Surrogate models and Costly Optimization with Mixed Categorical Variables

Sujet de thèse

Rodolphe Le Riche, CNRS (Director) ; Sanaa Zannane, Julien Pelamatti, Merlin Keller, EDF (advisors)
In collaboration with Safran Tech, IFPEN, CEA, polytechnique Montréal

Context of the thesis:

The ANR SAMOURAI is an ambitious research project funded in part by the French National Research Agency. This project starts in 2021 for a duration of four years. It brings together several partners from the industrial (CEA, EDF, IFP Energies Nouvelles and SAFRAN) and academic (CentraleSupélec, Ecole Nationale des Mines de Saint-Etienne and Polytechnique Montréal) worlds, around the themes of optimization, uncertainty analysis and reliability, based on numerical simulation and surrogate models. The proposed thesis is part of this project, and as such will benefit from a rich and stimulating scientific environment.

General problem:

Designing complex industrial systems often gives rise to a trade-off between the expected technical performances of the system (e.g.: power of an energy production system, efficiency of a distribution network) and the related costs (e.g.: initial investment costs, maintenance costs, etc.), while complying with the system's constraints (e.g.: compatibility between implemented components, operating ranges, etc.). The general formulation of an optimization problem under constraints G, involving decision variables X and an objective function F is the following:

$$x^* \in arg \min_{x \mid G(x) < 0} F(x) \qquad (1)$$

When the objective function F is expensive to compute (e.g. when it is the result of a complex simulation code) and its analytical expression is not available, the use of classical mixed optimization methods (typically evolutionary methods or separation-evaluation methods) may be prohibitive or unsuitable.

Moreover, it is common for the optimization (or more broadly decision) variables, grouped under the X notation, to be of a mixed nature, continuous for some and discrete, or even categorical, for others. For example, the design of a neural network involves adjusting continuous weights, choosing the number of neurons per hidden layer, not to mention the categorical choices of the form of the activation functions or the general architecture of the network. Because mixed optimization corresponds to a very general formulation of this type of problem, examples can be found in disciplines as diverse as engineering, physical sciences or statistical learning.

The simultaneous presence of continuous and discrete/categorical variables presents a considerable challenge for surrogate model-based optimization algorithms, in terms of both surrogate model construction and refinement strategy, as the two issues are naturally intertwined. Moreover, additional practical challenges arise from the fact that the existing algorithms which allow to solve mixed optimization problems are typically specifically developed for a given discipline or community. For this reason, one of the main objective of this thesis is to propose a unifying problem formulation, which would allow to combine and better exploit the existing approaches.

Objectives of the thesis:

The main ambition of the thesis is to propose the most generic solution possible to the problem of costly mixed optimization, overcoming in particular the following difficulties:

- **Combinatorial explosion and computational cost:** the presence of discrete variables (ordinal or nominal) in the absence of any notion of convexity leads to a number of possible combinations for the discrete variables that increases exponentially with the search space dimension. This is particularly problematic when the associated problem functions (objectives and constraints) are costly to evaluate. The development of Gaussian process surrogate models and strategies for refining numerical designs of experiments adapted to mixed variables seems at present a very promising prospect ;
- **Genericity**: Mixed optimization problems have long been studied by the operations research community, and have led to the creation of a large number of specialized approaches, adapted to various cases. The emergence of mixed surrogate models and adapted refining criteria enables the possibility of developing more generic methods. The demonstration of this genericity requires in particular the possibility of testing the new methods on different industrial applications; this is why four main test cases, coming from different industrial sectors, are considered for this thesis: design of a wind power plant, a turbo-machine, and offshore wind turbine floats, as well as the dimensioning of an electrical network (see details below) ;

Work program:

The following avenues were identified as the most promising to address the above challenges:

1. **Development of covariance kernels for mixed-variable surrogate models**: The recent works of [Hutter, 2011] and [Pelamatti, 2019], among others, open large fields of exploration of new possibilities to build surrogate models with mixed inputs. These may include sets of optimization variables, the dimension of which is itself part of the variables to be optimized! Joint kernels can be defined by combining independent kernel choices for continuous, discrete and dimensional variables. Although many kernels are theoretically applicable in this context, only a few special cases have been explored so far. This is true for the choice of continuous, exponential-quadratic covariance kernels in [Pelamatti, 2019], Matérn in [Muñoz & Sinoquet, 2020]. This is all more true for discrete kernels, for which a large literature already exists and needs to be explored, from latent variable representation in [Zhang et al. 2019; Cuesta Ramirez et al. 2019] to variable-size kernels in [Pelamatti, 2020]. This is why, during the surrogate model construction, choosing the different kernel models and the way of combining them, based on model selection criteria (AIC, BIC, Bayes factor, cross-validation predictive scores, ...), can be a challenge if we consider the combinatorial explosion of the number of available mixed kernel models. This results in an auxiliary mixed optimization problem of the model choice criterion, depending on categorical variables representing the different choices of continuous, categorical, dimensional kernels, and the way of combining them. The Gaussian process approach seems a very promising approach. However, comparison with other classes of surrogate models used in mixed problems, such as radial function bases (RBF) [Müller et al., 2013; Costa and Nannicini, 2018] or random forests [Hutter et al., 2011], will help validate their operational interest.

2. **Choice of a global optimization method for mixed variables**: Bayesian optimization, like the EGO algorithm [Jones et al., 1998] seems a natural choice, especially if the surrogate model built at the previous step is a Gaussian process. It depends on an initial design of experiment, as well as on an refining criterion, typically the expected improvement (EI) and variants based on the two first moments of the surrogate model. The prospect of adding several points at a time to the current design of experiment, based on parallel computing architectures, could to some extent limit the explosion of the number of configurations. In addition, other approaches popular in the optimization community could be considered, such as the adaptive quadratic confidence zone method in [Conn et al. 2009]. An extension of this method to the design of a turbomachine [Tran et al., 2020], in which mixed-binary variables are involved, naturally provides a reference approach. Finally, hybrid approaches, as proposed in [Régis, 2015] for mixed variables, could allow combining the flexibility of Gaussian processes with the good observed behavior of large

confidence region approaches. An example of such approaches is provided in [Audet et al., 2019], with an implementation in the NOMAD software [Le Digabel, 2011] which could serve as a reference for the development of new algorithms.

3. **Development of enrichment strategies for Bayesian optimization**: Optimizing the acquisition criterion once again amounts to solving an optimization problem with the same variables as the original problem, but this time with a criterion characterized by a negligible computational cost. The very large literature on this subject provides many solutions that should be compared on different applications: hybrid evolutionary algorithms [Cauwet et al., 2019], adaptation of the MADS algorithm [Audet and Dennis, 2006] to mixed problems in [Munoz Zuniga and Sinoquet, 2020], evolutionary budget allocation strategies [Pelamatti, 2020], or Bayesian-inspired strategies in [Bergstra et al., 2011], to name but a few. Comparing all of these approaches across multiple cases will help to better understand the benefits of each.

Use-cases:

This thesisaims to propose a generic solution to mixed-variable problems arising from non-linear learning and optimization for expensive models. For this reason a wide variety of application cases are considered in order to stimulate the development of the most generic approaches.

More specifically, two application cases will concern design problems, one for a turbomachine and the other for a wind turbine float. They will each involve discrete choices of components (through different choices of technologies, number of components installed, etc.).

Two other cases will deal with larger scale problems, through the design of a wind farm and the management of an electricity network involving discrete choices related to sizing (number and type of wind turbines, production units), which influence the number of continuous optimization variables (positioning of the wind turbines).

In addition to these four main application cases, others may be added, depending on the available time, in order to demonstrate the methodological developments of the thesis, such as for example the optimization of the hyperparameters of an artificial intelligence system (neural network) and the design of an innovative process for flaw detection on an industrial component using eddy currents.

Advisors:

- The thesis will be directed by Rodolphe Le Riche, director of research at CNRS (LIMOS laboratory at Mines St-Etienne), in collaboration with EDF R&D, represented by Sanaa Zannane, Julien Pelamatti and Merlin Keller, research engineers in the PRISME department at EDF Lab Chatou;
- This thesis is part of the ANR SAMOURAI (Simulation Analytics, Meta-model-based solutions for Optimization, Uncertainty and Reliability Analysis), which starts in March 2021. This ANR comprises 3 theses (including the present one) and 2 post-docs on 4 different research axis: large-scale surrogate models, sequential methods in reliability, mixed surrogate models and optimization dealing with hidden constraints. Moreover, support from the IFPEN (D. Sinoquet) and the GERAD laboratory (S. Le Digabel) is planned for actions related to the NOMAD platform;
- The four application cases considered for the thesis are provided by the different partners: EDF (wind farm design), Safran Tech (turbomachinery design), IFPEN (float design for offshore wind turbines), CEA (management of electricity production and distribution networks). This opens the opportunity to interact with leading industrial players ;

Desired profile:

- Probability/statistics/operational research student, with a master degree or equivalent
- Good mastery of the foundations of statistical learning and optimization
- Ease in scientific programming, with a good knowledge of R, Python

Practical arrangements:

- Start : third quarter of 2021
- Location : EDF Labs in Chatou (6, quai Watier 78401 Chatou) or Ecole de Mines (158 Cours Fauriel, 42023 Saint-Étienne), depending on the candidate preferences. Long work sessions will be arranged in the other lab.

Contacts:

leriche@emse.fr
sanaa.zannane@edf.fr
julien.pelamatti@edf.fr
merlin.keller@edf.fr

References:

[Hutter et al., 2011] doi:10.1007/978-3-642-25566-3_40
[Pelamatti et al., 2019] doi:10.1007/s10898-018-0715-1
[Müller et al., 2013] doi:10.1016/j.cor.2012.08.022
[Costa and Nannicini, 2018] doi:10.1007/s12532-018-0144-7
[Jones et al., 1998] doi:10.1023/A:1008306431147
[Conn et al., 2009] doi:10.1137/1.9780898718768
[Tran et al, 2020] doi:10.1007/978-3-030-38364-0
[Regis, 2015] doi:10.1080/0305215X.2015.1082350
[Audet et al., 2019] doi:10.1137/18M1175872
[Le Digabel, 2011] doi:10.1145/1916461.1916468
[Cauwet et al., 2019] hal-02170283
[Audet and Dennis, 2006] doi:10.1137/040603371
[Munoz Zuniga and Sinoquet, 2020] doi:10.1080/03155986.2020.1730677
[Pelamatti et al., 2020] arxiv:2003.03300
[Bergstra et al., 2011] hal-00642998