SAND99-0363C

A Resampling Based Approach to Optimal Experimental Design for Computer Analysis of a Complex System

Brian Rutherford (bmruthe@sandia.gov)
Statistics and Human Factors Department (12323)
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0829

Key Words: Computer-aided design; Experimental design; Large-scale simulation; Uncertainty analysis.

Abstract:

The investigation of a complex system is often performed using computer generated response data supplemented by system and component test results where possible. Analysts rely on an efficient use of limited experimental resources to test the physical system, evaluate the models and to assure (to the extent possible) that the models accurately simulate the system under investigation. The general problem considered here is one where only a restricted number of system simulations (or physical tests) can be performed to provide additional data necessary to accomplish the project objectives. The levels of variables used for defining input scenarios, for setting system parameters and for initializing other experimental options must be selected in an efficient way. The use of computer algorithms to support experimental design in complex problems has been a topic of recent research in the areas of statistics and engineering. This paper describes a resampling based approach to formulating this design. An example is provided illustrating in two dimensions how the algorithm works and indicating its potential on larger problems. The results show that the proposed approach has characteristics desirable of an algorithmic approach on the simple examples. Further experimentation is needed to evaluate its performance on larger problems.

Introduction

Many applications in science and engineering have turned to computational simulation -- the use of large physics-based computer codes -- to supplement physical test data for systems analysis. The use of simulation codes introduces a host of new error sources into an analysis. These sources range from conceptual and mathematical modeling uncertainties to convergence and roundoff errors resulting from discretization of the (generally) continuous conservation equations, and initial and boundary conditions (see Oberkampf, et al 1999). Included in this list is the modeling uncertainty introduced when a response surface is constructed to model the system response at points in the input space (that aren't evaluated through the computational simulation model). It is the efficient reduction of this "response modeling uncertainty" through selection of simulation runs that provides the focus of the computer experimental design. A method for accomplishing this selection is discussed and illustrated in this paper.

In An Ideal Computer World

Consider Figure 1. This figure illustrates two general problems often approached using computational simulation and Figure 2 shows how response model uncertainty might impact each type of analysis.

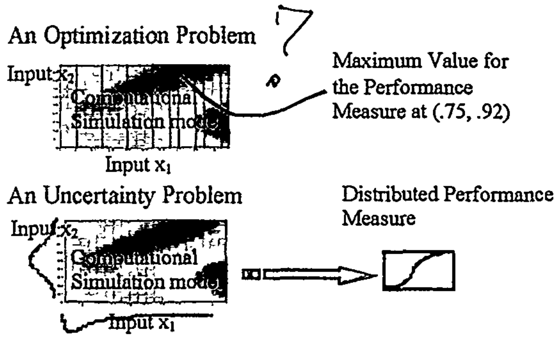
An Optimization Problem
Input x_2
Computational Simulation model
Input x_1
Maximum Value for the Performance Measure at (.75, .92)
An Uncertainty Problem
Input x_2
Computational Simulation model
Input x_1
Distributed Performance Measure

Figure 1

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Figure 1 indicates how, in an ideal computing world, problems involving optimization (used in engineering design) or involving uncertainty analyses (used for prediction) could be solved by setting up a grid across the input space, making the necessary calculations and either choosing the optimal value (design) or integrating numerically across the response with respect to the input distributions to calculate the distribution of the desired performance criteria (prediction).

In A Less Ideal Computer World

Computational simulation using large physics-based codes is, however, an expensive process. A "high fidelity" code can require days of run time, even when executed on the fastest computer system. As the computer systems become more sophisticated, so do the simulation codes – incorporating more detail of the physical processes and their interactions. Because the codes are so expensive to run, the number of simulations that can be performed is limited. Very often, the system responses are modeled using data from these computational simulation runs and it is the response model that is then used through the remainder of the analysis. Figure 2 illustrates how additional uncertainty might be introduced in this situation. For the design problem, the exact optimum would not be known. For the prediction problem, an additional component of uncertainty would be introduced.

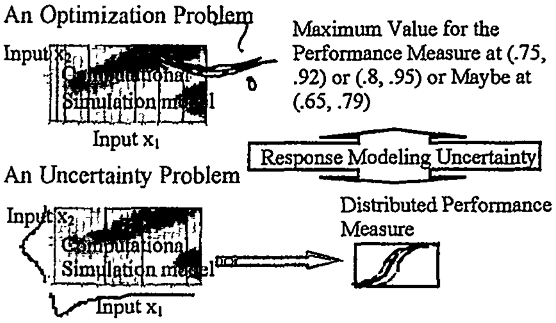
An Optimization Problem
Input x₂
Computational Simulation model
Input x₁
Maximum Value for the Performance Measure at (.75, .92) or (.8, .95) or Maybe at (.65, .79)
Response Modeling Uncertainty
An Uncertainty Problem
Input x₂
Computational Simulation model
Input x₁
Distributed Performance Measure

Figure 2

Taking this view, the computational simulation model is executed using inputs that provide appropriate computer data for constructing the response surface. The response surface is a spatial model, defined throughout the input space, that is based on previous simulations and physical test results. The response surface is used together with inputs that may or may not be probabilistically described to perform the remainder of the analysis and evaluate the performance criterion. It is precision in the performance criterion that should dictate the additional simulation runs to perform (i.e. the computer experimental design). Note three clear features of a good choice of design:

1. The computed responses should have an impact on the performance criterion -- for example, in a reliability problem, regions of the input space where the response is clearly "safe" or clearly "failed" are of little interest, even if these regions have high response model uncertainty.
2. Runs in regions of high modeling uncertainty provide more information -- computer runs, "close" to previous experimentation where the response is known, provide mostly redundant information.
3. Runs in regions of higher relative probability in uncertainty analyses are more likely to impact the performance measures -- hence, resolving response modeling uncertainty in these areas will generally increase performance measure precision.

These criteria can (and will in the Example Section) be used on a problem of low dimension to evaluate the selected experimental design.

The Resampling Experimental Design Methodology

The approach to computer experimental design pursued here consists of two basic steps. First, the response is characterized by a probability measure based on data from previous computer runs and physical testing. Second, an algorithm is executed for candidate experimental designs that evaluates their potential for yielding relevant information and increasing precision in the estimate of the performance measure. The design that indicates the most potential is selected for the computational simulation runs.

Estimating a Probability Measure over Responses

Data from computer analyses and physical test data give response values at specific locations in the input space. In general, however, the performance criteria are evaluated using information throughout this space. The step of constructing a probabilistic representation of response values throughout the input space in a manner consistent with the sample data is a focus of our current research. The approach that is taken for results provided in this paper is to use a model of the form:

$$r^*(\mathbf{X}) = \beta_o + \sum_i \beta_i x_i + \sum_i \sum_j \beta_{ij} x_i x_j + \varepsilon(\mathbf{X}), \qquad (1)$$

where i and j range over the inputs. Here, ε is a stochastic process over the inputs **X** defined through a spatial covariance function:

$$\gamma(\mathbf{x}_i,\mathbf{x}_j) = \gamma(\|\mathbf{x}_i - \mathbf{x}_j\|) \text{ for any } i \text{ and } j.$$

This model has been used successfully for applications in geology and hydrology for several years. It was introduced as a model for engineering analyses in Sacks, et al (1989), and is discussed in detail in O'Connell and Wolfinger (1997). The model (1) describes a response surface defined over the input space. Our objective is to use response surfaces like this to construct a discrete approximation to a probability measure. This is accomplished here by constructing an ensemble of 'realizations' of the form (1) where the fraction of realizations in any interval of the response space is roughly consistent with the probability of a response in this interval according to relative likelihoods of the model parameters based on the original simulation and test data.

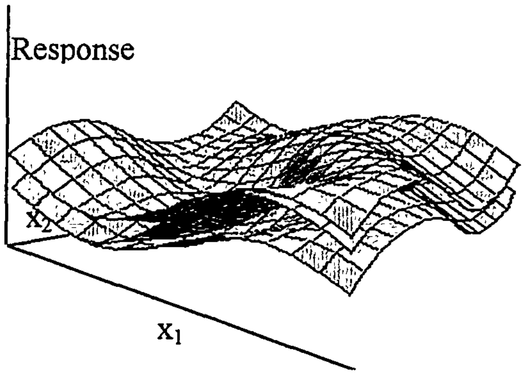

Figure 3

Figure 3 shows three possible members of such an ensemble. Note that individual evaluation of each realization provides an estimate of the performance criterion yielding an optimal value (for the optimization problem in Figure 1) or a performance measure distribution (for the uncertainty problem). In the latter case, considering this distribution for each realization separately, one can partition the uncertainty between response modeling and input-based uncertainty propagated through the analysis.

The method used here to generate the realizations is also borrowed (in part) from the geosciences. First, a polynomial is generated randomly using the fitted polynomial coefficients and the coefficient correlation structure assuming normality. Next, a stochastic surface is added to this estimate to complete the realization. The stochastic surfaces are constructed using the Sequential Gaussian Approach, see Deutch and Journel (1998), with a covariance model fitted to the residuals of the polynomial. The realizations are conditioned to fit exactly the residual values of the generated quadratic surface, and consequently, all realizations provide an exact fit to the original data. This approach to estimating a probabilistic representation of the response is one of several being considered -- it is the approach used in the example problem presented in the final section of this paper.

A Resampling-Based Experimental Design Algorithm

The algorithmic portion of the proposed methodology is illustrated through the flowchart in Figure 4. Inputs to the algorithm are the probabilistic representation of the response discussed in the previous section and a method for selecting candidate designs. Output from the algorithm is the experimental design -- sets of inputs to be run using the computational simulation model. The selection of candidate designs is not discussed in detail here. At worst, the designs could be generated randomly using the input probability distributions where applicable. Efficiency, however, is greatly improved when random search methods or branch and bound methods are employed to select candidate designs. For the example problems, an evolutionary algorithm was used to generate the candidates.

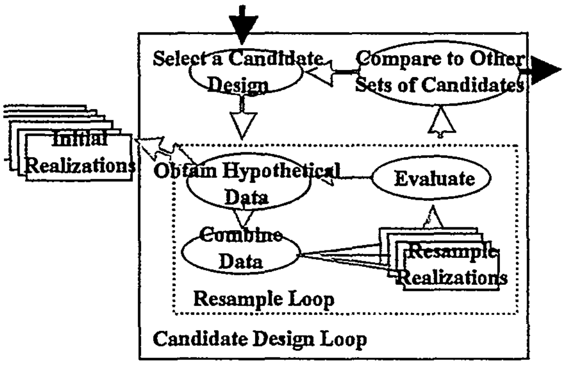

Figure 4

The outer loop in Figure 4 is the candidate-design loop. Candidate designs are selected, evaluated through the inner loop and compared to determine their "information potential". A discussion of the inner loop is required for a meaningful description of the metrics used for comparison. The inner loop is the resample loop. The processes involved here are similar to the construction of an approximate probability measure discussed in the previous section. Each time around this loop, one realization from the initial ensemble is selected. Hypothetical data at the candidate design points are taken from the selected response surface and combined with the actual data. The augmented data set is evaluated in the same way the initial data were, and the resulting distribution(s) of the performance measure are computed. This provides an estimate of the (remaining) response model uncertainty, valid if that design were chosen, and the response was as indicated by the selected initial realization.

This process is repeated using the other realizations in the ensemble to obtain the hypothetical data at the design points, each time yielding an estimate of (post computer runs) response model uncertainty. Each time around the loop provides an estimate of the performance measure distribution conditioned on the design data resampled on that loop. Differences between these distributions on different loops provide an indication of the discriminatory potential of the design.

For the example given in the next section it is an expectation that is of interest when evaluating the performance criterion. Hence, as described above, we can talk about the within resample variance of the expectations:

$$\hat{\sigma}_w^2 = \frac{1}{k'}\sum_{i=1}^{k'}\left(E(pm_i) - \overline{E(pm)}\right)^2$$

where there are k' realizations in the inner loop evaluation. We can also estimate the between resample variance of the expectations:

$$\hat{\sigma}_b^2 = \frac{1}{k}\sum_{j=1}^{k}\left(\overline{E(pm_j)} - \overline{\overline{E(pm)}}\right)^2$$

where there are k realizations in the original ensemble or where that ensemble is resampled k times. Note that an uninformative design would yield roughly $\hat{\sigma}_w^2 = \hat{\sigma}^2/k'$ and $\hat{\sigma}_b^2 = \hat{\sigma}^2/k'k$ where $\hat{\sigma}^2$ is the variance of the performance measure attributed to input uncertainty. The ratio $\hat{\sigma}_b^2/\hat{\sigma}_w^2 = 1/k$ for the uninformative design provides an indication of the information provided by the design. A good design, however, would yield a substantially higher value for $\hat{\sigma}_b^2$ and a (slightly) lower value for $\hat{\sigma}_w^2$. It is this ratio that is used to select the best design in the outer loop for the example problem. Different (but usually fairly simple) metrics are required for other types of problems and performance measures.

Example

A simple analytical example in two dimensions is provided here to demonstrate how this approach works. Figure 5 shows the true (but generally unknown) response $r(x_1, x_2)$ and a plot of the average realization based on a sample taken (from the true surface) at ten input locations.

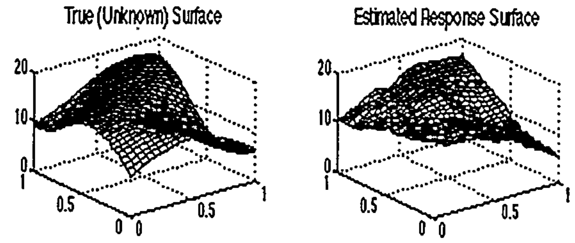

Figure 5

For any value of x_2, the performance measure was computed as:

$$pm(x_2) = \int I\left(r^*(x_1, x_2) > 10\right) * \left(r^*(x_1, x_2) - 10\right)^2 dx_1 .$$

The distribution function of the performance measure was then computed assuming a uniform distribution for x_2 as:

$$F_{pm}(z) = \int I\left(pm(x_2) > z\right) dx_2$$

Figure 6 shows mean and standard deviation contours computed from the fifteen initial generated realizations.

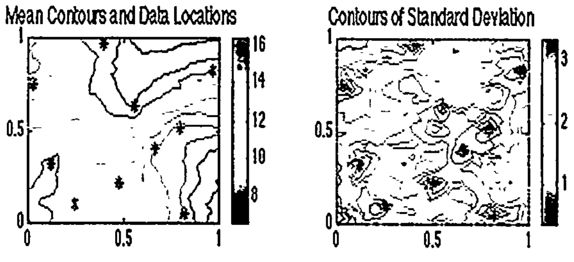

Figure 6

Six of these realizations are plotted in Figure 7.

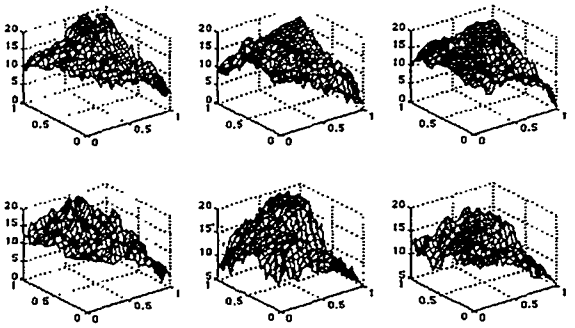

Figure 7

Figure 8 shows the fifteen probability distribution functions associated with these realizations together with the density function calculated for their equally weighted mixture. This latter (darker) cumulative distribution function includes uncertainty propagated through to the performance measure from the distributed input x_2 and response modeling uncertainty. From these density functions, we compute $\hat{\sigma}_m^2 = ??$, where $\hat{\sigma}_m^2$ is the estimated mixture variance -- a quantity similar to $\hat{\sigma}_w^2$, but based only on the initial data.

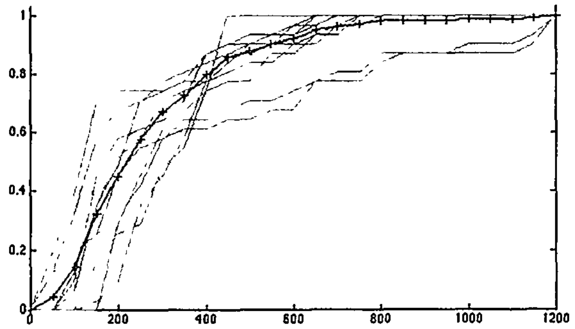

Figure 8

The algorithm described in the previous section was run to select a 3-point design. Figure 9 illustrates this design on the contour maps of mean and standard deviations together with data from the initial runs. Note that this selection tends to support the first two criterion listed earlier for a good experimental design. The design points are at input locations that are of relatively high value in modeling uncertainty and appear to have a significant impact on the performance measure. The third criterion is not tested in this example because of the uniform distribution selected for x_2.

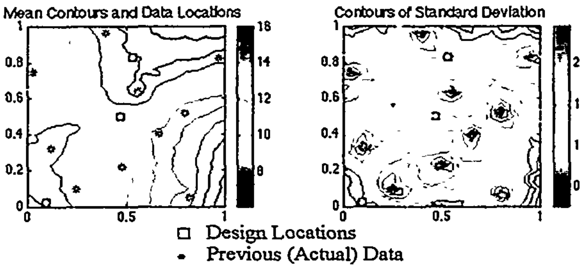

Figure 9

Summary

A resampling-based approach to the algorithmic selection of input locations for computer experimentation was described and illustrated. Results on a simple analytical example demonstrate how the methodology might be used. These encouraging results suggest that this approach might be of significant value for higher dimensional problems where the complexities make a solution based entirely on analyst's intuition impossible.

References

Deutsch C. V. and Journel, A. G. (1998), "GSLIB Geostatistical software Library and Users Guide Second Edition", Oxford University Press, New York.

Oberkampf W. L., DeLand S. M., Rutherford B. M. Diegert K. V., and Alvin K. F. (1999) "A New Methodology for the Estimation of Total Uncertainty in Computational Simulation", *Proceedings of the AIAA Non-Deterministic Approaches Forum*, St. Louis, MO, 1999.

O'Connell, M. A. and Wolfinger, R. D. (1997), "Spatial Regression Models, Response Surfaces, and Process Optimization, *Journal of Computational and Graphical Statistics*, Vol. 6, No. 2, pp. 224-241.

Sacks, J., Schiller, S. B., and Welch, W. J. (1989a). "Designs for Computer Experiments" *Technometrics*, Vol. 31, No. 1, pp. 41-47.