CHARACTERIZING PROBABILITY-BASED SAMPLING FOR HIGH-FIDELITY SURROGATE MODELING

By

Junqiang Zhang

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: MECHANICAL ENGINEERING

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Prof. Achille Messac, Thesis Adviser

Prof. Prabhat Hajela, Member

Prof. Nikhil Koratkar, Member

Prof. Thomas C. Sharkey, Member

Rensselaer Polytechnic Institute
Troy, New York

April 2012
(For Graduation December 2012)

ABSTRACT

The fidelity of surrogate models remains one of the primary concerns in their application to represent complex system behavior. Appropriate sampling of training points is one of the main factors affecting the fidelity of surrogate models. This dissertation investigates the criteria that regulate the relative advantage of probability-based sampling over distance-based sampling, for systems where the inputs follow a distribution.

This dissertation first explores the important characteristics of probability-based sampling. In this context, the probability of occurrence of physical parameters are assumed to be known, assumed, or predefined. Conventional representations of sample density (or sample crowding) involve metrics defined in terms of coordinate-distances in the actual variable/parameter space. This dissertation instead defines novel (sampling) metrics in terms of the probability of occurrence of the concerned parameters. For generating sample points, a sequence of numbers between 0 and 1 can either be scaled to the values of coordinates, or be expressed as the values of probabilities. Owing to the similarities between the metrics defined by coordinates and probabilities, a sequence of probabilities is characteristically similar to that of coordinates. The properties of low-dispersion and low-discrepancy sequences of probabilities are studied. This dissertation also compares the probability densities of probability-based sample points and distance-based sample points, when they are transformed or scaled from the same sequence of numbers between 0 and 1. It is proved that, in each dimension, if the Probability Density Function (PDF) is monotonic, the sample points obtained using inverse transform sampling by solving the inverse Cumulative Distribution Function (CDF) have higher PDF values than those of the corresponding points directly scaled from the sequence. This conclusion also holds for joint PDF values if the distributions in a multi-dimensional space are independent.

To study the suitability of probability-based sampling for surrogate modeling, Mean Squared Error (MSE) of a monomial form is formulated based on the relation-

ship between the squared error of a surrogate model and the volume or hypervolume per sample point (equivalently, the volume or hypervolume of the region between sample points). To generate training points with the same distribution as the test points, inverse transform sampling is used to evaluate the coordinates of a sequence of probabilities using the inverse CDF function of the test points. The obtained coordinates are used as training points to develop the surrogate model. This sampling approach is probability-based sampling. The same sequence is also directly scaled to the coordinates of training points to develop another surrogate model. This sampling method is distance-based sampling. If the sequence is uniform, the fidelities of the two surrogate models can be compared using the monomial loss function. The exponent of the monomial function indicates which of the two surrogate models has higher fidelity. When the exponent of the monomial function is between 0 and 1, the fidelity of the surrogate model trained using probability-based sampling is higher than that trained using distance-based sampling. When the value of the exponent is greater than 1, the fidelity of the surrogate model developed using distance-based sampling is higher than that using probability-based sampling. This theoretical conclusion is verified using test functions.

The probability-based sampling is applied to the development of surrogate models for window performance evaluation. Sobol sequences and uniform grid sampling are used to generate the training points. The comparison of the fidelities of surrogate models developed using probability-based sampling to that using distance-based sampling agrees with the hypothesized conditional advantage of probability-based sampling.

Surrogate modeling is used for the optimal control of an active thermoelectric window operating under varying climatic conditions. Typical climatic conditions and the optimal operations of the window under these conditions are used as training points for surrogate models. The operation of the window is maintained energy efficient using the developed surrogate models.