International Journal of
Mathematical Modelling & Computations
Vol. 02, No. 01, 2012, 35 - 44

Application Of Kriging Method In Surrogate Management Framework For Optimization Problems

B. Azarkhalili[a], M. Rasouli[b], P. Moghadas[c], and B. Mehri[a*]

[a]*Mathematics Department, Sharif University of Technology, Azadi Ave, Tehran, Iran.*
[b]*Electrical Engineering Department, Sharif University of Technology, Azadi Ave, Tehran, Iran.*
[c]*Aerospace Engineering Department, Sharif University of Technology, Azadi Ave, Tehran, Iran.*

Received: 17 August 2011; Accepted: 6 December 2011.

Abstract. In this paper, Kriging has been chosen as the method for surrogate construction. The basic idea behind Kriging is to use a weighted linear combination of known function values to predict a function value at a place where it is not known. Kriging attempts to determine the best combination of weights in order to minimize the error in the estimated function value. Because the actual function value is not known, the error is modeled using probability theory and then minimized. The result is a linear system of equations that can be solved to find a unique combination of weights for a given point at which interpolation is to be performed.

Keywords: Surrogate Management Framework, Kriging, Computational Order, Convergence, Gradient.

Index to information contained in this paper

1. Introduction

In optimizing Problem a computationally expensive function in an engineering application, it is advantageous to obtain predictions of the function behavior without performing many costly function evaluations. In this paper, we present methods that take advantage of surrogates to predict and approximate function behavior for use in optimization. The term "surrogate" as applied to optimization is an umbrella term referring to any instance in which the true function is replaced by a

[*] Corresponding author. Email: mehri@sharif.edu.

stand-in. A surrogate can be an approximation of the true function, a simplified physics model, a ”yes or no” answer, or anything else that substitutes for the true function during optimization.

Surrogate functions have been successfully used in a trust region method by Chung & Alonso [5]. In this work, gradient information was used to construct Cokriging approximations, Kriging approximations which incorporate derivative information. Their work showed that the use of gradients allows for a more accurate model with many fewer points compared to a standard Kriging surrogate. In evolutionary algorithms, surrogate functions have been used to reduce the cost of optimization by Ong et al. [14]. A nice overview of the use of surrogate methods in engineering is given in Guinta [7].

The surrogate management framework (SMF) was developed to increase the efficiency of pattern search methods for expensive problems that may have little or no gradient information. The surrogate management framework falls into both the categories of approximation modeling methods and pattern search methods. However, the convergence analysis for the expensive problem is independent of the accuracy of the approximate modeling approach used. The SMF method provides a robust and efficient alternative to traditional gradient method such as gradient methods which use an adjoint the use of surrogate functions.

2. Construction of Surrogate Models Using Kriging

One of the important features of SMF is the use of a surrogate to predict the minimum of the cost function. This section is meant to be a tutorial in which the construction of surrogate models using Kriging is discussed in detail. Other types of surrogate models which we do not discuss here include polynomials (response surfaces) and splines. Comparisons of response surface and Kriging models are presented by Simpson et al. (1998) and Guinta & Watson [8]. Kriging originated in the field of geostatistics, as presented in Isaaks & Srivastava [10], and is named for South African geologist Krige. It is a statistical method based on the use of spatial correlation functions.

In this work, Kriging has been chosen as the method for surrogate construction for several reasons. First, it is easily extended to multiple dimensions, making it attractive for construction of surrogates in optimization with several parameters. As the dimension increases, polynomials and splines both become problematic and may produce spurious oscillations.

Construction of surrogate models for computer generated data calls for different statistical techniques than an experiment done in a lab. Unlike lab measurements, computer simulations are deterministic, meaning that they are repeatable with no random error. This difference has ramifications when constructing surrogate models. One is that the adequacy of the model fitted through the observed data is determined by systematic bias and not by random error. Another is that the usual measures of least squares uncertainty have no obvious statistical meaning when applied to deterministic data. The first work that looked at experimental design specific to deterministic computer codes was McKay et al. In their paper Latin hypercube sampling (LHS) was introduced as a method of choosing well distributed data sets for computer experiments.

2.1 *Intuitive Construction Of A Kriging Surrogate*

The basic idea behind Kriging is to use a weighted linear combination of known function values to predict a function value at a place where it is not known. Kriging attempts to determine the best combination of weights in order to minimize the error in the estimated function value. Because the actual function value is not known, the error is modeled using probability theory and then minimized. The result is a linear system of equations that can be solved to find a unique combination of weights for a given point at which interpolation is to be performed.

We first demonstrate this concept using a more intuitive explanation, and then present a formal derivation. Let us assume that we wish to approximate a function v at location x_0 given a set of n known data points $v_i(x_i)$, $i = 1, \ldots, n$. To do this, we must choose a vector of weights wi to act on the known points. The resulting predicted value of the function $v_0(x_0)$ will be a weighted sum of known values

$$v_0(x_0) = \sum_{j=1}^{n} w_i v_i(x_i) \tag{1}$$

To find the values of the weights, we model the covariance C_{ij}, which depends on the distribution of the known data points. The covariance describes, in terms of probability, the degree to which a function value at a given point is similar to values nearby. An example of Gaussian covariance is

$$C_{ij} = exp(-((x_i - x_j)/a)^2) \tag{2}$$

Where a is a constant to be chosen. We note that the dimension of the problem only appears when taking the Euclidean norm in the above expression. For this reason, the method is easily extended to high dimensional problems with no change in complexity. Using the covariance, we construct the following linear system

$$\begin{pmatrix} C_{11} & C_{12} & \ldots & C_{1n} & 1 \\ C_{21} & C_{22} & \ldots & C_{2n} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ C_{n1} & C_{n2} & \ldots & C_{nn} & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ \lambda \end{pmatrix} = \begin{pmatrix} C_{10} \\ C_{20} \\ \vdots \\ C_{n0} \\ 1 \end{pmatrix} \tag{3}$$

Row $n + 1$ in the above matrix ensures that the weights sum to one, and λ is a Lagrange multiplier introduced in the Kriging error minimization. The entries in the correlation matrix on the left hand side of Equation (3) represent how well correlated a given known point is with another known point. The entries in the vector on the right hand side gives the correlation between the unknown data point x_0 and each of the known data points x_i. The system of equations in (3) can be inverted to find the weights w_i, and then substituted into Equation (1) to give the interpolated function value v_0 at the location x_0. The following section formalizes these concepts.

2.2 *Formal Derivation of Kriging Surrogates*

Following Sacks et al. [16] and Lophaven et al. we derive an expression for a general Kriging approximation. Kriging is also developed in detail by Koehler & Owen . We wish to approximate the function value at an unknown location $x \in R^n$ based on a set of known data points. We start with m known data points $\{s_i\} \in R^n$ and define $y_s \in R^m$ to be the column vector whose elements are the corresponding function values, i.e. $[y_s]_i = \{y(s_i)\}$, $i = 1, 2, \dots, m$.

If the number of dimensions is n, each s_i is a vector of n elements and each $y(s_i)$ is a scalar function value. We wish to predict the value of the function based on the values of known points, so we consider the linear predictor

$$y(x) = c^T(x) y_s \tag{4}$$

Where $c(x) \in R^m$ (hereafter referred to as c for simplicity) is a vector of weights applied to the known functions values ys. By determining a set of weights c, we will be able to find an approximation of the function at any location x given a set of known data points. We may assume that the deterministic function $y(x)$ can be modeled as the realization of a stochastic process $Y(x)$, which is the sum of a regression model having basis functions $f_j : R \to R$ and coefficients β_j , $j = 1, 2, \dots, k$, and a random function $Z : R^n \to R$, giving

$$Y(x) = \sum_{j=1}^{n} \beta_i f_j(x) + Z(x) \text{ or}$$
$$Y(x) = \beta^T f(x) + Z(x) \tag{5}$$

Where $\beta = (\beta_1, \beta_2, \dots, \beta_k)^T$ and $f(x) = (f_1(x), f_2(x), \dots, f_k(x))^T$. For any point $x \in R^n$, the random process $Z(x)$ is assumed to have zero mean, variance σ^2, and correlation $R(w, x)$ between x and any other point w. The covariance of Z is then

$$E\left[z_1(w) z_1(x)\right] = \sigma^2 R(w, x) \tag{6}$$

Choice of regression model $f(x)$ and correlation function $R(w, x)$ will be discussed at the end of this section. Because we are modeling y_s as a random process, the predictor becomes

$$y(x) = c^T(x) Y_s \tag{7}$$

We can now compute the mean square error (MSE) of the predictor averaged over the random process. The best choice of Kriging weights will be determined by minimizing the MSE of the predictor, which is the error between the predicted value and the actual value at location x,

$$\text{MSE}[y(x)] = E\left[c^T(x) Ys - Y(x)\right]^2 \tag{8}$$

Where $Ys \in R^m$ is the vector defined by $[Y_s]_i = Y(s_i)$, $i = 1, 2, \dots, m$. Letting $F = [f(s_1), \dots, f(s_m)] \in R^{k \times m}$ and $Z = [z_1, \dots, z_m]$ we have

$$Y_s = F\beta + Z \tag{9}$$

and

$$Y(x) = f^T(x)\beta + z \tag{10}$$

Then,

$$c^T Y_s - Y(x) = c^T(F\beta + Z) - (f^T(x)\beta + z) = c^T Z - z + F^T c - f(x)\beta \tag{11}$$

Where z is a realization of $Z(x)$, and $z_1, \ldots, z_m$ are realizations of $Z(s_i)$, $i = 1, 2, \ldots, m$. We impose an unbiasedness constraint, ensuring the weights c must sum to one,

$$F^T c(x) - f(x) = 0 \tag{12}$$

so that (15) becomes

$$c^T Y_s - Y(x) = c^T Z - z \tag{13}$$

Now, the MSE is

$$\text{MSE}[y(x)] = E[(c^T Z - z)^2] = E[z^2 + c^T Z Z^T c - 2c^T Z z] \tag{14}$$

From the covariance of Z, we have $E[z^2] = \sigma^2$, $E[Zz] = \sigma^2 r$ and $E[ZZ^T] = \sigma^2 R$, where $r \in R^m$ is a vector of correlations between the known points and an untried point x,and $R \in R^{m\times m}$ is the matrix of correlations between the known points. With this, the MSE becomes

$$\text{MSE}[y(x)] = \sigma^2(1 + c^T Rc - 2cr) \tag{15}$$

We wish to find the weights c that minimizes the MSE subject to the constraint (16). To do this, we use the method of Lagrange multipliers with the Lagrangian function

$$L(c, \lambda) = \sigma^2(1 + c^T Rc - 2cr) - \lambda^T(F^T c - f) \tag{16}$$

The gradient of the Lagrangian (20) with respect to c is

$$L'_c(c, \lambda) = 2\sigma^2(Rc - r) - F\lambda \tag{17}$$

Setting the gradient to zero to find the minimum, we obtain the set of equations

$$\begin{aligned} Rc + F\lambda* &= r \\ F^T c &= f \end{aligned} \tag{18}$$

where $\lambda* = -\lambda/2\sigma^2$. Or, in matrix form,

$$\begin{pmatrix} R & F \\ F^T & 0 \end{pmatrix} \begin{pmatrix} c \\ \lambda* \end{pmatrix} = \begin{pmatrix} r \\ f \end{pmatrix} \tag{19}$$

Solving this system of equations yields

$$\begin{aligned} \lambda^* &= (F^T R^{-1} F)^{-1}(F^T R^{-1} r - f) \\ c &= R^{-1}(r - F\lambda^*) \end{aligned} \tag{20}$$

The correlation matrix R and also R^{-1} are symmetric. So, substituting the solution (24) Into the predictor (12), we have

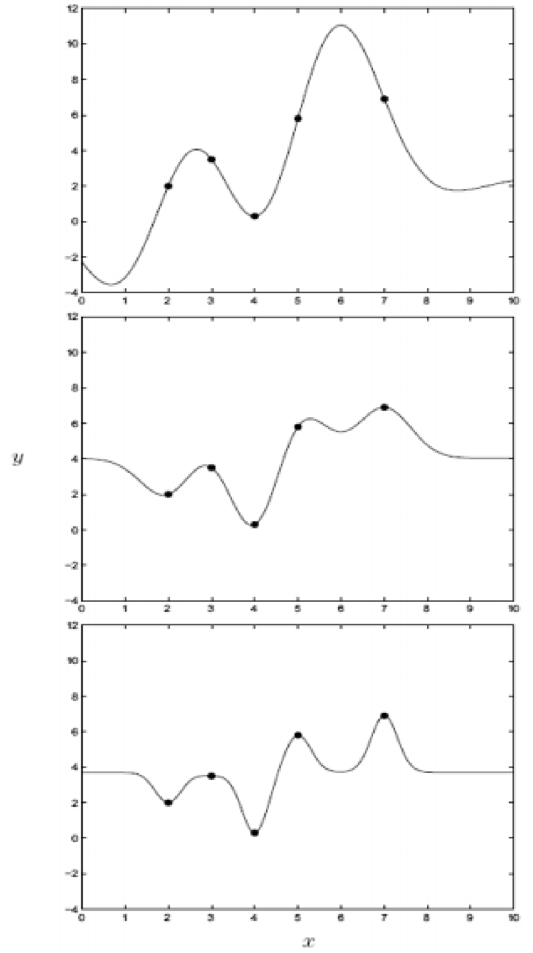

Figure 1. Examples of a Kriging Fit Using Values of $\theta = 1$ (top), $\theta = 5$ (Middle) and $\theta = 20$ (Lower) for the Same data set. Function $y(x)$ is Known at Five Data Points (Dots) and Approximated with a Kriging Function (Solid Line).

$$\begin{aligned} y(x) &= (r - F\lambda^*)^T R^{-1} Y_s = r^T R^{-1} Y_s - \lambda^{*T} F^T R^{-1} Y_s \\ &= r^T R^{-1} Y_s - (F^T R^{-1} r - f)^T (F^T R^{-1} F)^{-1} F^T R^{-1} Y_s \end{aligned} \tag{21}$$

Defining

$$\begin{aligned} \beta^* &= (F^T R^{-1} F)^{-1} F^T R^{-1} Y_s \\ \gamma^* &= R^{-1}(Y - F\beta^*) \end{aligned} \tag{22}$$

We have

$$y(x) = f(x)^T\beta^* + r(x)^T R^{-1}(Y_s - F\beta^*) \tag{23}$$

and

$$y(x) = f(x)^T\beta^* + r(x)^T\gamma^* \tag{24}$$

For a given set of data and choice of regression and correlation functions, β^* and γ^* are fixed and need not be recomputed for each new point x.

To check that the Kriging predictor exactly interpolates the known data, we let $x = s_i$, one of the known data points. Then $R^{-1}r(x) = e_i$ the unit vector and (2.26) gives

$$y(s_i) = f(s_i)^T\beta^* + e_i^T(Y_s - F\beta^*) = f(s_i)\beta^* + y_i - F_i, \quad \beta^* = yi \tag{25}$$

2.3 *Choosing Kriging Correlation Function*

To complete our description of Kriging surrogate models we must choose a regression model and a correlation function. The most common choice of regression model is simply $f(x) = 1$ so that Equation (2.14) becomes

$$Y = \beta + Z \tag{26}$$

With this regression function, the Kriging predictor (2.26) becomes

$$\begin{aligned} y(x) &= \beta^* + r(x)^T R^{-1}(Y_s - 1.\beta^*) \\ \beta^* &= (\sum_j Y_s(j)\sum_j R_{ij}^{-1})(\sum_{i,j} R_{ij}^{-1})^{-1} \end{aligned} \tag{27}$$

Other common choices for the regression model are first or second order polynomials. The correlation function is chosen to be the product of stationary one dimensional correlation. This makes the model easily extendable to multiple dimensions. The correlation between two points x and w is then

$$R(\theta, w, x) = \prod_{j=1}^{n} R_j(\theta, w_j - x_j) \tag{28}$$

A common choice of correlation function is to express the correlation between two points x and w in terms of a Gaussian process

$$R(\theta, w, x) = \prod_{j=1}^{n} \exp(-\theta_j(w_j - x_j)^2) \tag{29}$$

The Kriging surrogate in (29) is completed with the matrix of correlations between the values of z at any two known design sites, which is defined by

$$R_{ij} = R(\theta, s_i, s_j), \quad i, j = 1, 2, \ldots, m \tag{30}$$

And the vector of correlations between the value of z at a known design site and any point x.

$$r(x) = [R(\theta, s_1, x), \ldots, R(\theta, s_m, x)] \tag{31}$$

Other commonly used correlation functions include exponential, spline, cubic, spherical and linear. These are discussed in greater detail in Lophaven et al.

Assuming a Gaussian correlation function, the only remaining piece left to define is the choice of the parameter θ. The optimal value θ^* of θ is found using maximum likelihood estimation in each dimension, so that θ^* solves

$$\min_{\theta} \left\{ \psi(\theta) \equiv |R|^{\frac{1}{m}} \sigma^{*2} \right\} \tag{32}$$

Where $|R|$ is the determinant of R, and

$$\sigma^{*2} = \frac{1}{m}(y_s - F\beta^*)^T R^{-1}(y_s - F\beta^*) \tag{33}$$

In practice, the value of the parameter θ determines the smoothness of the Kriging approximation. The value of θ can be viewed as a knob that is dialed up or down to change the radius of influence of a data point on the surrounding approximation. Smaller values of theta will result in a smoother surface, in which the radius of influence is large, whereas larger values of theta makes the surface approximation less smooth and the radius of influence smaller. Examples of one dimensional Kriging function fits using progressively increasing values of θ are shown in Figure 1. In the top plot, we see that a low value of $\theta = 1$ gives the smoothest fit, but may also result in large overshoots of the data. In the lower plot, we see that using a high value of $\theta = 20$ results in a function fit that only deviates from the mean value in the immediate neighborhood of the data points. In general, it is best to choose a moderate value of θ, as shown in them iddleplot.

3. Well-Conditioned Kriging

Another modification to Kriging models offers a solution to the well-known problem of “pileup” of points. It has been observed that the similarity of Kriging surrogates to the true function tends to deteriorate as points become clustered close together, especially nearing convergence. In addition, the problem of determining parameters for Kriging models often becomes ill-conditioned as points pile up due to the very small distances between points. Improvement to surrogate models in the presence of clustering is important for maintaining the usefulness of surrogates as the optimization proceeds. An elegant solution to this problem has been proposed and tested by Booker [4] and discussed further in Audet et al. Booker proposes to model the output as the sum of two stochastically independent Gaussian processes. The first uses the original correlation parameters, estimated from the first set of known points. The second uses a finer correlation structure. The Kriging model is thus constructed by the following sum

$$Y(x) = \beta + Z_1(x) + Z_2(x), \tag{34}$$

Where Z_1 and Z_2 are independent of each other. The resulting correlation function for Y is then

$$R(x, w) = \lambda R_1(x, w) + (1 - \lambda)R_2(x, w) \tag{35}$$

where

$$\lambda = \sigma_1^2(\sigma_1^2 + \sigma_2^2)^{-1} \tag{36}$$

and R_i and σ_i correspond to Z_i. This method potentially offers increased accuracy and cost savings and should be studied further in future engineering optimization problems.

4. Conclusion

There are several variations of the surrogate management framework that can offer savings in computational cost, and improvements in cost function reduction. In problems with full or even partial gradient information, large gains in efficiency and cost function reduction are often possible. Automatic differentiation (Bischof et al.) and adjoint solvers (Jameson, [11] b,a; Jameson et al. [12]) are promising methods for obtaining gradient information, even for complex problems. The SMF method is general enough to incorporate gradient information in a number of ways.

Another use of gradient information is in the construction of Kriging models. Gradients can be incorporated into Kriging models using a method called Cokriging. Use of models constructed with Cokriging has been demonstrated by Chung & Alonso [5]. In their work, it was demonstrated that a highly accurate surrogate can be constructed using very few points with their corresponding gradient information.

References

[1] Abramson, M. A. Pattern search algorithms for mixed variable general constrained optimization problems. PhD thesis, Department of Computational and Applied Mathematics, Rice University, Houston, TX, (2002).

[2] Abramson, M. A., Audet, C., Dennis, Jr., J. E. Generalized pattern searches with derivative information. Mathematical Programming, Series B, **100** (2004) 3–25.

[3] Booker, A. J. Well-conditioned Kriging models for optimization of computer models. Mathematics and Computing Technology Report 002. Boeing Phantom Works, Seattle, WA, (2000).

[4] Booker, A. J., Dennis, Jr., J. E., Frank, P.D., Serafini, D.B., Torczon, V., Trosset, M. W. A rigorous frameworks for optimization of expensive functions by surrogates. Structural Optimization **17 (1)** (1999) 1–13.

[5] Chung H., Alonso, J. Design of a low-boom supersonic business jet using cokriging approximation models. AIAA, (2002) 65–98.

[6] Gill, P. E., Murray, W., Wright, M. H. Practical Optimization. San Diego: Academic Press, (1981).

[7] Guinta, A. A. Use of data sampling, surrogate models, and numerical optimization in engineering design. AIAA Paper, (2002) 05–38.

[8] Guinta, A. A., Watson, L. T. A comprasion of approximation modelling techniques: polynomial versus interpolating models. AIAA Paper, (1998) 47–58.

[9] Hansen, N., Ostermeier, A. Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation. Proc. of the 1996 IEEE Intl. Conf. on Evolutionary Computation, (1996) 312–317.

[10] Isaaks, E. H., Stivastava M. An introduction to applied geostatistics, Oxford University Press, (1989).

[11] Jamesom A. Optimum aerodynamic design using CPD and control theory. AIAA, **95** (1995) 17-29.

[12] Jamesom A., Martinelli L., Pierce N. A., Optimum aerodynamic design using the Navier-Stokes equation, theoret. Comp, Fluid Dynamics, **10** (1998) 213-283.

[13] Lewis, R. M., Torczon, V. A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds. SIAM J. Optim. **12** (2002) 1075–1089.

[14] Ong, C., Wan, D., Ong, K. An exploratory study on interlacking directions in listed firms in Singapore, Corporate Governance: An International Review, **11(4)** (2003) 323–833.

[15] Ong, Y. S., Nair, P. B., Keane, A. J. Evolutionary optimization of computationally expensive problems via surrogate modeling. AIAA J. **41 (4)** (2003) 687–696.

[16] Sacks, J., Welch, W. J., Mitchell, J. J., Wynn H. P. Design and analysis of computer experiments, Statistical Science, **4(4)** (1989) 409-485.

[17] Simpson, T. W., Korte, J. J., Mauery, T. M., Mistree, F. Comparison of response surface and kriging models for multidisciplinary design optimization. AIAA Paper, (1998) 47–55.

[18] Azarkhalili B., Rasouli M., Moghadas P., Mehri B. Introduction And Development Of Surrogate Management Framework For Solving Optimization Problems, International Journal of Mathematical Modeling & Computations, **1(4)** (2011) 235-244.