Universal distribution for surrogate models uncertainty assessment

M. Ben Salem
École des Mines de St-Etienne - ANSYS, Inc.

Supervisor(s): Olivier Roustant (École des Mines de St-Etienne), Fabrice Gamboa (Institut de Mathématiques de Toulouse) and Lionel Tomaso (ANSYS, Inc.)

Ph.D. expected duration: 2015-2018

Adress: 11 Avenue Albert Einstein, 69100 Villeurbanne, France

Email: malek.ben-salem@emse.fr

Abstract: The use of surrogate models is very convenient in engineering. Their main purpose is to replace an expensive-to-evaluate function s by a simple response surface also called surrogate model. They are based on a given training set of n observations $z_j = (x_j, y_j)$ where $1 \leq j \leq n$ and $y_j = s(x_j)$. Using this training set, we build a surrogate model $\hat{s}_n$ that mimics the behavior of s. The accuracy of the surrogate model relies, *inter alia*, on the relevance of the training set. The aim of surrogate modeling is to estimate some features of the function s using $\hat{s}$. Of course one is looking for the best trade-off between a good accuracy of the feature estimation and the number of calls of s. Consequently, the design of experiments (DOE), that is the sampling of $(x_j)_{1 \leq j \leq n}$, is a crucial step and an active research field.

There are two ways to sample: either drawing the training set $(x_j)_{1 \leq j \leq n}$ at once or building it sequentially. Some sequential techniques are based on probabilistic surrogate models. The main advantage of probabilistic approaches is that they provide a measure of uncertainty associated with the surrogate model in the whole space. This uncertainty is an efficient tool to construct strategies for various problems such as prediction enhancement, optimization or inversion. For instance, the Expected Improvement (EI) [5] or the Expected Feasibility (EF) [2] are computed within the Gaussian frame.

Nevertheless, several methods are generally not naturally embeddable in some stochastic frame. Hence, they do not provide any prediction error distribution. To overcome this drawback, several empirical design techniques have been discussed in the literature. These techniques are generally based on resampling methods such as bootstrap, jackknife, or cross-validation. Among these techniques, we can cite [3, 4, 6]. However, most of these resampling method-based design techniques lead to clustered sets of points [1, 4].

We propose a universal method to define a measure of uncertainty suitable for any surrogate model either deterministic or probabilistic. It relies on Cross-Validation (CV) sub-models ($(\hat{s}_{n,-i})$, $i = 1, \ldots, n$) predictions. This empirical distribution may be computed in much more general frames than the Gaussian one. So that it is called the Universal Prediction distribution (UP distribution). It allows the definition of many sampling criteria for global refinement, optimization, and inversion problems.

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

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

Where the weights are defined below

$$w_{i,n}(\mathbf{x}) = \frac{1 - e^{-\frac{d((\mathbf{x},\mathbf{x_i}))^2}{\rho^2}}}{\sum_{j=1}^{n} \left(1 - e^{-\frac{d(\mathbf{x},\mathbf{x_j})^2}{\rho^2}}\right)} \tag{2}$$

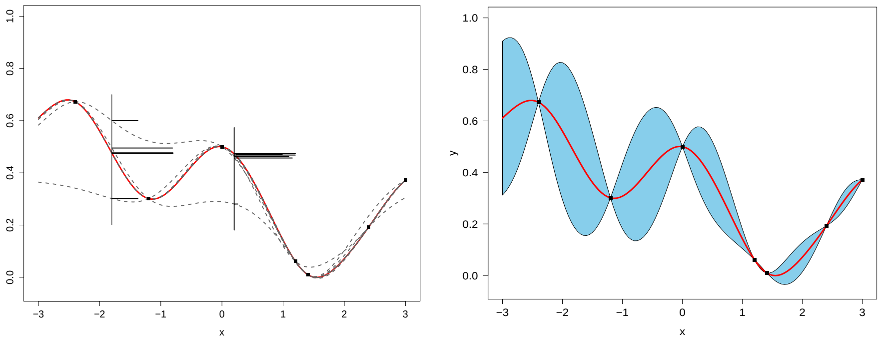

Figure 1: Illustration of the UP distribution. Left: Dashed lines: CV sub-models predictions, solid red line: master model prediction, horizontal bars: local UP distribution at $x_a = -1.8$ and $x_b = 0.2$, black squares: design points. Right: Uncertainty quantification based on the UP distribution. Red solid line: master model prediction $\hat{s}_n(\mathbf{x})$, blue area: region delimited by $\hat{s}_n(\mathbf{x}) \pm 3\hat{\sigma}_n(\mathbf{x})$.

References

[1] V. Aute, K. Saleh, O. Abdelaziz, S. Azarm, and R. Radermacher. Cross-validation based single response adaptive design of experiments for kriging metamodeling of deterministic computer simulations. *Struct. Multidiscipl. Optim.*, 48(3):581–605, 2013.

[2] B. J. Bichon, M. S. Eldred, L.P. Swiler, S. Mahadevan, and J. M. McFarland. Efficient global reliability analysis for nonlinear implicit performance functions. *AIAA journal*, 46(10):2459–2468, 2008.

[3] S. Gazut, J.-M. Martinez, G. Dreyfus, and Y. Oussar. Towards the optimal design of numerical experiments. *IEEE Trans. Neural. Netw.*, 19(5):874–882, May 2008.

[4] R. Jin, W. Chen, and A. Sudjianto. On sequential sampling for global metamodeling in engineering design. In *Proc. ASME Des. Autom. Conf.*, pages 539–548. American Society of Mechanical Engineers, 2002.

[5] 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.

[6] J.P.C. Kleijnen, W. van Beers, and I. van Nieuwenhuyse. Expected improvement in efficient global optimization through bootstrapped kriging. *J. Glob. Optim.*, 54(1):59–73, 2012.

Short biography – After an engineering diploma from the french engineering school ISIMA (Clermont-Ferrand) and a Research Masters Degree from the Blaise Pascal university (Clermont-Ferrand), Malek Ben Salem started a PhD thesis at École des Mines Saint-Étienne. He is funded by ANSYS, Inc and ANRT under a CIFRE contract and he works on surrogate models aggregations, prediction uncertainty quantification and sequential design.