10th World Congress on Structural and Multidisciplinary Optimization
May 19 - 24, 2013, Orlando, Florida, USA

Characterizing Probability-based Uniform Sampling for Surrogate Modeling

Junqiang Zhang[1], Souma Chowdhury[2], Achille Messac[3]

[1] Syracuse University, Syracuse, NY, USA, jzhang62@syr.edu
[2] Syracuse University, Syracuse, NY, USA, sochowdh@syr.edu
[3] Syracuse University, Syracuse, NY, USA, messac@syr.edu

1. Abstract
Appropriate sampling of training points is one of the primary factors affecting the fidelity of surrogate models. This paper investigates the relative advantage of probability-based uniform sampling over distance-based uniform sampling in training surrogate models whose system inputs follow a distribution. Using the probability of the inputs as the metric for sampling, the probability-based uniform sample points are obtained by the inverse transform sampling. To study the suitability of probability-based uniform sampling for surrogate modeling, the Mean Squared Error (MSE) of a monomial form is formulated based on the relationship between the squared error of a surrogate model and the volume or hypervolume per sample point. Two surrogate models are developed respectively using the same number of probability-based and distance-based uniform sample points to approximate the same system. Their fidelities are compared using the monomial MSE function. When the exponent of the monomial function is between 0 and 1, the fidelity of the surrogate model trained using probability-based uniform sampling is higher than that of the other one trained using distance-based uniform sampling. When the exponent is greater than 1 or less than 0, the fidelity comparison is reversed. This theoretical conclusion is successfully verified using standard test functions and an engineering application.
2. Keywords: Surrogate modeling, Probability-based sampling, Distance-based sampling

3. Introduction
Surrogate modeling is a statistical approach used to develop approximation functions that adequately represent the relationship between inputs and outputs based on known data[1]. To alleviate the burden of high experimental or computational costs resulting from complex engineering design problems, surrogate modeling has been frequently used as an efficient approach[2, 3]. To develop the surrogate model of a system, sample points of the inputs to the system need to be generated. These sample points and their corresponding system outputs are used to train the surrogate model. Distance-based sampling uses coordinate-distances as the metrics between points in the sample space. If the inputs to the system are physical parameters whose probability of occurrence is known or predefined, probability-based sampling can use the difference of probability as the metrics between points. These sampling approaches can generate sample points uniformly in terms of distance and probability, respectively.
Radial Basis Function (RBF) and Kriging are expressed as combinations of basis functions[1]. Since their overall error is related to the sample density, or equivalently the volume or hypervolume per sample point, this paper formulates the Mean Squared Error (MSE) of a monomial form based on the relationship between the squared error of a surrogate model and the volume or hypervolume per sample point. The overall MSEs of the two surrogate models of the same system developed respectively using the probability-based uniform sampling and the distance-based uniform sampling are compared.
Section 4 introduces the probability-based sampling approach. Section 5 presents the estimation of the MSE of a surrogate model. Section 6 illustrates the measure of a region associated with a sample point. Section 7 formulates the MSE of the monomial form used to analyze the fidelity of surrogate models. The suitability of probability-based uniform sampling is studied in Sec. 8. Section 9 discusses how to fit the MSE of the monomial form. Section 10 tests the conclusion of the suitability of probability-based sampling. Section 11 applies probability-based sampling to the development of surrogate models for window performance evaluation. The concluding remarks are presented in Sec. 12.

4. Probability-based Sampling
The popular approaches to generate sample points from a probability distribution include inverse transform sampling[4], rejection sampling[5], importance sampling[6], Markov Chain Monte Carlo methods[7]. In this paper, inverse transform sampling is used to evaluate the values of random variables corresponding

to their designated probabilities. It generates sample points from a probability distribution given the Cumulative Distribution Function (CDF). In an n-dimensional space R^n, $x^{(k)}$, $k = 1, 2, \ldots, n$, is used to represent the k^{th} random variable. Suppose the n random variables are independent. Function $F_k(x^{(k)})$ is the CDF of the variable $x^{(k)}$. A set of numbers $c_i^{(k)} \in [0, 1]$, $i = 1, 2, \ldots, m$, are the values of probability of $x^{(k)}$. The probability-based sample points $x_i^{(k)(p)}$ corresponding to the set of probabilities $c_i^{(k)}$ are evaluated by

$$x_i^{(k)(p)} = F_k^{-1}\left(c_i^{(k)}\right). \tag{1}$$

Suppose the lower and upper boundaries of $x^{(k)}$ are $x_{\min}^{(k)}$ and $x_{\max}^{(k)}$, respectively. The distance-based sampling scales the set of numbers $c_i^{(k)}$ to the coordinates $x^{(k)(d)}$ by

$$x_i^{(k)(d)} = x_{\min}^{(k)} + c_i^{(k)}\left(x_{\max}^{(k)} - x_{\min}^{(k)}\right). \tag{2}$$

Many sampling sequences have been developed to address different design space exploration demands. Full factorial sampling sequence[1], low-dispersion sequence[8], and low-discrepancy sequence[9] are popular uniform sequences used for optimization, surrogate modeling, and numerical integration. Scaled or inversely transformed from the sequences, the distance-based and probability-based sample points are uniform in terms of distance and probability, respectively.

5. Estimation of MSE

5.1. Integrated Estimation of MSE

The output y of a system is approximated as a function $h(x)$ of the system input x. The CDF of x is $F(x)$. The integrated MSE of the surrogate model $h(x)$ is given by[10]

$$MSE_I = \int (y - h(x))^2 \, dF(x). \tag{3}$$

5.2. Empirical Estimation of MSE

In practical engineering problems, the output y of a system is usually only known at a limited number of test points. The integrated estimation of the MSE is not readily available. Using a set of test points (x_j, y_j), $j = 1, 2, \ldots, s$, the empirical estimation of the MSE is evaluated by[10]

$$MSE_E = \frac{1}{s}\sum_{j=1}^{s}\left(y_j - h(x_j)\right)^2. \tag{4}$$

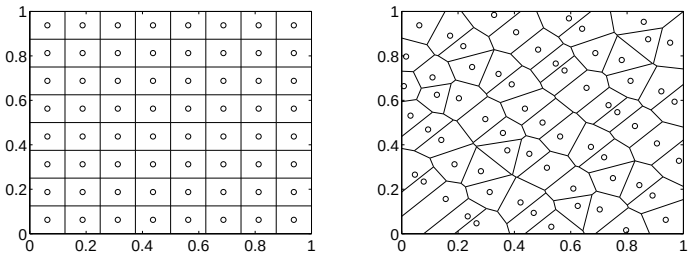
If the probability distribution of the test points x_j follows the distribution of x, the empirical estimation of the MSE should be close to the integrated estimation as the number of test points becomes sufficiently large.

6. Measure of a Region

The measure of a one-dimensional, two-dimensional, three-dimensional, or higher-dimensional region is a length, an area, a volume, or a hypervolume, respectively. In this paper, volume is used to represent the measure regardless of the number of dimensions. The measure of a region and the probability of a region are two important concepts used in this paper. Their values are always nonnegative.

The study in this paper involves the relationship between the overall error of a surrogate model and the sample density of its inputs. In a sample space, its sample density can be equivalently represented by the measures of regions associated with individual sample points. One approach to divide a sample space into regions is the Voronoi diagram[11]. Before a sequence c_i in $[0, 1]^n$ is scaled into distance-based sample points, or inversely transformed into probability-based sample points, the domain $[0, 1]^n$ can be partitioned into regions by the Voronoi diagram of the sequence. For any point in the Voronoi region of c_i, c_i is its closest sample point using the Euclidean distance. Figures 1(a) and 1(b) show the Voronoi diagrams for the Sukharev grid (a low-dispersion sequence)[12] and the Sobol sequence (a low-discrepancy sequence)[13], respectively. A sample point is approximately in the center of each Voronoi region. The points on an edge have equal distances to the sample points whose Voronoi regions are bounded by

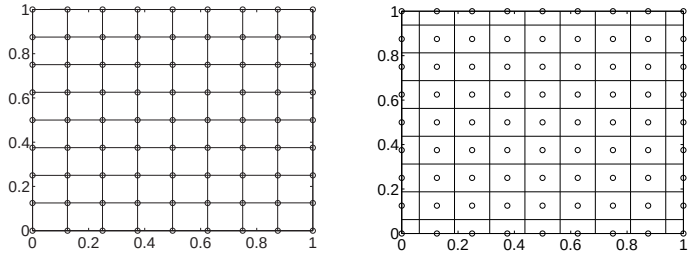
the edge. When the sequence is scaled into distance-based sample points in the x space, or inversely transformed into probability-based sample points in the x space, the edges of the cells in $[0,1]^n$ are also converted into edges in the x space. The whole region of the x sample space is separated into subregions by the edges. The measure of each region associated with a sample point in the x space is used in the formulation of the monomial MSE function in Sec. 7.



(a) 64-point Sukharev grid (b) 64-point Sobol sequence

Figure 1: Voronoi diagrams

The sample density of the inputs to a system can also be equivalently represented by the measures of the regions confined by sample points. The circles in Fig. 2 are the 64 full factorial sample points equally distributed from edge to edge. In Fig. 2(a), the squares that are confined by sample points at their vertices and have no inside sample points are the regions that represent sample density. If the points on the boundaries are viewed as half points, and the points at the four vertices ($(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$) are viewed as quarter points, the area of each square is 1 divided by the equivalent number of sample points. It is the same as the area of each Voronoi region in Fig. 2(b). The Voronoi regions bordering the boundaries are viewed as half regions, and the regions bordering the four vertices are viewed as quarter regions. The measure of the confined region or the Voronoi region is also the same as that in Fig. 1(a), since the number of sample points is the same. Therefore, the Voronoi regions and the regions confined by sample points can be viewed equivalently for the purpose of analyzing the relationship between the overall error of a surrogate model and its sample density.



(a) Confined regions (b) Voronoi diagram

Figure 2: 64-point full factorial sampling equally from edge to edge

The Sobol sequence has a desirable property that the sequence of m points in an n-dimensional space is a subset of $m+1$ or more points in an $n+1$ or higher-dimensional space. The number of regions in an x sample space can increase continuously. The full factorial sampling and the Sukharev grid do not have this property. The Sobol sequence is recommended for the study of the monomial MSE of surrogate models.

The considered region in an x sample space is A. It consists of subregions A_i, $i = 1, 2, \ldots, m$, associated with sample points x_i, $i = 1, 2, \ldots, m$. Any two subregions A_i and A_j in A satisfy $A_i \cap A_j = \emptyset$ for $i \neq j$. Therefore, the region $A = \cup A_i$. The measure of the region A_i is V_i. The measure of the entire region A is $V = \sum V_i$. The probability of the region A_i is F_i. The probability of the region A is $F = \sum F_i$. If the entire x sample space is considered, its probability F is 1.

Distance-based uniform sampling divides a sample space into m equal regions. The measure of each

region, $V_i^{(d)}$, is given by

$$V_i^{(d)} = \frac{V}{m}. \tag{5}$$

Probability-based uniform sampling divides a sample space into m regions with equal probability. The probability of each region, $F_i^{(p)}$, is given by

$$F_i^{(p)} = \frac{F}{m}. \tag{6}$$

7. MSE of a Monomial Form

RBF and Kriging are expressed as combinations of basis functions[1]. A basis function is a function of the distance between a sample point and the input to RBF or Kriging. If the number of sample points increases, the number of basis functions will also increase, and the distance between the input and its closest sample point will decrease. Meanwhile, the overall error of the surrogate model is expected to decrease. The error of a surrogate model is expected to be related to the distances between sample points, or equivalently the volume per sample point. This paper formulates the Mean Squared Error (MSE) of a monomial form based on the relationship between the squared error of a surrogate model and the volume per sample point.

Since the change of the MSE is related to the change of the volume per sample point, the MSE of the subregion A_i associated with x_i is statistically approximated as a monomial function of its volume V_i, which is given by

$$\text{MSE}(A_i) = aV_i^l. \tag{7}$$

Equation 7 statistically reflects the relationship between the MSE and the measure of a region. It is not necessarily accurate for each individual subregion. When V_i changes, a and l can also change.

Since $V_i > 0$, then $V_i^l > 0$. Since MSE is nonnegative, the parameter a is also nonnegative. The parameter a and the exponent l determine how the MSE changes as the measure V_i changes. Generally, the exponent $l > 0$. It indicates that, as V_i becomes smaller, the error also becomes smaller. It is not common that the exponent is $l < 0$.

For different values of the exponent l, the shapes of the MSE are different. For the same value of a, Fig. 3 shows the shapes for five different scenarios: (1) $l = 0$, (2) $l = 1$, (3) $0 < l < 1$, (4) $l > 1$, and (5) $l < 0$.

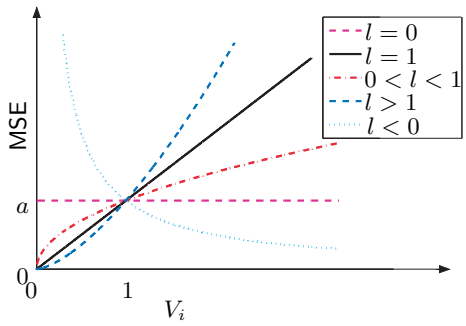


Figure 3: Monomial MSE for different l

As described in Sec. 6, in a sample space A with a measure of V and probability of F, $V = \sum V_i$ and $F = \sum F_i$. The Mean Square Error (MSE) of the surrogate model in region A_i is evaluated by

$$\text{MSE}(A_i) = \frac{\int_{A_i} (y - h(x))^2 \, dF(x)}{\int_{A_i} dF(x)} = \frac{\int_{A_i} (y - h(x))^2 \, dF(x)}{F_i}. \tag{8}$$

Therefore,

$$\text{MSE}(A_i)F_i = \int_{A_i} (y - h(x))^2 \, dF(x). \tag{9}$$

The MSE over the entire space A is given by

$$\text{MSE} = \int_A (y - h(x))^2 \, dF(x) = \sum_{i=1}^{m} \int_{A_i} (y - h(x))^2 \, dF(x) = \sum_{i=1}^{m} \text{MSE}(A_i)F_i. \tag{10}$$

Substituting MSE(A_i) by the monomial MSE expressed by Eq. 7, the overall MSE is given by

$$\text{MSE} = \sum_{i=1}^{m} aV_i^l F_i. \tag{11}$$

8. Fidelity Comparison using the MSE of a Monomial Form

8.1 Power Mean Inequality

Power mean[14]: If q is a nonzero real number, the weighted power mean with exponent q of positive real numbers b_i, $i = 1, 2, \ldots, m$ is defined as

$$M_q\left(b_1, \ldots, b_m\right) = \left(\frac{1}{m}\sum_{i=1}^{m} b_i^q\right)^{1/q}. \tag{12}$$

Power mean inequality[14]: For power means, if $q_1 < q_2$, then $M_{q_1}(b_1, \ldots, b_m) \leq M_{q_2}(b_1, \ldots, b_m)$, and the two means are equal if and only if $b_1 = \cdots = b_i = \cdots = b_m$.

8.2. Comparison of Overall MSE

Using the MSE of a monomial form, the fidelities of the two surrogate models of the same system developed respectively using probability-based and distance-based uniform sampling are compared.

For distance-based uniform sampling, substituting V_i in Eq. 11 by $V_i^{(d)}$ in Eq. 5, the overall MSE is

$$\text{MSE}^{(d)} = a\sum_{i=1}^{m}\left(\frac{V}{m}\right)^l F_i = a\left(\frac{V}{m}\right)^l \sum_{i=1}^{m} F_i = aF\left(\frac{V}{m}\right)^l. \tag{13}$$

For probability-based uniform sampling, substituting F_i in Eq. 11 by $F_i^{(p)}$ in Eq. 6, the overall MSE is

$$\text{MSE}^{(p)} = a\sum_{i=1}^{m}\left(V_i^{(p)}\right)^l \frac{F}{m} = \frac{aF}{m}\sum_{i=1}^{m}\left(V_i^{(p)}\right)^l. \tag{14}$$

The comparison between $\text{MSE}^{(d)}$ and $\text{MSE}^{(p)}$ are performed for five different scenarios: (1) $l = 0$, (2) $l = 1$, (3) $0 < l < 1$, (4) $l > 1$, and (5) $l < 0$.

Scenario 1: $l = 0$.

$$\text{MSE}^{(d)} = aF\left(\frac{V}{m}\right)^0 = aF. \tag{15}$$

$$\text{MSE}^{(p)} = \frac{aF}{m}\sum_{i=1}^{m} V_i^0 = aF. \tag{16}$$

When $l = 0$, the overall MSE is aF for both distance-based and probability-based uniform sampling. If the entire x space is considered, its probability is $F = 1$, and the overall MSE is a.

Scenario 2: $l = 1$.

$$\text{MSE}^{(d)} = \frac{aFV}{m}. \tag{17}$$

$$\text{MSE}^{(p)} = \frac{aF}{m}\sum_{i=1}^{m} V_i = \frac{aFV}{m}. \tag{18}$$

When $l = 1$, the overall MSE is aFV/m for both distance-based and probability-based uniform sampling. If the entire x space is considered, the probability is $F = 1$, and the overall MSE is aV/m.

To compare $\text{MSE}^{(d)}$ and $\text{MSE}^{(p)}$ under scenarios 3, 4, and 5, rewrite the two equations 13 and 14 as follows.

$$\left(\frac{\text{MSE}^{(d)}}{aF}\right)^{\frac{1}{l}} = \left(\frac{1}{m}\sum_{i=1}^{m}\left(V_i^{(p)}\right)^1\right)^1. \tag{19}$$

$$\left(\frac{\text{MSE}^{(p)}}{aF}\right)^{\frac{1}{l}} = \left(\frac{1}{m}\sum_{i=1}^{m}\left(V_i^{(p)}\right)^l\right)^{\frac{1}{l}}. \tag{20}$$

Scenario 3: $0 < l < 1$.
Using the power mean inequality, when $0 < l < 1$,

$$\left(\frac{\text{MSE}^{(p)}}{aF}\right)^{\frac{1}{l}} \leq \left(\frac{\text{MSE}^{(d)}}{aF}\right)^{\frac{1}{l}}. \tag{21}$$

Since $l > 0$,

$$\frac{MSE^{(p)}}{aF} \leq \frac{MSE^{(d)}}{aF}. \tag{22}$$

Since $a > 0$ and $F > 0$, $\text{MSE}^{(p)} \leq \text{MSE}^{(d)}$. $\text{MSE}^{(p)} = \text{MSE}^{(d)}$ only and only if $V_1^{(p)} = \cdots = V_i^{(p)} = \cdots = V_m^{(p)} = V/m$.
Scenario 4: $l > 1$.
Using the power mean inequality, when $l > 1$,

$$\left(\frac{\text{MSE}^{(p)}}{aF}\right)^{\frac{1}{l}} \geq \left(\frac{\text{MSE}^{(d)}}{aF}\right)^{\frac{1}{l}}. \tag{23}$$

Since $l > 1$,

$$\frac{\text{MSE}^{(p)}}{aF} \geq \frac{\text{MSE}^{(d)}}{aF}. \tag{24}$$

Since $a > 0$ and $F > 0$, $\text{MSE}^{(p)} \geq \text{MSE}^{(d)}$. $\text{MSE}^{(p)} = \text{MSE}^{(d)}$ only and only if $V_1^{(p)} = \cdots = V_i^{(p)} = \cdots = V_m^{(p)} = V/m$.
Scenario 5: $l < 0$.
Using the power mean inequality, when $l < 0 < 1$,

$$\left(\frac{\text{MSE}^{(p)}}{aF}\right)^{\frac{1}{l}} \leq \left(\frac{\text{MSE}^{(d)}}{aF}\right)^{\frac{1}{l}}. \tag{25}$$

Since $l < 0$,

$$\frac{\text{MSE}^{(p)}}{aF} \geq \frac{\text{MSE}^{(d)}}{aF}. \tag{26}$$

Since $a > 0$ and $F > 0$, $\text{MSE}^{(p)} \geq \text{MSE}^{(d)}$. $\text{MSE}^{(p)} = \text{MSE}^{(d)}$ only and only if $V_1^{(p)} = \cdots = V_i^{(p)} = \cdots = V_m^{(p)} = V/m$.
Conclusion to fidelity comparison: Two surrogate models of the same system are developed respectively using probability-based and distance-based uniform sampling with the same number of sample points. Suppose the MSE of the volume per sample point has the form of aV_i^l. The MSEs of the two surrogate models, $\text{MSE}^{(p)}$ and $\text{MSE}^{(d)}$, have the following relations for different values of exponent l.

1. $l = 0$. $\text{MSE}^{(p)} = \text{MSE}^{(d)} = aF$.

2. $l = 1$. $\text{MSE}^{(p)} = \text{MSE}^{(d)} = aFV/m$.

3. $0 < l < 1$. $\text{MSE}^{(p)} \leq \text{MSE}^{(d)}$. $\text{MSE}^{(p)} = \text{MSE}^{(d)}$ only and only if $V_1^{(p)} = \cdots = V_i^{(p)} = \cdots = V_m^{(p)} = V/m$.

4. $l > 1$. $\text{MSE}^{(p)} \geq \text{MSE}^{(d)}$. $\text{MSE}^{(p)} = \text{MSE}^{(d)}$ only and only if $V_1^{(p)} = \cdots = V_i^{(p)} = \cdots = V_m^{(p)} = V/m$.

5. $l < 0$. $\text{MSE}^{(p)} \geq \text{MSE}^{(d)}$. $\text{MSE}^{(p)} = \text{MSE}^{(d)}$ only and only if $V_1^{(p)} = \cdots = V_i^{(p)} = \cdots = V_m^{(p)} = V/m$.

9 Fitting Monomial MSE
The expression of the MSE (Eq. 13) for distance-based uniform sampling provides an approach to fit the parameter a and the exponent l. The probability F and the entire volume V of a sample space are known for a specific problem. If pairs of $\text{MSE}^{(d)}$ and m are available, a and l can be obtained by regression.
Distance-based uniform sampling can generate a series of sets of sample points. The numbers of points in these sets are $m_1, m_2, \ldots, m_t$. The corresponding volumes per point are $V/m_1, V/m_2, \ldots, V/m_t$. The values of $\text{MSE}^{(d)}$ of the surrogate models developed using these sets of points are $MSE_1^{(d)}, MSE_2^{(d)}, \ldots, MSE_t^{(d)}$.

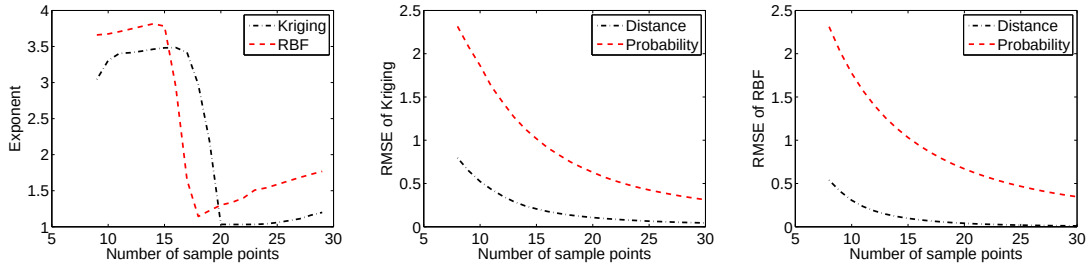
The parameter a and the exponent l in Eq. 13 are fitted using the pairs of $V/m_1, V/m_2, \ldots, V/m_t$ and $MSE_1^{(d)}, MSE_2^{(d)}, \ldots, MSE_t^{(d)}$.
Since the parameter a and the exponent l can change for different values of V_i, the series of the numbers of sample points should be appropriately selected. If many numbers are used to fit one set of a and l, the fitted values cannot accurately show how the value of l changes when the number of sample points is slightly changed. The details of the change of exponent l are lost. However, if the series is too small, the fitted result will have considerable noises. If the volumes of subregions for probability-based uniform sampling are not very different from $V/m_1, V/m_2, \ldots, V/m_t$, the fidelity comparison conclusion is expected to be accurate.

10. Testing the Fidelity Comparison Conclusion

In this section, RBF and Kriging are developed for test functions. The probability distributions of all the variables are assumed as independent Gaussian distributions. The full factorial sequence is scaled to distance-based uniform sample points, and also inversely transformed to probability-based sample points. For each test function, two series of RBF and Kriging models are developed respectively using these two sampling approaches. The parameter a and the exponent l are fitted using the surrogate models developed using the distance-based uniform sampling. The Root Mean Squared Error (RMSE) is the root of the MSE. The overall RMSE of a surrogate model is evaluated using test points.
Test function 1: 1-Variable function

$$f(x) = (6x-2)^2 \sin(2(6x-2)), \qquad (27)$$
$$\text{where} \qquad x \in [0,1]$$

The probability distribution of x is a Gaussian distribution with a mean of 0.5 and standard deviation of 0.15. Figure 4(a) shows the fitted values of exponent l for different numbers of sample points. Figures 4(b) and 4(c) show the RMSE of the surrogate models developed using Kriging and RBF, respectively. The value of exponent l is consistently greater than 1. The RMSE of the surrogate models developed using probability-based uniform sampling is consistently larger than that developed using distance-based uniform sampling.



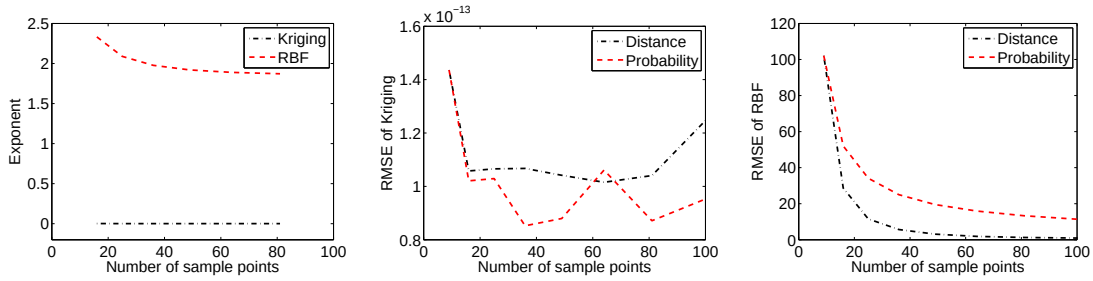
(a) Fitted exponent l (b) RMSE of Kriging (c) RMSE of RBF

Figure 4: Test function 1

Test function 2: Booth function in two-dimensional space

$$f(x) = (x_1 + 2x_2 - 7)^2 + (2x_1 + x_2 - 5)^2, \qquad (28)$$
$$\text{where} \qquad x_1 \in [-10, 10], \qquad x_2 \in [-10, 10]$$

The probability distribution of x_1 and x_2 is a bivariate Gaussian distribution with both means of 0 and standard deviations of 3.5. Figure 5(a) shows the fitted values of exponent l for different numbers of sample points. These fitted values for Kriging and RBF are significantly different. Figures 5(b) and 5(c) show the RMSE of the surrogate models developed using Kriging and RBF, respectively. For the surrogate models constructed using Kriging, the exponent l is very close to 0. The RMSE of Kriging for probability-based uniform sampling is generally lower than that for distance-based uniform sampling. The RMSE values of Kriging for both sampling approaches are close to 0. For the surrogate models constructed using RBF, the exponent l is larger than 1. The RMSE of RBF for probability-based uniform sampling is higher than that for distance-based uniform sampling.



(a) Fitted exponent l (b) RMSE of Kriging (c) RMSE of RBF

Figure 5: Test function 2

Test function 3: Hartmann function in three-dimensional space

$$f(x) = -\sum_{i=1}^{4} c_i \exp\left\{-\sum_{j=1}^{n} A_{ij}(x_j - P_{ij})^2\right\}, \tag{29}$$

$$\text{where} \qquad x = (x_1, x_2, \ldots, x_n) \qquad x_i \in [0, 1]$$

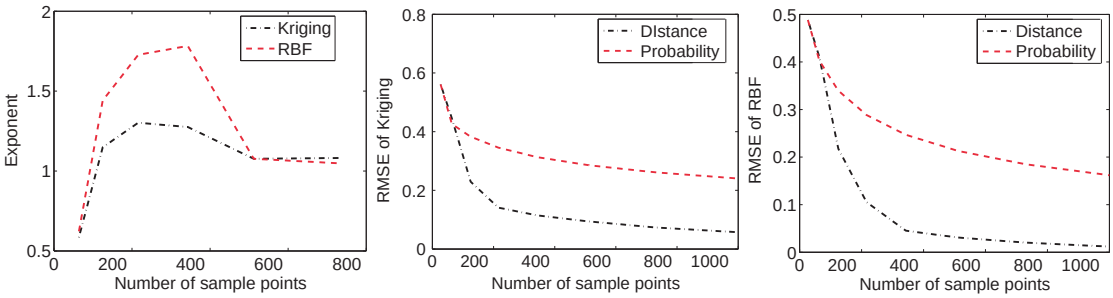
The parameter vector, c, is given by $c = [1, 1.2, 3, 3.2]^T$. The parameters, A and P, are given by

$$A = \begin{bmatrix} 3.2 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \end{bmatrix},$$

and

$$P = \begin{bmatrix} 0.3689 & 0.1170 & 0.2673 \\ 0.4699 & 0.4387 & 0.7470 \\ 0.1091 & 0.8732 & 0.5547 \\ 0.03815 & 0.5743 & 0.8828 \end{bmatrix}.$$

The probability distribution of x_1, x_2, and x_3 is a trivariate Gaussian distribution with all three means of 0.5 and standard deviations of 0.15. Figure 6(a) shows the fitted values of exponent l for different numbers of sample points. Figures 6(b) and 6(c) show the RMSE of the surrogate models developed using Kriging and RBF, respectively. When the number of sample points is small, l is smaller than 1. The RMSE of both Kriging and RBF developed using probability-based uniform sampling are smaller than those developed using distance-based uniform sampling. When the number of sample points becomes large, l becomes larger than 1. The RMSE of the surrogate models developed using probability-based uniform sampling are larger than those developed using distance-based uniform sampling.



(a) Fitted exponent l (b) RMSE of Kriging (c) RMSE of RBF

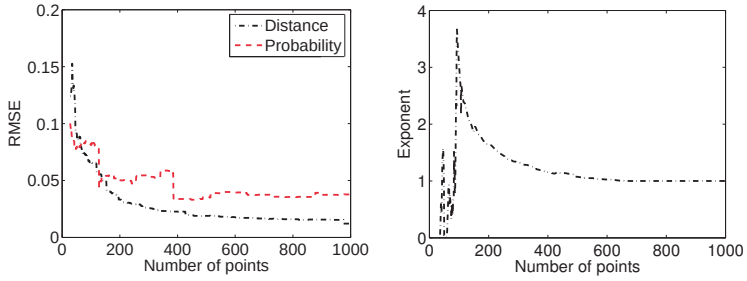
Figure 6: Test function 3

11. Surrogate Models Used for Window Performance Evaluation

Distance-based and probability-based uniform sampling approaches are used to develop surrogate models

representing the heat transfer rates of a triple pane window under varying climatic conditions[15]. Three climatic conditions, namely, the air temperature, the wind speed, and the solar radiation, are the inputs to the surrogate models. The heat transfer rate of the window is the output of each of the surrogate models. January and August are chosen as the typical months in winter and summer, respectively. The heat transfer rates under sampled climatic conditions are evaluated by Fluent[16] simulations.
The target location for window performance evaluation is Michigan, ND. Its climatic data are obtained from the North Dakota Agricultural Weather Network[17]. The distributions of air temperature, wind speed, and solar irradiance are assumed to be a Gaussian distribution, a Weibull distribution, and a Gamma distribution, respectively. All the distributions are fitted using the maximum likelihood estimation method[18]. In either January or August from 2006 to 2010, there are 3720 observations of climatic conditions. They are used as test points to evaluate the MSE and RMSE of the surrogate models. The fitted value of l is constrained to be grater than 0 in this section. The Kriging models are developed using the DACE Matlab Kriging toolbox[19].

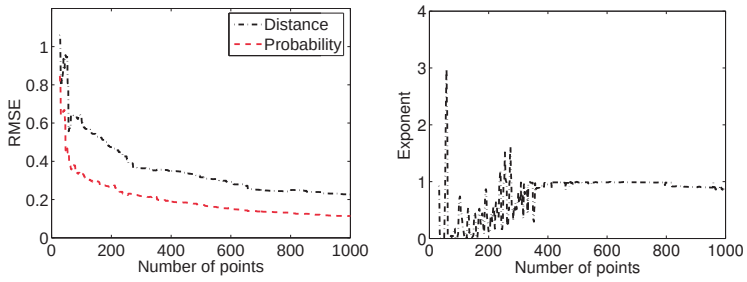
11.1. Sample Points Transformed from the Sobol Sequence
Probability-based and distance-based sample points are generated from Sobol sequences for both January and August. Four series of Kriging surrogate models are developed respectively using probability-based sample points for January, distance-based sample points for January, probability-based sample points for August, and distance-based sample points for August. The numbers of training points for Sobol sequences used for these four surrogate models are 27 to 1000. The exponent of the monomial MSE is fitted using the $\text{MSE}^{(d)}$ of the surrogate models developed using consecutively increasing numbers of distance-based sample points. To fit the exponent for a specific number of sample points, 9 surrogate models trained by consecutively-increasing numbers of sample points are used. The fitted value of the exponent is for the middle number of each group of 9 numbers.
The RMSEs of the two surrogate models for January are shown in Fig. 7(a). Figure 7(b) shows the values of the exponent of the monomial MSE function for January. The RMSEs of the two surrogate models for August are shown in Fig. 8(a). Figure 8(b) shows the values of the exponent of the monomial MSE function for August. Generally, when the exponent is between 0 and 1, the RMSE of the surrogate models developed using probability-based sampling are less that of the surrogate models developed using distance-based sampling. Generally, when the exponent is greater than 1, the comparison result is reversed.



(a) RMSE (b) Exponent

Figure 7: Kriging models for January using Sobol sequences

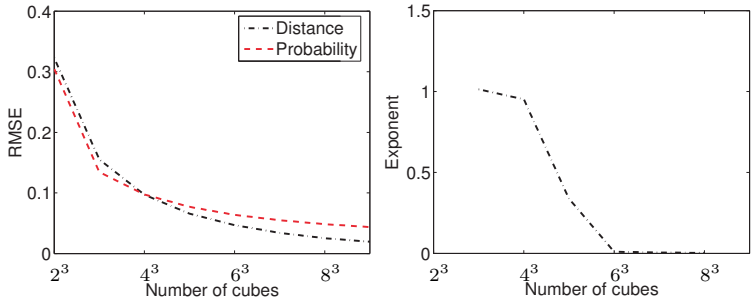
11.2. Sample Points Transformed from Full Factorial Sampling Sequence
Probability-based and distance-based sample points are generated from full factorial sampling sequences equally distributed from edge to edge for both January and August. The numbers of points in each dimension are 3, 4, 5, 6, 7, 8, 9, and 10. Correspondingly, the total numbers of training points in the three-dimensional sample space are 3^3, 4^3, 5^3, 6^3, 7^3, 8^3, 9^3, and 10^3; and the total numbers of confined cubes are 2^3, 3^3, 4^3, 5^3, 6^3, 7^3, 8^3, and 9^3. Four series of Kriging surrogate models are developed respectively using probability-based sample points for January, distance-based sample points for January, probability-based sample points for August, and distance-based sample points for August. The exponent of the monomial MSE is fitted using the $\text{MSE}^{(d)}$ of surrogate models trained by consecutively increasing numbers of distance-based sample points. To fit the exponent for a specific number of sample points, 3 consecutively-increasing numbers of training points are used. The fitted value of the exponent is used for the middle number of the 3 numbers of training points.



(a) RMSE (b) Exponent

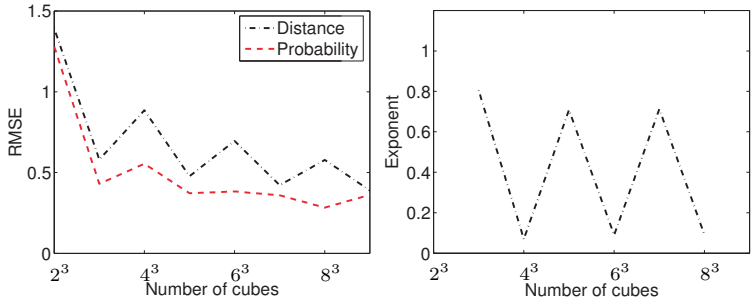
Figure 8: Kriging models for August using Sobol sequences

The RMSEs of the two surrogate models for January are shown in Fig. 9(a). Figure 9(b) shows the values of the exponent of monomial MSE function for January. The RMSEs of the two surrogate models for August are shown in Fig. 10(a). Figure 10(b) shows the values of the exponent of monomial MSE function for August. The RMSE values of the two on Fig. 9(a) are very close. Figure 10(b) shows that the values of exponent for the surrogate models of August are between 0 and 1. Figure 10(a) shows that the RMSE values of surrogate models developed using inverse transform sampling are less than those using direct full factorial sampling.



(a) RMSE (b) Exponent

Figure 9: Kriging models for January using full factorial sampling



(a) RMSE (b) Exponent

Figure 10: Kriging models for August using full factorial sampling

12. Concluding Remarks

The mean squared error of a monomial form is formulated in this paper based on the relationship between the mean squared error of a surrogate model and the volume or hypervolume per sample point. Probability-based and distance-based uniform sampling approaches generate points uniformly distributed in sample space in terms of probability and distance, respectively. Using these two sampling approaches with the same number of points, two surrogate models are developed to approximate the same system.

Their fidelities are compared using the monomial MSE function. When the exponent of the monomial function is between 0 and 1, the fidelity of the surrogate model trained using probability-based uniform sampling is higher than that of the other one trained using distance-based uniform sampling. When the value of the exponent is greater than 1 or less than 0, the fidelity comparison is reversed. This theoretical conclusion is successfully verified using standard test functions and an engineering application.

13. References

[**1**] A. I. J. Forrester, A. Sobester, and A. J. Keane, Engineering Design via Surrogate Modelling: A Practical Guide, Wiley, Chichester, West Sussex, UK, 1st edition, 2008.

[**2**] J. I. Madsen, W. Shyy, and R. T. Haftka, Response surface techniques for diffuser shape optimization, AIAA Journal, 38(9):1512-1518, 2000.

[**3**] J Zhang, A Messac, J Zhang, and S Chowdhury, Adaptive Optimal Design of Active Thermoelectric Windows Using Surrogate Modeling, Optimization and Engineering (Accepted).

[**4**] F. P. Miller, A. F. Vandome, and M. B. John, Inverse Transform Sampling, VDM Publishing, Saarbrcken, Germany, 2010.

[**5**] J. von Neumann, Various techniques used in connection with random digits, Nat. Bureau Stand. Appl. Math. Ser., 12:36-8, 1951.

[**6**] A. W. Marshall, The use of multi-stage sampling schemes in Monte Carlo computations, H. A. Meyer (ed.), Symposium on Monte Carlo Methods, Wiley, Hoboken, NJ, 1956.

[**7**] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Markov Chain Monte Carlo in Practice, Interdisciplinary Statistics. Chapman & Hall, London, 1996.

[**8**] S. M. LaValle. Planning Algorithms, Cambridge University Press, Cambridge, UK, 2006.

[**9**] H. Niederreiter, Point sets and sequences with small discrepancy, Monatshefte fr Mathematik, 104:273-337, December 1987.

[**10**] E. L. Lehmann and G. Casella, Theory of Point Estimation, Springer, New York, 2nd edition, 1998.

[**11**] F. Aurenhammer, Voronoi diagrams - a survey of a fundamental geometric data structure, ACM Computing Surveys, 23(3):345-405, 1991.

[**12**] A. G. Sukharev, Optimal strategies of the search for an extremum, Computational Mathematics and Mathematical Physics, 11(4):910-924, 1971.

[**13**] I. M. Sobol, Uniformly distributed sequences with an additional uniform property, USSR Computational Mathematics and Mathematical Physics, 16:236-242, 1976.

[**14**] P. S. Bullen, Handbook of Means and Their Inequalities, Mathematics and Its Applications, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

[**15**] J. Zhang, A. Messac, J. Zhang, and S. Chowdhury, Improving the accuracy of surrogate models using inverse transform sampling, 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, number AIAA 2012-1429, Honolulu, Hawaii, April 2012.

[**16**] ANSYS Fluent getting started guide, 2012, http://www1.ansys.com.

[**17**] NDAWN, The North Dakota agricultural weather network, http://ndawn.ndsu.nodak.edu.

[**18**] A. Hald, On the history of maximum likelihood in relation to inverse probability and least squares, Statistical Science, 14(2):214-222, 1999

[**19**] S. N. Lophaven, H. B. Nielsen, and J. Sondergaard, DACE: A Matlab Kriging toolbox, Technical University of Denmark, http://www2.imm.dtu.dk.