**Engineering Letters, 14:1, EL_14_1_25 (Advance online publication: 12 February 2007)**

# Uncertainty Modeling for Expensive Functions: A Rank Transformation Approach

J. Umakant , K. Sudhakar, P.M. Mujumdar, and C. Raghavendra Rao

***Abstract--*An uncertainty model for an expensive function greatly improves the effectiveness of a design decision based on the use of a less accurate function. In this paper, we propose a method that exploits the concept of rank transformation to aggregate high fidelity information in a cost effective manner. This information is used to develop an empirical cumulative probability distribution function for residue such that it models the uncertainty with greater precision in the regions where the expensive function is potentially attractive as compared to other regions in the design space. The performance and robustness of the algorithm are demonstrated for univariate synthetic functions.**

***Index Terms----* model fidelity, sequential sampling, attractive zone**

## I. INTRODUCTION

Design of any system involves trade-off among various options and selecting one that meets the requirements best. In general, it may be desirable to assess the various options using a model that accurately evaluates the performance metrics of the system. In a computer based design framework these models are generally, computationally expensive functions or High Fidelity (HF) models. However, such functions have been usually used to examine or analyze a particular design in great detail rather than to evaluate large number of designs. This is mainly due to practical difficulties like unrealistic cycle time and issues related to integration of several HF models as may be needed in the design of a complex multidisciplinary system.

Hence the designer is confronted with the challenge of taking decisions in an environment wherein uncertainty is ever present. Development of strategies to model the expensive function in order to reduce the risk in the design decisions is the topic of many current researches. Design of experiments theory has been combined with regression techniques to create response surface models (RSM). These models approximate the behavior of the expensive function and have been used in Multidisciplinary Design Optimization of complex systems like aerospace vehicles [1]. However, the curse of dimensionality hinders the use of RSM for problems of large dimensions. Approximate/model management optimization [2] is another approach that uses approximations to alleviate the expense of relying exclusively on high fidelity models. In this method, at a current design point the values of the function and its derivative obtained using the approximate model is corrected to match the values obtained using the high fidelity model. This ensures that first order consistency conditions are satisfied. The correction is updated once the design point is outside the trust region. Soft computing methods like rough sets [3] have been applied to address the problem from a slightly different perspective. Rough set based approach seek to identify multiple, sub-regions in a design space, within which all of the design points are expected to have a performance value equal or less than a given level and is applied iteratively on a growing sample set. Optimization of the computationally intensive function is then carried out in the smaller sub-regions of the original design space. Yet another approach that has been widely adopted in engineering design is the use of Medium/Low fidelity (LF) models that are globally valid over the design space and are computationally light, albeit with less accuracy. Complementing such tools with an uncertainty model for expensive functions can greatly enhance the effectiveness of design decisions. Several probabilistic approaches have been suggested towards this. It is important here to distinguish between variability and uncertainty. Variability [4,5] is inherent randomness in the system. The use of probability theory to represent variability is well-established. Uncertainty is defined as “a potential deficiency in any phase or activity of the modeling process that is due to lack of knowledge” [4, 5]. Uncertainty may also arise when there is a scarcity of high fidelity information. This is true when new classes of systems are being developed and

J. Umakant is with CASDE, Department of Aerospace Engineering, IIT Bombay, Mumbai-400076, India. & Directorate of Aerodynamics, DRDL, Kanchanbagh, Hyderabad-500058, India. (Email: umakantj@yahoo.com)
K. Sudhakar and P.M. Mujumdar are with CASDE, Department of Aerospace Engineering, IIT Bombay, Mumbai-400076, India.(Email: sudhakar@aero.iitb.ac.in, mujumdar@aero.iitb.ac.in
C. Raghavendra Rao[c] is with [c]Department of Mathematics and Statistics, University of Hyderabad, Hyderabad-500047, India (Email: crrsm@uohyd.ernet.in)

no historical database exists. Probabilistic approaches to handle uncertainty in lieu of replacing the expensive functions with low/medium tools and their application in design scenarios have been demonstrated [6,7]. In most of these studies, an uncertainty model for the expensive function was assumed and the focus was to propagate its effect in a multi-disciplinary design scenario and take design decisions. Alternate frameworks like Evidence theory [8], Possibility theory [9], Info-Gap theory [10] seek to address the fundamental issue of developing uncertainty model. Applications to design scenarios are, at present, however, few [11]. Thus the process of building the uncertainty model itself to bridge the gap between HF model and LF model continues to be a challenge in uncertainty modeling.

In this paper we present a novel approach that utilizes the concept of ranks for the purpose of uncertainty modeling. Ranking refers to the process of ordering a sample (say of size 'N') with respect to a system performance metric. For minimization problems, the best observation receives the highest rank (rank N) while the worst observation is given the lowest rank (rank 1). Ranking procedures are one approach in multiple decision theory [12] where a simple loss function (zero-one) is used and the risk is an incorrect decision. Dell and Clutter [13]show that using relative cheap methods, like "judgment ordering", where it is difficult or expensive to obtain the characteristic of interest, may still increase the efficiency of using the sample mean as an estimator for the population mean. Cronan et.al.,[14] report that for small sample sizes, the rank regression technique produces a model with better estimates of residential property value as compared to the model based on multiple regression analysis technique.

However to the best of the author's knowledge there has been no study in the area of uncertainty modeling based on rank transformation approach. Rather than trying to create an uncertainty model of the expensive function for the entire design space, we propose a novel method, which exploits the concept of ranks, to sequentially aggregate a high fidelity sample from the regions where the expensive function is potentially attractive. A stopping criterion is used to limit the sample size to a reasonable number. Based on this information, we estimate the residues as the difference between the high fidelity response value and the corresponding low fidelity response value. A probabilistic model for the residues is then constructed. Uncertainty model for the expensive function is now defined as the low fidelity model complemented with probabilistic model of the residue. It may be noted that in the context of optimization (for minimization), the inaccuracy of the model in the regions where the function is relatively higher is not of much interest. Rank transformation of the response enables to introduce the preferential characteristics in the uncertainty model.

The rest of the paper is arranged as follows: in section II, the notations used are described. Details of the rank transformation approach are presented in section III. Some limitations of the approach with suggested means to circumvent are discussed in section IV. Residue analysis and uncertainty modeling are described in section V. Criteria for verifying the robustness of the algorithm is specified in section VI. Results for univariate synthetic functions are presented in section VII. Section VIII concludes with a summary of the work.

## II. NOTATIONS

In this article we will denote $Z = F(\mathbf{X})$, to represent the high fidelity model and $z = f(\mathbf{X})$ to represent the low fidelity model that describe the system behavior. $\mathbf{X} \in D \subseteq \mathbf{R}^n$, denotes the input vector for both the models. D is the design space in n dimensional real space $\mathbf{R}^n$. F(.) is the high fidelity or expensive function and Z is the high fidelity response. F(.) is the low fidelity function and z is the low fidelity response. $\mathbf{X}$ typically describes the parameterization of the system while Z or z describes a performance metric of the system.

Then for one dimensional space $\mathbf{R}$ we use the following notations

$D=[a\ \ b]$ for design space

$X_1,\ldots\ldots X_N,\ldots\ldots$ for design points in D

$Z_1,\ldots\ldots Z_N,\ldots\ldots$ for high fidelity responses corresponding to the design points

$z_1,\ldots\ldots z_N,\ldots\ldots$ for low fidelity responses corresponding to the design points

$S_X[N,M] = \{X_1,\ldots\ldots\ldots X_N, X_{N+1},\ldots\ldots\ldots X_M\}$ for set of design points

$S_R[N,M] = \{Z_1,\ldots Z_N, z_{N+1},\ldots\ldots\ldots z_M\}$ for set of high fidelity and low fidelity responses corresponding to the points at $S_X[N,M]$

$r_i$ for rank of the $i^{th}$ response in $S_R[N,M]$

$P_i$ for empirical probability at $X = X_i$ in $S_X[N,M]$

$C_i$ for empirical cumulative probability at $X = X_i$ in $S_X[N,M]$

$e_i$ for residue $Z_i - z_i$ at the point $X = X_i$

$X_0 = a < X_1 < \ldots < X_{p-1} < X_p = b$ for partition of D with size p

$f(D) = [f_L\ \ f_U]$ for low fidelity response space for D

$f_L = z_0 < z_1 < \ldots < z_{p-1} < z_p = f_U$ for partition of f(D) with size p

$S_L = \{ (X,z) \mid X \in D , z = f(X) \}$ for low fidelity system state descriptor

$S_H = \{ (X,Z) \mid X \in D , Z = f(X) \}$ for high fidelity system state descriptor

$S_A = \{ (X,Z) \mid (X,Z) \in S_H \text{ and } Z \le z_1 \}$ for system state descriptor in attractive zone

$n_A$ for cardinality of $S_A$ indicating number of hits in attractive zone

We wish to construct a model such that $\mathbf{Z} \sim \mathbf{z} + U(\mathbf{Z})$, where $U(\mathbf{Z})$ is a probability distribution that represents the uncertainty in the estimation of $\mathbf{Z}$.

## III. Rank Transformation Approach

In this section we present a novel approach that utilizes the concept of ranks to collect high fidelity information for the purpose of constructing an uncertainty model.

### *A. Selection of Initial Sample*

We assume that to begin with there are 'K' (K≥2) high fidelity responses available. These responses correspond to arbitrarily selected points in the design space. Alternately expert opinion can be solicited as to where to perform the initial expensive function evaluations. A typical table consisting of three initial points is shown in table1a.

### B. Augmentation of the sample with low fidelity responses

Low fidelity responses are now used to augment the sample. These responses are evaluated at design points resulting from uniform gridding of the design space. Augmenting is done to encourage global representation of the function and the design space, while ranking. In case of multimodal functions this helps to avoid gathering high fidelity information that is restricted to a local valley. However, caution must be exercised such that at any point of time, the number of low fidelity observations is not very significantly higher as compared to the number of high fidelity observations. Otherwise the trend of the high fidelity information, as predicted from the sample will be dominated by the low fidelity model. Table1b shows the data after addition of six low fidelity observations. It may be argued that low fidelity responses may be added sequentially. However, adding one low fidelity observation at a time may not encourage global representation in the sample, especially in the initial stages when the available number of observations is small.

| S.No. | X | Z |
|---|---|---|
| 1 | 1.21 | 46.09 |
| 2 | 6.92 | 55.69 |
| 3 | 9.07 | 88.22 |

Tab 1a: Typical Sample of High Fidelity Responses

| S.No. | X | Z |
|---|---|---|
| 1 | 1.21 | 46.09 |
| 2 | 6.92 | 55.69 |
| 3 | 9.07 | 88.22 |
| 4 | 0.50 | 65.20 |
| 5 | 2.50 | 22.96 |
| 6 | 4.50 | 39.60 |
| 7 | 6.50 | 72.58 |
| 8 | 8.50 | 119.41 |
| 9 | 10.50 | 179.49 |

Tab1b. Typical Augmented data

| S.No. | X | Z | Rank | Probability Density | Cumulative Probability |
|---|---|---|---|---|---|
| 4 | 0.50 | 65.20 | 5 | 0.11 | 0.10 |
| 1 | 1.21 | 46.09 | 7 | 0.16 | 0.26 |
| 5 | 2.50 | 22.96 | 9 | 0.20 | 0.46 |
| 6 | 4.50 | 39.60 | 8 | 0.18 | 0.63 |
| 7 | 6.50 | 72.58 | 4 | 0.09 | 0.72 |
| 2 | 6.92 | 55.69 | 6 | 0.13 | 0.86 |
| 8 | 8.50 | 119.41 | 2 | 0.04 | 0.90 |
| 3 | 9.07 | 88.22 | 3 | 0.07 | 0.97 |
| 9 | 10.50 | 179.49 | 1 | 0.02 | 0.99 |

Tab1c. Typical data showing rank transformation of responses and cumulative probabilities of design points

### C. Rank Transformation of responses

The high fidelity responses and low fidelity responses are combined to form the dataset $S_R$ and the elements of the set are sorted in a descending order. Rank transformation is now applied to the responses by assigning ranks to them in a serially increasing manner, starting from one to the value of the maximum rank. The maximum rank has a value equal to M, the total number of high fidelity and low fidelity responses. Thus the response with the minimum value receives the maximum rank while the response with maximum value receives a rank of one. If a response value occurs more than once, then same rank is assigned to all its occurrences. Typical rank transformation of the responses is given in the fourth column of table1c.

### D. Mapping of the ranks onto design support

The ranks for the responses are now mapped to their corresponding design points in $S_X$. Thus the design points are now given an additional attribute, namely rank that defines its preference for selection.

### E. Computation of the empirical PDF and CDF of the design support

We sort the design points of the information table in an ascending order with respect to their respective values together with their associated ranks and define the empirical probability density of the $i^{th}$ point as the ratio of its rank to the summation of ranks of all the points, shown as below:

$$P_i = \frac{r_i}{\sum_{i=1}^{M} r_i}$$

For the lower bound and upper bound of the design support we assign respectively zero and one as their empirical probability densities. Since the points are sorted in ascending order, the empirical cumulative density of the $i^{th}$ point in the table is the ratio of sum of ranks of all the points above it to the total summation of ranks and is denoted as below:

$$C_i = \frac{\sum_{j=1}^{i} r_j - 0.5}{\sum_{i=1}^{M} r_i}$$

The factor 0.5 is used for continuity correction to encompass the entire design support. Piece-wise linear interpolation yields the cumulative density at any other point in the design space. Thus we have constructed an empirical probability distribution function (PDF) and cumulative distribution function (CDF) for the design support. The fifth and sixth columns of table 1c depict the probability and cumulative probability of the design points.

### F. Selection of new point for HF evaluation

The CDF of the design support can now be used to choose a new point where the expensive function can be evaluated. This is accomplished by applying a random inverse transformation on the CDF. Points where the response is likely to be relatively lower have greater probability of getting selected than the points where the function responses are likely to be higher. However, allowing some chance for unfavorable points to also get selected enables the algorithm to scan the whole design support. This feature is useful when the expensive function exhibits multimodal characteristics. We have now K+1 high fidelity responses. An updated table is shown at table1d.

| S.No. | X | Z |
|---|---|---|
| 1 | 1.21 | 46.09 |
| 2 | 6.92 | 55.69 |
| 3 | 9.07 | 88.22 |
| 4 | 5.10 | 36.61 |

Tab1d. Typical High Fidelity sample after selecting one point from CDF of design space

The steps of augmenting the sample with low fidelity observations, ranking the collection of the responses, mapping the responses onto the design support are repeated to update the CDF of the design support. The updated distribution is used for selection of the next design point for high fidelity evaluation. This process is repeated till we have the desired number of high fidelity responses. Alternatively a heuristic stopping criteria described in the next section can be used.

### *G. Stopping Criterion*

A heuristic criterion is used to decide when to stop the process of collecting high fidelity information. We monitor the trajectory of high fidelity responses and the design points with respect to the number of trials and stop the process when the trajectory shows repetitions in the path.

Thus at the end of this process we have accumulated 'N+K' high fidelity responses, where 'N' is the number of trials performed.

## IV. LIMITATIONS OF THE RANK TRANSFORMATION AND MEANS TO CIRCUMVENT THEM

The rank transformation procedure described in the previous section has the following limitations:

i) The procedure may be sensitive to the initial sample and consequently influence the selection of subsequent high fidelity responses

ii) Since the inverse transformation for selecting new design point is random, there is a possibility of clustering of the design points.

The first limitation can be largely circumvented by starting with an initial sample that spans the entire design support and the range of the response. We propose a strategy based on stratification to implement this. The second limitation is addressed by defining a minimum distance criterion.

### *A. Stratification Algorithm*

Stratification refers to classifying or coding the design space D and f(D) into 'p' number of levels or strata with the following property:

$X_s = i$ , iff $X_{i-1} < X \leq X_i$ and

$z_s = i$ , iff $z_{i-1} < z \leq z_i$ ; i= 1, 2,….p

A typical system state descriptor $S_L = (X, z)$ is coded as $(X_s, z_s)$.

Each level can be interpreted as an isocontour. Our purpose is to choose the initial points such that there is a representation of all the strata in X and z, at least once. In other words, p points are chosen such that each point is representative of one strata of X and z. This problem can be formulated as an assignment problem (or an integer programming problem), in Boolean space. However, for the purposes of our requirements, we have chosen to implement the same as described below.

| S.No. | z | X |
|---|---|---|
| 1 | 17.60 | 0.00 |
| 2 | 11.76 | 0.35 |
| 3 | 9.86 | 0.71 |
| 4 | 10.16 | 1.07 |
| 5 | 11.29 | 1.42 |
| 6 | 12.29 | 1.78 |
| 7 | 12.56 | 2.14 |
| 8 | 11.94 | 2.50 |
| 9 | 10.61 | 2.85 |
| 10 | 9.16 | 3.21 |
| 11 | 8.58 | 3.57 |
| 12 | 10.25 | 3.92 |
| 13 | 15.92 | 4.28 |
| 14 | 27.75 | 4.64 |
| 15 | 48.28 | 5.00 |

Tab.2a. Typical data in physical domain for stratification

| S.No. | z | $X_s$ |
|---|---|---|
| 1 | 2 | 1 |
| 2 | 1 | 1 |
| 3 | 1 | 1 |
| 4 | 1 | 1 |
| 5 | 1 | 1 |
| 6 | 1 | 2 |
| 7 | 1 | 2 |
| 8 | 1 | 2 |
| 9 | 1 | 2 |
| 10 | 1 | 2 |
| 11 | 1 | 3 |
| 12 | 1 | 3 |
| 13 | 2 | 3 |
| 14 | 3 | 3 |
| 15 | 3 | 3 |

Tab. 2b Typical Coded data in after stratification

| S.No. | z | $X_s$ |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 2 |
| 3 | 1 | 3 |
| 4 | 2 | 1 |
| 5 | 2 | 3 |
| 6 | **3** | **3** |

Tab.2c Typical Coded data after deleting duplicate points

| | z | $X_s$ |
|---|---|---|
| Level 1 | 3 | 2 |
| Level 2 | 2 | 1 |
| Level 3 | 1 | 3 |

Tab..2d Frequency table of coded data

| S.No. | z | $X_s$ |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 2 |
| 3 | **2** | **1** |

Tab.2e Typical Coded data after selecting first point

We uniformly grid the entire support space into large number of points and use the low fidelity tool for evaluating the response at each of the grid points. Typical representation for 15 grid points is shown at table 2a.The grid points and the corresponding response values are then coded or stratified according to the definition above. We have thus set up stratification table of coded values for the design points and their respective response values. Table2b illustrates this process for three stratification levels. A consequence of the stratification is that the points falling in the same strata lose their distinct identity, resulting in duplicate points that do not possess any additional information. For further analysis, such points are not retained in the stratification table. Table2c highlights such data. A frequency table indicating the number of occurrences of $X_s$ and $z_s$ is then set up and a typical result is shown in table 2d. Strata level corresponding to the maximum number of occurrences either in $X_s$ or $z_s$ is identified and a point from the set of coded values corresponding to this strata level is selected as highlighted with an arrow in table 2c. Any tie that occurs is resolved randomly. Other points having the same stratification identity either in support space or in the response space are then removed and the stratification table is updated as illustrated in table 2e. The process is repeated for the specified number of strata levels, resulting in 'p' distinct coded points with their corresponding response codes also being distinct. An illustration of the updated table is shown in table 2f. Table 2g summarizes the selected strata levels for the support points and the corresponding response. The selected stratified levels are then mapped to the physical domain and the points are randomly selected from the respective strata. At these initial points we perform the expensive function evaluations. For example, the entry at serial number 3 in table 2g corresponds to a design point in strata level 2 with its response in strata level 1. In the physical domain, the entries at serial number 6 through 10 of table 2a correspond to these strata levels and we randomly select any one of these points.

| S.No. | z | $X_s$ |
|---|---|---|
| 1 | **1** | **2** |

Tab.2f Typical Coded data after selecting second point

| S.No. | z | $X_s$ |
|---|---|---|
| 1 | **3** | **3** |
| 2 | **2** | **1** |
| 3 | **1** | **2** |

Tab. 2g Summary of Coded data selected for three levels

## B. *Minimum Distance Criterion*

Random inverse transformation of the cumulative distribution function to select the new design points can sometimes result in the new point to lie within the neighborhood of existing points. To avoid this, we calculate the $L_1$ distance of the new point with respect to the existing points and retain the point only if the norm is greater than a specified tolerance bound; otherwise we sample another point from the distribution. It may be noted that we perform this check before the expensive function is evaluated.

# V. UNCERTAINTY MODELING

We now construct an empirical cumulative distribution function for the residue. Residue, defined as the difference between the high fidelity response and the corresponding low fidelity response, is used to characterize the uncertainty in the estimation of the expensive function. Diagnosis of the trajectory of the residues with respect to the low fidelity response enables to infer whether or not the residues are correlated with low fidelity response. In case the trajectory exhibits a random path, then we infer that the differences are random. However, if a systematic variation is observed then we may infer that the low fidelity responses are correlated with the high fidelity responses.

## A. *CDF for the Residue*

The residues, denoted by '*e*', for the N+K high fidelity responses available are sorted in ascending order. We assume homogenous distribution for the residues.
The cumulative density for the $i^{th}$ residue is defined as

$$C_i(e) = \frac{(i-0.5)}{N+K} \quad i = 1,2,\ldots,N+K$$

The lower bound of the residues is defined as the minimum value of the observed residues decremented by one unit. Similarly the upper bound of residue is defined as the maximum value of the observed residue incremented by one unit. Cumulative densities of zero and one are assigned respectively to the lower and upper bounds. We have now defined empirical cumulative densities for the residues at discrete points. A fit for this data will give a smooth representation. Linear interpolation may also be used to estimate the cumulative density of for any intermediate values of residue. This type of distribution is also referred to as non-parametric distribution.

## B. *Restricted CDF of residue for attractive zone*

It may be recalled that we had stratified the range of the response into three contour levels. We denote the zone defined by the first contour level as the attractive zone, since the response value corresponding to the first level is smaller as compared to that for the second and third levels. In the context of minimization, we are interested only in those high fidelity observations that are contained in the first contour level. Hence we filter the N+K responses and retain only those responses that are within the attractive zone. We can now construct the cumulative distribution function of the residue for the attractive zone in a similar manner as described in the previous section. It may be noted that the uncertainty bounds of the residue will now be typically lower than that obtained in the previous section and hence the estimate of the expensive function is not unnecessarily conservative. This is consistent with our philosophy that we wish to create an uncertainty model for the expensive function that is more appropriate in the regions of interest, rather than trying to characterize the uncertainty for the entire design support.

## VI. VALIDATION OF ALGORITHM

In the above procedure, the new design point for high fidelity estimation has been chosen randomly using inverse transformation of the cumulative distribution function of the design support. Hence the repeatability or robustness of the algorithm is verified by conducting large number of simulations and examining the results. Verification is carried out in the following ways:

i) The probability of number of observations in the attractive zone should be non-trivial with at least 95% confidence.

ii) The empirical cumulative distribution of the design support and residue are examined to check if the same type of distribution is obtained in most of the simulations.

iii) Coefficient of variation, defined as the percentage ratio of standard deviation to mean, is examined for the lower and upper bounds of the residue. The algorithm can be considered to be stable [15] if the coefficient of variation is less than 33%.

## VII. RESULTS FOR SYNTHETIC EXAMPLES

For the purpose of illustrating the performance of above algorithm, we present results for two synthetic examples. The expensive function is represented by a univariate function. In the first example the function has unimodal variation while in the second example the function exhibits a bimodal variation. The low fidelity representation of the unimodal function is achieved through perturbations in the coefficients of the function, while, for the bimodal function the low fidelity model is represented by a quadratic function. Three levels of stratification were adopted for both the cases. Low fidelity tool was used to partition the expected range of the response into three contour levels. The first level has the lowest contour value. Attractive zone for the response then corresponds to first contour level. The details regarding the functions are given in tables 3a and 3b.

| | Function | Design Space | Attractive Zone |
|---|---|---|---|
| High Fidelity | $Z = X^2 + \frac{54}{X}$ | $D=[0.5\ \ 10.5]$ | $Z \le 51$ |
| Low Fidelity | $z = 1.6 * X^2 + 0.6 * \frac{54}{X}$ | $D=[0.5\ \ 10.5]$ | |

Tab 3a. Univariate Unimodal Function

| | Function | Design Space | Attractive Zone |
|---|---|---|---|
| High Fidelity | $Z=(X-0.5)(X-2)(X-4)(X-3.25)+10$ | $D=[0\ 5]$ | $Z \le 8.3$ |
| Low Fidelity | $z = 1.8785 * X^2 + 7.8888 * X + 15.731$ | $D=[0\ 5]$ | |

Tab 3b. Univariate Bimodal Function

### A. *Univariate Unimodal Function*

Figure1 depicts the high fidelity and low fidelity contours of the function. The effect of number of trials on the trajectory of the design variable in the design support and on the high fidelity response values is shown respectively in figures 2a and 2b. Each trial augments the sample size by one. Ignoring the three initial samples, we can observe that the ten trials are sufficient to ensure that the design support is well represented.

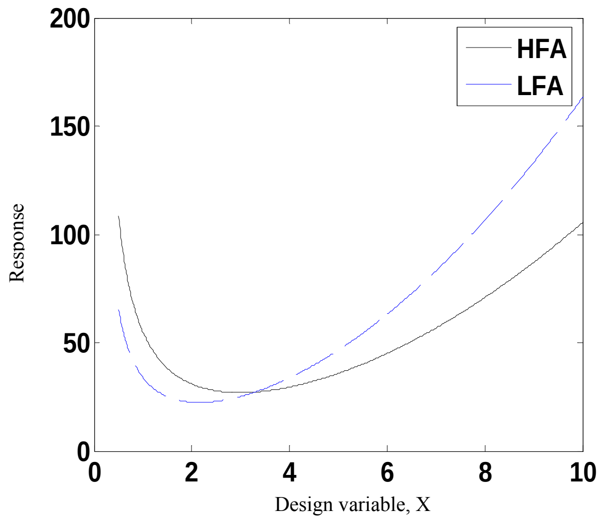



Fig.1 High fidelity and Low fidelity contours of function F1

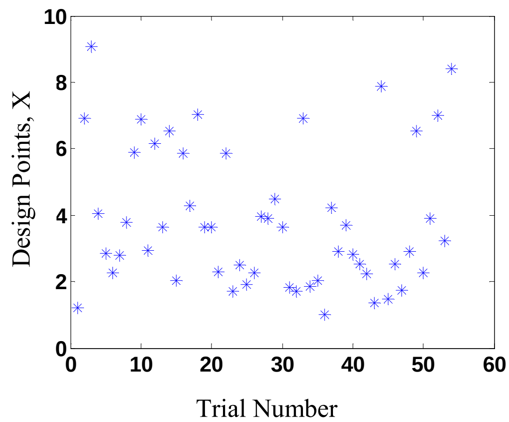



Fig.2a. Trajectory of Design Variable during aggregation of design points

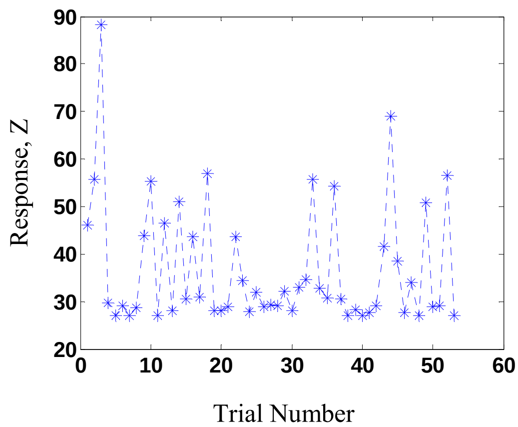



Fig.2b. Trajectory of High fidelity response during aggregation of design points

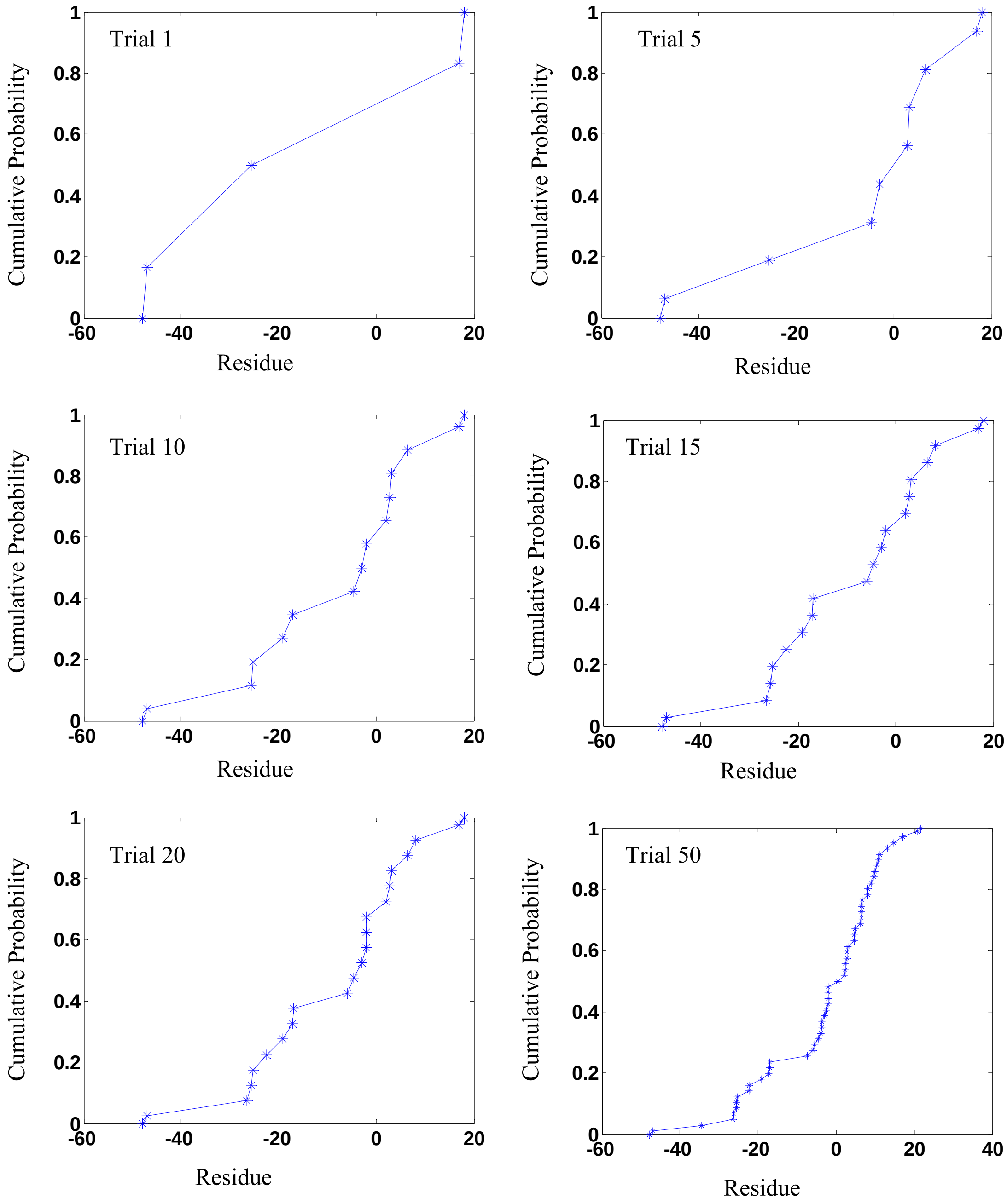



Fig.3 History of Cumulative Distribution Function of Residue

When the number of trials exceeds ten clustering in the trajectory of the design variable is noticed. Similarly the trajectory of high fidelity responses gets repeated for more than ten trials. The cumulative distribution of the residue for various numbers of trials is shown in figure 3. The CDF for the first trial corresponds to that based on the initial three points selected by stratification. It can be noticed that nearly the same cumulative distributions is obtained when the number of trials is ten or more. Based on this empirical evidence, we restrict the number of trials to ten. In practical scenarios, a budgetary constraint on the number of high fidelity evaluations may also restrict the sample size.

The design points selected based on the algorithm for high fidelity evaluations are shown in figure 4a. The symbols shown in this figure represent the selected points. The initial points selected by stratification are distinguished by the symbol '*'. The value of first response level based on the stratification is 51, i.e., the attractive zone corresponds to the design space where response of the expensive function is less than 51. The cumulative distribution function for X is shown in figure 4b. Two distinctive features can be noticed in this distribution. One, there is a rather steep increase in the cumulative probability between X=1 and X=4, indicating a preference for selection of points from this region. Referring back to figure4a, it can be observed that the high fidelity function is minimum at X=3. Two, the CDF exhibits a large tail for X greater than 6.8. In our experiment, we notice that eight of the ten points selected based on rank transformation are from the attractive zone. Considering the stratification points also yields a total of nine points out of thirteen in the attractive zone. If we had naively chosen K+N grid points, in this case thirteen, only seven points would have been in this region. Further, only three of the selected points were from the region X greater than 6.8, whereas grid points would have resulted in five high fidelity evaluations in this region. This justifies that rank transformation encourages selection of new design points from the region where the function is likely to be relatively lower. Thus we can state that, in the context of optimization, aggregation of the points selected based on rank transformation have superior attributes than simple grid selected points.

The CDF of residue considering all the thirteen high fidelity responses is shown in figure 4c. It can be observed that the lower bound of the residue is -49 while upper bound of the residue is +45. The effectiveness of the low fidelity model when used in conjunction the uncertainty model is demonstrated in figure 4d. The upper and lower bounds in the estimation of the expensive function are denoted by the dashed lines, while the solid line represents the expensive function. It can be observed that the uncertainty model has successfully characterized the uncertainty in the estimation of the expensive function for almost the entire design support.

Figure 4e highlights the CDF of residue constructed by restricting the sample to only those observations that are within the attractive zone. The width between the lower and upper bound is now considerably reduced. The effectiveness of the low fidelity model when used in conjunction with restricted CDF is shown in figure 4f. It can be observed that the estimation of the expensive function is considerably less conservative in the attractive zone. Thus the model characterizes the uncertainty in the expensive function with better accuracy in the attractive zone than in the other regions.

Histogram of the number of high fidelity responses that are in the attractive zone for 100 Monte-Carlo simulations is shown in figure 5a. The total number of samples was 13. It is observed that 97 out of the 100 simulations record at least 6 samples within the attractive zone. Thus the non-trivial probability is 0.46 with 97% confidence. Figure 5b shows that CDF of the design support exhibit similar characteristics in all the simulations. The simulation results for CDF of the residue and restricted CDF are shown in fig. 5c and fig. 5d respectively. The coefficient of variation of the bounds is given in table 4a. It can be seen that the coefficient of variation of the bounds is less than 33%.

| | Residue | | Restricted Residue | |
|---|---|---|---|---|
| | Lower Bound | Upper Bound | Lower Bound | Upper Bound |
| Mean | -50.30 | 37.90 | -16.70 | 17.00 |
| Standard Deviation | 6.03 | 8.30 | 4.16 | 0.23 |
| Coefficient of Variation , % | 12.00 | 21.90 | 25.70 | 1.00 |

Tab. 4a. Metrics from Monte-Carlo Simulations of Univariate Unimodal Function

| | Residue | | Restricted Residue | |
|---|---|---|---|---|
| | Lower Bound | Upper Bound | Lower Bound | Upper Bound |
| Mean | -3.70 | 5.70 | -3.11 | -0.59 |
| Standard Deviation | 0.49 | 2.35 | 0.50 | 0.36 |
| Coefficient of Variation , % | 13.20 | 41.20 | 16.00 | 60.00 |

Tab. 4b. Metrics from Monte-Carlo Simulations of Univariate Bimodal Function

### B. *Univariate Bimodal function*

Typical results for a bimodal function are presented in this section. Figure 6a shows the high fidelity and low fidelity function contours. It may be noted that the function has a global minima at X=1.05 and another local minima at X=3.75. The design points selected based on the ranking algorithm are represented in the figure by the symbols. It can be observed that the algorithm has selected design points near both the valleys present in the high fidelity function and three points out of the sample size of thirteen are in the attractive zone. However, only two points would have been in the attractive zone if we had based the selection on simple grid of the design support. It may be noted that the algorithm has successfully

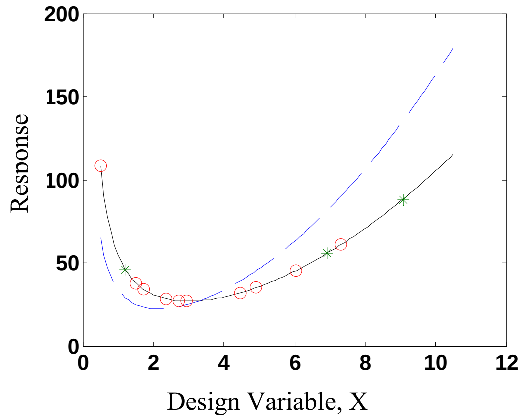



Fig.4a High fidelity and Low fidelity contours with selected design points

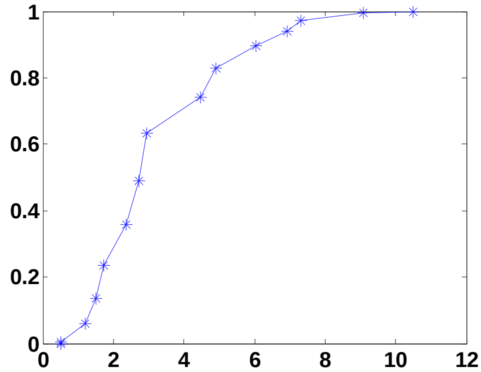



Fig.4b. Cumulative Distribution Function of Design Support

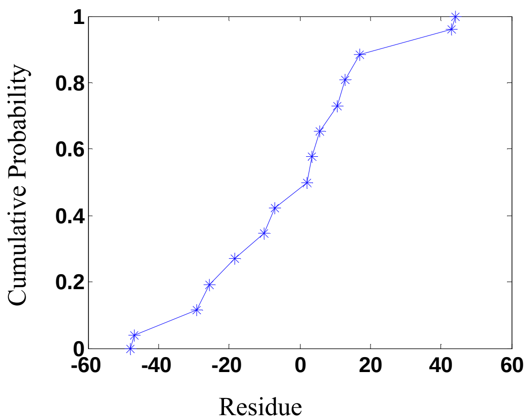



Fig.4c. Cumulative Distribution Function of Residue

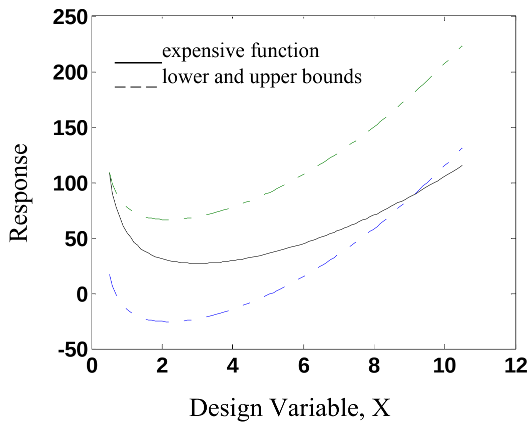



Fig.4d. Uncertainty in estimation of Expensive Function

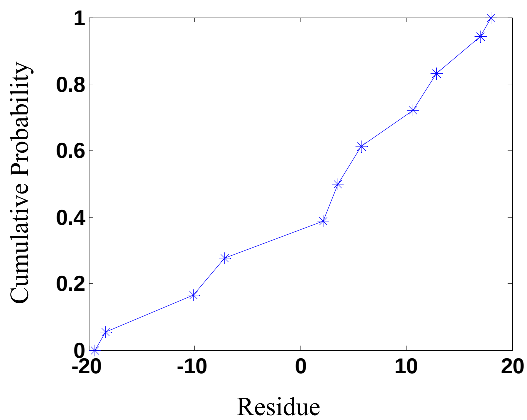



Fig.4e. Restricted Cumulative Distribution Function of Residue

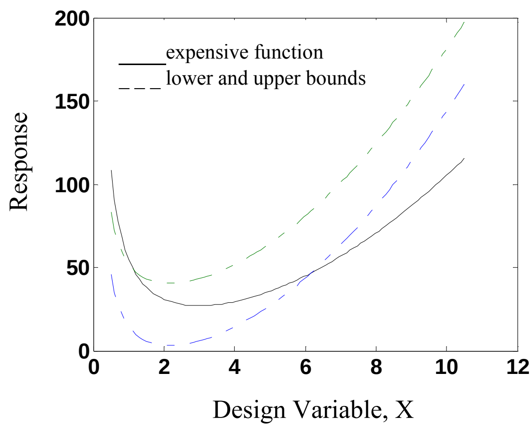



Fig.4f. Uncertainty in estimation of Expensive Function using restricted CDF

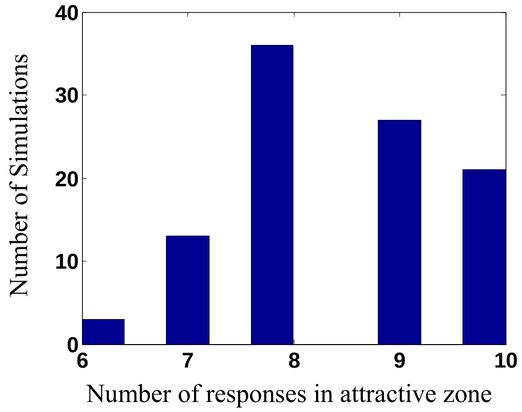



Fig.5a. Histogram of number of hits in the attractive zone

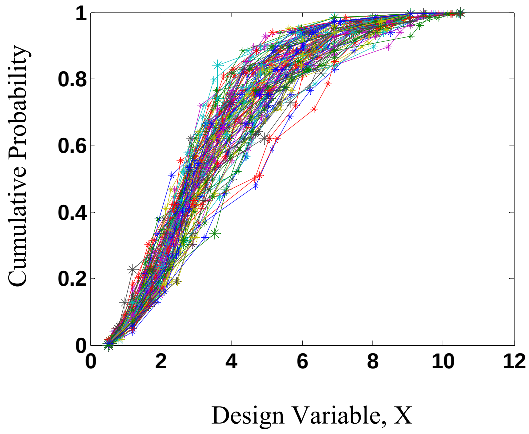



Fig.5b. Cumulative Distribution Function of design support for various simulations

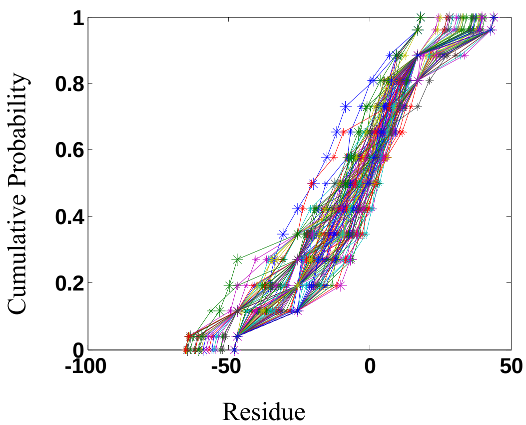



Fig.5c. Cumulative Distribution Function of Residue for various simulations

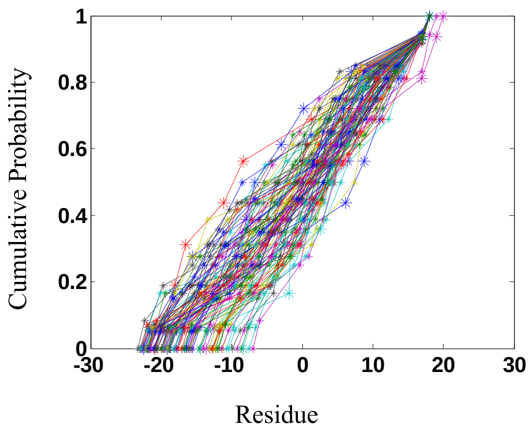



Fig.5d. Restricted Cumulative Distribution Function of Residue for various simulations

selected points in the favorable regions even though the low fidelity model is unimodal whereas the expensive function is bimodal. The CDF for the design support is shown in figure 6b. It can be noticed that the probability of selecting a point from the unfavorable region ie., X greater than 4 is less than 0.02 while it is only 0.2 for the other unfavorable region ie.,X less than 1. The CDF of the residue based on the total sample of high fidelity information is given in figure 6c and the corresponding prediction of the expensive function using this model is shown in figure 6d. Again it can be seen that the uncertainty has been characterized over most of the design support. Figure 6e shows the restricted CDF based on restricting the sample to the observations within the attractive zone. This enabled better characterization of uncertainty in the expensive function in favorable regions than in other regions and the result is visualized in figure 6f. Results of Monte-Carlo simulations are shown in figures 7a-7d. In this case the non-trivial probability of number of observations in the attractive zone is 7.7% with 97% confidence. The CDF for the design support and residue exhibits similar characteristics in most of the simulations. The coefficient of variation for the bounds is tabulated in table 4b. It can be noticed that the upper bounds of the residue have coefficient of variation greater than 33%. This is mainly due to the fact that the low fidelity model is unimodal whereas the high fidelity is bimodal. Still, the algorithm has characterized the uncertainty better in the attractive zone.

## VIII. SUMMARY

In this paper, we have showcased the utility of ranks to develop a method for modeling uncertainty of an expensive function. This has been accomplished in two stages. Instead of aggregating high fidelity information based on a naive uniform grid of the design support, we exploited the concept of ranks. The high fidelity responses, augmented with limited number of low fidelity observations, were assigned ranks based on their value and these were mapped onto the design

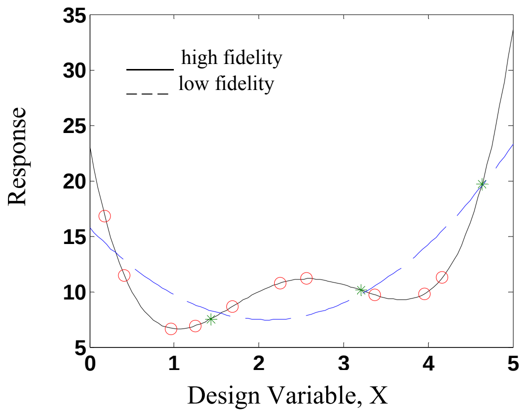



Fig.6a. High Fidelity and Low fidelity contours with selected design points

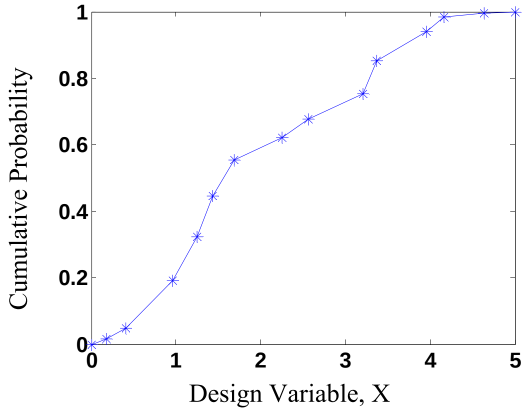



Fig.6b. Cumulative Distribution Function of Design Support

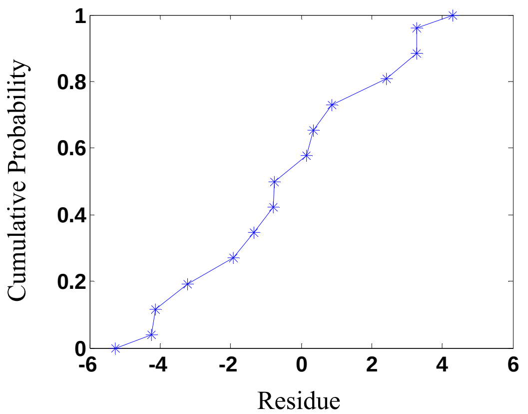



Fig.6c. Cumulative Distribution Function of Residue

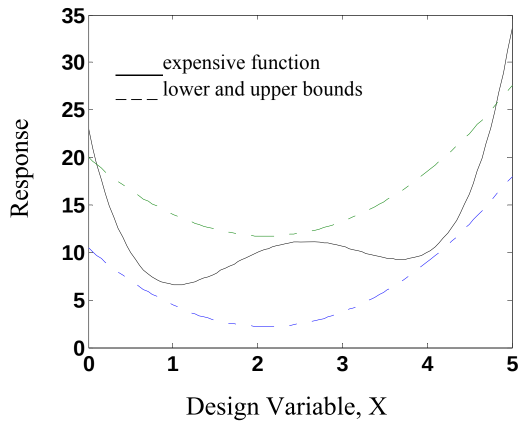



Fig.6d. Uncertainty in estimation of Expensive Function

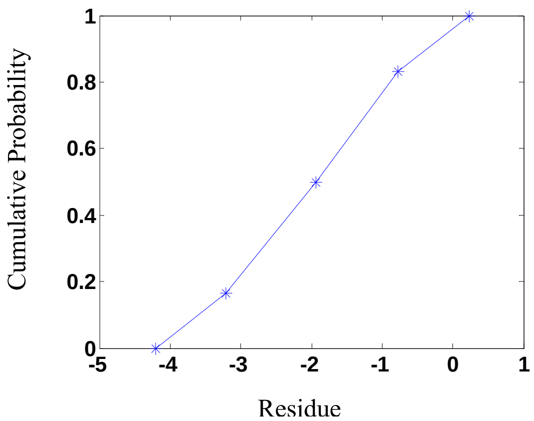



Fig.6e. Restricted Cumulative Distribution Function of Residue

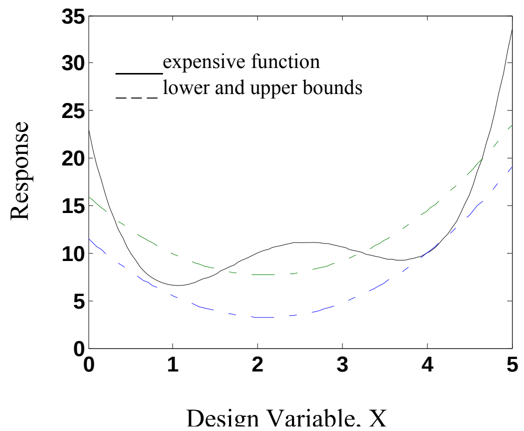
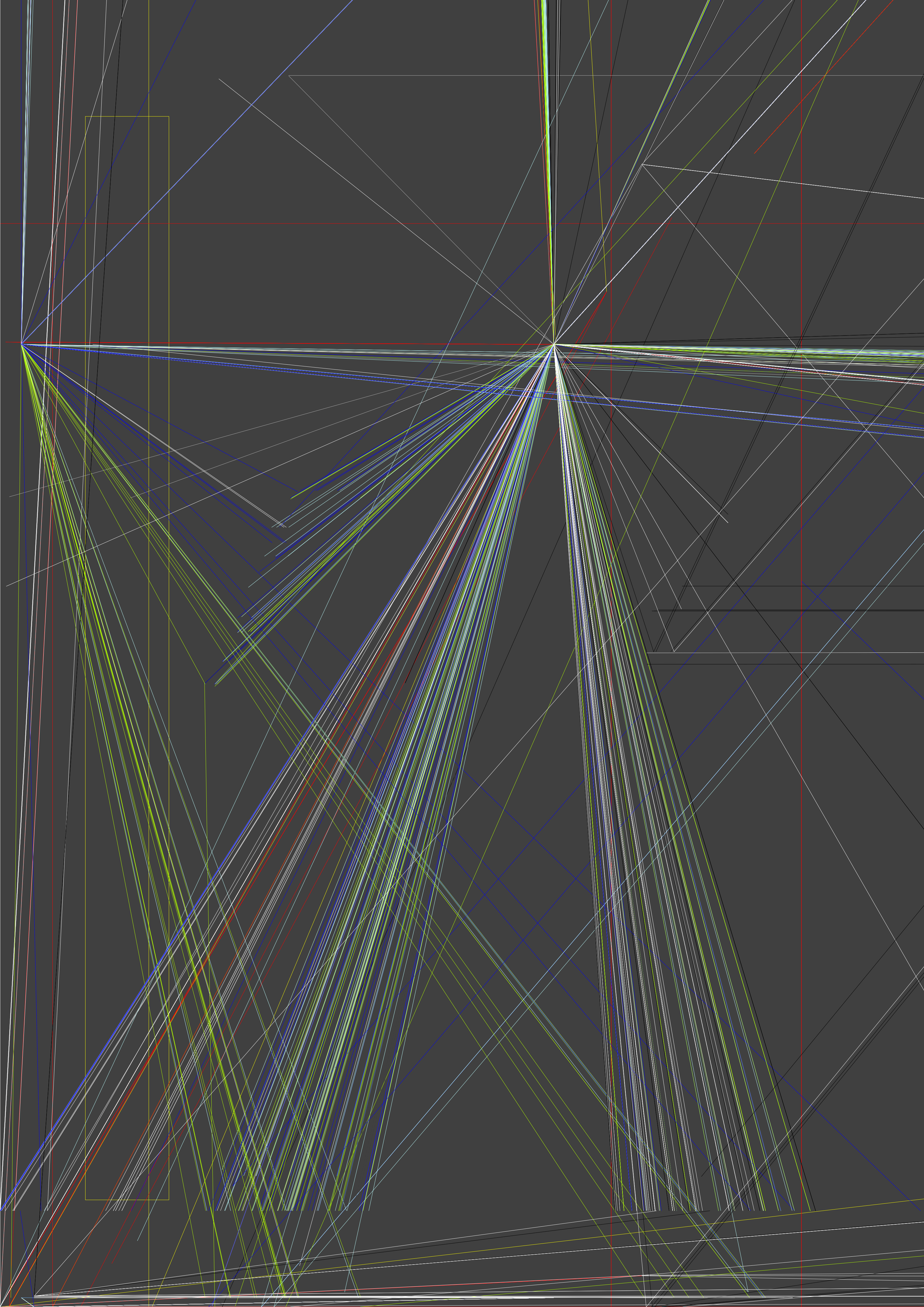



Fig.6f. Uncertainty in estimation of Expensive Function using restricted CDF



Optimization. 1st International workshop on Surrogate Modelling and Space Mapping for Eng. Opt, 2000

[3] Shan, Songqing and Wang, Gary G. Space Exploration and Global Optimization for Computationally Intensive Design Problems: A Rough Set Based Approach. Journal of Structural and Multidisciplinary Optimization, 2004.

[4] Zang, Thomas A., Hemsch, Michael J., Hilburger, Mark W., Kenny, Sean P., Luckring, James M., Maghami Peiman, Padula, Sharon L., Stroud, Jefferson W. Needs and Opportunities for Uncertainty Based Multidisciplinary Design Methods for Aerospace Vehicles. NASA-TM-2002-211462, July 2002

[5] Oberkampf,W.L., Bibliography for Verification and Validation in Computational Simulation. SAND 98-2041.

[6] DeLaurentis, Daniel A. A Probabilistic Approach to Aircraft Design Emphasizing Stability and Control Uncertainties, Ph D thesis, Georgia Institute of Technology, 1998

[7] Umakant, J., Sudhakar, K., Mujumdar, P.M., Panneerselvam, S., Configuration Design of Air-Breathing Hypersonic Vehicle considering Fidelity Uncertainty. AIAA-2004-4543, 10th AIAA / ISSMO Multi-disciplinary Analysis and Optimization Conference, 30th August-1st September, 2004. Albany, New York

[8] Sentz, Kari., Ferson, Scott., Combination of Evidence in Dempster-Shafer Theory SAND 2002-0835 Sandia National Laboratories

[9] Dubois D, Prade H. Possibility theory: an approach to computerized processing of uncertainty, 1st ed. New York;Plenum Press;1998

[10]Ben-Haim, Yakov. Uncertainty, probability and information-gaps. Reliability Engineering and System Safety 85(2004) 249-266

[11] Agarwal, Harish., Renaud, John E., Preston, Evan L., Padmanabhan, Dhanesh. Uncertainty quantification using evidence theory in multidisciplinary design optimization. Reliability Engineering and System Safety 85(2004) 281-294

[12] Gupta, S. S., Huang, D.Y. Multiple Statistical Decision Theory. Springer-Verlag, New York, 1981

[13] Dell,T.R. and Clutter,J.L. Biometrics, 28, 545-553, 1972

[14] Cronan, Timothy P., Epley, Donald R., Perry, Larry G. The Use of Rank Transformation and Multiple Regression Analysis in Estimating Residential Property Values with a Small Sample. The Journal of Real Estate Research Vol.1No.1 Fall 1986

[15] Agarwal, B.L., Basic Statistics. Third Edition, New Age International (P) Limited Publishers, 1996.

J. Umakant was born in Chennai, India in 1965. B.Sc in Mathematics from Madras Christian College,1986. B.Tech in Aeronautical Engineering from Madras Institute of Technology and M.E. in Aerospace Engineering from Indian Institute of Science, Bangalore in 1991.

He has been working in the Defence Research & Development Laboratory, Hyderabad since 1991. Currently he is Scientist 'E' in the Directorate of Aerodynamics. His research work in configuration design of aerospace vehicles, optimization and uncertainty modeling have been reported in AIAA conferences.

J. Umakant is a Member of the Aeronautical Society of India.

K. Sudhakar was born in Kerala in 1948. B. Tech and M. Tech in Aerospace Engineering from IIT Madras in 1971 and 1973 respectively. Ph.D. in Aerospace engineering from IIT Bombay in 1989.

He worked as design engineer at Hindustan Aeronautics Limited during 1973 to 1984. He joined the Department of Aerospace Engineering, IIT Bombay, Mumbai as an Assistant Professor in 1984 where he presently is a Professor. His research work in the areas of Aerodynamics & Flight Dynamics and Design Optimization have been reported in AIAA Conferences and Journals.

Prof. Sudhakar is a senior member of the American Institute of Aeronautics & Astronautics, Fellow of the Aeronautical Society of India and Member of Systems Society of India.

P.M. Mujumdar was born at Nagpur, India in 1958. B. Tech and M. Tech in Aerospace Engineering from IIT Bombay in 1981 and 1983 respectively. Ph. D. in Aerospace engineering from IIT Bombay in 1988.

He joined the Department of Aerospace Engineering, IIT Bombay, Mumbai as an Assistant Professor in 1988 where he presently is a Professor. His research work in the areas of Structural Dynamics, Aeroelasticity & Design Optimization have been reported in AIAA Conferences and Journals.

Prof. Mujumdar is a member of the American Institute of Aeronautics & Astronautics, the Aeronautical Society of India and the Indian Society of Advancement in Materials and Processes Engineering.

C. R. Rao was born at Guntur, India in 1958. B. Sc and M.Sc in Statistics from Andhra Univerisity and Osmania Univeristy in 1978 and 1980 respectively. Ph. D. in Statistics from Osmania University in 1986. The author is a member of IAENG ( Membership number 61460)

He worked as a lecturer in Statistics at Osmania University from 1984 to 1986. Since 1986, he is working in Department of Mathematics and Statistics, University of Hyderabad where presently he is a Reader. His current research interests are Simulation & Modeling and Knowledge Discovery. His research work in Knowledge Discovery has been reported in Journal of Bipolar Disorders, Journal of Electronic Modeling and Journal of Information.

Dr Rao is a member of the Operation Research Society of India, Indian Mathematical Society and a Fellow of IETE.