Surrogate Based Optimization Using Kriging Based Approximation

Abdellatif Lghali
VU BA paper
vanlghali@gmail.com

1. Introduction

In many engineering design problems, processes are so complex to the point to make experiments either time consuming or computationally expensive. As a challenging issue in most optimization procedures, the objective function usually has a large number of local minima, a large number of local maxima and is represented as a 'Black box' function (a black box function is a function that without been explicitly described and given a list of a finite number of points in the input space, corresponding outputs can be obtained). Consequently, existing techniques applied to nonlinear optimization problems often require a large number of function evaluations. Therefore, solution methodologies need to be custom developed for computationally expensive analysis and unknown function properties (i.e., black-box function).

Recent developments in the field of optimization have lead to an increasing interest in approximation models or surrogate models as alternatives that may help to solve those problems. Surrogate models or response surfaces have been shown to be effective approaches in constructing fast models that mimic the behavior of computationally expensive and complex systems. Within the optimization area, surrogate-models both speed optimization processes of problems with non-smooth or noisy responses and provide insight into the relationship between output responses, y, and input design, x.

Numerous methods exist to generate surrogate models, each with their relative merits. Examples include: rational functions, Kriging models, Artificial Neural Networks (ANN), spline, and Support Vector Machines (SVM). Some of the different real-world applications of surrogate models are documented in [11]. For instance, least square support vector machine (LSSVM) has been used in the decision-making processes associated with supply chain management [12]. Recently, a radial basis neural network (RBNN) has been employed in optimization of Wire-Wrapped Fuel Assembly [13]. So far most frequently used models in engineering designs and aerospace design problems have been Gaussian processes [4, 14, 15]. For example, kriging or Gaussian process has been utilized in finding the optimal values of reorder point and the maximum inventory level in an inventory optimization problem [14].

The aim of this paper is to review the literature concerning surrogate models, highlighting concepts, techniques and other methods used within surrogate-based optimization approaches. In particular, this review is centered around Kriging-an approximation technique made popular due to its ability to model complex landscape and provide error estimate.

This paper has been divided into two parts. The first part deals with the key stages of the surrogate-based optimization processes putting more emphasis on the surrogate model building process.

In the second part, I will summarize the Efficient Global Optimization (EGO) approach proposed by Jones et al. (1980). This approach is based on a Gaussian process based method of Kriging (first proposed by, and named after, Danie Krige (1951)). Finally, the problems associated with failed design evaluations and noisy data will be discussed, before drawing conclusions in the final section.

2. Overview of Surrogate-based optimization

Alexander I.J. Forrester [4], has suggested that most optimization-based search using surrogate models requires similar steps as those indicated in the following figure:

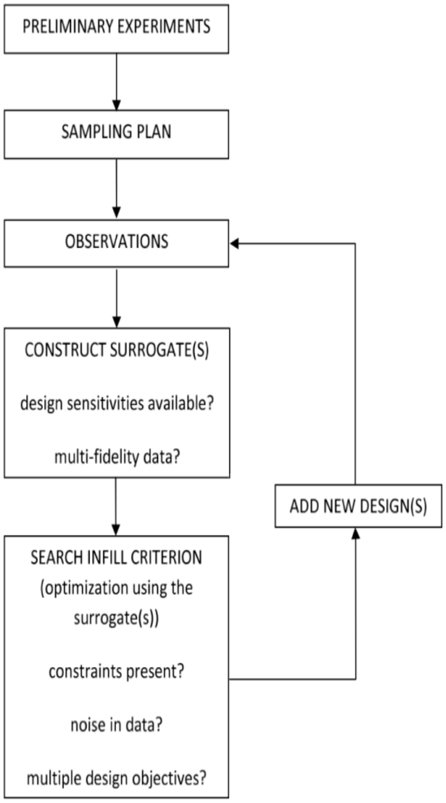


Figure 1: a surrogate-based optimization framework. Forrester [4].

Figure 1 illustrates the essential steps involved in most surrogate-based optimization approaches. There are several options for each of these steps, as well as several advantages and disadvantages of each option. In the following sections I will briefly describe the first two steps and then will focus on the model choice and model fitting steps of the framework.

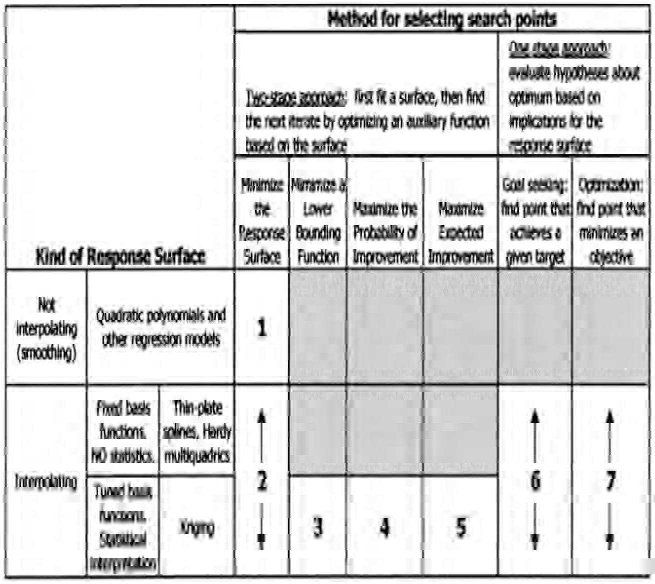
2.1 Sampling Plan

The first step from the above framework suggests the design and analysis of some preliminary experiments. These preliminary experiments allow choosing the variables that will be taken forward to optimize the design space. That is, choosing a subset of design variables from all existing variables. However, there are some challenges faced in selecting the right number of variables that provide enough information to achieve good performance in prediction. Some bottlenecks in this stage might be the curse of dimensionality of some design spaces and the relative sparsity of the observations.

After identifying the design space, we must then choose which designs we wish to evaluate in order to construct the surrogate model. That is designing the plan in such a way that the resulting data will contain a representative sample of the parameters of interest. The most frequently used techniques in this stage of the process are Latin Hypercube Design, Full-factorial Design, Orthogonal Arrays and Box-Behnken Design. Each of these methods has its own advantages and disadvantages depending on the characteristics of the design problem. However, Forrester [5] favors the Latin Hypercube Design technique, which efficiently samples large design spaces and provides a sample of points whose projections onto each variable axis are uniform. In fact Latin Hypercube Design algorithm divides uniformly the design space for each factor and combines randomly these levels to specify n points defining the final design matrix. I will not cover all methods here and the reader may wish to consult Forrester for an in-depth description, including Matlab code.

2.2. Modeling Approaches

In order to predict accurately the function landscape we are trying to imitate, an initial surrogate must be constructed using a limited number of intelligently chosen data points. Again, there are a variety of options to accomplish this task and several benefits and limitations of each option. Donald R. Jones [1], has proposed a more specific discussion of response models, where he contrasts seven existing modeling approaches using response surfaces for global optimization. In the table below the seven approaches have been represented and classified into two major categories. The first category contains non-interpolating surfaces (method 1) and the second category contains interpolating models, where the interpolator goes through all available data points (methods 2,3,4,5,6 and 7).



Kind of Response Surface			Method for selecting search points					
			Two-stage approach: first fit a surface, then find the next iterate by optimizing an auxiliary function based on the surface				One stage approach: evaluate hypotheses about optimum based on implications for the response surface	
			Minimize the Response Surface	Minimize a Lower Bounding Function	Maximize the Probability of Improvement	Maximize Expected Improvement	Goal seeking: find point that achieves a given target	Optimization: find point that minimizes an objective
Not interpolating (smoothing)	Quadratic polynomials and other regression models		1					
Interpolating	Fixed basis functions. NO statistics.	Thin-plate splines, Hardy multiquadrics	↑ 2 ↓				↑ 6 ↓	↑ 7 ↓
	Tuned basis functions. Statistical interpretation	Kriging		3	4	5		

Figure2. Taxonomy of response-surface-based global optimization methods. [1]

As can be deduced from figure 2, almost all the approaches are based on the procedure of fitting a surface, finding the minimum and iterating. However, the implemented function and the methods used to select search points differ from one approach to another. As Donald R Jones [1] notes, simple approach that are implemented using quadratic response surface such as method 1 fails in finding the global minimum and still adding more points will not improve the finding. In addition, approaches that used another type of function such as splines perform well for local optimization, although they can easily miss the global minimum (i.e. method 2). However, Alexandrov [18] has shown the possibility to insure a locally convergence by obliging the gradient of the surface to match the gradient of the function where the search stops. Methods 3-7 were found promising to converge to the goal by exploiting the kriging's ability to estimate potential error in its predictions. Although, methods 6 and 7 are a one stage approach in selecting search points and can be computationally very intensive if Kriging is used for the surface. Overall, method 5 stands out as a most promising method that uses Kriging as a response surface and selects search points by maximizing the expected Improvement function. In what follows, a powerful algorithm based on method 5 from figure 2 will be discussed, its benefits and so its possible limitations.

3. Efficient Global Optimization algorithm

Jones et al. [2], proposed the Efficient Global Optimization (EGO) approach based on Kriging model and the Expected Improvement method. This approach consists of the following steps:
Step 1. Build an initial Kriging model for the objective function.
Step 2. Use cross validation to ensure that the Kriging prediction and measure of uncertainty are satisfactory.
Step 3. Find the location that maximizes the expected Improvement (EI) function. If the maximal EI is sufficiently small, stop.
Step 4. Add an evaluation at the location where the EI is maximized. Update the Kriging model using the new data point. Go to Step 3
The decision of whether using the Kriging method or not will depend upon various factors. One important factor is the lack of random error that makes computer experiments different from physical experiments, calling for other methods. The other basic requirement of using Kriging is to have a sufficient number of data points and the data being estimated are stationary.

3.1 Overview of Kriging

The Kriging approach treats the function of interest as a realization of a random function (stochastic process)$Y(x)$. For this reason the mathematical model of Kriging has been presented as a linear combination of a global model plus departures:

$$y(x) = f(x) + Z(x) \qquad (1)$$

Where y(x) is the unknown deterministic response, $f(x)$ is a known (usually polynomial) function of x, and $Z(x)$ is a realization of a stochastic process with mean zero, variance σ^2, and non-zero covariance.

3.1.1 The stochastic process model

The response surface methodology described in this paper is based on modeling the function of interest with stochastic processes. A stochastic process $\{X(t), t \in T\}$ is a set of random variables where the index set T may be discrete ($(T = \{0,1,2...\})$ or continuous $(T = [0,\infty))$. In a stochastic model the change of the system is at least partially random and if the process is run several times, it will not give identical results. Moreover, fitting a stochastic process to data provides us with an insight on how the function typically behaves and how much the function tends to change as we move by different amounts in each coordinate direction [2].

3.1.2 The mathematics of kriging

In contrast to linear regression models, the stochastic process approach assumes that the errors are dependent. In other words, the correlation between errors is related to the distance between the corresponding points. In this model, the distance used is the spatial weighted distance obtained using the distance formula:

$$d(x^{(i)},x^{(j)})=\sum_{h=1}^{k}\theta_h\left|x_h^{(i)}-x_h^{(j)}\right|^{p_h}\ (\theta_h\ge 0,p_h\in[1,2]) \qquad (2)$$

(The meaning and the roles played by the parameters θ_h and p_h will be discussed shortly)
In kriging model, the uncertainty about a value function at a new point is modeled as a realization of a random variable $Y(x)$ that is normally distributed with mean μ and variance σ^2 [1] . In addition, the correlation between the random variables in the Kriging model is given by

$$Corr\left[Y(x_i),Y(x_j)\right]=\exp(-\sum_{h=1}^{k}\theta_h\left|x_h^{(i)}-x_h^{(j)}\right|^{p_h} \qquad (3)$$

From this, one can represent the uncertainty about the function's values at the n points using a random vector with a mean equal to1μ, where 1 is a n$\times 1$ vector of ones, and covariance matrix equal to

$$Cov(Y)=\sigma^2R \qquad (4)$$

Where R is a $n\times n$ matrix with (i,j) element given by Eq. (3). The parameter θ_h in the distance formula (2) can be interpreted as measuring the activity of the variable x_h . For instance, if θ_h is very large, then small values of $\left|x_h^{(i)}-x_h^{(j)}\right|$ translate to large 'distance' and hence low correlation.

Therefore, in order to predict the value of the function at some new points x*, we need first to estimate the parameters of the likelihood formula:

$$\frac{1}{(2\pi)^{\frac{\pi}{2}}(\sigma^2)^{\frac{\pi}{2}}|R|^{\frac{1}{2}}}\exp\left[-\frac{(y-1\mu)'R^{-1}(y-1\mu)}{2\sigma^2}\right] \qquad (5)$$

Then, if we maximize the log of the likelihood formula, set the derivatives with respect to $\mu\ and\ \sigma^2$ to zero and solve, we will get the optimal values expressed as function of R:

$$\hat{\mu}=\frac{1'R^{-1}y}{1'R^{-1}1} \qquad (6)$$

$$\hat{\sigma}^2=\frac{(y-1\hat{\mu})'R^{-1}(y-1\hat{\mu})}{n} \qquad (7)$$

Substituting (6) and (7) on (5) we get a concentrated log-likelihood function, as called in the literature, which depends only on R [1]:

$$-\frac{n}{2}\log(\hat{\sigma}^2)-\frac{1}{2}\log(|R|) \qquad (8)$$

In the first place the maximization of this function provides the estimated parameters$\widehat{\theta_l}\ \ and\ \widehat{p_l}\ \ (l=1,\dots d)$,, in the second place these parameters will be used to compute the estimates $\hat{\mu}$ and $\hat{\sigma}^2$.
Overall the basic idea behind Kriging prediction at some point x*, is guessing a function value y*, adding the point (x*, y*) to the data as the $(n+1)^{th}$ and computing the 'augmented 'likelihood function using parameter values obtained in the maximum likelihood estimation. Consequently, the augmented log likelihood will become a function of y* and demonstrates how consistent the point (x*, y*) with the observed pattern of variation. The value of y* that maximizes this augmented log likelihood function proved to be the Kriging predictor and is given by:

$$\hat{y}(X_{new})=\hat{\mu}+rR^{-1}(y-1\hat{\mu}) \qquad (9)$$

Where r denotes the vector of correlation $Y(x_{new})$ with $Y(x_i)$ for $i=1,\dots n.$

Moreover, kriging is more attractive because of its ability to provide error estimates in the predictor. This mean-squared error is derived using the standard stochastic-process approach and can be computed using:

$$s^2(x^*)=\hat{\sigma}^2\left[1-r'R^{-1}r+\frac{(1-r'R^{-1}r)^2}{1'R^{-1}1}\right] \qquad (10)$$

This formula has a property that reflects no uncertainty about the point we have already sampled. To see this for every x^* = x_i the formula should be equal to zero [1].

The possibility of missing the global optimum when exploring the optimum point, introduce the need of a criterion value called the expected improvement (EI).
This approach involves computing how much improvement we expect to achieve if we sample at a giving point. In what follows, the Expected improvement approach will be described as well as the mathematics behind this method will be highlighted.

3.2 Optimization using the surrogate

Once the Kriging model is built using a set of training data (as is any surrogate model), the parameters of the model have to be estimated to give the best fit to the training data. After finding an optimum design by the Kriging model, this design evaluation (infill point) has to be added to the training data set. Then the Kriging parameters have to be re-estimated and again re-search the model. This process is iterated until we reach some convergence criterion.

In the following section, the researcher will discuss the Expected Improvement method (EI), the mathematics behind this formula and how we can imply (EI) to find the global minimum/maximum.

3.2.1 The Expected Improvement Approach

As described in the previous section, using Kriging technique in optimization requires fitting the kriging model, finding the point that maximizes expected improvement, evaluating the function at this point, and –ultimately- iterating. The second step of this procedure is based on the fact that Kriging helps in estimating the model uncertainty and stresses on exploring points where we are uncertain. This uncertainty is demonstrated and measured by the standard error of the predictor from the previous section.
The ‘expected improvement approach EI’ is a method that incorporates both local and global search to find an optimum. In fact, the EI function is an infill criterion that balances local and global search and helps computing how much improvement we will expect if we sample at a giving point. In order to do this, kriging treats the value of the function at x as if it were the realization of a stochastic process Y(x), with the mean giving by the predictor$\hat{y}(x)$ and variance $s^2(x)$.

Furthermore, if we assume that the current best function value is C_{min} , one can make an improvement denoted by I if $Y(x) = C_{min} -$ I.

To put this in mathematical perspective, the likelihood of achieving this improvement is given by the normal density function

$$\frac{1}{\sqrt{2\pi}s(x)}\exp\left[-\frac{(c_{\min}-I-\hat{y}(x))^2}{2s^2(x)}\right] \qquad (11)$$

Finally, If we integrate over this density function, we will find the expected value of the improvement [1]

$$E(I)=\int_{I=0}^{I=\infty}\left\{\frac{1}{\sqrt{2\pi}s(x)}\exp\left[-\frac{(c_{\min}-I-\hat{y}(x))^2}{2s^2(x)}\right]\right\} \qquad (12)$$

Using integration by parts we can solve this integration and express it as:

$$E(I(x)=(c_{\min}-\hat{y}(x)\Phi(\frac{c_{\min}-\hat{y}(x)}{s})+s\phi(\frac{c_{\min}-\hat{y}(x)}{s}) \qquad (13)$$

Where Φ and Θ are respectively the normal cumulative distribution function and density function.
In short, the EI approach is based on employing an infill criterion that balances local exploitation of $\hat{y}(x)$ and global exploration using $s^2(x)$ by maximizing the expectation of improving upon the current best solution. As a result, using (EI) permits the selection of update points where the model needs more improvement and adding those points as elements of exploration. Using expected improvement can be advantageous and may guarantee global convergence in optimization problems. However, this approach treats the estimated standard error founded in the modeling step as if it is correct. Consequently, points that are close to the current best points have high-expected improvement and the algorithm will require exhaustive search around the initial best point before it begins to search more globally. Moreover, Kleijnen has demonstrated bootstrapped EI as an alternative for the classic EI, and estimates the effect of the initial sample size in some applications bootstrapped EI can find the global optimum faster than classic EI does [17].

3.2 Missing data

When failures occur at any stage of the design evaluation process, no infill point can be added to the surrogate and hence the model will stay unchanged. Unfortunately, most of optimization processes tend to stall and require enhanced techniques to deal with missing data. In order to solve this problem via the medium of surrogate-based optimization, Forrester [5] proposed an imputation model that interpolates the feasible data and provide significant time saving over direct global search methods. In this model he uses a penalized imputation (prediction + error estimate) to add some information or perturb the model

3.3 Noisy data

Similarly to missing data, most data sets are corrupted by noise due to experiments errors or human errors. However, in the context of surrogate-based optimization, we are interested in deterministic noises. Deterministic noise is a non-random noise such that if we repeat experiments we will get the same results.

Although the Kriging stands out other regression models, this approach fails when dealing with noisy data and is unable to approximate noisy function. However, Forrester (2006) introduced an approach to solve this problem by allowing the Kriging model to regress the data. In this approach, adding a regression constant to the diagonal of the Kriging correlation matrix R, will oblige the predictor to deviate from the sample points ensuring an improvement of the likelihood of the data. Furthermore, he also suggested using the re-interpolation method, which guarantees the global convergence of the maximum expected improvement criterion while benefiting from the regression model.

4. Closing remarks and discussion

In general, therefore, it seems that surrogate-based optimization is an extremely promising area for further research and has made significant progress in addressing the analysis and optimization of a variety of complex and expensive systems. In this paper, I have covered the EGO approach, which is based on the Kriging method and the Expected improvement criterion and have noted the advantages of this approach regarding the prediction and the optimization stages.

While Kriging is an exciting and promising technique to build a model for the function of interest, we should bear in mind that the applicability of this method may be problem dependent and must be chosen carefully. Furthermore, the choice of which surrogate to use should be based on the problem size. Unfortunately, Fitting a Kriging model can be beneficent only for relatively low dimensional problems due to the expense of training model. Moreover, since prediction with the kriging model requires the inversion and multiplication of many matrices, prediction may become computationally expensive when the number of sample points increases.

However, there are other extensions of the ordinary Kriging method that may enhance the processes of prediction and optimization. For instance, Sluiter (2009) lists many extensions that have been made to Ordinary Kriging for particular applications-Cokriging, Universal Kriging, Residual Kriging, Blind Kriging, Probability Kriging, and Disjunctive Kriging. Co-kriging is a form of Kriging that involves multiple variables and the estimations of this method are based on a linear weighted sum of all examined variables. Residual kriging is widely used in meteorology where the residuals from a previously fitted regression are interpolated using ordinary Kriging [10]. Disjunctive Kriging is non-linear procedure we assume the all data pairs originate from a bivariate normal distribution and where the data set must be transformed using a series of additive functions. In universal Kriging [4] the mean term is expressed as a set of function of x,

$$\hat{u} = \hat{u}(x) = \sum_{i=0}^{m} \mu_i v_i(x)$$

Where v_i's are some known function and the u_i's are unknown parameters. The idea behind universal kriging is that the model can be tuned regarding the trends in the data, giving better accuracy. In blind kriging, some data-analytic procedures used to identified the v_i's which enhance the accuracy of the model [4].

These extensions need to take into account the adaptive and iterative nature of the optimization problem, which can vary from trivial to impossible.

5. References

[1] D. R. Jones. A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization, 21:345-383, 2001.

[2] W. J. Welch. D.R. Jones, M. Schonlau. Efficient global optimization of expensive black-box function. Journal of Global Optimization, 13:455-492, 1998.

[3] A. I. J. Forrester, A.J. Keane, and N.W. Bressloff. Design and analysis of “noisy” computer experiments. AIAA Journal, 44:2331-2339, 2006.

[4] A. I. J. Forrester and A. J. Keane. Recent advances in surrogate-based optimization. Progress in Aerospace Sciences, 45:50-79, 2009.

[5] A. I. J. Forrester, A. Sobester, and A.J. Keane. Engineering Design via Surrogate Modelling. John Wiley and Sons, 2008.

[6] A. I. J. Forrester, A. Sobester, and A.J. Keane. Optimization with missing data. Proc. R.Soc. A, 462(2067):935-945, 2006.

[7] J. M. Parr, C. M. E. Holden, A. I. J. Forrester, and A.J. Keane. Review of Efficient Surrogate Infill Sampling Criteria with Constraint Handling. International Conference on Engineering Optimization. 2010.

[8] N. V. Queipo, R. T. Haftka, W. Shyy, T. Goel, R. Vaidyana than, and P.K. Tucker. Surrogate-based analysis and optimization. Progress in aerospace sciences, 41:1-28, 2005.

[9] J. Sacks, W.J. Welch, T. J. Mitchell, and H. Wynn. Design and analysis of computer experiments. Statistical Sciences, 4(4): 409-423, 1989.

[10] R. Sluiter. Interpolating methods for climate data. Literature review. 2009

[11] T. W. Simpson, J. D. Peplinski, P. N. Koch, and J. K Allen. On the use of statistic in design and the implications for deterministic computer experiments. 1997.

[12] X. Wang, J. F. Pekny and G. V. Reklaitis. Simulation-based optimization with surrogate models-Application to supply chain management. 29(6): 1377-1328, 2005.

[13] W. Raza and K. Y Kim. Evaluation of surrogate models in optimization of wire-wrapped fuel assembly. Journal of Nuclear Science and Technology, vol. 44, 819-822 (2007).

[14] M. Zakerifar, W. E. Biles and G. W. Evans. Kriging metamodeling in multi-objective simulation optimization. 2115-2122, 2009.

[16] A. Ratle. Kriging as a surrogate fitness landscape evolutionary optimization. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 15,37-49, 2001.

[17] J. Kleijnen, E. Mehdad. Kriging in multi-response simulation including a monte carlo laboratory. 2012

[18] N. M. Alexandrov, R. M. Lewis, C. R. Gumbert, L. L. Green and P. A Newman. Optimization with variable-fidelity models applied to wing design. In Proceedings of the 38th Aerospace Sciences Meeting & Exhibit AIAA paper 2000-0841, 2000.