Journal of Global Optimization, **30:** 39–58, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

A Derivative Based Surrogate Model for Approximating and Optimizing the Output of an Expensive Computer Simulation

STEPHEN J. LEARY, ATUL BHASKAR and ANDY J. KEANE
Computational Engineering and Design Centre, School of Engineering Sciences, The University of Southampton, Highfield, Southampton, SO17 1BJ, UK (e-mail: s.j.leary@soton.ac.uk)

(Received 28 August 2001; accepted in revised for 10 August 2003)

Abstract. Approximation methods have found an increasing use in the optimization of complex engineering systems. The approximation method provides a 'surrogate' model which, once constructed, can be called instead of the original expensive model for the purposes of optimization. Sensitivity information on the response of interest may be cheaply available in many applications, for example, through a pertubation analysis in a finite element model or through the use of adjoint methods in CFD. This information is included here within the approximation and two strategies for optimization are described. The first involves simply resampling at the best predicted point, the second is based on an expected improvement approach. Further, the use of lower fidelity models together with approximation methods throughout the optimization process is finding increasing popularity. Some of these strategies are noted here and these are extended to include any information which may be available through sensitivities. Encouraging initial results are obtained.

Key words: derivatives, kriging, multifidelity models, optimization, surrogate model.

1. Introduction

Expensive computer codes based on mathematical models of some system of interest are commonly used throughout the engineering industry, for example, a finite element analysis in structural engineering or a Navier-Stokes model in a CFD analysis. If the goal is to perform optimization with respect to the model, we are often overcome by the model's computational expense. Direct optimization, which requires many calls to the model of interest, is more often than not, unrealistic.

As a result, cheap approximations termed 'surrogates' to the expensive model are sought. These are based on only a limited number of calls to the expensive model, which we term the high fidelity model. Once the surrogate model is constructed, it replaces the original model for the purposes of optimization.

When we consider optimization, such approximations may be defined locally or globally. Local approximations, see for instance Myers and Montgomery (1995), are defined over a specific region of interest, usually about the current best design.

These approximations are often based on some low order polynomial approximation of the model's response and are only valid in some small neighbourhood surrounding the current best design. The optimization proceeds using a move-limit or trust-region strategy: we optimize only over the region where the approximation is valid. Then a new approximation is sought and the process is repeated until either the maximum allowable number of calls to the model is reached or an optimum is found.

Global approximations on the other hand try to capture the model's behaviour over the entire domain of interest. Many different approximations can be considered, for example, artificial neural networks (White et al., 1992) or kriging (Sacks et al., 1989). This paper considers global approximations; nevertheless the ideas considered could be implemented locally using, say, a trust-region approach. We consider global optimization based on an advanced kriging model in this paper. The use of such stochastic processes for global optimization (commonly known as Bayesian global optimization) dates as far back as 1964 (Kushner, 1964).

Traditionally the information available to construct the surrogate is in terms of the response only. Nowadays, however, gradient information (i.e. derivatives of the response with respect to the independent variables or inputs) may also be cheaply available. For instance, a pertubation analysis of a finite element solution can lead to a very good approximation of the derivatives, whereas in an adjoint CFD analysis, all the derivatives are directly available. This information can be incorporated into the current model. See for example Morris et al. (1993) for a full description of a kriging model using derivative information.

Section 2 reviews this approach, then Section 3 considers its use in optimization. Two models are considered, direct optimization and the Efficient Global Optimization algorithm of Jones et al. (1998), which has been adapted here to incorporate derivatives. Two simple examples demonstrate the approach in Section 4.

One concern with these global approximations is their level of accuracy. These models are simply forms of curve fitting built using selective calls to the high fidelity model and do not attempt to incorporate further information on the problem in hand. As a consequence there has been a growing interest in the use of simpler low fidelity models in overcoming this burden.

Low fidelity models, while being less accurate than the original high fidelity models, are generally much cheaper to compute. As an example a low fidelity model may use a considerably coarser mesh than the high fidelity model. We may also consider, for example, using an Euler code to approximate an expensive Navier-Stokes model in CFD or only partially converging a solution. The low fidelity model is used to obtain some 'rough' information as to the global behaviour of the response of interest and selective calls to the high fidelity model provide 'corrections' to this response.

The low fidelity model may be included in the approximation in numerous ways. Perhaps the simplest way of utilizing information provided by the low fidelity model is to consider the difference between the models. Examples of this approach include its use in structural optimization Leary et al. (to appear), application to wing design in Balabanov et al. (1998) as well as to problems in microwave design in Watson and Gupta (1996).

An alternative to considering the difference in low and high fidelity response would be to consider their ratio. Chang et al. (1993) calculate the ratio and derivatives at one point in order to construct local approximations and demonstrate the approach on a wing box model of a high speed civil aircraft. Alexandrov et al. (1998) combine this approach with a classic trust-region methodology to produce a local approximation algorithm which alters the size of the trust region (the portion of the domain on which the local approximation is valid) as the optimization proceeds.

The general approach, termed a correction response surface model, has been applied to aerodynamic problems by Hutchinson et al. (1994) and to structural problems by Vitali et al. (1999).

Another example of the use of low fidelity models in approximations is the space mapping algorithm, see for instance Bandler et al. (1994), or Bakr et al. (1998, 1999). Here the idea is to establish a mapping on the inputs such that the response of the low fidelity model with the mapped parameter agrees with the response of the high fidelity model. This algorithm is typically applied locally; nevertheless, a global approximation is also possible.

Recently Wang and Zhang (1997) introduced a knowledge-based neural network model for microwave design; this model used information provided by empirical functions as knowledge to improve the accuracy of approximations. Leary et al. (submitted for review) consider a similar model where the empirical function is replaced by a low fidelity model: this approach also falls into the multifidelity approximation framework. The approximations can be based on a neural network model as in Wang and Zhang or on a kriging model as in Leary et al.

Once more, these multifidelity strategies are typically based on response values only, the above approaches are easily developed to include derivative information. One approach is described in detail in Section 5 and an example is given in Section 5.1. Finally, in Section 6 some conclusions are drawn.

2. Methodology

A kriging model incorporating derivative information (Morris et al., 1993) is first used to generate fast approximations to the high fidelity response. Before we build the approximation, however, we need some information on the problem in hand. This we obtain by evaluating the response of interest $y(\mathbf{x})$, $\mathbf{x} \in \mathcal{R}^k$ for several combinations of the inputs $\mathbf{x}$. We require a systematic means of selecting the set of points $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\}$ called a design of experiments (DOE for short) within the k dimensional space at which to perform computational analyses.

The 2^k vertices formed by the upper and lower bounds of the components of $\mathbf{x}$ form the design bounding box within which the experimental design is created. Simple experimental designs include 2^k full factorial designs that are created by specifying each design variable at two levels, the lower and upper bounds of the design bounding box. 3^k full factorial designs additionally include the midpoint of each input. These experimental designs prove expensive, (particularly for large k) so fractional factorial designs, or alternatively S-optimal designs could be considered. Many examples of these approaches exist in the response surface literature, the interested reader should consult Myers and Montgomery (1995) or alternatively Mead (1988) for further details.

In this paper latin hypercube sampling (Mackay et al., 1979) is used to construct a space filling set of inputs. The strategy adopted here is the following: many latin hypercube designs are generated and the one that minimizes

$$\sum_{i=1}^{N}\sum_{j=i+1}^{N}\frac{1}{d_{ij}^2} \tag{1}$$

is chosen. Here d_{ij} refers to the distance between points i and j. An example of choosing three design points in two dimensions is shown in Figure 1.

We note that our design of experiments requires a search over many latin hypercube designs. When N and k are small we are able to generate good space filling designs in a computationally efficient manner. However, for large N and/or k, such a search may result in an expensive process. In this case it may be better to use special techniques that do not require a potentially expensive search to generate a good design. One such example are LP_τ sequences, see Sobol (1979) for details.

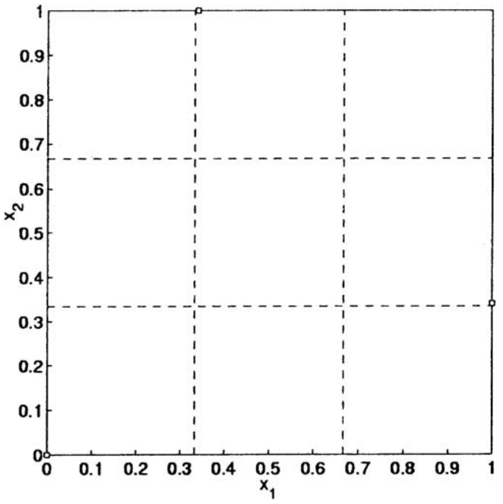

Figure 1. Latin hypercube design with $N = 3$ and $k = 2$ that minimizes the function in (1).

Following Morris et al. (1993), we observe the response y and its derivatives $y_1, y_2, \dots, y_k$ at the N design points $D = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}\}$ and store these in the $N(k+1)$ vector $\mathbf{y}$. We wish to use this information to obtain predictions to $y(\mathbf{x}^*)$ where $\mathbf{x}^*$ is some previously unsampled input.

We represent the uncertainty in y (and hence $y_1, y_2, \dots, y_k$) by a Gaussian stochastic process. We do not assume errors are uncorrelated as in regression, but that the errors are correlated, the correlation between errors being related to some distance measure.

We use the correlation function

$$R_l(x_l^{(i)}, x_l^{(j)}) = \mathrm{e}^{-\theta_l (x_l^{(i)} - x_l^{(j)})^2}, \quad \theta_l > 0, \ l = 1, \dots, k, \tag{2}$$

where θ_l, $l = 1, \dots, k$ are parameters yet to be determined. This function is a partial case of that introduced by Jones et al. (1998). The overall correlation, again following the arguments of Jones et al. (1998), is defined by the product correlation

$$R(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \prod_{l=1}^{k} R_l(x_l^{(i)}, x_l^{(j)}). \tag{3}$$

Desirable properties of this correlation function are given in Jones et al. (1998). For the derivative based approximations we also note that another requirement of this function is that it is at least twice differentiable, clearly the correlation function in (2) satisfies this property.

An $N(k+1) \times N(k+1)$ correlation matrix C of the sampled data D is then defined as in Morris et al. (1993). This matrix is defined in terms of the correlation functions R_l, $l = 1, \dots, k$ as well as their first and second derivatives R_l', R_l'', $l = 1, \dots, k$. Similarly an $N(k+1)$ vector of correlations, $\mathbf{r}$ between a new point $\mathbf{x}^*$ and the previously sampled data D can be defined. For full details of this procedure we refer the reader to Morris et al. (1993).

The hyperparameters $\theta_1, \dots, \theta_k$ are chosen to maximize the likelihood of the sample. The log-likelihood is expressed as

$$L(\theta) = -N(k+1)\ln\sigma^2 - \ln|C_\theta| - \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{v}\mu)^T C_\theta^{-1}(\mathbf{y} - \mathbf{v}\mu), \tag{4}$$

Here μ and σ^2 represent the mean and variance of the data, and $\mathbf{v}$ is an $N(k+1)$ binary vector with 1 in position $(i-1)(k+1)+1$, $i = 1, \dots, N$ and 0 everywhere else. The dependence on the parameters $\theta_1, \dots, \theta_k$ through the matrix C is indicated, these parameters here being collectively denoted by θ.

For fixed θ, maximization of L over μ and σ^2 is obtained by

$$\hat{\mu}_\theta = \frac{\mathbf{v}^T C_\theta^{-1} \mathbf{y}}{\mathbf{v}^T C_\theta^{-1} \mathbf{v}} \tag{5}$$

and

$$\hat{\sigma}_\theta^2 = \frac{1}{N(k+1)}(\mathbf{y}-\mathbf{v}\mu)^T C_\theta^{-1}(\mathbf{y}-\mathbf{v}\mu). \tag{6}$$

Substituting (5) and (6) into (4) we obtain a function of θ_l, $l=1,\ldots,k$ only, this we maximize to obtain $\hat{\theta}$ and hence an estimate of the overall correlation matrix C (n.b., the function $L(\theta)$ is often highly multi-modal and so this maximization has to be carried out with some care).

Formulas (5) and (6) then provide us with an estimate of $\hat{\mu}$ and $\hat{\sigma}^2$.

The predictor at an unsampled point is then given by (Morris et al., 1993)

$$\hat{y}(\mathbf{x}^*) = \hat{\mu} + \mathbf{r}^T C^{-1}(\mathbf{y}-\mathbf{v}\hat{\mu}). \tag{7}$$

If $\mathbf{x}^*$ corresponds to a sampled point $\mathbf{x}^{(i)}$, then $\mathbf{r}$ corresponds to the ith row of C, as a result, the model interpolates the data. The advantage of this model over ordinary kriging models is that the model also gives the correct gradient at a sampled point, and as a result, predictions can be far more accurate. It's disadvantage is that the correlation matrix of the sampled data is far larger and hence the maximization of (4) is much more expensive, particularly for large k.

The mean squared error of the predictor is now

$$s^2(\mathbf{x}^*) = \sigma^2\left[1 - \mathbf{r}^T C^{-1}\mathbf{r} + \frac{(1-\mathbf{v}^T C^{-1}\mathbf{r})^2}{\mathbf{v}^T C^{-1}\mathbf{v}}\right]. \tag{8}$$

This measure provides us with an estimate of the accuracy in our model.

3. Strategies for Optimization

The first strategic decision that must be made in following any of the approaches described here concerns the size of the initial DOE, usually as some fraction of the total number of expensive evaluations that may be made – this is far from trivial, but lies beyond the scope of this paper. Having carried out the initial run of expensive calculations and before performing optimization over our approximate surface, it should be noted that some sort of model validation could be considered; for example leave one out cross validation. Strategies for doing this can be found in Jones et al. (1998).

3.1. DIRECT OPTIMIZATION

The simplest algorithm one could possibly conceive is to optimize over the approximate surface, then evaluate the response at the optimum found, then add this point to D by running a new expensive calculation and then form a new approximation. The process may then be repeated until either convergence is in some sense achieved or the maximum number of allowable expensive evaluations is reached. This process is shown in Figure 2.

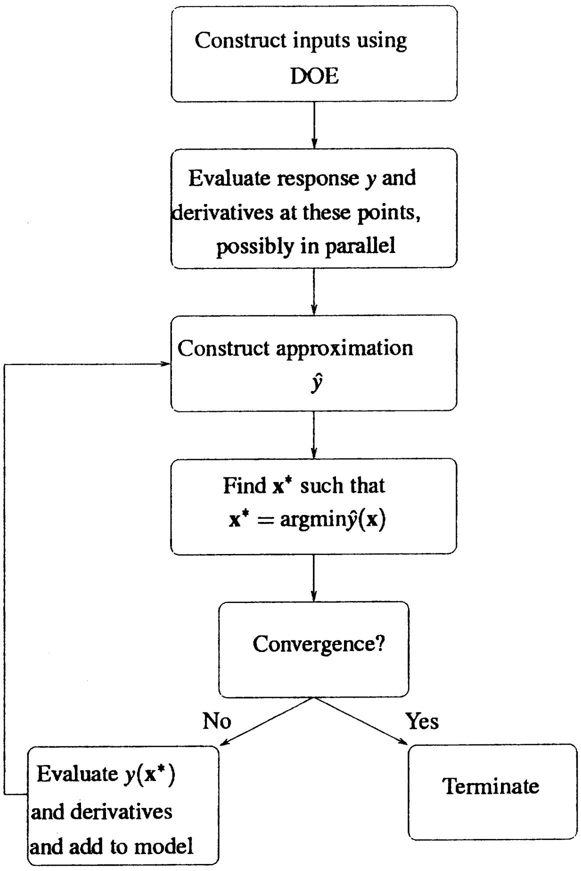
Construct inputs using DOE
Evaluate response y and derivatives at these points, possibly in parallel
Construct approximation $\hat{y}$
Find $\mathbf{x}^*$ such that $\mathbf{x}^* = \operatorname{argmin}\hat{y}(\mathbf{x})$
Convergence?
No
Yes
Evaluate $y(\mathbf{x}^*)$ and derivatives and add to model
Terminate

Figure 2. The direct strategy.

3.2. EXPECTED IMPROVEMENT

In Jones et al. (1998) an expected improvement algorithm is presented for the ordinary kriging model, here the algorithm is applied to the derivative based model.

The expected improvement is a figure of merit that balances local and global search. It does so by defining an objective which balances the goodness of the prediction together with the predictions uncertainty. An early example of the concept can be found in Mockus et al. (1978). The expected improvement is computed as follows: first the current best function value $y_{\min} = \min(y_1, y_2, \ldots, y_N)$ is computed. Consider the example in Figure 3, the function is sampled at five points. The prediction is shown in Figure 4 where $y_{\min} = 2.062$ at $x = 4.2$. Let us consider a Gaussian distribution with mean given by the prediction and standard deviation given by its standard error. This is also shown in Figure 4 for $x = 9$. The tail of this distribution will fall below $y = y_{\min}$ that is, there is some probability that the function's value here will improve upon $y_{\min}$ shown shaded in the figure.

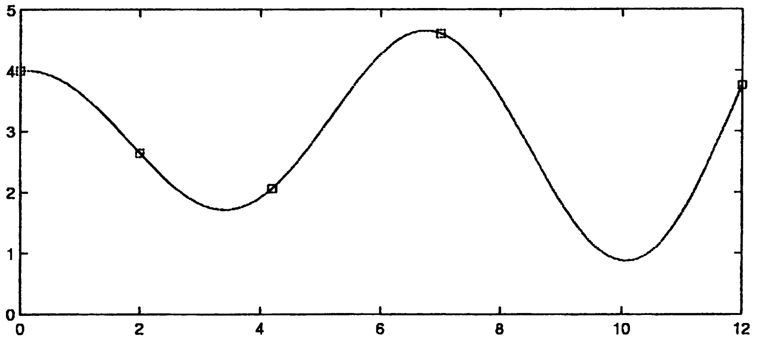

Figure 3. Actual response and sampled points.

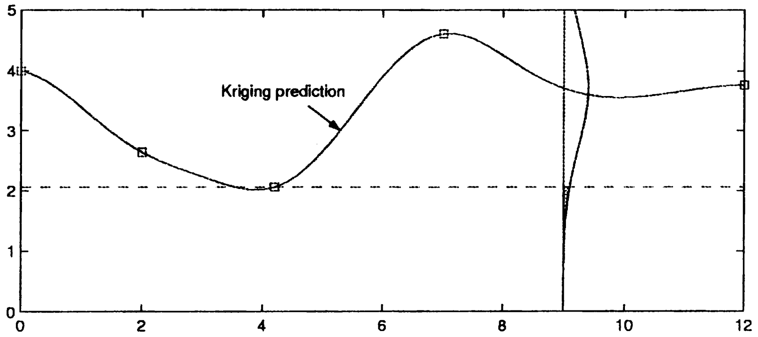

Figure 4. Kriging prediction.

Formally, the expected improvement at $\mathbf{x}$ is defined as

$$E[I(\mathbf{x})] = (y_{\min} - \hat{y})\Phi\left(\frac{y_{\min} - \hat{y}}{s}\right) + s\phi\left(\frac{y_{\min} - \hat{y}}{s}\right) \tag{9}$$

where ϕ and Φ are the standard normal density and distribution functions respectively and $\hat{y}$ and s are defined using ordinary kriging, see Jones et al. (1998). The strategy in this case is to resample where the expected improvement is maximized, then form a new approximation and repeat.

Here we consider derivative based approximations: the strategy remains identical to the above, the difference being that the prediction comes from (7) and the standard error comes from (8). The prediction is shown in Figure 5 and is seen to be much more accurate. The Gaussian distribution is again shown at $x = 9$, this time using (7) and (8). It is clear that improving the prediction will, in general, increase the rate of convergence of the procedure.

The overall strategy in this case is given in Figure 6. This approach has been termed Efficient Global Optimization or EGO for short.

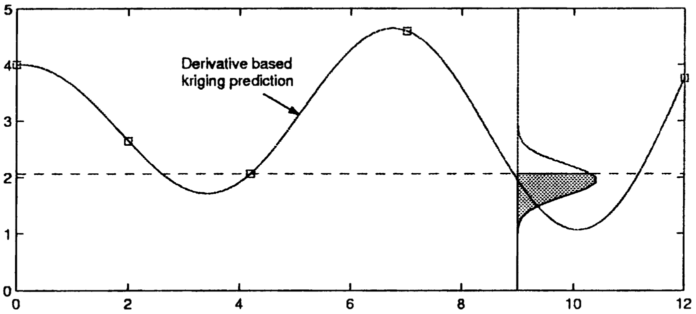

Figure 5. Kriging prediction (including derivative information).

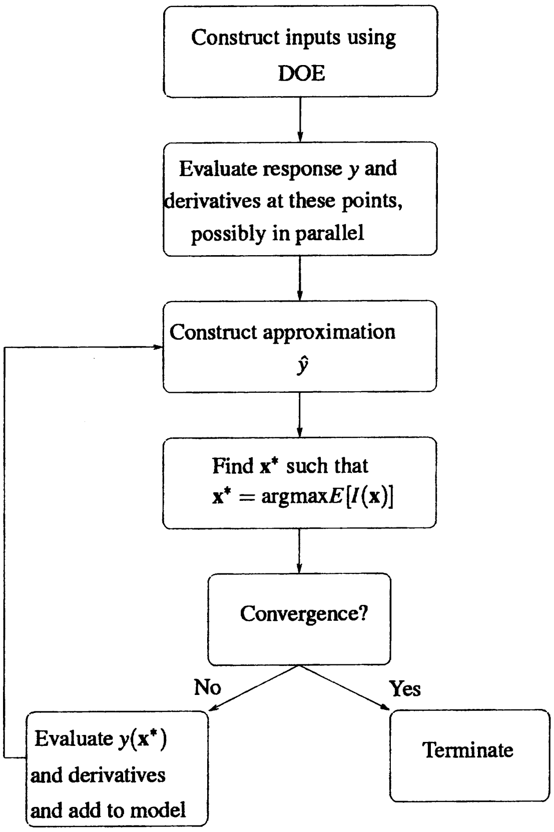
Construct inputs using DOE
Evaluate response y and derivatives at these points, possibly in parallel
Construct approximation $\hat{y}$
Find $\mathbf{x}^*$ such that $\mathbf{x}^* = \mathrm{argmax}E[I(\mathbf{x})]$
Convergence?
No
Yes
Evaluate $y(\mathbf{x}^*)$ and derivatives and add to model
Terminate

Figure 6. The expected improvement strategy.

4. Results

4.1. EXAMPLE 1

The first example considered is the Branin function (Dixon and Szego, 1978), a well known test problem in global optimization.The function is defined as

$$y(x_1,x_2)=a(x_2-bx_1^2+cx_1-d)^2+e(1-f)\cos(x_1)+e \tag{10}$$

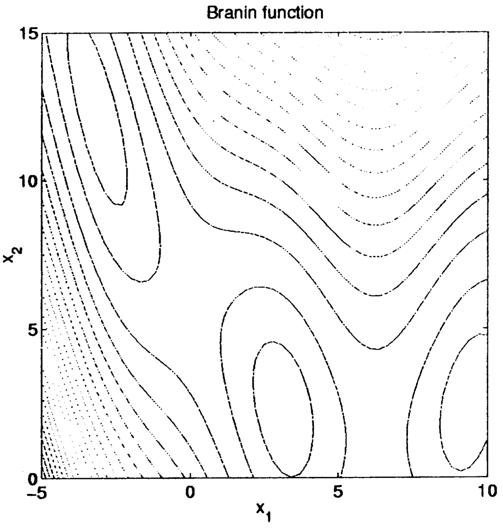

Figure 7. The Branin function.

where $a=1, b=5.1/(4\pi^2), c=5/\pi, d=6, e=10$, and $f=1/(8\pi)$. The function has three global minima occuring at $(x_1,x_2)=(9.42478,2.475),(-\pi,12.275), (\pi,2.275)$ with function value $y(x_1,x_2)=0.397887$ at all three locations. A contour plot of this function in shown in Figure 7.

We consider the initial latin hypercube design shown in Figure 8 (left) consisting of seven computer experiments. Using the direct optimization algorithm (Figure 2) we then perform ten further function evaluations: results are as shown in Table 1. For the derivative based EGO algorithm (Figure 6) we also perform 10 additional function evaluations and the results are as shown in Table 2. Contour plots of the resulting approximations and successive optima for these approaches are given in Figure 9.

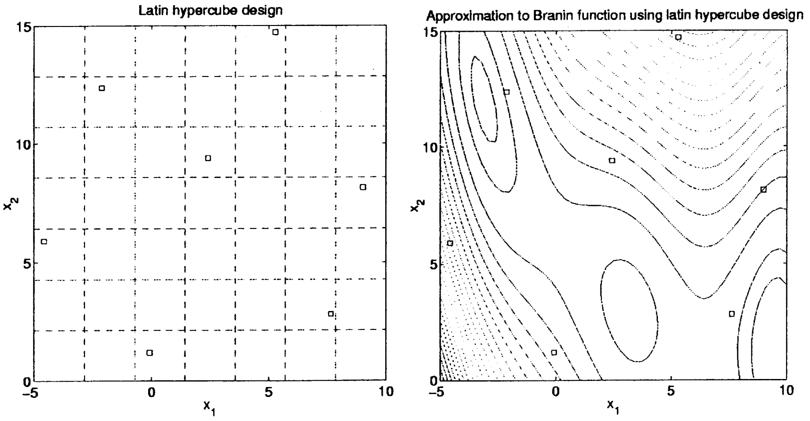

Figure 8. Experimental design and initial approximation.

Table 1. Results of direct optimization strategy with derivative information.

Iteration	x_1	x_2	$y(x_1,x_2)$
1	9.5368	1.3163	2.0328
2	−3.0007	11.6592	0.5713
3	9.3993	2.4642	0.4011
4	9.4161	2.4814	0.3984
5	9.4214	2.4841	0.3981
6	9.4214	2.4811	0.3980
7	−3.1984	12.5480	0.4319
8	−3.1376	12.2568	0.3980
9	−3.1183	12.2094	0.4006
10	9.4238	2.4750	0.3979

Clearly this process could be continued until some rigorous convergence criterion was met. For example, when using direct optimization, we may stop when either $||y^n_{\text{min}} - y^{n+1}_{\text{min}}|| < \varepsilon_y$ or when $||\mathbf{x}^n_{\text{min}} - \mathbf{x}^{n+1}_{\text{min}}|| < \varepsilon_{\mathbf{x}}$ where n and $n+1$ refer to the results of two consecutive optimizations and ε_y, $\varepsilon_{\mathbf{x}}$ are small positive constants, whereas in the EGO algorithm we may stop when the expected improvement is less than 1% of the current best function value as in Jones et al. (1998). Note that in Table 2 this condition has already been met and that this number of evaluations (17 in total) is less than the 28 evaluations reported in Jones et al. (1998). Of course, this is to be expected since we are now incorporating derivative information into our approximation. Note also that even if the derivative calculation that provides all the gradient information at each iteration costs the same as a single function evaluation this approach is actually more expensive than the basic strategy.

The results (Figure 9, left) show that in the direct optimization, the approximation quickly focuses on two of the global minima, however it misses the third. This algorithm potentially suffers from the drawback that it will become stuck in a local minimum. The algorithm is most probably useful when a good solution is required quickly.

Table 2. Results of EGO with derivative information.

Iteration	x_1	x_2	$y(x_1,x_2)$
1	9.5682	1.2335	2.3600
2	−2.9945	11.6441	0.5801
3	9.3946	2.4666	0.4025
4	−4.1000	15.0000	4.5723
5	3.1317	2.1300	0.4217
6	3.1004	2.4003	0.4147
7	−3.2000	12.6000	0.4482
8	3.1215	2.3638	0.4052
9	3.1500	2.2711	0.3982
10	9.4233	2.5125	0.3994

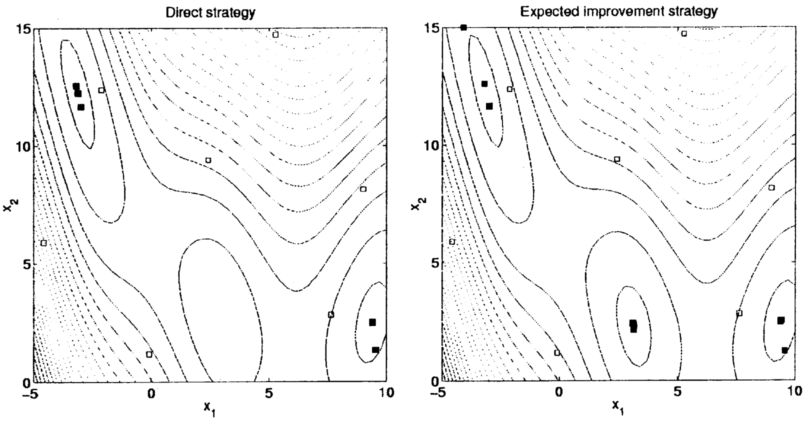

Figure 9. Successive minima and final approximations.

On the other hand the expected improvement algorithm will sample at all points it thinks are interesting (in Figure 9, right, we see the algorithm sampling around each of the global minima) and this approach is more likely to obtain the global minimum, but at the expense of extra function evaluations. One should note here that the expected improvement function is by definition more strongly multimodal than the underlying function: a plot of the expected improvement function for the initial approximation (seven data points) is given in Figure 10. As a result, a suitable optimization algorithm should be considered. Here simulated annealing is used to find the point that maximizes expected improvement.

We could even consider a hybrid approach, begin with EGO and later switch to direct optimization.

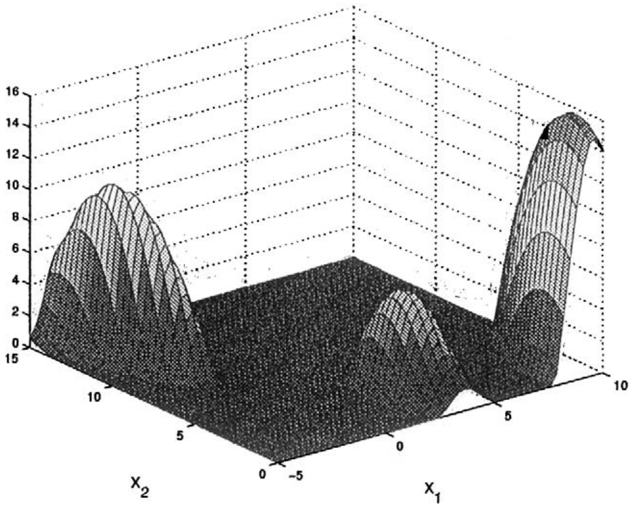

Figure 10. Plot of the expected improvement for seven data points.

4.2. EXAMPLE 2

Structural optimization is of great engineering interest, here the problem is one of performing optimization subject to certain constraints. How these constraints are treated within an approximation method is now addressed. Several potential strategies exist: one possibility would be to reduce the problem to one of unconstrained optimization using penalty functions, see e.g., Fox (1971).

In this paper, however, the objective and constraints are modelled separately and we carry out the optimization with respect to the approximate objective using, e.g., the direct or the EGO approach subject to the approximate constraints. As the iteration proceeds, the accuracy of both the objective and the constraints in interesting areas of the design space is increased.

The example we consider is a simple beam problem. The cross sectional parameters (breadth and height) of a cantilever beam subjected to a uniformly distributed loading are varied, the requirement here is to minimize the volume of the cantilever beam subject to certain constraints. Let x_1 define the breadth and x_2 define the height of the cross section. The objective is therefore given as

$$y = Lx_1x_2 \tag{11}$$

where L is the length of the beam. This is to be minimized while satisfying the stress constraint

$$\sigma_{\max} \leqslant 250 \text{ N/mm}^2. \tag{12}$$

The objective here is unimodal and there is unlikely to be any difference between the direct approach and EGO, we therefore arbitrarily choose to perform optimization using the direct approach. Here we evaluate the objective (and its derivatives) using the analytic expression given in (11). Stress evaluations (and derivatives) come from a finite element model.

We consider a relatively fine mesh (100 element) finite element model of the above. The objective is cheaply available and can always be evaluated exactly; only the stress constraint requires solving the resulting system of equations at some expense. Note that, in general, only expensive responses should be approximated; this issue is addressed in an example later in the paper. The purpose at this stage is to demonstrate the general algorithm assuming that both the objective and constraints are to be modelled.

The actual objective and constraint are shown in Figure 11 (left). We note the minimum of 6.9281 occurs at (0.5, 0.69281). A latin hypercube design consisting of three samples is taken as our initial design, a point is then added at the minimum of the current approximation. Figure 11 (right) shows the initial approximation. Table 3 shows the result obtained using the derivative based approach.

As the algorithm proceeds, we approach the $\mathbf{x}^*$ corresponding to the minimum y and the approximations to volume and stress become more accurate. This is shown in Figure 12.

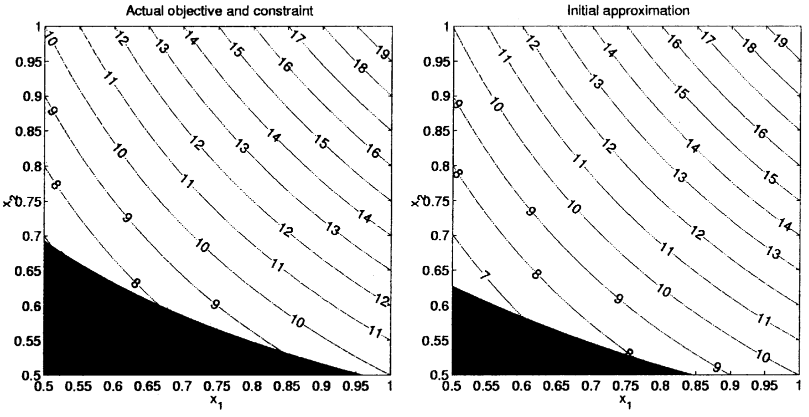

Figure 11. Objective and constraint for the beam problem (actual and first approximation).

Table 3. Results for example 2.

Minimum	x_1	x_2	$\hat{\sigma}$	$\widehat{V}$	σ	V
1	0.50011	0.62722	250.00	6.254	304.96	6.274
2	0.50062	0.69117	250.00	6.913	250.89	6.920
3	0.50004	0.69256	250.00	6.922	250.17	6.926
4	0.50003	0.69263	250.00	6.924	250.12	6.927
5	0.50000	0.69269	250.00	6.927	250.09	6.927

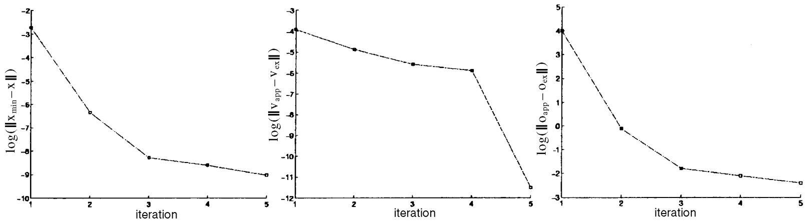

Figure 12. Convergence towards the predicted minimum and accuracy of volume and stress approximations as the iteration proceeds.

Once again, the use of derivatives critically depends on their cost: if the derivatives are relatively cheap then the approach is justified. If, however, the derivatives are obtained by finite differencing the expensive function their use will rarely make sense.

We note that when considering the EGO algorithm it is important to note that, in the constrained case, $y_{\min}$ should be taken as the minimum of the feasible sampled responses (to see this consider the objective and constraint shown in Figure 11 (left), if we sample at (0.5, 0.5) then this point is infeasible, however, $y_{\min}$ will occur here and as a result the expected improvement would be virtually zero over the entire feasible region).

5. Multifidelity Modelling

We now turn our attention to the use of approximation techniques in combining a large number of low fidelity analyses together with a small number of high fidelity analyses for constructing approximations that are both cheap and accurate. The general strategy is to consider a low fidelity model f_a together with only selective calls to the high fidelity model f_e. The inputs for the selective calls to f_e are once again chosen using some design of experiments approach.

The low fidelity model provides some rough global information as to the response of the high fidelity model. Selective calls to f_e are then made to build corrections to the low fidelity model. We here once again consider the case when derivative information is cheaply available. The general strategy is shown in Figure 13 where it is assumed that f_a can be calculated at negligible cost.

There is a choice on how to proceed in steps 3 and 4 of this algorithm; consider the latter first. If the information provided by the low fidelity model is relatively

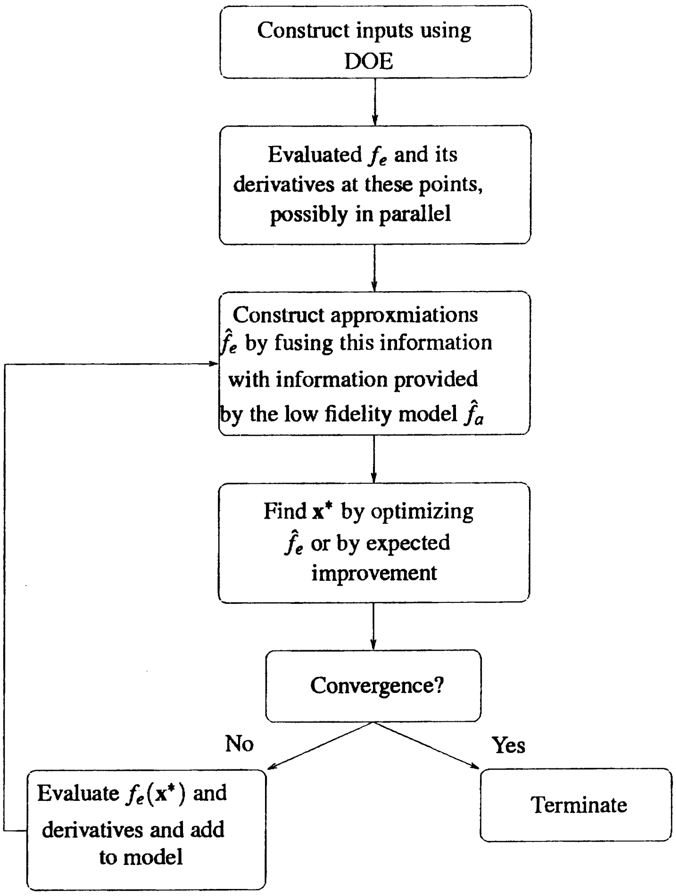
Construct inputs using DOE
Evaluated f_e and its derivatives at these points, possibly in parallel
Construct approxmiations $\hat{f}_e$ by fusing this information with information provided by the low fidelity model $\hat{f}_a$
Find $\mathbf{x}^*$ by optimizing $\hat{f}_e$ or by expected improvement
Convergence?
No
Yes
Evaluate $f_e(\mathbf{x}^*)$ and derivatives and add to model
Terminate

Figure 13. The multifidelity modelling strategy.

accurate, then direct optimization would suffice, as expected improvement would most likely give a similar result because the difference or ratio of the models will be of low multimodality. If there is some lack of correlation between the two models then an expected improvement approach would be more likely to produce a global minimum. We expect reasonable correlation between the two models f_e and f_a if they are representing the same physical system. Hence a direct approach should suffice. Nevertheless, the reader should be aware that some lack of correlation between these models would increase the possibility of the expected improvement approach producing a better solution.

There are many ways in which the low fidelity model can be included inside the approximation. We here opt for a ratio model, rather than approximate f_e alone we approximate f_e/f_a. This is known at the sampled points, as are the derivatives

$$\frac{\partial r}{\partial x_l} = \frac{f_a \frac{\partial f_e}{\partial x_l} - f_e \frac{\partial f_a}{\partial x_l}}{f_a^2}, \quad l = 1, \ldots, k. \tag{13}$$

Hence a gradient based approximation $\hat{r}$ can be constructed and $\hat{r} f_a$ can be used as a surrogate for f_e. When there is good correlation between f_a and f_e (as there should be, the models represent the same physical system) this can dramatically improve the accuracy of the surrogate model, particularly at extrapolated points. We assume here that $f_a \neq 0$: if this situation should occur then a vertical shift of the cheap response is necessary. This does not affect the correlation between the two models.

Other possibilities for incorporating a low fidelity in the approximation include a difference approach (Leary et al., to appear; Watson and Gupta, 1996), an approach including weighted low fidelity models (Leary et al., submitted for review) and a global space mapping approach (Leary et al., to appear).

These models incorporating derivatives will obviously be more accurate than standard multifidelity strategies that exclude derivative information. Whether they will yield more efficient strategies for finding global optima than those that do not make use of gradients is by no means obvious. Any of the above approaches could be used to demonstrate the effectiveness of utilizing the low fidelity model in our prediction of f_e. We arbitrarily choose to model the ratio in the example that follows.

5.1. DEMONSTRATIVE EXAMPLE

As a final example we consider the mechanical structure shown in Figure 14. In this example, we consider the length of each beam to be 1m. The upper horizontal beams are subjected to a uniformly distributed loading and we wish to minimize the weight of the structure by varying the thickness of the beams. We here consider a three variable problem where the beams are of square cross-section. The three design variables relate to the cross-sectional thicknesses of the lower

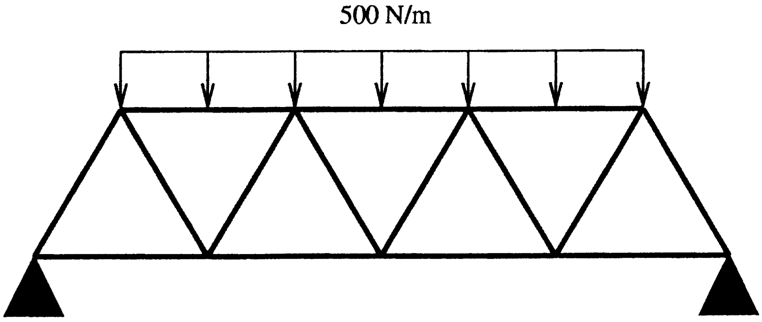

Figure 14. The problem.

horizontal beams, diagonal beams and upper horizontal beams and these are all varied between 0.05 and 0.1 m.

The minimization would normally be carried out subject to stress and stiffness constraints, we here consider modelling the stress which is output from a finite element program. We note that in a static finite element analyses such as those in the examples presented here, the derivatives are available at a fraction of the cost of one finite element solution.

The problem is analysed using a simple finite element beam model. Two levels of complexity are considered, a 'low fidelity' model consisting of just two elements per beam and a 'high fidelity' model consisting of 10 elements per beam. The objective (volume) is cheap to calculate and can be evaluated exactly. The stress which forms the constraint is more expensive to evaluate, requiring solution of the finite element model. It is this variation in stress that we attempt to model using various approaches.

The low fidelity model response is well correlated with the high fidelity response, nevertheless, there is an error between the models, hence direct optimization of the low fidelity model would yield an inaccurate result. Therefore improved approximations are sought next.

To assess the overall accuracy, approximations are constructed using a design of experiments using only eight calls to the high fidelity model. Table 4 shows the average percent error and maximum percent error in the stress for the various

Table 4. Errors in stress for various surrogate models of the structure in Figure 14 based on approximating the response at an eight point design of experiments.

Model	Av. % error	Max. % error
Low. fid.	18.04	26.12
Kriging	45.73	92.01
Kriging (inc. deriv.)	23.87	63.04
Multifidelity model	3.17	11.99
Multifidelity model (inc. deriv.)	2.66	11.81

approaches. These were evaluated on a $6 \times 6 \times 6$ grid of points spanning the domain of interest.

These results highlight the use of multifidelity approximations within the optimization process: if the low fidelity model response is reasonably correlated with the high fidelity model response (which we hope it is – they represent the same physical system) then great improvements in the accuracy of the approximations can be made. This would be particularly true in high dimensions when training data are sparse. Further, if these approximations are to be used for optimization, the increased accuracy will lead to further improvements in the efficiency of this process.

6. Conclusions

In this paper, several strategies for approximating outputs from complex engineering codes are described. If sensitivity information is cheaply available, this can be included in the approximation. A kriging model including derivative information is thus used for defining the approximation. If this sensitivity information is available, including it within the approximation allows comparable results (to approaches that exclude this information) to be achieved with fewer training data. Whether or not including such information is justified crucially depends on the relative costs of calculating the functions involved and their derivatives.

Two strategies for optimization have been described, the first of these updates the approximate model based on the best point in the current approximation. This is useful if a relatively good solution is to be found quickly, but in general it will not find a global minimum in a highly multimodal problem. The second procedure is based on the notion of expected improvement and provides an approach which, whilst involving more calls to the original model, is more likely to find the global minimum. As noted earlier, some hybrid approach may also be appropriate perhaps using multiple population based updates if suitable parallel processing hardware is available.

The incorporation of low fidelity models in these approximations is also described. Once again, optimization can be based on either of the two criteria defined. Including these low fidelity models can dramatically increase the accuracy of the approximation, hence it can allow the optimization process to terminate after fewer calls to the high fidelity model, increasing the efficiency of the approach.

It has been assumed here that derivatives are available at very little cost, this is the case in the static finite element analyses considered. In, e.g, an adjoint CFD calculation, to calculate all the derivatives requires a similar computational cost to that of one expensive function evaluation. The effectiveness of the algorithms under these circumstances remains to be seen: further work on this type of problem will be undertaken in the future. These approaches have been demonstrated on small scale problems where the problems have been kept simple for ease of visualization. We plan to apply the approaches proposed herein to more complex engineering problems (particularly in higher dimensions) in the near future.

Acknowledgements

We thank Rolls-Royce Plc for the financial support received as a part of the University Technology Partnership (UTP) in design.

References

Alexandrov, N.M., Dennis Jr, J.E., Lewis R.M. and Torczon, V. (1998), A trust-region framework for managing the use of approximation models in optimization, *Structural Optimization*, 15, 16–23.

Bakr, M.H., Bandler, J.W., Biernacki, R.M., Chen, S.H. and Madsen K. (1998), A trust region aggressive space mapping algorithm for EM optimization, *IEEE Transactions on Microwave Theory and Techniques*, 46, 2412–2425.

Bakr, M.H., Bandler, J.W., Georgieva, N. and Madsen, K. (1999), A hybrid aggressive space mapping algorithm for EM optimization, *IEEE Transactions on Microwave Theory and Techniques*, 47, 2440–2449.

Balabanov, V., Haftka, R.T., Grossman, B., Mason, W.H. and Watson, L.T. (1998), Multifidelity response surface model for HSCT wing bending material weight, *7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, St. Louis, MO, AIAA Paper 98-4804*, pp. 778–788.

Bandler, J.W., Biernacki, R.M., Chen, S.H., Grobelny, P.A. and Hemmers, R.H. (1994), Space mapping technique for electromagnetic optimization, *IEEE Transactions on Microwave Theory and Techniques*, 42, 2536–2544.

Chang, K.J., Haftka, R.T., Giles, G.L. and Kao, P.-J. (1993), Sensitivity-based scaling for approximating structural response, *Journal of Aircraft*, 30, 283–287.

Dixon, L.C.W. and Szego, G.P. (1978), The global optimization problem: an introduction, In: Dixon, L.C.W. and Szego, G.P. (eds.), *Towards Global Optimization*, Vol. 2, North Holland, Amsterdam, 1–15.

Fox, R.L., (1971), *Optimization Methods for Engineering Design*, Reading, MA: Addison-Wesley.

Hutchinson, M.G., Unger, E.R., Mason, W.H., Grossman, B. and Haftka, R.T. (1994), Variable-complexity aerodynamic optimization of a high speed civil transport wing, *Journal of Aircraft*, 31, 110–116.

Jones, D.R., Schonlau, M. and Welch, W.J. (1998), Efficient global optimization of expensive black-box functions, *Journal of Global Optimization*, 13, 455–492.

Kushner H.J., (1964), A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise, *Transactions of the ASME. series D Journal of Basic Engineering*, 86, 97–106.

Leary, S.J., Bhaskar, A. and Keane, A.J. A constraint mapping approach to the structural optimization of an expensive model using surrogates, *Optimization and Engineering*, to appear.

Leary, S.J., Bhaskar, A. and Keane, A.J. A knowledge-based approach to response surface modelling in multifidelity optimization, *Journal of Global Optimization*, submitted for review.

Mackay, M.D., Beckman, R.J. and Conover, W.J. (1979), A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, 21, 239–245.

Mead, R. (1988), *The Design of Experiments*, Cambridge University Press, Cambridge.

Mockus, J., Tiesis, V. and Zilinskas, A. (1978), The application of Bayesian methods for seeking the extremum. In: Dixon, L.C.W. and Szego, G.P. (eds.), *Towards Global Optimization*, Vol. 2, North Holland, Amsterdam, pp. 117–129.

Morris, M.D., Mitchell, T.J. and Ylvisaker, D. (1993), Bayesian design and analysis of computer experiments: use of derivatives in surface prediction, *Technometrics*, 35, 243–255.

Myers, R.H. and Montgomery, D.C. (1995), *Response Surface Methodology: Process and Product Optimization using Designed Experiments*, John Wiley and Sons, New York.

Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, H.P. (1989), Design and analysis of computer experiments (with discussion), *Statistical Science*, 4, 409–435.

Sobol, I. M., (1979), On the systematic search in a hypercube, *SIAM Journal of Numerical Analysis*, 16, 790–793.

Vitali, R., Haftka, R.T. and Sankar, B.V. (1999), Multifidelity design of a stiffened composite panel with a crack, *4th World Congress of Structural and Multidisciplinary Optimization*, Buffalo, NY.

Wang, F. and Zhang, Q. (1997), Knowledge-based neural models for microwave design, *IEEE Transactions on Microwave Theory and Techniques*, 45, 2333–2343.

Watson, P.M. and Gupta, K.C. (1996), EM-ANN models for microstrip vias and interconnects in dataset circuits, *IEEE Transactions on Microwave Theory and Techniques*, 44, 2495–2503.

White, H., Gallant, A.R., Kornik, K., Stinchcombe, M. and Wooldridge, J. (1992), *Artificial Neural Networks: Approximation and Learning Theory*, Blackwell, Oxford.