Split, doubt and design in high dimension with general surrogates.

M. BEN SALEM
Mines Saint-Étienne - Ansys, Inc

Supervisor(s): O. Roustant (Mines Saint-Étienne), F. Gamboa (Institut de mathématiques de Toulouse) and L. Tomaso (Ansys, Inc)

Ph.D. expected duration: May. 2015 - April. 2018

Adress: 11 Avenue Albert Einstein, 69100 Villeurbanne, France

Email: malek.ben-salem@emse.fr

Abstract: In sequential design problems, the main goal is generally to estimate a feature of an expensive function (optimum, probability of failure, ...). Several methods have been proposed to achieve this goal. Some of these techniques are based on surrogate models, specifically the probabilistic ones. The main advantage of a probabilistic approach is that it provides a measure of uncertainty associated with the surrogate model in the whole space. Nevertheless, its usage is limited to functions depending on a moderate number of variables. In this work, we consider two limitations of these methods: the restriction to the Gaussian case and the curse of dimensionality.

First, there are several available and useful surrogate models. Nevertheless, they are not all naturally embeddable in some stochastic frame. Hence, they do not all provide a prediction distribution. To overcome this drawback, several empirical design techniques have been discussed in the literature. They are generally based on resampling methods such as bootstrap, jackknife, or cross-validation. However, most of them lead to clustered sets of points. We propose a universal method to define a measure of uncertainty suitable for any surrogate model: deterministic, probabilistic and ensembles. It relies on Cross-Validation sub-models predictions. This empirical distribution may be computed in much more general frames than the Gaussian one. For this reason, we call it Universal Prediction distribution [1]. It allows the definition of many sampling criteria. We investigate particularly adaptive sampling techniques for global refinement and an extension of the so-called Efficient Global Optimization (EGO) [2] for all types of surrogate models.

Definition 1 *The Universal Prediction distribution (UP distribution) is the weighted empirical distribution*

$$\mu_{(n,\mathbf{x})}(dy) = \sum_{i=1}^{n} w_{i,n}(\mathbf{x}) \delta_{\hat{f}_{n,-i}(\mathbf{x})}(dy) \tag{1}$$

where $w_{i,n}(\mathbf{x}) = \frac{1-e^{-\frac{d(\mathbf{x},\mathbf{x_i})^2}{\rho^2}}}{\sum\limits_{j=1}^{n}\left(1-e^{-\frac{d(\mathbf{x},\mathbf{x_j})^2}{\rho^2}}\right)}$, $\mathbf{x_i}$ *for* $i = 1, \ldots, n$ *the set of design points and* $\hat{f}_{n,-i}$ *the Leave-One-Out sub-surrogate leaving the* i^{th} *design point.*

Second, several real life problems involve a large number of variables. Here, we consider the high-dimensional case with a moderate number of influential variables. A classical approach is two-stage. First, sensitivity analysis is performed to reduce the dimension of the input variables. Second, a feature is estimated by considering only the selected influential variables. This approach can be computationally expensive and may lack flexibility since dimension reduction is done once and for all. In this work, we propose a so called Split-and-Doubt algorithm that performs sequentially both dimension reduction and feature oriented sampling. The 'split' step identifies influential

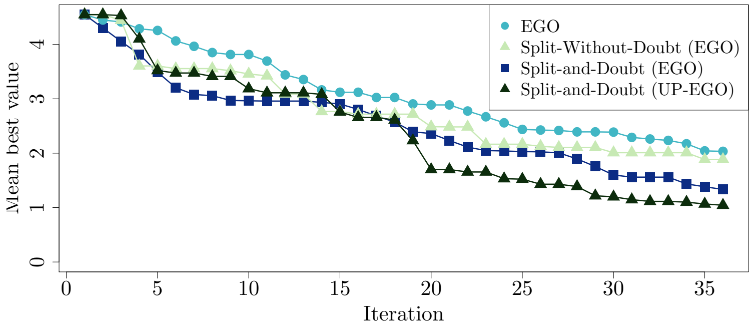

Figure 1: Mean over 10 different initial designs of the best (minimum) discovered value.

variables. This selection relies on new theoretical results on Gaussian process regression. We prove that large correlation lengths of covariance functions correspond to inactive variables. Then, in the 'doubt' step, a doubt function is used to update the subset of influential variables. A summary of the method is given in Algorithm 1.

```
Data: Function of interest $f$
Result: Estimate of a feature of $f$ (optimum, probability of failure, ...)
Create an initial design and define the stopping conditions;
while Stopping conditions are not met do
    1. Split the variables into the major subspace and the minor subspace;
    2. Design in the major variables subspace: Optimize a goal-oriented criterion;
    3. Doubt the variable splitting: Find challenger correlation lengths within a likelihood region;
    4. Design in the minor variables subspace: Optimize a contrast criterion between the initial
       estimation and the challenger correlation lengths;
    5. Evaluate the new point and update the current design.
end
```

Algorithm 1: Split-and-Doubt summary

Numerical tests show the efficiency of the Split-and-Doubt algorithm especially when used with the UP distribution (Definition 1). For instance, we compared the performances of 4 optimization algorithms: EGO [2], the Split-and-Doubt (EGO) that uses EGO in step 2, the Split-and-Doubt (UP-EGO) that uses UP-EGO in step 2, and the Split-without-Doubt (EGO) that replaces step 3 and 4 of Split-and-Doubt (EGO) by a uniform random sampling of the minor variables. We display in Figure 1 the results for the 6-dimensional Ackley function embedded in dimension 20.

References

[1] M. Ben Salem, O. Roustant, F. Gamboa, and L. Tomaso. Universal prediction distribution for surrogate models. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):1086–1109, 2017.

[2] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. *J. Global Optim.*, 13(4):455–492, 1998.

Short biography – After an engineering diploma from the French engineering school ISIMA (Clermont-Ferrand) and a Research Master Degree from the Blaise Pascal university (Clermont-Ferrand), Malek Ben Salem started a PhD thesis at Mines Saint-Étienne. He is funded by a CIFRE grant from the ANSYS company, subsidized by the French National Association for Research and Technology (ANRT, CIFRE grant number 2014/1349) and he works on surrogate model aggregation, prediction uncertainty quantification and sequential design.