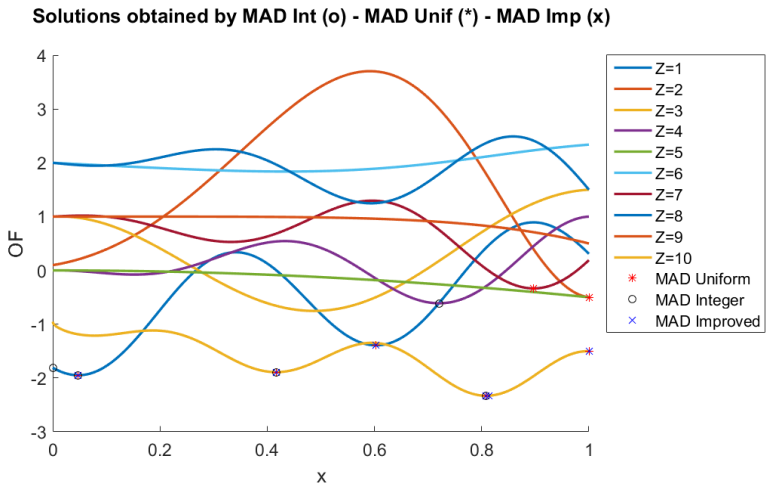
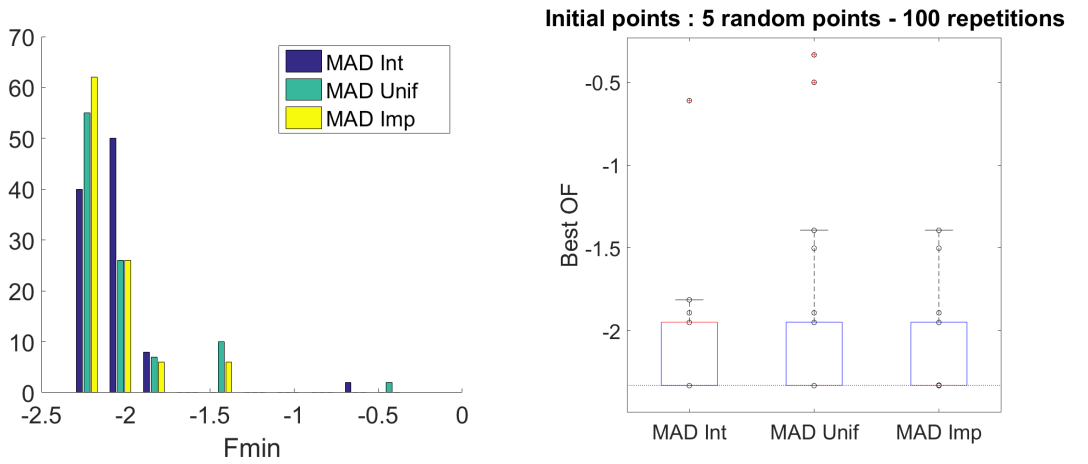


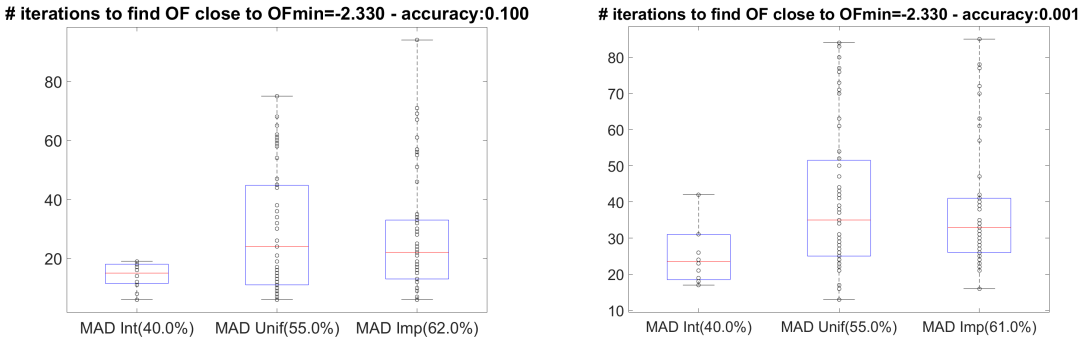
Figure 1. Toy problem with one continuous variable and one categorical variable with 10 levels. The global minimum is located at $x = 0.808$ on level 10 of the categorical variable, the associated objective function value is -2.329. Solutions obtained by 100 runs of NOMAD with 3 different poll step strategies (adapted to integer, uniform sampling, "improved" sampling).



(a) Histograms of minimal objective functions

(b) Box-and-whisker plots of minimal objective functions

Figure 2. Minimal objective functions for 100 runs of NOMAD with 3 different poll step strategies (adapted to integer, uniform sampling, "improved" sampling). (a) Histograms of minimal objective functions. (b) Box-and-whisker plots of minimal objective functions, red lines indicate the medians (middle quartiles), the boxes include 50% of the values, the whiskers cover 99.3% of the values (under Gaussian distribution assumption), the red crosses being considered as outliers. The circles are the 100 minimal objective functions.



(a) Accuracy of 0.1 (percentiles of success)

(b) Accuracy of 0.001 (percentiles of success)

Figure 3. Box-and-whisker plots of the number of simulations necessary to reach the minimal objective functions with two given accuracies (0.1 and 0.001) for 100 runs of NOMAD with 3 different poll step strategies (adapted to integer, uniform sampling, "improved" sampling). The accuracy is on the absolute error of minimal objective function compared to global optimum. See legend of Figure 2 for details on Box-and-whisker plots.

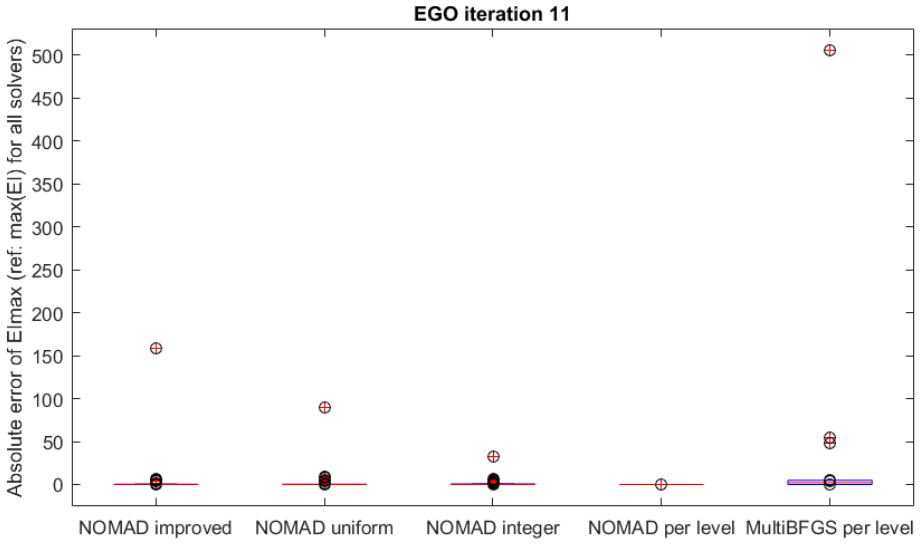


Figure 4. Absolute errors between the maximum EI of each method and the overall maximum EI. The results are presented in a Box-and-whisker plot accounting for 800 repetitions obtained by NOMAD with 3 different poll step strategies (adapted to integer, uniform sampling, "improved" sampling), one NOMAD per level and one multi-start BFGS per level.

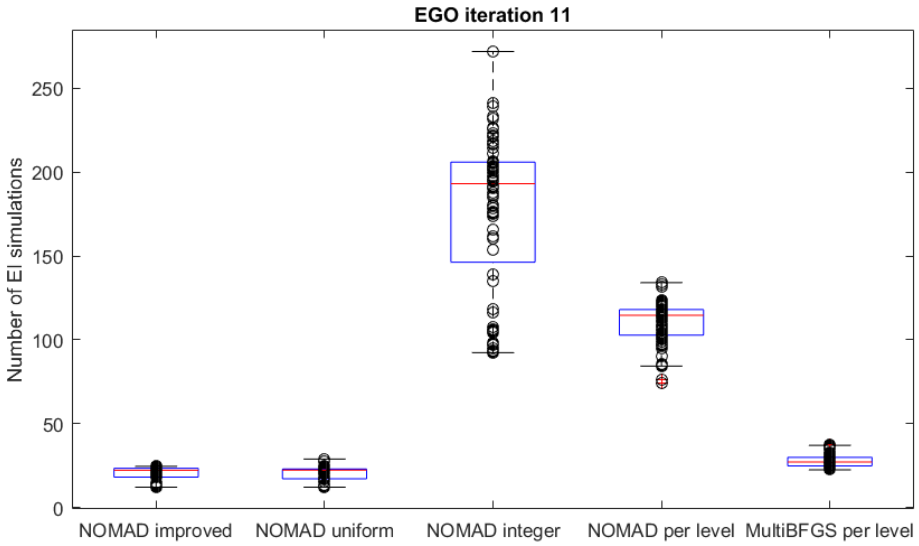
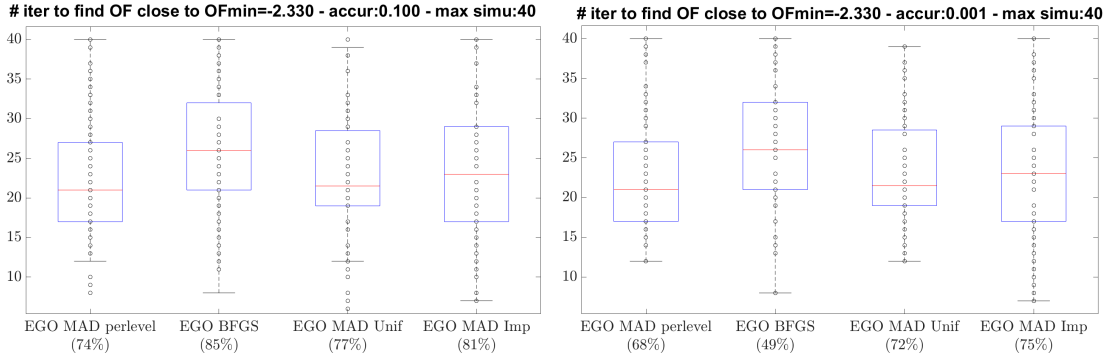


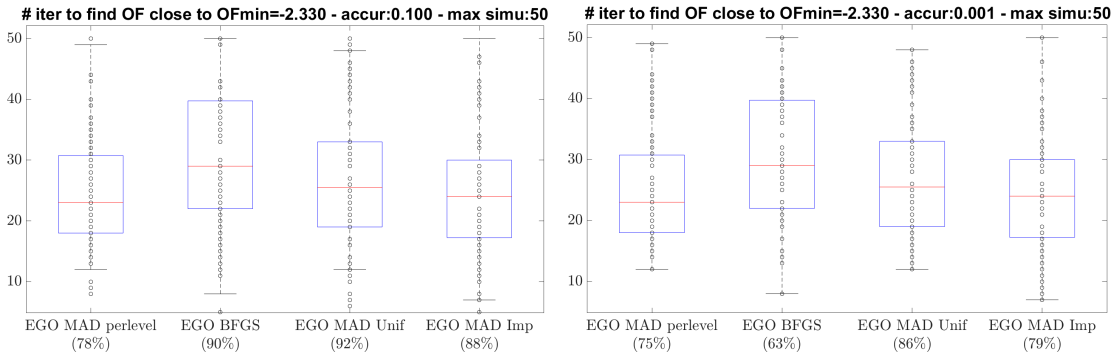
Figure 5. Number of simulations necessary to reach the maximal EI for 800 repetitions obtained by NOMAD with 3 different poll step strategies (adapted to integer, uniform sampling, "improved" sampling), one continuous NOMAD per level and one multi-start BFGS per level.



(a) Accuracy of 0.1 (percentiles of success)

(b) Accuracy of 0.001 (percentiles of success)

Figure 6. Number of simulations necessary to reach the maximal objective function for Cat-EGO with 2 given accuracies for a fixed simulation budgets of 40 simulations. Starting from 100 initial design of experiments of 5 points, 4 EI sub-optimization methods are evaluated: NOMAD with the 2 randomized poll step strategies (uniform sampling and "improved" sampling) and 2 "per level" strategies: Multi-start BFGS and NOMAD.



(a) Accuracy of 0.1 (percentiles of success)

(b) Accuracy of 0.001 (percentiles of success)

Figure 7. Number of simulations necessary to reach the maximal objective function for Cat-EGO with 2 given accuracies for a fixed simulation budgets of 50 simulations. Starting from 100 initial design of experiments of 5 points, 4 EI sub-optimization methods are evaluated: NOMAD with the 2 randomized poll step strategies (uniform sampling and "improved" sampling) and 2 "per level" strategies: Multi-start BFGS and NOMAD.

6.3. Benchmark: global skills of Cat-EGO

Note that in the sequel the NOMAD solver is run from an initial point which gives the highest EI value within the LHS sample used by Cat-EGO method (of size 3 or $5 \times p \times N_{\mathrm{GL}}$ as described further).

We test our method on 15 box-constrained problems from the literature (Hock and Schittkowski (1981), Lukšan and Vlček (2000), and GECCO benchmark COCO (2017)) listed in Table 1. These test problems for continuous optimization are transformed into mixed integer problems. For Hock and Schittkowski (1981) and COCO (2017), following Liuzzi et al. (2012), we define integer variables z_i by restricting some continuous variables (every even index of variable vector) to take a finite number of values, m_i; i.e.,

$$\forall \text{ even } i = 1, 2, ..., q,\ x_i \in \left\{ \underline{x}_i + h\frac{(\bar{x}_i - \underline{x}_i)}{z_i - 1} \right\},\ \text{for } z_i = 0, 1, \ldots, m_i - 1, \qquad (20)$$

with $\underline{x}_i$ and $\bar{x}_i$ the respective lower and upper bounds of the original variable x_i.

Minimax problems

$$\min_{x \in [\underline{x}; \bar{x}]} F(x) := \max_{1 \leq z \leq m} (f_z(x)),$$

defined in Lukšan and Vlček (2000), are transformed into mixed categorical-continuous problems

$$\min_{x \in [\underline{x}; \bar{x}], 1 \leq z \leq m} \tilde{F}(x, z) := \begin{cases} f_1(x) \text{ if } z = 1 \\ f_2(x) \text{ if } z = 2 \\ \ldots \\ f_m(x) \text{ if } z = m \end{cases}$$

Functions from Hock and Schittkowski (1981) are smooth functions whereas the two

Test names	ncont (p)	ncat (q)	nlevels (m)	ref
EDV52	3	1	6	Lukšan and Vlček (2000)
RosenSuzuki	4	1	4	Lukšan and Vlček (2000)
SPIRAL	2	1	2	Lukšan and Vlček (2000)
Wong1	7	1	5	Lukšan and Vlček (2000)
HS2	1	1	4	Hock and Schittkowski (1981)
HS2 rand	1	1	4	Hock and Schittkowski (1981)
HS229log	1	1	4	Hock and Schittkowski (1981)
HS229log rand	1	1	4	Hock and Schittkowski (1981)
HS2	1	1	11	Hock and Schittkowski (1981)
HS229	1	1	11	Hock and Schittkowski (1981)
HS3	1	1	11	Hock and Schittkowski (1981)
bbob 10 3	2	1	4	COCO (2017)
bbob 21 3	2	1	4	COCO (2017)
bbob 22 3	2	1	4	COCO (2017)
bbob 21 5	3	2	4	COCO (2017)

Table 1. Benchmark functions: number of continuous variables, number of categorical variables and number of levels for each categorical variable.

other benchmark functions are more complex with several local minima (see, for instance, Figures 9 and 8).
The procedure to test the Cat-EGO method on these benchmark functions consists of

- an initial design of experiments built from concatenation of N_{GL} Latin hypercube designs, one for each global-level; the size of each design is $k \times p$, with $k = 3$ and 5, depending on the total number of levels,
- a limited budget of simulations, which is a common stopping criterion in practical applications of blackbox optimization for expensive simulators (Moré and Wild, 2009). Here, we chose a budget of 300 simulations.

Figure 10 illustrates the behavior of cat-EGO during the iterations: starting from an initial simulation set of 40 points built from concatenated Latin Hypercube designs of 10 points per level, the maximization of the expected improvement criterion leads to a compromise between space exploration and local minimization. This criterion relies strongly on the learnt model and especially on the learnt correlations between global-levels: Figure 11 displays the evolution of the correlations $T_{c_i c_j}$ between the 4 levels during the iterations. A strong correlation has been detected between levels 1 and 4, between 2 and 3 and between 3 and 4 whereas levels 2 and 5 are anti-correlated. We observe then in Figure 10 that the exploration of the continuous domain is complementary within the correlated levels whereas same zones of continuous space may be explored when levels are not correlated: *e.g.* levels 1 and 3. The adapted structure of our model allows then to save some simulations thanks to the correlation information learnt from the simulated data, as shown on Table 2. In comparison, RBFOpt and MISO methods explore much more at each level leading to a larger number of simulations, as shown in Figures 12 and 13.

The results on the 15 functions are summarized in Table 2. Figures 14 and 15 display the mean relative error of objective functions (compared to the best value found by the 4 optimizers) for the 15 functions of the benchmark. The Cat-EGO method is the most robust, leading to a mean error of 1%, whereas the other optimizers fail to obtain

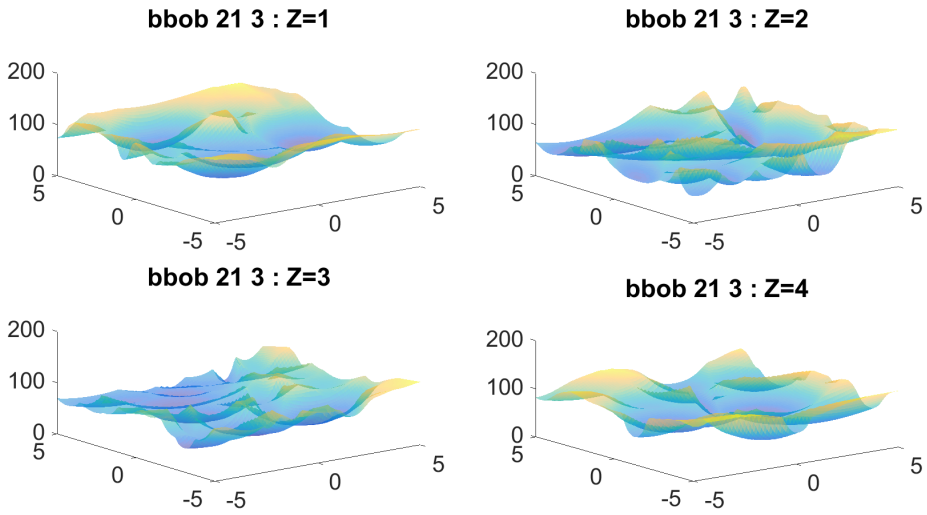


Figure 8. Function bbob21 of benchmark from COCO (2017) in 3 dimensions: 2 continuous variables and 1 categorical variable with 4 levels.

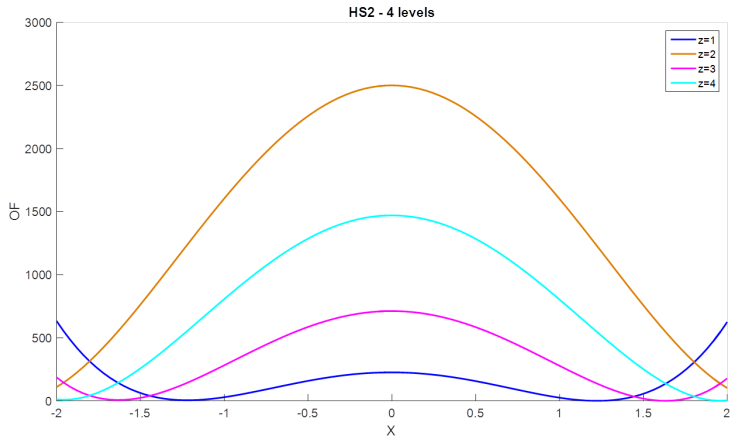


Figure 9. Function HS2 from Hock and Schittkowski (1981) in 2 dimensions: 1 continuous variable and 1 categorical variable with 4 levels.

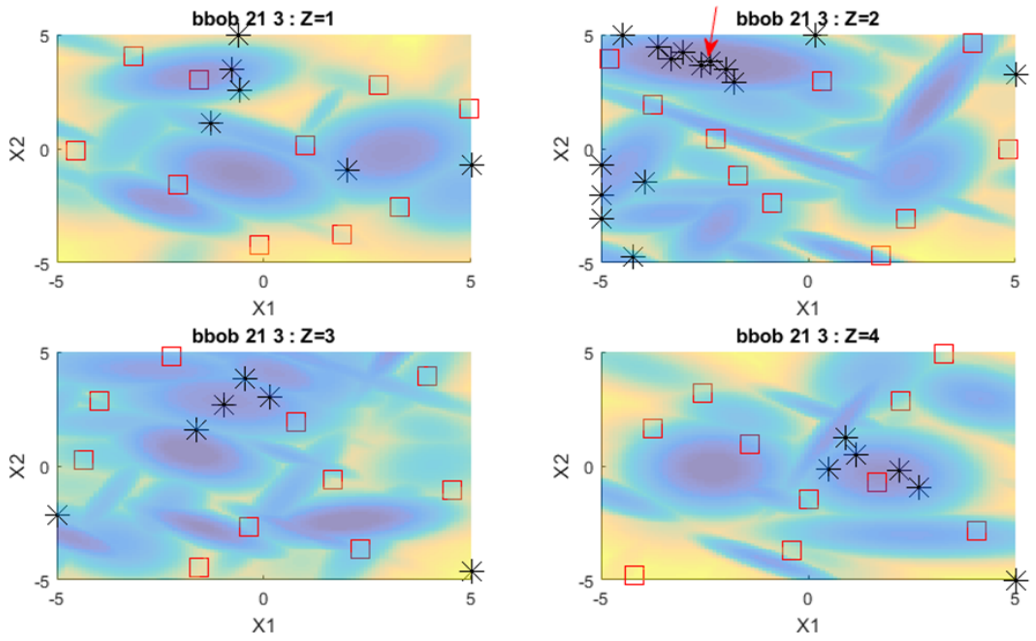


Figure 10. Simulation locations along EGO optimization iterations for test case bbob_21_3. The initial design of experiments is indicated with red squares. Black crosses are the additional simulated points determined by Expected Improvement maximization. The red arrow indicates the global optimum.

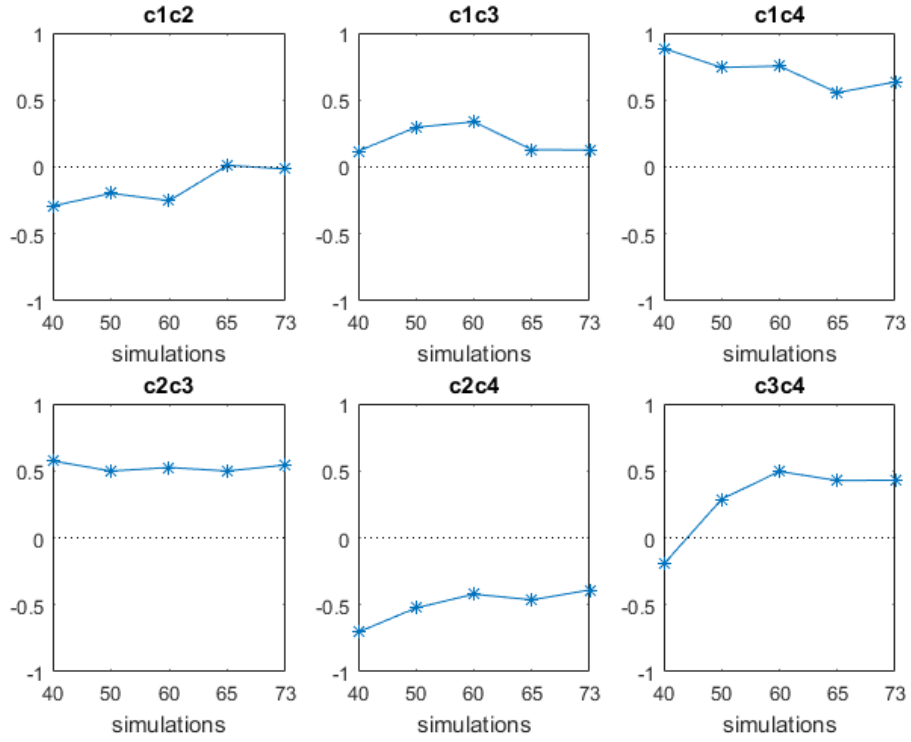


Figure 11. Evolution of EGO model correlations of categorical variables with simulations for test case bbob_21_3.

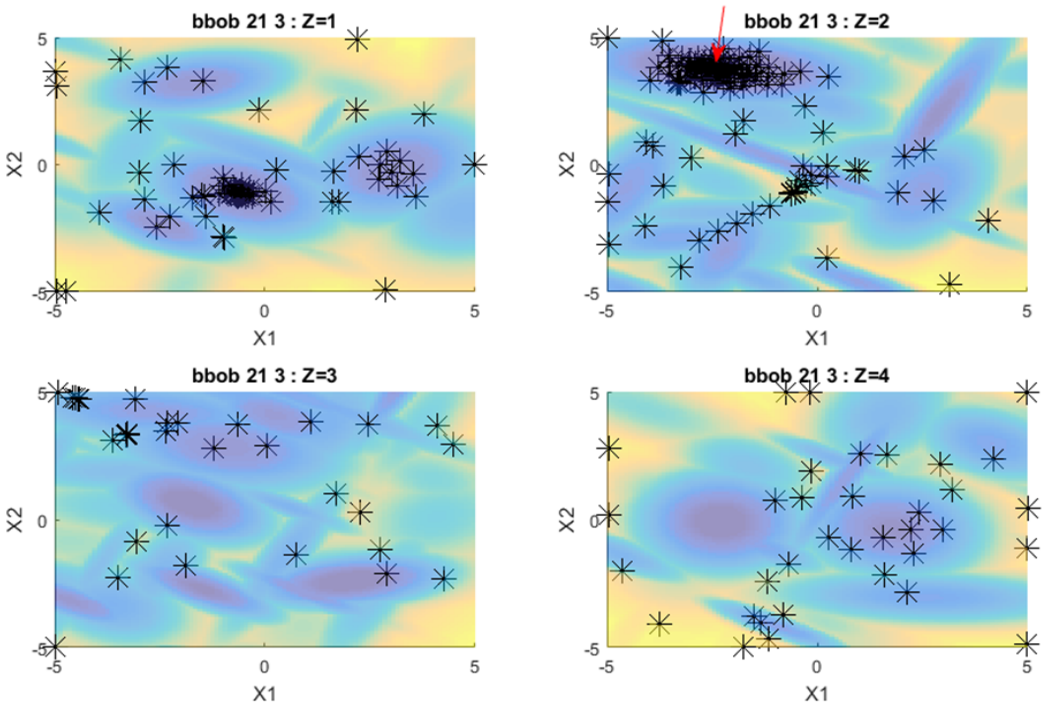


Figure 12. Simulation locations along RBFOpt optimization iterations for test case bbob_21_3. The red arrow indicates the global optimum.

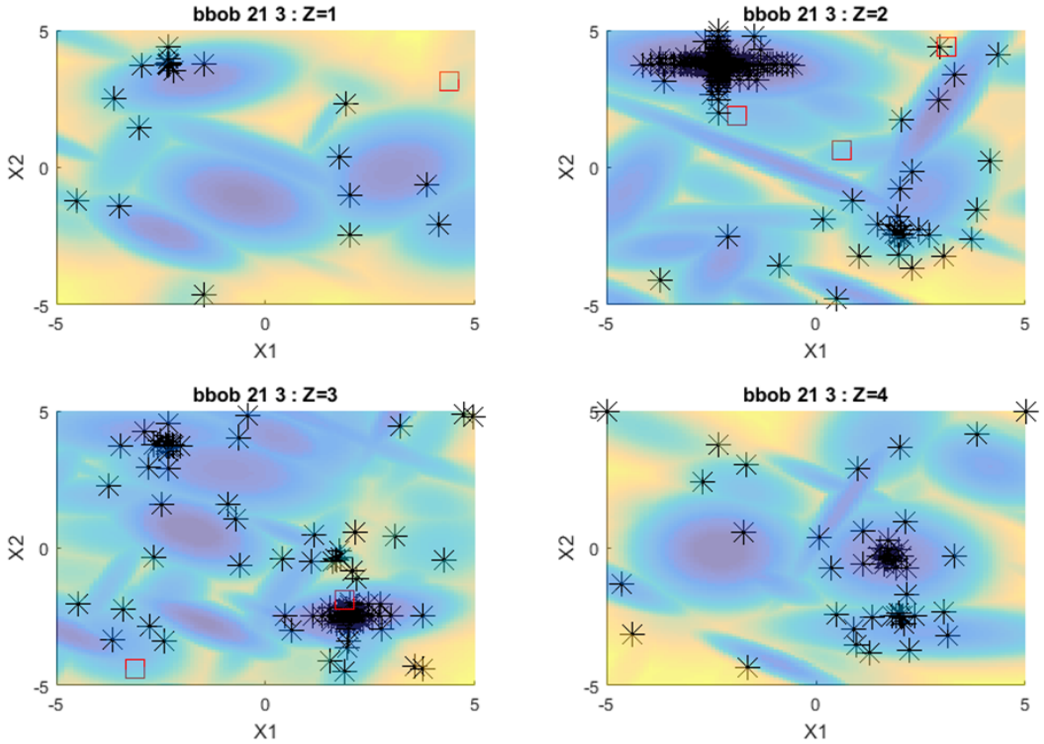


Figure 13. Simulation locations along MISO optimization iterations for test case bbob_21_3.

	Cat-EGO		NOMAD		RBFOpt		MISO	
EDV52	106	-1457.98	300	-99.00	166	-1455.51	300	***-1458.00***
RosenSuzuki	184	***-113.13***	300	-79.88	68	-79.87	300	-79.88
SPIRAL	235	***0.00***	300	***0.00***	297	***0.00***	300	0.03
Wong1	231	-1387.95	300	***-3375.62***	290	-2377.18	300	-3369.54
HS2 (10 levels)	124	***0.05***	177	6.16	104	***0.05***	242	***0.05***
HS2	38	***0.05***	244	0.92	217	0.40	220	***0.05***
HS2 rand	34	***0.05***	257	***0.05***	78	0.40	116	4.94
HS229 (10 levels)	52	***0.00***	211	0.09	215	***0.00***	167	***0.00***
HS229 log	56	-0.73	140	-0.75	86	***-0.77***	124	-0.75
HS229 log rand	35	-0.75	253	***-0.77***	63	***-0.77***	116	***-0.77***
HS3	45	***0.00***	15	***0.00***	296	***0.00***	11	***0.00***
bbob 10 3	151	-47.55	300	***-54.62***	194	***-54.62***	300	***-54.62***
bbob 21 3	68	***40.78***	169	42.75	221	***40.78***	300	***40.78***
bbob 22 3	154	-992.84	99	-995.06	296	***-998.74***	300	-995.06
bbob 21 5	164	***41.01***	300	44.61	177	49.07	300	42.61

Table 2. Comparison of Cat-EGO, NOMAD, RBFOpt and MISO optimizers on benchmark functions. Bold values indicate the runs which reach the best values in less than 300 simulations.

an acceptable accuracy in the allocated simulation budget (300) on several test cases, as shown with the outliers on box-and-whisker plots of Figure 14.

6.4. Behavior of the method for a larger size problem

In this section we apply the 4 methods on a larger size test case with 2 continuous variables and 4 categorical variables with 3 feasible values for each, that leads to 81 feasible categorical combinations. The test case is a modified version of "bbob 21" function from GECCO benchmark COCO (2017) (see Table 1). The 4 categorical variables are build from arbitrary discrete values of the 4 last original variables. The global optimum is 0 at point $x_{1,\ldots,6} = 1$ ($z_{3,\ldots,6} = 2$).

Cat-EGO method is applied with two sizes of initial design of experiments: one of 162 points (a concatenated Latin Hypercube designs of 2 points per level) and one global Latin hypercube design of 70 points (5 times the number of hyper-parameters).

Figure 16 illustrates the results obtained with cat-EGO, MISO, RBFOpt and NOMAD for a maximal budget of 600 simulations. The cat-EGO method obtains a smaller value of the objective function in a smaller number of evaluations.

Figures 17, 18 illustrate the learning ability of cat-EGO models: the evolution of the correlation matrix during the iterations is displayed. Starting from a larger design of experiments (162 points) leads at the first iteration to a better estimate of the correlations: Figure 18 illustrates that the correlation matrix at the first iteration is very similar to the final one (after 600 simulations), whereas the correlations are different between the first and the last iterations of cat-EGO run started from a design of experiments of 70 points. When starting with 70 initial points, even with an imperfect correlation matrix but informative enough, we observe that cat-EGO reaches very quickly the optimal value of the objective function. The difference between the 70 points and 162 points initial DoE results (Figure 16) is probably due to an exploration-exploitation trade-off but this is not a definitive conclusion since this is only one run result.

in Figures 19 and 20 we observe how the function value evolves during the iterations

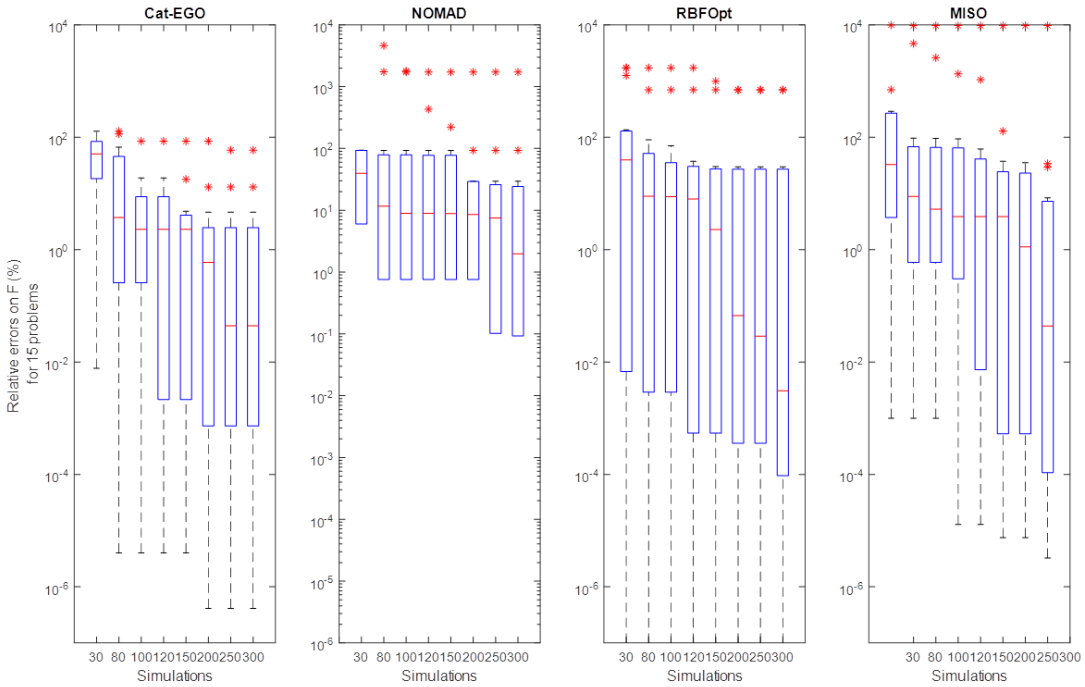


Figure 14. Box-and-whisker plots of mean relative error over the 15 benchmark functions versus simulations for Cat-EGO, NOMAD, RBFOpt and MISO optimizers.

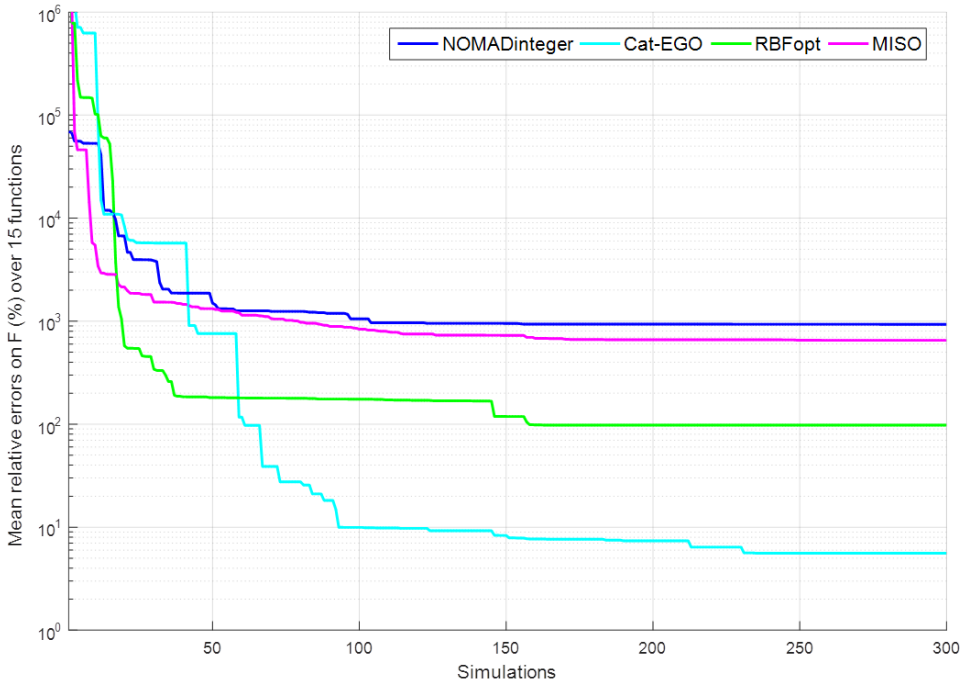


Figure 15. Mean relative error over the 15 benchmark functions versus simulations for Cat-EGO, NOMAD, RBFOpt and MISO optimizers.

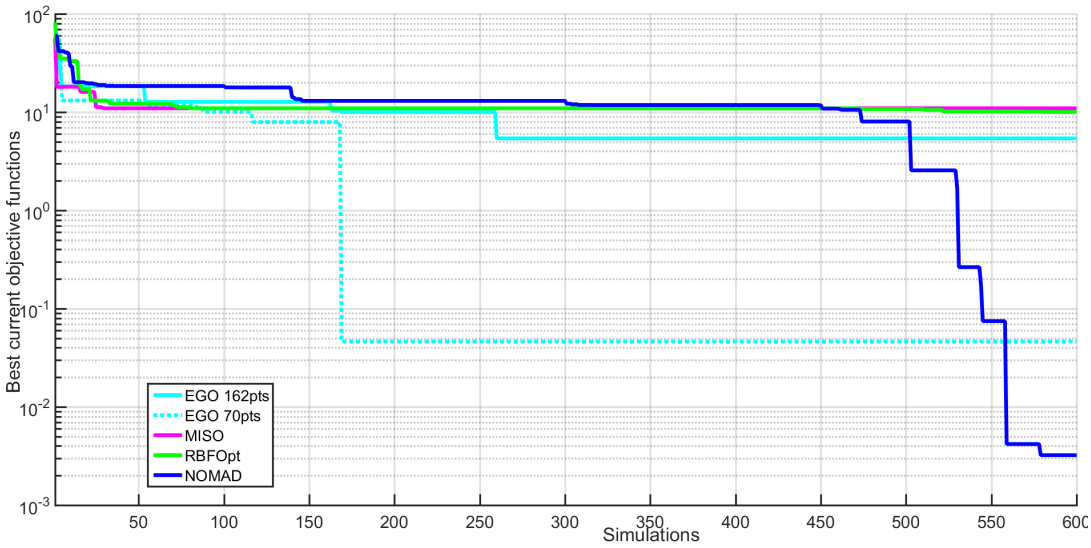


Figure 16. Evolution of the best current objective function value obtained with cat-EGO, MISO, RBFOpt and NOMAD for a maximal budget of 600 simulations.

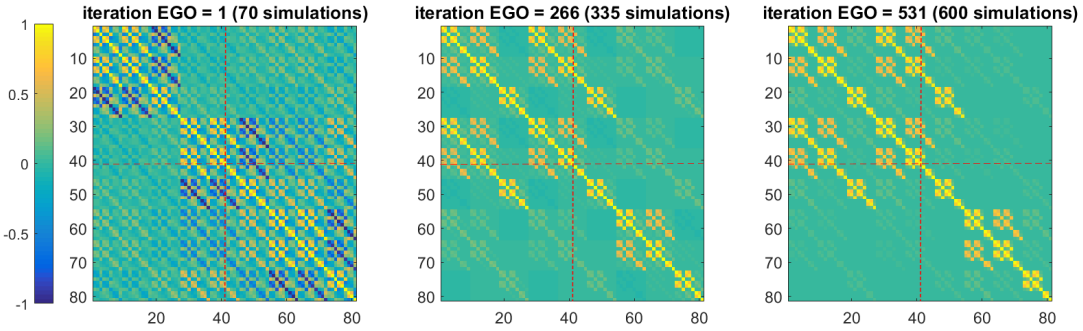


Figure 17. Model correlations learnt for categorical variables - between 81 combinations (the global-levels) taken pairwise - from the available simulation data points for 3 iterations of cat-EGO method started from a design of experiments of 70 points. The red lines indicate the row and the column associated with the global optimum.

with respect to the two different initial DoE sizes and when this value corresponds to points on the level containing the global minimum. The initial 70-points DoE has a very promising point on the optimal level which can explain the fast exploitation of this level. On the other hand, the initial 162-points DoE does not have much direct information on the optimal level but the correlations, being well estimated, push the exploitation of the optimal level quickly after the initial DoE. Indeed, the estimated correlations between the levels can drastically accelerate the exploration of the mixed variable space. Figures 21 and 22 illustrate how, at each iteration, the added point on a given global-level gives information on levels that are correlated with the currently explored one.

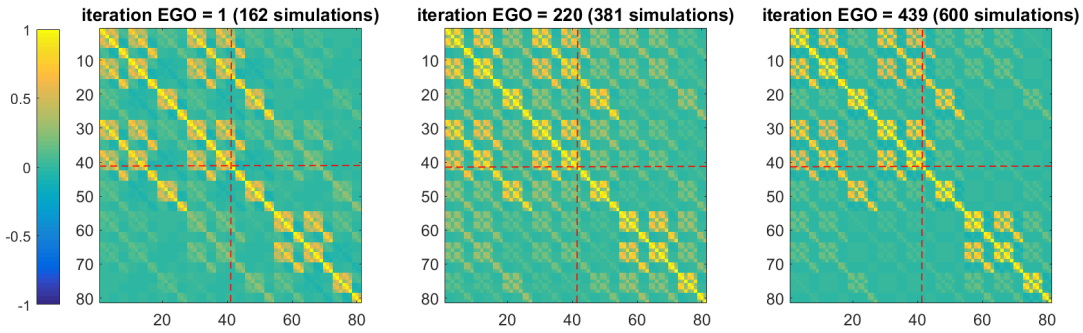


Figure 18. Model correlations learnt for categorical variables - between 81 combinations (the global-levels) taken pairwise - from the available simulation data points for 3 iterations of cat-EGO method started from a design of experiments of 162 points. The red lines indicate the row and the column associated with the global optimum.

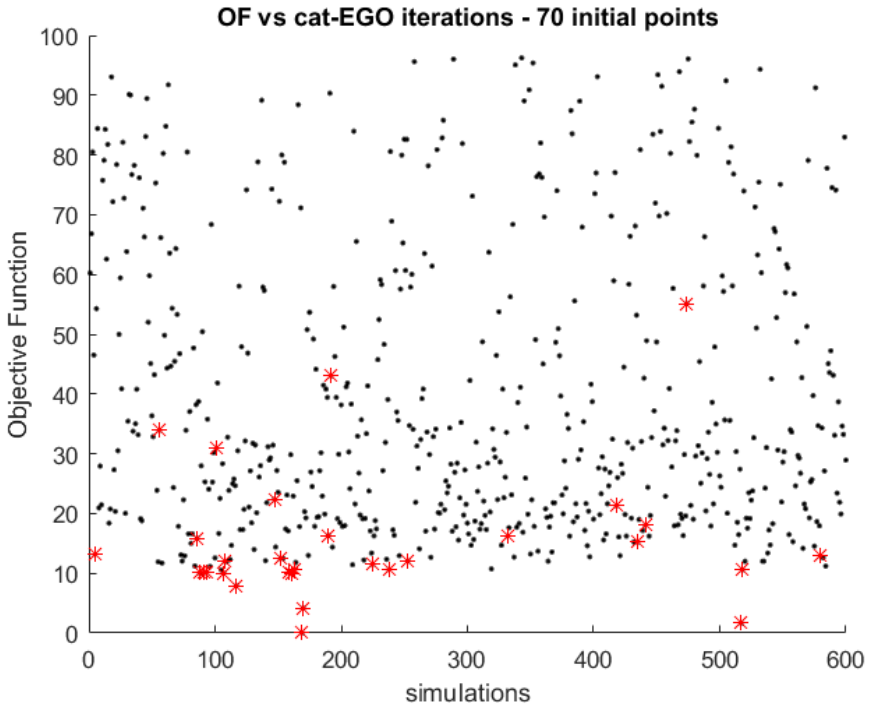


Figure 19. Objective function values during the iterations of Cat-EGO started with an initial design of 70 points. The red crosses highlight the values corresponding to a point on the global-level containing the minimum.

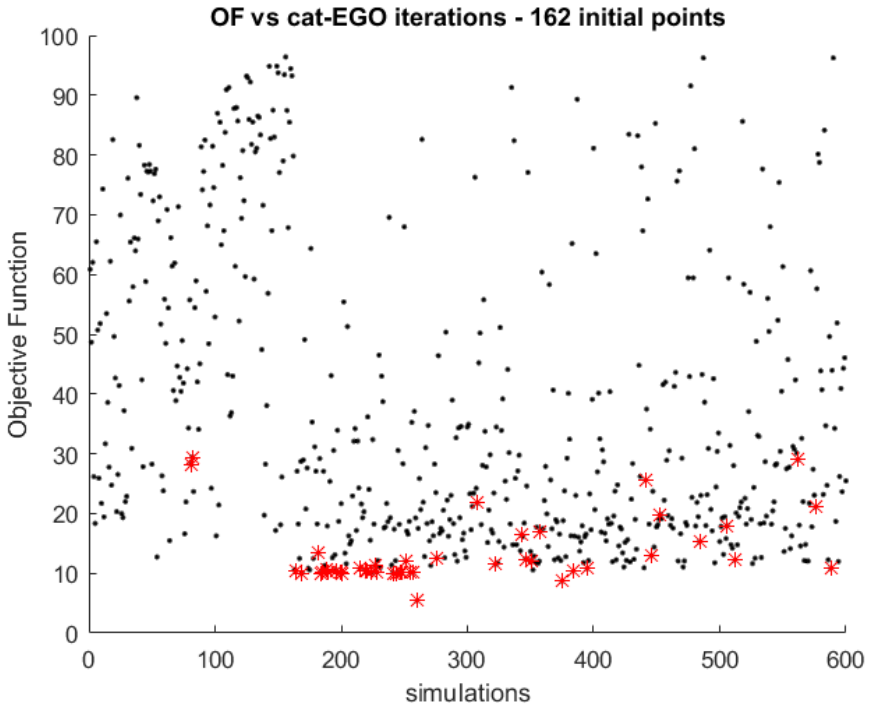


Figure 20. Objective function values during the iterations of Cat-EGO started with an initial design of 162 points. The red crosses highlight the values corresponding to a point on the global-level containing the minimum.

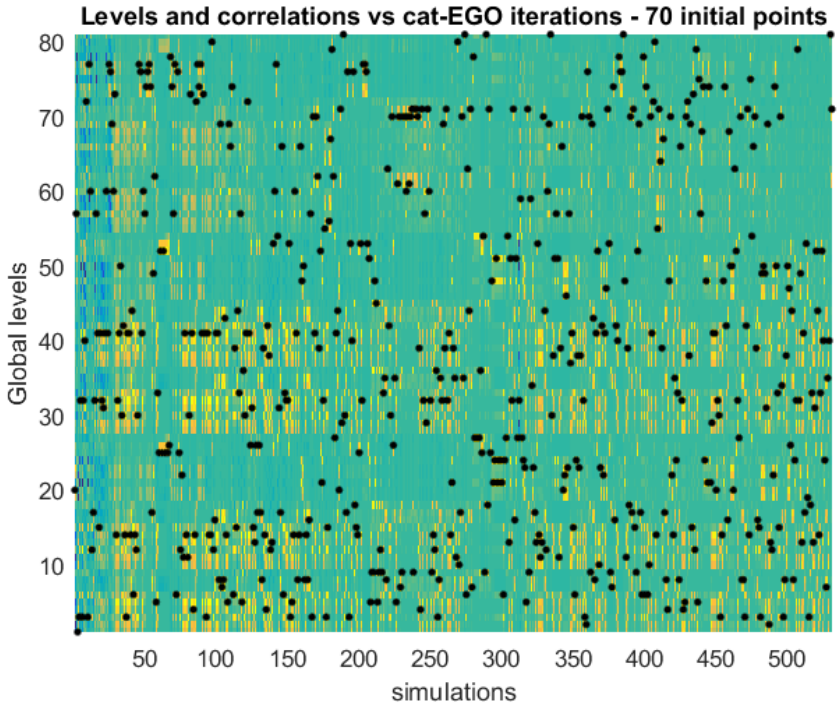


Figure 21. Global-levels visited during the iterations of cat-EGO with a 70 points initial DoE. The black dots indicate which global-level is visited at the corresponding iteration. At each iteration, the colors correspond to the degree of correlation between the visited global-level (where the black dot lies) and the other global-levels.

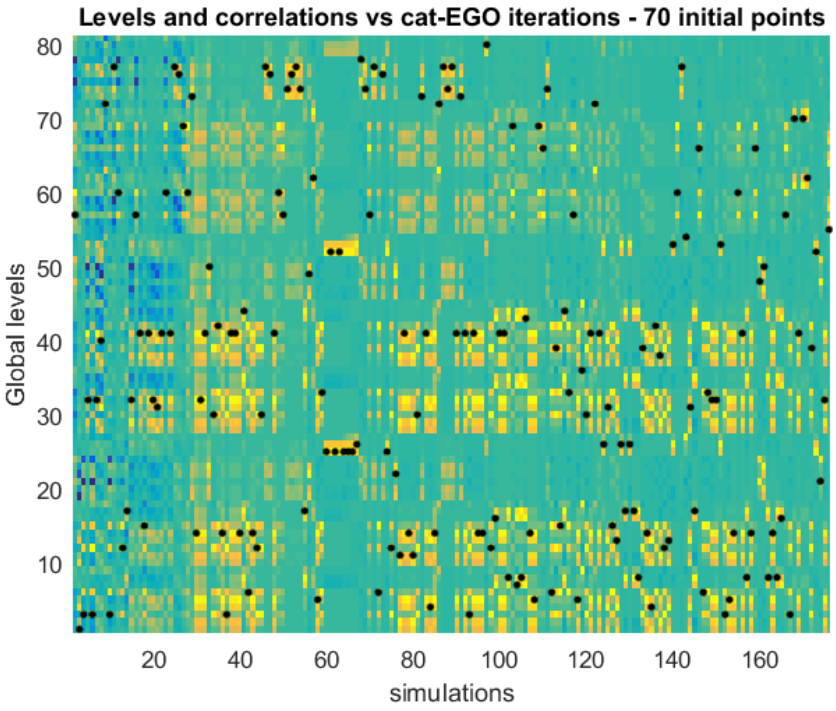


Figure 22. Zoom on 21 for the first simulations.

7. Conclusions and perspectives

We presented a strategy to tackle box-constrained mixed-integer global optimization problems involving expensive black box models and moderate number of input variables. The proposed approach appears to be a robust method for finding global solutions of optimization problems. This success comes from the use of a probabilistic surrogate model (Gaussian processes) flexible enough (thanks to its hyper-parameter structure and their estimation) to capture relevant information on the optimized function with respect to the continuous and the categorical variables. The GP approach also offers a quantification of the uncertainties on the objective function, enabling the construction of a built-in, optimization oriented, improvement criterion: the expected improvement (EI). For the EI sub-maximization task, we introduced a random exploration of the categorical variable space via a data based probability distribution. This latter enabled a faster recovery of an optimal solution of the main problem. On the other hand, the introduced flexibility comes with the price of a larger DoE as initialization and a larger computational time for the choice of the new points in the iterative scheme. Nevertheless, the method is still affordable when dealing with expensive-to-evaluate simulators and gives more robust results. In particular, we demonstrated the efficiency of the Cat-EGO strategy on a serie of test examples. We obtained on these test functions more robust results compared to RBFOpt and MISO. We are convinced that, on other test functions for which the kernel used by RBFOpt and MISO are well adapted, these latters will perform better in terms of minimal number of simulations required to reach the global optimum. This is explained because our method requires and uses a significant part of the simulations to learn the hyper-parameters. If the objective function is simple enough for the RBFOpt or MISO to be relevant then our

hyper-parameters learning stage is not necessary. In total generality, faced with an unknown objective function, we emphasize that it is more reasonable to tackle it with a flexible kernel (with more hyper-parameters), as the one we used.
At this stage we suggest the following future research directions:

- On the one hand, reducing the number of categorical hyper-parameters with a learning strategy such as the one introduced in (Roustant et al., 2018) and, on the other hand, enabling different correlation lengths (one per global-level) in the continuous kernel part as in (Han et al., 2009),
- Further refinement of the discrete probability used in the NOMAD poll step in particular the adaptive calibration of the weight in (17),
- Penalizing the log-likelihood with the norm of the hyper-parameters with an adaptive penalization parameter driving the flexibility of the kernel. At the beginning of the method the size of the DoE is small with respect to the number of hyper-parameters. At this stage, the hyper-parameter optimization problem is not well posed (strongly non-convex). We thus propose to penalize their norm. This indeed inflates the hyper-parameters values, which leads to a very regular approximation of the function (with a small mean number of up-crossings for the continuous variables and high correlations for the categorical ones) and also a better posed hyper-parameter optimization problem.

Appendix A

We define the following rough global-level potential of improvement measure as

$$S_{k,i} = \bar{f}_{k,i} - 2\sigma_{k,i},$$

where $\bar{f}_{k,i}$ is the mean of the objective function values in the global-level c_i and $\sigma_{k,i}$ the corresponding standard deviation. This measure $S_{k,i}$ takes into account the mean value of function evaluations in the global-level but also a measure of the variability of the continuous part within the global-level ($\sigma_{k,i}$). A small value of $S_{k,i}$ corresponds to a global level with high minimization potential. $S_{k,i}$ can be seen as an approximation of $\bar{M}(z)$ in (16) with $z = c_i$. We then calculate the quantity

$$S^R_{k,i} = \frac{1}{1 + \exp(-b_{k,i}\frac{f_{min} - S_{k,i}}{\hat{\sigma}_{k,i}})} \tag{21}$$

which is a sigmoid function parametrized by $b_{k,i}$ evaluated at $(f_{min} - S_{k,i})/\hat{\sigma}_{k,i}$. This quantity approximates $\Psi_z(\frac{f_{min} - \bar{M}(z)}{\sigma_M(z)})$ in (16) for $z = c_i$. The coefficient $b_{k,i}$ should be selected in order to approximate the cumulative distribution Ψ_{c_i} and $\hat{\sigma}_{k,i}$ should be an approximation of $\sigma_M(c_i)$.

As presented, the proposed exploration scheme depends on the parameters α_k, $b_{k,i}$ and $\hat{\sigma}_{k,i}$. For the numerical results we directly set $\alpha_k = 0$ so that only $p^m_{k,i}$, the probability that the global-level c_i has high potential of containing the minimum, is considered in (17). We imposed $b_{k,i} = 1$ for all k and i since estimating the distribution seems too expensive. We also imposed $\hat{\sigma}_{k,i} = 1$ for all k and i, since an accurate approximation of this term is also expensive to compute and by definition $S_{k,i}$ already integrates some insight on the standard deviation of the corresponding global-level. Nevertheless, setting $\hat{\sigma}_{k,i}$ constant implies that the probability $p^m_{k,i}$ will

not converge to 1 for the level containing the minimum and 0 otherwise. This is not an issue here since the limit in k will not be reached within the limited number of iterations to be expected (a few hundreds), and furthermore, the probability will still be large in the global-level containing the objective function minimum. A finer analysis of the tuning of the parameters α_k, $b_{k,i}$ and $\hat{\sigma}_{k,i}$ is postponed to further work.

References

Abramson M, Audet C, Chrissis J, Walston J. 2009. Mesh adaptive direct search algorithms for mixed variable optimization. Optimization Letters. 3(1):35. Available from: https://doi.org/10.1007/s11590-008-0089-2.

Adler RJ. 1981. The geometry of random fields. Wiley, Chichester.

Audet C, Dennis JE. 2000. Pattern search algorithms for mixed variable programming. SIAM Journal on Optimization. 11:573–594.

Audet C, J E Dennis J. 2006. Mesh adaptive direct search algorithms for constrained optimization. SIAM Journal on Optimization. 17(1):188–217. Available from: https://doi.org/10.1137/040603371.

Byrd RH, Lu P, Nocedal J, Zhu C. 1995. A limited memory algorithm for bound constrained optimization. SIAM J Scientific Computing. 16:1190–1208.

COCO. 2017. Black-box optimization benchmarking. Available from: http://coco.gforge.inria.fr/.

Comola F, Janna C, Lovison A, Minini M, Tamburini A, Teatini P. 2016. Efficient global optimization of reservoir geomechanical parameters based on synthetic aperture radar-derived ground displacements. GEOPHYSICS. 81(3):M23–M33. Available from: https://doi.org/10.1190/geo2015-0402.1.

Costa A, Nannicini G. 2018. RBFOpt: an open-source library for black-box optimization with costly function evaluations. Mathematical Programming Computation. 10(4):597–629. Available from: https://doi.org/10.1007/s12532-018-0144-7.

Gramacy RB, Taddy M. 2010. Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an r package for treed gaussian process models. Journal of Statistical Software. 33(6).

Gutmann HM. 2001. A radial basis function method for global optimization. J of Global Optimization. 19(3):201–227. Available from: http://dx.doi.org/10.1023/A:1011255519438.

Hamza K, Shalaby M. 2014. A framework for parallelized efficient global optimization with application to vehicle crashworthiness optimization. Engineering Optimization. 46(9):1200–1221. Available from: https://doi.org/10.1080/0305215X.2013.827672.

Han G, Santner T, Notz W, Bartel D. 2009. Prediction for computer experiments having quantitative and qualitative input variables. Technometrics. 51(3):278–288. Available from: https://doi.org/10.1198/tech.2009.07132.

Helbert C, Dupuy D, Carraro L. 2009. Assessment of uncertainty in computer experiments: From universal kriging to bayesian kriging. Applied Stochastic Models in Business and Industry. 25:99–113.

Hock W, Schittkowski K. 1981. Test examples for nonlinear programming codes. vol. 187. Lecture Notes in Economics and Mathematical Systems, Berlin: Springer.

Jones DR, Schonlau M, Welch W. 1998. Efficient global optimization of expensive black-box functions. Journal of Global Optimization. 13(4):455–492. Available from:

https://doi.org/10.1023/A:1008306431147.
Kanazaki M, Matsuno T, Maeda K, Kawazoe H. 2015. Efficient global optimization applied to wind tunnel evaluation-based optimization for improvement of flow control by plasma actuators. Engineering Optimization. 47(9):1226–1242. Available from: https://doi.org/10.1080/0305215X.2014.958733.
Liuzzi G, Lucidi S, Rinaldi F. 2012. Derivative-free methods for bound constrained mixed-integer optimization. Computational Optimization and Applications. 53(2):505–526. Available from: https://doi.org/10.1007/s10589-011-9405-3.
Lukšan L, Vlček J. 2000. Test problems for nonsmooth unconstrained and linearly constrained optimization. Technical report VT798-00, Institute of Computer Science, Academy of Sciences of the Czech Republic.
McKay MD, Beckman RJ, Conover WJ. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics. 21:239–245.
Moré JJ, Wild SM. 2009. Benchmarking derivative-free optimization algorithms. SIAM J on Optimization. 20(1):172–191. Available from: http://dx.doi.org/10.1137/080724083.
Muller J. 2016. MISO: mixed-integer surrogate optimization framework. Optimization and Engineering. 17(1):177–203. Available from: https://doi.org/10.1007/s11081-015-9281-1.
Muller J, Shoemaker CA, Piché R. 2013. SO-MI: A surrogate model algorithm for computationally expensive nonlinear mixed-integer black-box global optimization problems. Computers & Operations Research. 40(5):1383 – 1400. Available from: http://www.sciencedirect.com/science/article/pii/S0305054812001967.
Pelamatti J, Brevault L, Balesdent M, Talbi EG, Guerin T. 2018. Efficient global optimimzation of constrained mixed variable problems. Journal of Global Optimization. 25. Available from: https://doi.org/10.1007/s10898-018-0715-1.
Pinheiro J, Bates D. 1996. Unconstrained parametrizations for variance-covariance matrices. Statistics and Computing. 6(3):289296.
Pinheiro J, Bates D. 2009. Mixed-effects models in s and s-plus. Statistics and Computing Springer New York.
Potdar K, Pardawala TS, Pai CD. 2017. A comparative study of categorical variable encoding techniques for neural network classifiers. International Journal of Computer Applications. 175(4):7–9. Available from: http://www.ijcaonline.org/archives/volume175/number4/28474-2017915495.
Qian P. 2012. Sliced latin hypercube designs. Journal of the American Statistical Association. 107(497):393–399. Available from: https://doi.org/10.1080/01621459.2011.644132.
Qian P, Wu H, Wu J. 2008. Gaussian process models for computer experiments with qualitative and quantitative factors. Annals of Statistics. Technometrics, Vol. 50, No.3:383–396.
Rasmussen CE. 2006. Gaussian processes for machine learning. MIT Press.
Roustant O, Ginsbourger D, Deville Y. 2012. Dicekriging, diceoptim: Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Journal of Statistical Software. 51(1).
Roustant O, Padonou E, Deville Y, Clément A, Perrin G, Giorla J, Wynn H. 2018. Group kernels for Gaussian process metamodels with categorical inputs. ArXiv e-prints.
Sacks J, Welch W, Mitchell T, Wynn H. 1989. Design and analysis of computer ex-

periments. Statistical Science. 4(4):409 – 435.
Santner T, Williams B, Notz W. 2003. The design and analysis of computer experiments. New York: Springer.
Schonlau M. 1997. Computer experiments and global optimization [dissertation]. University of Waterloo.
Swiler L, Hough P, Qian P, Xu X, Storlie C, Lee H. 2014. Surrogate models for mixed discrete-continuous variables. Cham: Springer International Publishing. p. 181–202.
Taddy M, Lee HKH, Gray GA, Griffin JD. 2009. Bayesian guided pattern search for robust local optimization. Technometrics. 51:389–401.
Zhang Y, Notz W. 2015. Computer experiments with qualitative and quantitative variables: A review and reexamination. Quality Engineering. 27(1):2–13. Available from: https://doi.org/10.1080/08982112.2015.968039.
Zhou Q, Qian PZG, Zhou S. 2011. A simple approach to emulation for computer models with qualitative and quantitative factors. Technometrics. 53(3):266–273. Available from: https://doi.org/10.1198/TECH.2011.10025.