Adaptive Sampling based Sampling Strategies for the DACE Surrogate Model for Expensive Black-box functions

AE 497 B.Tech. Project

By

Ankur Kulkarni

02001003

Under the guidance of

Prof. P. Mujumdar

Department of Aerospace Engineering,

Indian Institute of Technology, Bombay

April 2006

Table of Contents

Abstract **iii**

List of Figures **iv**

List of Tables **v**

Nomenclature **vi**

Certificate **ix**

1. Introduction **1**

1.1 Motivation 1

1.2 Aim and Scope of the Project 2

1.3 Layout of Report 2

2. Surrogate Modeling **3**

2.1 Computer Experiments and Surrogate Models 3

2.2 Surrogate models in literature 3

2.3 Functional Surrogate Models 4

2.3.1 Kriging 4

3. Universal Kriging and DACE **6**

3.1 Universal kriging structure 6

3.2 Kriging Predictor 7

3.3 Characteristics of the Kriging Predictor 7

3.4 Understanding the *MSE* 9

4. **Adaptive Sampling** **12**

4.1 The need for Adaptive Sampling 12

4.2 Issues arising in Adaptive Sampling 13

4.2.1 Initial Sampling Strategy 13

4.2.2 Infill Sampling Criteria 13

4.2.3 Convergence 15

4.2.4 Solution of the Infill Sampling Criteria 15

5. Infill Sampling Criterion 16

5.1 Framework for testing results 16

5.1.1 Measure of accuracy 16

5.1.2 Method of testing 17

5.2 Global Maximum of MSE as the ISC 17

5.3 Dual criteria adaptive sampling 23

5.3.1 Some comments about Dual Criteria ISC 27

5.4 Further research and exploration 27

6. Conclusions 28

References 29

Acknowledgements 32

Appendix 1 33

Appendix 2 35

Appendix 3 36

Abstract

Conducting physical experiments constitutes many practical difficulties. It has hence become practice to replace physical experiments by sophisticated computer codes. CFD, FEM are examples of such codes. High fidelity computer codes have now entered many areas of scientific research. With increase in fidelity, there has been a steady increase in the expense of running these codes. Hence the need has arisen to find a way of using cheap surrogates to replace the expensive computer codes. These surrogates approximate the behavior of the computer output.

In this report we survey a wide variety of surrogate models and study in depth the universal Kriging surrogate model and its so called DACE implementation. The objective of this report is to develop a sampling strategy for expensive black-box functions for the DACE surrogate model. DACE uses a probabilistic linear model to model a deterministic function, which in our case is an expensive computer code. The maximum likelihood estimation method is used to determine parameters of this model. The predictor, which is to serve as the approximation to the original computer code is taken to be linear and unbiased. The sampling strategy we have proposed is called adaptive sampling. Past uses of adaptive sampling in the Kriging context has been surveyed in the report. The sampling strategy is characterized by its infill sampling criterion. We demonstrate results obtained by using the maximum of the mean square error as the infill sampling criterion. We also present results obtained by the use of a new and original sampling criterion that we have developed called the Dual Criteria Infill Sampling Criterion. Satisfying the infill sampling criterion requires the solving of a global optimization problem. We have assumed in this report that the global optimum is obtainable, and it has been calculated by finely griding the domain.

List of Figures

1.1 Surrogate modeling philosophy 1

3.1 The function $y(x) = x\sin(x)$ and DACE

3.1a The function $y(x) = x\sin(x)$, Kriging predictor and sites in S 8

3.1b The *MSE* for $y(x) = x\sin(x)$ and its Kriging predictor. 9

3.2 Problem of aliasing and insufficient sampling

3.2a The function, the DACE predictor and sampled data 10

3.2b MSE 10

3.3 Effect of adding a new sample to an aliased function

3.3a DACE predictor after adding new sample 17

3.3b MSE 17

5.1 Sampling on peaks function with 3x3 uniform initial samples

5.1a $|\, y(x) - y_{\mathrm{p}}(x) \,|$ with samples 19

5.1b MSE with samples 19

5.2 Progress of adaptive sampling on peaks function with 3x3 initial sampling

5.2a NAD with number of injected samples 19

5.2b IE with number of injected samples 19

5.3 Percentage reduction in number of samples with MSE as ISC 21

5.4 Effect of Adaptive sampling on the Goldstein Price function

5.4a NAD with number of samples injected 21

5.4b IE with number of samples injected 21

5.5 MSE for Goldstein Price function

5.5a With 25 initial samples 22

5.5b With 25 samples + 1 injected sample 22

5.6 $|\, y(x) - y_{\mathrm{p}}(x) \,|$ for Goldstein Price 22

5.7 NAD for the peaks function with 81 initial samples with different ISCs

5.7a With DCISC 25

5.7b With only MSE 25

5.8 Percentage reduction in number of samples with DCISC 26

List of Tables

5.1 Adaptive sampling on the Peaks function using MSE as ISC 18

5.2 Dual criteria adaptive sampling on Peaks function 24

5.3 Comparative performance of DCISC and MSE based criterion 26

(a) Rosenbrock's valley function 26

(b) Branin's rcos function 26

(c) Goldstein Price function 26

Nomenclature

SYMBOLS

C_d	Coefficient of Drag
$c(x)$	m Vector of predictor functions
D	Design space subset of $\mathbb{R}^n$
det	Determinant
$f_i(x)$	The ith regression polynomial
$f(x)$	Vector of containing regression polynomials
E	Expectation
F	$m \times p$ matrix of regression polynomials evaluated at sites in S
g	superEGO parameter
I	Improvement function
L	Lagrangian
m	Number of sampled sites
n	Dimension of design space
p	Total number of regression polynomial; exponent of Gaussian correlation model
P	Probability
q	Dimension of output space
r	Vector of correlations between the untried point and design sites
$\mathbb{R}$	Set of Real numbers
R	Correlation matrix
s	A site in the design space
S	Set of m sites in design space

T	Tunneling function
x	A point in $\mathbb{R}^{\mathrm{n}}$; a general point in the design space
x_{i}	i^{th} component of x
x^*	Point of maximum discrepancy
X_{i}	Data points
y	A function $\mathbb{R}^{\mathrm{n}} \rightarrow \mathbb{R}^{\mathrm{q}}$ defined over D
y_{p}	DACE predictor
Y	Vector of outputs of y at sites in S
z	Stochastic process
Z	Vector of errors between the output Y and regression model

β	Vector of regression parameters; parameters of neural network
β^*	MLE of β
ϕ	Radial basis function; probability density function
Φ	Cumulative distribution function
ψ	Probability density function
γ	A m vector
Γ	Vector that holds the information of $f\left(x\right), x_{\mathrm{i}}^{*} and \lambda$
λ	Lagrange multipliers; pole strength in tunneling
ρ	Correlation function
σ	Square root of process variance; activation function for neural network
σ^{*2}	MLE of σ^2
θ	n Vector of correlation parameters; parameter of radial basis functions

ACRONYMS

argmax	Argument that maximizes
CFD	Computational Fluid Dynamics
BLUP	Best Linear Unbiased Predictor
DACE	Design and Analysis of Computer Experiments
DOE	Design of Experiments
EGO	Efficient Global Optimization
FEM	Finite Element Method
GA	Genetic Algorithms
iid	Independent and identically distributed
ISC	Infill Sampling Criterion or Criteria
IE	Integrated error
min	Minimize
MLE	Maximum Likelihood Estimation or Estimator
MSE	Mean Square Error
NAD	Normalized absolute deviation
superEGO	Super Efficient Global Optimization

Certificate

Certified that this B.Tech Project Report titled “**Adaptive Sampling based Sampling Strategies for the DACE Surrogate Model for Expensive Black-box functions**” by Ankur Kulkarni is approved by me for submission. Certified further that, to the best of my knowledge, the report represents work carried out by the student.

Date: Prof P. Mujumdar

Chapter 1

Introduction

Study of physical phenomena by simulating them using computer codes (computer models) has now become common among researchers. In the engineering context, experiments are always a series of tests carried out by changing system variables that are expected to have a bearing on the phenomenon being studied [1]. Physical experiments, like wind tunnel testing for instance, require large infrastructure, careful handling and also man power. Some physical experiments, like a study of weather trends, are impossible to conduct manually. With the greater availability of computing power computer codes started being used as a convenient replacement for physical experiment. Most engineering problems are not amenable to analytical solutions. Codes are also being used in such domains.

1.1 Motivation

With time the fidelity of computer models to nature has also steadily increased. Even with the fastest computers, computing expense is a problem in design optimization with high fidelity simulation. A need has now arisen to replace expensive computer simulations with alternative cheap surrogates in this arena [2]. These surrogates are approximate models which replace the behavior of the original high fidelity code. Surrogate modeling is now an active area of research. Figure (1.1) shows the entire philosophy of surrogate modeling in a flow chart.

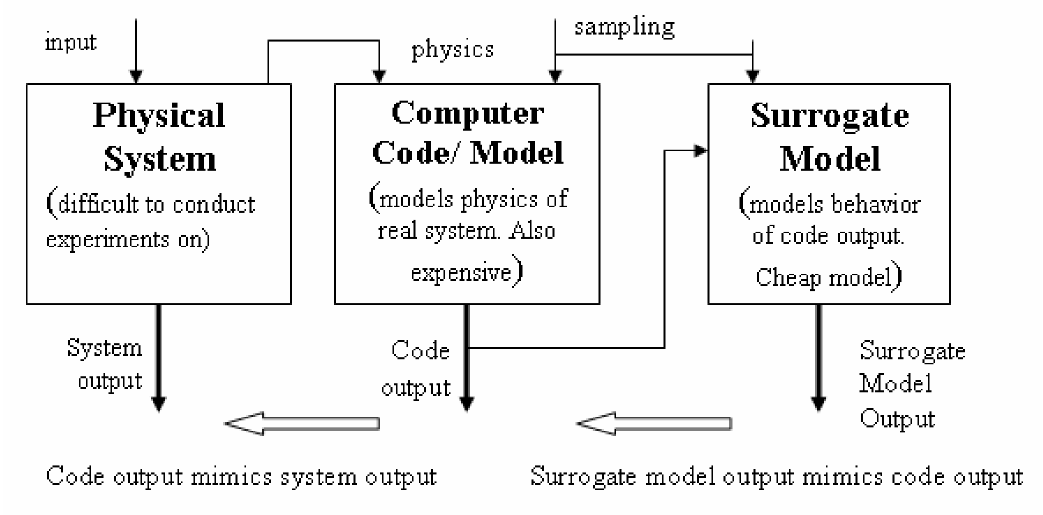

input
physics
sampling
Physical System
(difficult to conduct experiments on)
Computer Code/ Model
(models physics of real system. Also expensive)
Surrogate Model
(models behavior of code output. Cheap model)
System output
Code output
Surrogate Model Output
Code output mimics system output
Surrogate model output mimics code output

Fig (1.1) Surrogate modeling philosophy

1.2 Aim and Scope of the Project

This project deals with surrogate modeling and our aim is to develop a sampling strategy for the use of the *DACE surrogate model* [3]. The scope of the project is limited to surrogates for computer models only. Computer codes have some peculiarities which drive the methodology behind surrogate modeling and due to which they also warrant a different sampling strategy. Computer codes are black boxes i.e. usually no analytical expression is available for their output. Hence no a priori knowledge of the output variation with change in input is available. Computer codes are deterministic, i.e. there is no systematic, random, or human error involved in running a computer code. The DACE surrogate model mentioned above creates an approximation from a given set of samples for such deterministic black box functions. As shown in Fig. (1.1) the original computer code has the information of the physics of the system, but the surrogate model gets information about the output of the code only by sampling it. It must be noted that for black box codes surrogate modeling involves creating a model for the code output without any prior knowledge of the output variation. Hence sampling the black box function intelligently as well as choosing the surrogate model framework is the key to obtaining a good approximation. We shall elaborate on these key issues further in the report.

1.3 Layout of Report

The layout of this report is as follows. In Chapter 2 we introduce the concept of surrogate modeling precisely and survey the various kinds of surrogate models in literature. We introduce concept Design and Analysis of Computer Experiments (DACE) [3]. Chapter 3 deals in depth with a type of surrogate modeling methodology known as universal Kriging, its so called “DACE implementation” and the peculiarities of the DACE predictor. In Chapter 4 we survey a methodology called adaptive sampling as a strategy for improving the accuracy of the Kriging predictor. As we shall see, this strategy of adaptive sampling hinges on its “infill sampling criterion”. Chapter 5 shows results obtained by using 2 infill sampling criteria- one which is conventional, and the other which is an original one that we have developed. This Chapter will also provide pointers for further research and exploration in this field.

Chapter 2

Surrogate Modeling

2.1 Computer Experiments and Surrogate Models

The objective of the surrogate model, as shown in Fig. (1.1) of Chapter 1 is to predict the values of deterministic function $y(x)$, $x \in \mathbb{R}^{\mathrm{n}}$, $y \in \mathbb{R}^{\mathrm{q}}$ over a variable-space $D, D \subset \mathbb{R}^{\mathrm{n}}$ when its values are known only at a limited, finite number of sites contained in $S = \{s_1, s_2, ..., s_{\mathrm{m}} \mid s_{\mathrm{i}} \in D\}$. To make a good surrogate model we are free to choose S as we like. *m* is usually bounded by operational constraints. Finding these design sites and a suitable surrogate model is the objective of DACE. According to ref. [4] the problem of creation of a surrogate model for a computer experiment can be subdivided into 2 parts:

1. *The Design problem:* At which sites in $S = \{s_1, s_2, ..., s_{\mathrm{m}} \mid s_{\mathrm{i}} \in D\}$ should the output data $Y = \left[y(s_1), y(s_2), ..., y(s_{\mathrm{m}}) \right]^{\mathrm{T}}$ be collected?
2. *The Analysis problem:* How should the data be used to make a surrogate model that will help us predict $y(x)$ for all $x \in D$ with reasonable accuracy?

Note that this methodology is applicable only when $y(x)$ is a deterministic function.

2.2 Surrogate models in literature

Taking a satisfactory set of sites *S* as given, we shall first see the types of surrogate models existing in literature. Surrogate models are of various "forms" and varying in complexity. They can be broadly classified into two types- *functional models* and *physical models* [2], [5]. Physical models are mathematical models obtained by the conventional ideology of modeling the actual process using physical laws. These models may be *physical approximations* (for e.g. modal analysis of a beam using finite modes) or *mathematical approximations* (for e.g. Taylor series and finite difference approximations) [2], [5]. Hence CFD codes and FEM codes also qualify as surrogate models. The computer code referred to in Fig. (1.1) of Chapter 1 is also by this definition, a surrogate model. Functional models are mathematical constructs that simply mimic the behaviour of the output of the process. In general they have no physical basis and can be constructed for any system without the knowledge of the governing equations. Hence functional models exist only in the context of prior sampled data. On the other hand, they are generic and can be applied to a wide class of problems. The surrogate model showed in Fig. (1.1) of Chapter 1 falls under this category of surrogate models. In this report, we shall be focusing on functional type of surrogate models.

2.3 Functional Surrogate Models

Surrogates can be classified on the basis of the methodology used for generating these models. Some functional models are pure interpolating approximations (for e.g. spline interpolation). Under certain conditions and given sufficient data points they interpolate the data over the entire domain. Reference [6] provides a list of some of these surrogate models. The other chief methodologies for surrogate modeling are [2]:

1. *Regression models***:** polynomial, response surface models, rational function approximations, Kriging, wavelets. They consist of the broadest class of surrogate models. They use algebraic expressions as basis functions to fit the sampled data.
2. *Radial functions***:** Kriging, neural networks, radial basis functions. They use combinations of basis functions localized around sampled points.
3. *Single point approximations***:** reciprocal approximations, conservative approximations, posynomial approximations.

2.3.1 Kriging

We mentioned Kriging above as a part of both regression models and radial basis functions. Indeed Kriging is in many ways a cross between the two. In this approach the underlying process is assumed to be a superposition of a linear model and departures from the linear model [4].

$$\text{Actual Process} = \text{Linear model} + \text{Systematic departures} \tag{2.1}$$

Such a model is also called a “probabilistic linear model” [9]. Based on the knowledge of the underlying process being studied, the systematic departures are modeled as stochastic process of a certain kind [3]. The approach of using a realization of a stochastic process has traditionally been used in geostatistics under the name of Kriging. Derived from a miner named Krige, Kriging now is synonymous with spatial prediction. The linear model is a taken as regression model. Thus,

$$\text{Linear model} = \sum_{i=1}^{\text{p}} f_i \beta_i \tag{2.2}$$

where $\beta_i\,'s$ are regression parameters and f_i’s are regression functions. β is found as the least squares solution of the regression problem. The stochastic process is taken to be a decaying function of the distance from a point. Hence Kriging also qualifies as a type of radial basis function surrogate model. And it indeed produces an approximation of the kind $y(x) \approx f(x)^{\mathrm{T}} \beta + r(x)^{\mathrm{T}} \gamma$ Depending on the choice of the stochastic process we get different kinds of Kriging approximations. For e.g. ref. [10] shows how Kriging can be related to an approximation of cubic splines with a certain choice of the stochastic process. Ref [11] also discusses how the stochastic process model is related to the traditional DOE methods. Ref [12] discusses the relationship between kringing and radial basis functions.

There are several types of Kriging methods- simple Kriging, ordinary Kriging, indicator Kriging, universal Kriging to name a few. Brief introductions to simple, ordinary and universal Kriging can be found in ref [3] and to indicator Kriging in [13].

In ref. [4] the entire framework of universal Kriging was put into a form that could be used to make approximations to expensive and complex computer models. In some sense this framework provides a “model of a model”- a function surrogate model of an expensive physical surrogate model. Such a model is often referred to in literature as a *metamodel.* This framework is also called *Design and Analysis of Computer Experiments* (DACE), after the paper in Ref. [4] by Sacks et al, and is available as a code in MATLAB [14], [15]. The only reason why DACE is amenable to computer models and not to physical systems is that it is applicable only to deterministic systems. Essentially DACE can be applied to all kinds of processes that are deterministic. Reference [16] provides a good discussion of the distinction between DOE and DACE and the distinction between deterministic and random process from the point of view of industry experiences and their impacts on designs. Henceforth in this report we shall use DACE only in the context of surrogates for black box computer models.

In the absence of any prior information about the black box computer model DACE requires the sampling be done in a space-filling way. Reference [14] mentions about 3 such sampling strategies, viz. random sampling, uniform sampling and Latin hypercube sampling.

In this report we shall be studying in depth the universal Kriging method and its implementation in the form of DACE described by [4], [3] and [14].

Chapter 3
Universal Kriging and DACE

3.1 Universal Kriging structure

1. As we had seen in Chapter 2, the Kriging structure assumes that the underlying process for scalar $y\left(x\right)$ is given by

$$y\left(x\right)=\sum_{i=1}^{\mathrm{p}} f_i\left(x\right)\beta_i+z\left(x\right) \tag{3.1}$$

where $z\left(x\right)$is a stochastic process, f_i's are known regression functions,β_i's are unknown regression parameters. In the case that $y\in\mathbb{R}^{\mathrm{q}}$ Eq. (3.1) becomes

$$y_{\mathrm{k}}\left(x\right)=\sum_{i=1}^{\mathrm{p}} f_{\mathrm{i}}\left(x\right)\beta_{\mathrm{i;k}}+z_{\mathrm{k}}\left(x\right),\ \mathrm{k=1,2,...,q} \tag{3.2}$$

2. $z\left(x\right)$is assumed to be a process of 0 mean and constant variance. That is,

$$E\left(z\left(x\right)\right)=0\,,E\left(z(x).z\left(x\right)\right)=\sigma^2 \tag{3.3}$$

Hence the covariance between the systematic departures at different sites x,x' is

$$E\left(z(x).z\left(x'\right)\right)=\sigma^2\rho\left(x,x'\right) \tag{3.4}$$

ρ is called the "correlation function" and σ^2 is called the process variance. ρ is assumed to be only a function of $(x-x')$ and its parameters can found such that they suit the data best. In this respect universal Kriging is different from other radial basis functions in which the parameters of ρ are specifiable by the user. Clearly, from Eq. (3.3) and Eq. (3.4) $\rho\left(x,x\right)=1$. This means that the model is deterministic. In ref [3], [14] $\rho\left(x,x'\right)$is given by

$$\rho\left(x,x'\right)=\prod_{1}^{\mathrm{n}}\exp\left(-\theta_{\mathrm{j}}\,|\,x_{\mathrm{j}}-x'_{\mathrm{j}}\,|^{\mathrm{p_j}}\right) \tag{3.5}$$

where θ_{j}>0 along with σ are parameters of the stochastic process. $0<\mathrm{p_j}\le 2$ are specified constants. Model in Eq. (3.5) is called the Gaussian correlation model. The variance-covariance matrix R is defined for $s_{\mathrm{i}},s_{\mathrm{j}}\in S$ as follows.

$$R_{\mathrm{i,j}}=\rho\left(s_{\mathrm{i}},s_{\mathrm{j}}\right) \tag{3.6}$$

The choice and the number of regression functions f_{i}'s depends on our understanding of the nature of underlying process. They are usually lower order polynomials.

In compact form Eq. (3.1) is written as

$$y(x) = f(x)^{\mathrm{T}} \beta + z(x) \tag{3.7}$$

$$f(x) = \left[f_1(x), f_2(x), ..., f_{\mathrm{p}}(x) \right]^{\mathrm{T}} , \beta = \left[\beta_1, \beta_2, ..., \beta_{\mathrm{p}} \right]^{\mathrm{T}}, \ \theta = [\theta_1, \theta_2, ..., \theta_{\mathrm{n}}]^{\mathrm{T}} \tag{3.8}$$

$z(x)$ is a function of σ^2 and θ**.** The task now is two find the parameters β , σ^2 and θ which is posed as maximum likelihood estimation problem. The maximum likelihood estimation technique and its use for Kriging is described in Appendix 1.

3.2 Kriging Predictor

The predictor of universal Kriging is, according to Ref. [3],

1. *Linear with respect to the output data*
2. *Unbiased.*

Linearity implies that the predictor must be of the form given below.

$$y_p(x) = c(x)^{\mathrm{T}} Y \tag{3.9}$$

Unbiasedness implies that $E\left(y_p(x)\right) = E\left(y(x)\right)$. For any given data, there exist infinite predictors that are both linear and unbiased. The "Best Linear Unbiased Predictor" (BLUP) is one that gives minimum mean square error between the predictor and the function. Thus finding the BLUP requires that we

1. Minimize $E\left(| y_p(x) - y(x) |^2\right)$ with respect to $c(x)$
2. subject to $E\left(y_p(x)\right) = E\left(y(x)\right)$ (3.10)

The set of conditions in (3.10) now form a constrained optimization problem, which is solved by a method of Lagrange multipliers in Appendix 2. Solution of this problem will yield us the BLUP for the data S .

$$y_p(x) = r(x)^{\mathrm{T}} R^{-1} Y - \left(F^{\mathrm{T}} R^{-1} r(x) - f(x)\right)^{\mathrm{T}} \left(F^{\mathrm{T}} R^{-1} F\right)^{-1} F^{\mathrm{T}} R^{-1} Y \tag{3.11}$$

Using the notation of [15], finally the Kriging predictor is given by

$$y_p(x) = f(x)^{\mathrm{T}} \beta^* + r(x)^{\mathrm{T}} \gamma^* \tag{3.12}$$

where $\gamma^* = R^{-1}\left(Y - F\beta^*\right)$ and β^* is as solved for in the Appendix.

3.3 Characteristics of the Kriging Predictor

It is interesting to observe that β^* and γ^* are independent of the untried point *x*. Hence as seen in Eq. (3.12) evaluation of $y_p(x)$ for every new x involves only the computation of

$f(x)^{\mathrm{T}}$ and $r(x)^{\mathrm{T}}$ which are much easier to compute than the original function. This is the huge benefit obtained from the entire exercise of creating a computer model. The mean square error associated with $y_p(x)$ is given by

$$MSE(x) = \sigma^{*2}\left(1 + u(x)^{\mathrm{T}}\left(F^{\mathrm{T}}R^{-1}F\right)^{-1}u(x) - r(x)^{\mathrm{T}}R^{-1}r(x)\right) \tag{3.13}$$

where $$u(x) = F^{\mathrm{T}}R^{-1}r(x) - f(x) \tag{3.14}$$

The mean square error is a measure of the uncertainty associated with the predicted value. A greater value of *MSE* at a point implies that the underlying process is inadequately represented by the samples in the region around that point. We shall investigate the reasons for this in Chapter 4. Fig. (3.1a) the shows function $y(x) = x\sin(x)$ and its Kriging predictor obtained by taking *S* as a set of 12 equi-spaced points in the domain *D*=[0,10π].

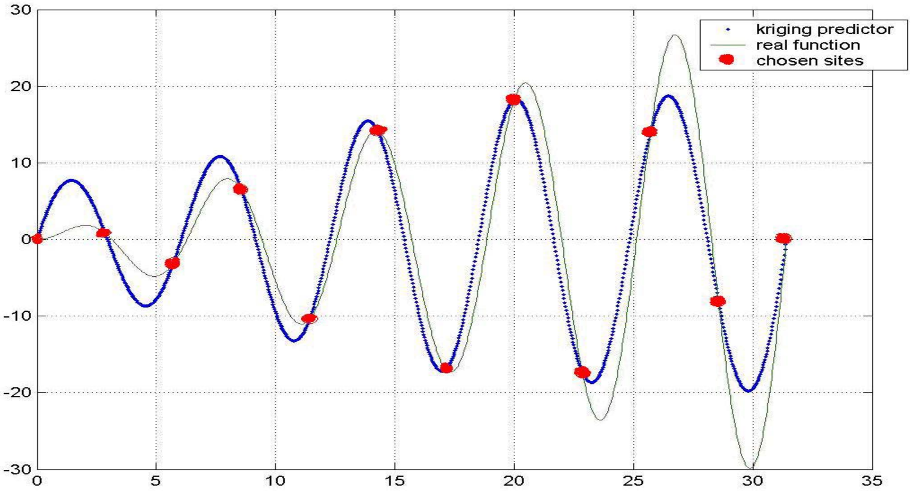


Fig. (3.1a)The function $y(x) = x\sin(x)$, Kriging predictor and sites in *S*

From Fig (3.1a) it can be seen that the Kriging predictor intersects $y(x) = x\sin(x)$ at precisely the12 points that belong to S. The corresponding *MSE* shown in Fig (3.1b) for these points is also exactly 0. This can also be seen by substituting $x = s_{\mathrm{i}}$ in Eq. (3.27) and Eq. (3.28). The universal Kriging predictor is hence a surrogate model that is exact for the specified sites. This is a powerful property of the universal Kriging predictor. As a result of this, in regions where the predictor shows a departure from the behaviour of the actual function it is possible to enforce the required behaviour by specifying the point through which the Kriging predictor must pass. This kind of fitting is not possible in the conventional regression based methods, since in those methods the least squares approximation does not necessarily pass through the specified sites and the introduction of new sites does not guarantee the expected change in shape of the predictor.

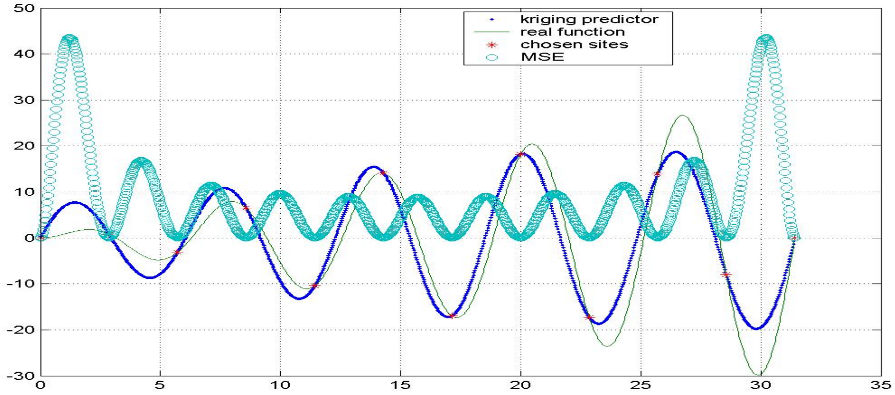


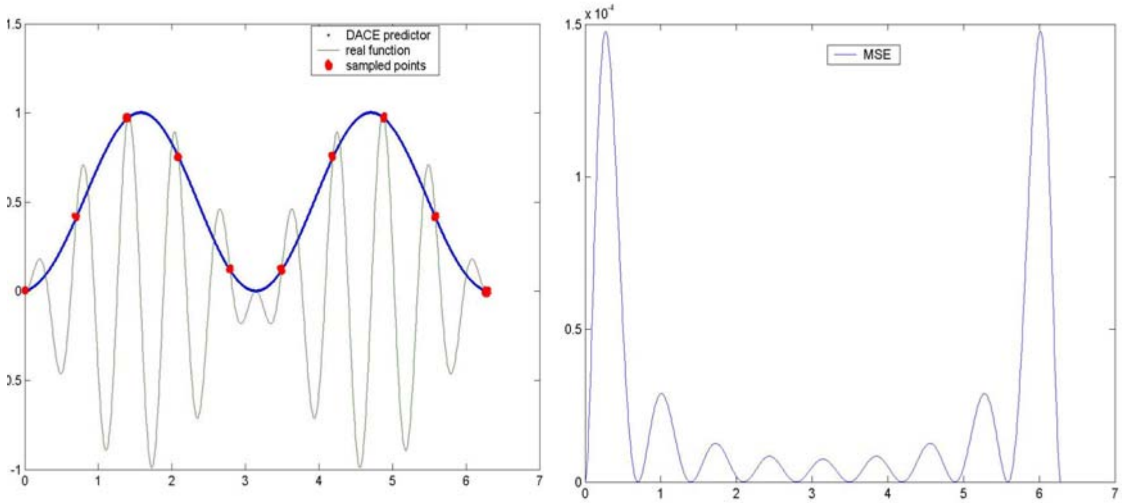
Fig (3.1b) The *MSE* for $y(x) = x\sin(x)$ and its Kriging predictor.

Kriging always provides a better approximation than traditional regression models for the same set of samples for a deterministic function [11]. Reference [11] provides an insightful discussion on this. Regression assumes that errors between the regression model and the process at any two points are independent and can be treated as random departures from the regression model. Hence it does not model them. But for deterministic models any such lack of fit is due improper modeling and not noise. Given two points $x \& x'$ that are very close the errors of the regression model $\in(x) \& \in(x')$ are also close and are hence correlated [11].

Fig (3.1b) also shows that the *MSE* associated with the Kriging predictor is highly multimodal. Since *MSE*(x) is 0 for $x \in S$, the *MSE* function has at least as many valleys as the number of points in *S*. Hence even in higher dimensions, the *MSE* function is multimodal.

3.4 Understanding the *MSE*

$MSE(x) = E\left(|\, y_p(x) - y(x)|^2\right)$. Hence *MSE* is regarded to be the confidence or the uncertainty that one can place in the values predicted by $y_p(x)$. For a black box function $y(x)$ whose analytical form is not known, the *MSE* is calculated by using the probabilistic linear model for $y(x)$ in Eq. (3.7). The parameters of this model are calculated on the basis of the existing samples. Hence the *MSE* found is not the "real error" between the predictor and the actual underlying process. It is simply the error between the predictor and the model in Eq. (3.7). How closely *MSE* resembles the real error depends on how well we have sampled the existing data. A low value of *MSE* does not necessarily imply that the real error is small. This peculiarity of the *MSE* can be seen in Fig. (3.2)



(a) **(b)**

Fig. (3.2) Problem of aliasing and insufficient sampling

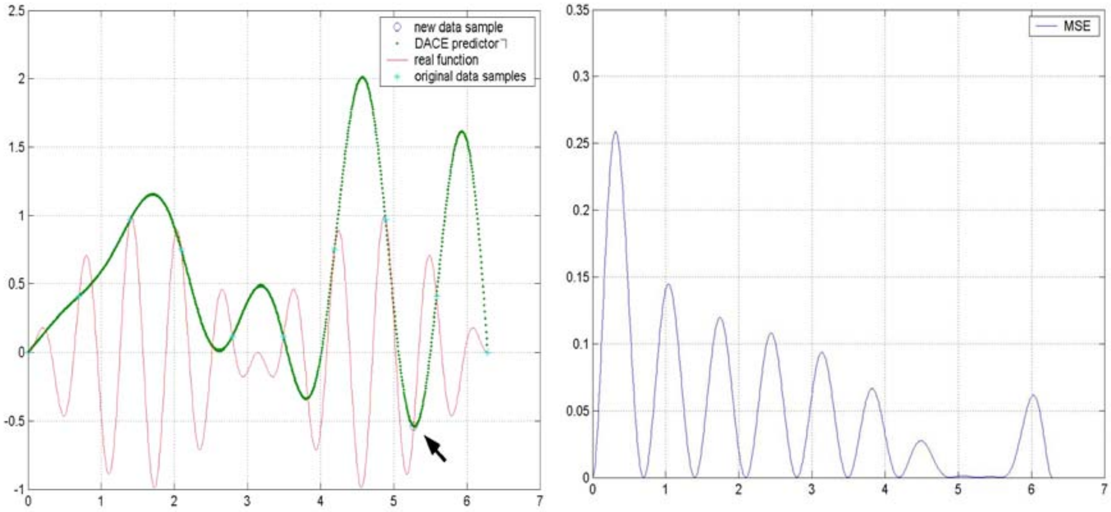
(a) The function, the DACE predictor and sampled data.

(b) MSE

In Fig. (3.2a) the real function is $\sin(10x)\sin(x)$ which has been sampled at 10 uniformly spaced points in $[0, 2\pi]$. We see that the *MSE* is ~10^{-4} times the maximum value of the real function. This however does not imply that the predictor is a good approximation. The real function consists of a fast varying component whose variations we have not been able to capture adequately through our current sampling. In order that our *MSE* provides a correct measure of the real error, it is necessary that we sample finely enough to capture all variations in the real function.

In signal processing parlance this phenomenon of higher frequency components appearing above lower frequency components is called *aliasing.* Nyquist criterion [18] states that the sampling frequency should be at least twice the highest frequency present in the signal in order to be able to capture it correctly. Reference [16] provides some discussion on aliasing in surrogate models.

Let us now add a sample (indicated by the small arrow) close to the 8th sample in Fig (3.3a). The effects are seen in Fig (3.3). The maximum value of *MSE* has increased to ~ 0.25 times the maximum value of the real function after the injection of the new sample. DACE has now encountered the new sample which was injected near another sample.



(a) **(b)**

Fig. (3.3) Effect of adding a new sample to an aliased function

(a) DACE predictor after adding new sample (b) MSE

As a result DACE could discover the sharp variation that exists in the neighbourhood of the sample. In order to capture this sharp variation the entire predictor has got reshaped. As a result the uncertainty in predictor is also higher. This reemphasizes the point that low values of *MSE* do not necessarily imply that the approximation is good. High values of *MSE* in a certain region mean that the real function is insufficiently sampled in it. That means that there are too few points in the region or that the points are not close enough to capture large gradients in the actual function.

Chapter 4

Adaptive Sampling

4.1 The need for Adaptive Sampling

In Chapter 3 we saw the effect sampling has on the DACE predictor. Let us now return to the objective of this report, that of developing a sampling strategy for expensive black-box functions. Suppose we have a fixed budget of *m* evaluations that can be done on the black-box function. Given that DACE will be used to create an approximation out of those points, the question that we are trying to answer is the design problem stated in Section 2.1: where should we sample the function?

In the case of a black box computer model, where the function topology is not known, no proactive sampling strategy can be relied upon for a good approximation. Hence none of the conventional space filling strategies like random sampling, uniform sampling, Latin hypercube sampling can be used effectively with all the *m* points. The only information we have about the actual function is from the existing samples and the DACE predictor created out of them. Hence it is essential to use a reactive strategy which "learns" from the information provided by previous samples to get the new set of samples. Such a process is commonly referred to as *adaptive sampling*.

Adaptive sampling as a technique often appears in the area of clinical research [19]. For example *Bandits problems* choose points from a finite set of alternatives with unknown yield to maximize some overall yield [20]. Adaptive sampling is an area of active research [19], [21], [22], [23]. The adaptive sampling has also been used in the framework for efficient global optimization (EGO) [11] and for superEGO [24], [25] on Kriging models to find the global optimum of the actual function. A general framework for DACE with adaptive sampling is as follows:

1. Use some space filling sampling strategy to sample some points in the variable-space
2. Use DACE to fit a surrogate model
3. Find the new sample by satisfying some infill sampling criteria (ISC) [19]
4. Repeat step 2 & 3 with new set of samples
5. Stop when some termination criterion is reached (e.g. maximum number of samples is overshot)

4.2 Issues arising in Adaptive Sampling

Some issues emerge out of the inspection of this framework.

1. What should be the initial space filling sampling strategy?
2. What should be the ISC?
3. What is the measure of accuracy of a model?
4. What is the guarantee of convergence of the algorithm vis-à-vis the initial sampling strategy and the ISC?

Let us look into these one by one.

4.2.1 Initial Sampling Strategy

The decision of the initial sampling strategy can be decomposed into two choices

a) The type of sampling, i.e. the distribution of points in the design space
b) Number of points (out of *m*) to be sampled with respect to the maximum number of samples

Both (a) and (b) are heavily dependent on the user’s knowledge of the underlying process in the computer model. In the case where no knowledge is available about the underlying process, the user is best placed to use any of the standard space filling strategies. Some popular space filling sampling strategies are [14]

a) Random sampling- randomly chooses points over the entire domain.
b) Uniform sampling- fills the space with uniformly spaced points
c) Latin hypercube sampling – chooses points such that no two of them have the same coordinate, and fills the space thus.

4.2.2 Infill Sampling Criteria

A variety of different ISC have been used and documented in literature. We shall survey them here in brief. Reference [11] uses adaptive sampling to find the global optimum of the black box function in his technique called EGO. It uses a function called *expected improvement* which is a probabilistic measure of the chances of that the new global minimum is lower than the current minimum sampled function value $f_{\min}$.

$$E\left(I\left(x\right)\right)=\left(f_{\min}-y_{\mathrm{p}}\left(x\right)\right)\Phi\left(\frac{f_{\min}-y_{\mathrm{p}}\left(x\right)}{\sqrt{MSE\left(x\right)}}\right)+\sqrt{MSE\left(x\right)}\phi\left(\frac{f_{\min}-y_{\mathrm{p}}\left(x\right)}{\sqrt{MSE\left(x\right)}}\right) \qquad (4.1)$$

where Φ is the cumulative distribution function and ϕ is the probability distribution function. The ISC is the global maximum of $E\left(I\left(x\right)\right)$. It is useful to study the idea behind choice of this ISC. When looking for the next sample there is always a dilemma about choosing between:

(a) searching in the neighborhood of $f_{\min}$ for a better minimum by sampling in the vicinity of $f_{\min}$ and

(b) searching in region that is sparsely sampled to check for possibilities of a better global minimum in that region [11].

Disregarding either (a) or (b) would be using the known information insufficiently. It is necessary to balance both (a) and (b) which is what $E\left(I\left(x\right)\right)$ does. Hence the expected improvement is large when $y_{\mathrm{p}}\left(x\right)$ is likely to be less than $f_{\min}$ or when the uncertainty itself is large. In the same vein as EGO, superEGO uses a *generalized expected improvement* [25]. By introducing a non-negative integer parameter *g* the generalized improvement is defined as

$$I_g\left(x\right) = \max\left\{0, \left(f_{\min} - y\left(x\right)\right)^{\mathrm{g}}\right\} \tag{4.2}$$

The value of *g* determines which of the trends (a) and (b) is given more preference.

Reference [26] specifies 3 criteria for infill in a problem for finding extremes. Each criterion solves a different problem viz. (i) to locate threshold-bounded extreme (ii) locate regional extreme or (iii) minimize surprises [25].

Note that though EGO and superEGO have been used to find optima of the black box function, their ISC are completely flexible and can be fine tuned to meet the objective we set for ourselves. Reference [19] is one such recent (2002) work in this direction. Reference [19] has used the adaptive sampling in the EGO framework with Kriging approximation to improve the design of ergonomics experiments. The ISC used by Ref [19] is to maximize the *MSE* of the DACE model. However this ISC is applied selectively only in certain regions of the design space that are most relevant to their experiment.

Our objective is similar to that of Ref. [19] and hence we shall be using maximizing *MSE* as our ISC. This is to say the new sample x_{new} is given by

$$x_{new} = \arg\max_{x \in D}\left(MSE\left(x\right)\right) \tag{4.3}$$

It is crucial that the ISC is well catered to the kind of problem we are trying to solve. Existence of a better ISC is always matter for further research. In Chapter 5 we shall also develop a new ISC that is suitable to our problem.

4.2.3 Convergence

According to Ref. [25] there are no rigorous convergence criteria for the adaptive sampling technique used by EGO and superEGO. Reference [3] suggests that the sampling may be stopped once the improvement of current best sample becomes sufficiently small. There is however no discussion of the guarantee of convergence of adaptive sampling strategies in literature. In our case where the function is a black box expensive function, the most prudent termination criterion is the maximum number of samples. In lieu of the problem of aliasing discussed in Section 4.1, it is necessary that while using adaptive sampling we first check for aliasing. Aliasing can be most easily be checked by sampling two points very close by and seeing its effect on the predictor.

4.2.4 Solution of the Infill Sampling Criteria

In all literature on adaptive sampling solving the ISC is a global optimization problem. We mentioned in Chapter 3 that the *MSE* of Kriging is highly multimodal and possibly highly dimensional. Hence solving Eq. (4.3) necessitates the use of a global optimization strategy that is amenable to multimodal functions. We shall assume in this report that the global optimum is obtainable, and in all our simulation results it has been found by griding the domain finely. Since the simulations are performed on the standard test case problems in optimization literature, griding is inexpensive.

Chapter 5

Infill Sampling Criterion

5.1 Frame work for testing results

In this Chapter we shall be testing the adaptive sampling algorithm with 2 different infill sampling criteria. Before we can get into the results of these tests, let us first establish a framework for testing and judging them.

5.1.1 Measure of accuracy

We saw in Chapter 3, that MSE cannot be taken to be a measure of accuracy for the DACE predictor. Hence we measure accuracy using the following 2 parameters.

1. Normalized absolute deviation (NAD): defined as

$$NAD = \max\left(|\, y(x) - y_{\mathrm{p}}(x)\,|\right) \big/ range \tag{5.1}$$

where, $$range = \left(\max\left(y(x)\right) - \min\left(y(x)\right)\right) \big/ 2 \tag{5.2}$$

2. Integrated error (IE): defined as

$$IE = \int_{domain} |\, y(x) - y_{\mathrm{p}}(x)\,|\, dx \tag{5.3}$$

NAD measures the maximum deviation that the predictor shows from the real function. It is normalized with respect to *range*, which is the half of the maximum variation in the values that the function shows. As a result of this normalization, NAD can be used to compare accuracies across different functions $y(x)$. IE is the integrated absolute difference between the predictor and the function. While NAD is a local measure of how well the predictor resembles the real function, IE is a global measure of how close the function and predictor are. If the predictor has its highest mismatch with the real function in only a small region and shows a good match everywhere else, NAD will be as high as the mismatch. But, depending on the size of this region and the extent of this mismatch, the value of IE will be either large or small. No or little change in NAD after the injection of a sample would imply that the new samples have not been put in the region of maximum mismatch. However even when NAD is constant, if IE shows decrease it means a significant match between the predictor and the real function has occurred in some other region of the domain. If a sample results in almost no decrease in NAD and a small decrease in IE, it is redundant to the approximation. The objective of the entire exercise of adaptive sampling is to bring $NAD \le 0.01$.

5.1.2 Method of testing

Approximation algorithms are usually developed for specific classes of functions only. In our case, since we are not targeting any specific type of function, we will have to carry out a more generic study. We shall be testing our algorithm for its effectiveness through a method of experimental exploration on various test functions. There are no standardized set of test functions available in literature for testing approximation algorithms. The test functions we have used are the ones used for testing optimization algorithms. The complete set is in shown in Appendix 3. The method of comparing results is the following:

We first find the number of uniformly distributed samples needed to have $NAD \le 0.01$. This number is compared with the number of samples we need for the same accuracy while sampling them using adaptive sampling. This comparison is carried out on every test function. When carrying out adaptive sampling the following choices are varied:

1. ISC
2. Number of initial samples

The ISC is a rule which specifies where the next sample is to be injected. As we shall see, the ISC is what characterizes the adaptive sampling algorithm. In the following sections we shall discuss and compare the use of 2 different ISCs.

5.2 Global Maximum of MSE as the ISC

The MSE of DACE as is a measure of the uncertainty in the value predicted by the predictor. We had seen in Section 3.5 that the MSE is large in regions that are insufficiently sampled or in those that have large gradients. As seen in Section 4.2, by taking inspiration from literature like ref. [19], the global maximum of the MSE of DACE is a plausible infill sampling criterion. In this section we shall present some results pertaining to the same.

1. **Peaks function:**

$$peaks(x_1,x_2) = 3(1-x_1)^2 \exp\left(-x^2-(y+1)^2\right) - 10\left(\frac{x}{5}-x^3-y^5\right)\exp\left(-x^2-y^2\right) - \frac{1}{3}\exp\left(-(x+1)^2-y^2\right)$$

$$-3 \le x_1 \le 3, -3 \le x_2 \le 3$$

It is found that uniform proactive sampling of 196 (14x14) samples results in $NAD \le 0.01$.

We carried out simulations on the peaks function with varying number of initial samples. The observations from these tests are included below. The complete results are in Table 5.1

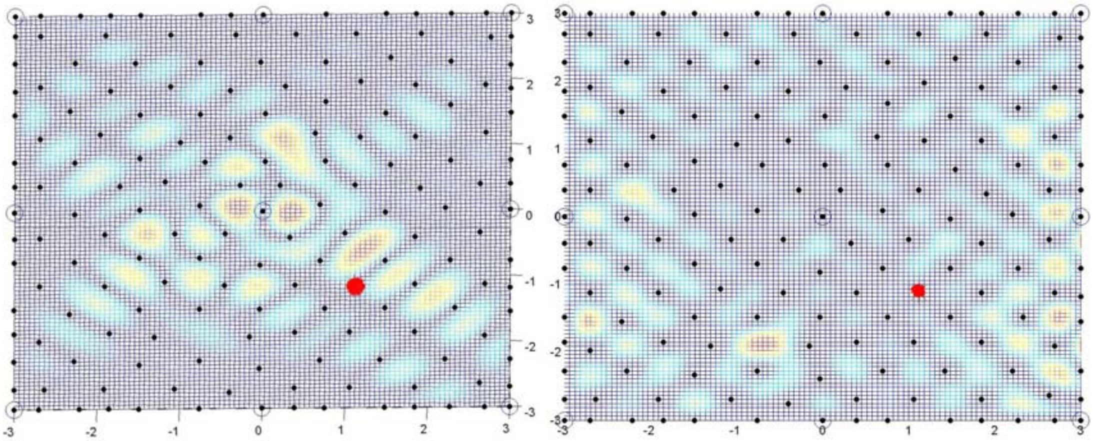
Observations:

- The first observation we see is that the total number of samples required with the adaptive sampling strategy is indeed lesser than the number required in a proactive sampling. Hence adaptive sampling indeed works better than any proactive sampling.
- With adaptive sampling, it is also found that the number of points required for $NAD \leq 0.01$ changes with the number of samples put in initially. The number of samples required also doesn't vary uniformly with the number initial samples, but shows a rather erratic behavior. The maximum effectiveness of this strategy (the minimum number of samples needed) is obtained when we sample about 121 points initially and get the rest from adaptive sampling. The maximum percentage reduction (from the number of points required in a proactive sampling) is about 18%.

Number of initial samples	Total number of samples required	Percentage reduction in number of samples
9	167	14.80
16	167	14.80
25	165	15.82
36	182	7.14
49	177	9.69
64	164	16.33
81	175	10.71
100	176	10.2
121	161	17.86
144	164	16.33
169	187	4.59

Table 5.1 Adaptive sampling on the Peaks function using MSE as ISC

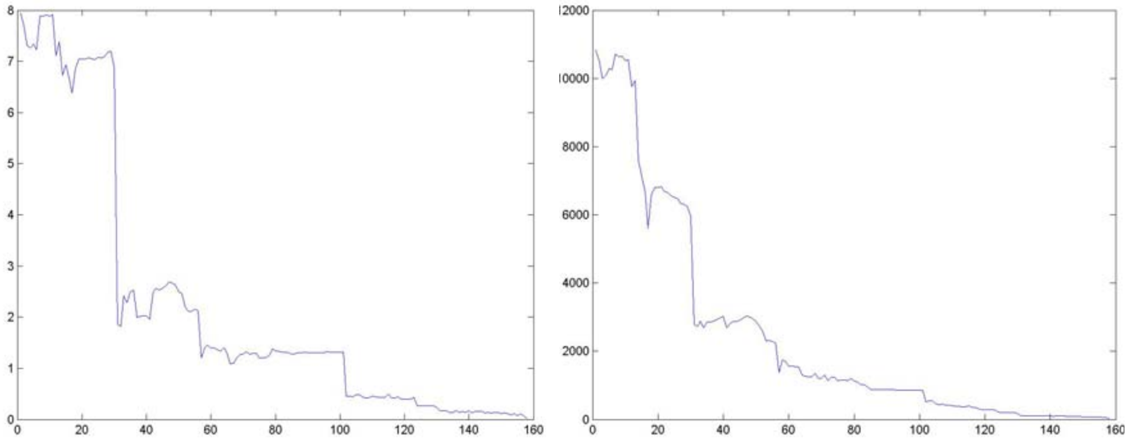
- It is also worthwhile seeing where the sampling has occurred for some of these cases. Figure 5.1a shows the function $|\, y(x) - y_{\mathrm{p}}(x) |$with the samples and Fig 5.1b shows the MSE with the samples. The samples are denoted by the black dots. The bright red dot is the last sample. The circled black dots are the initial samples. Regions that are red have higher value while blue ones have lower value of the function being plotted. Several observations can be made from these 2 figures. It is firstly seen that though the initial samples were far apart, the injected samples have filled the space almost uniformly. There are several samples injected on the boundary of the domain. The reason for this is that for wavy functions like the peaks function, the MSE is usually higher on the boundary. It is also seen that even after sampling has been stopped, MSE and $|y(x)-y_{\mathrm{p}}(x)|$ are dissimilar. The maxima of MSE do not coincide with the regions where $|y(x)-y_{\mathrm{p}}(x)|$is maximum. This is true not just at this stage of sampling, but at every stage of adaptive sampling.



(a) $|y(x)-y_{\mathrm{p}}(x)|$**with samples** **(b)MSE with samples**

Fig 5.1 Sampling on peaks function with 3x3 uniform initial samples

- Having seen that the maxima of $|y(x)-y_{\mathrm{p}}(x)|$and MSE do not coincide, it is important that we now look at the behavior of NAD and IE to see if all the samples that have been injected have indeed been useful in improving the approximation of the predictor. Figure 5.2 shows the plots of NAD and IE for 3x3 = 9 uniformly distributed initial samples.



(a) NAD with number of injected samples (b) IE with number of injected samples

Fig 5.2 Progress of adaptive sampling on peaks function with 3x3 initial sampling

- The plots show a step-like behavior for NAD. That is to say that during adaptive sampling, there are phases when NAD remains almost unchanged. That again reiterates the fact that the successive samples are not getting injected in the region where there is maximum mismatch between the function and the predictor. The corresponding IE shows small reduction in value. Which means that though in regions where the sample is being injected the approximation has improved, the injected sample is redundant and has not helped to improve the fit significantly. Such behavior is observed for the peaks function even with more initial samples.
- The reasons for this step-like behavior are as follows. When we sample the domain with, say 25 uniformly distributed points, we find several local maxima of the MSE function distributed over the domain. Several of these maxima are of similar height. When a sample is injected at one of those maxima, the MSE in its vicinity reduces, whereas the MSE in the rest of the domain remains almost unchanged. In successive iterations, the algorithm samples points one by one at these maxima. However not all of these points are useful in getting a good approximation, and several of them are redundant. But while sampling, when the algorithm does sample in a region where the mismatch is maximum, the NAD suddenly falls. This problem occurs because, as discussed in Section 3.5 and above, the MSE is not the real error between the function and the predictor. This problem is fundamental to DACE and is a significant weakness in using global maximum of MSE as the ISC. It hence warrants the development of a new and more effective ISC. The next section elaborates on the new ISC we have developed.

2. **Goldstein Price, Branin's rcos, Rosenbrock's valley function**

We also carried out similar tests on the Goldstein Price function, Branin's rcos function and the Rosenbrock's valley function. Please refer to Appendix 3 for the function definition. The waviest topology is of the Peaks function. The others mentioned above are moderately wavy. Functions have been simulated with various numbers of uniformly distributed initial samples Percentages are calculated w.r.t. the number of such samples required for $NAD \leq 0.01$.

Observations:

- The results obtained from these simulations show that for wavy functions, adaptive sampling is less effective. The extent of success of adaptive sampling though is different for different functions. As functions get flatter DACE is able to capture their variations with fewer samples and adaptive sampling proves effective in locating those samples. Figure 5.3 shows the percentage reduction in number of samples required for $NAD \leq 0.01$ plotted against the percentage of initial samples.

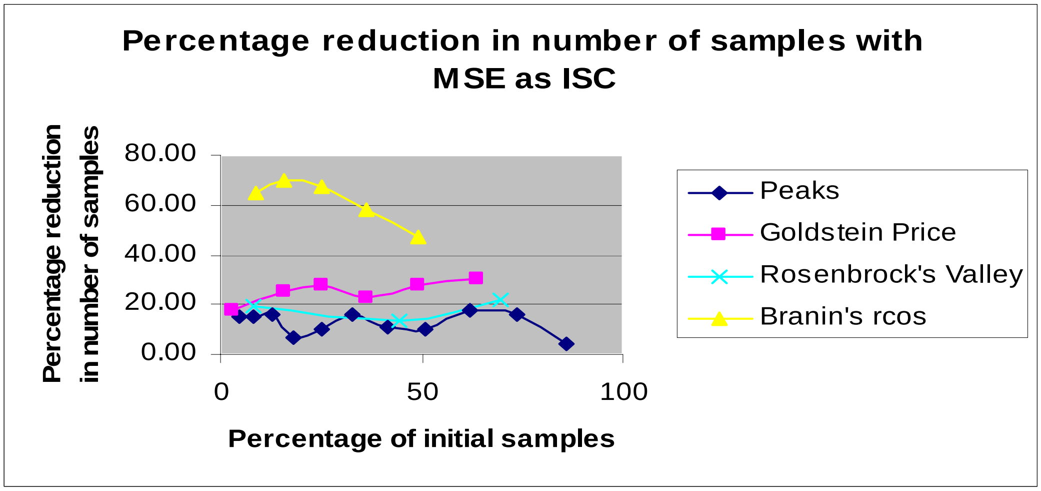
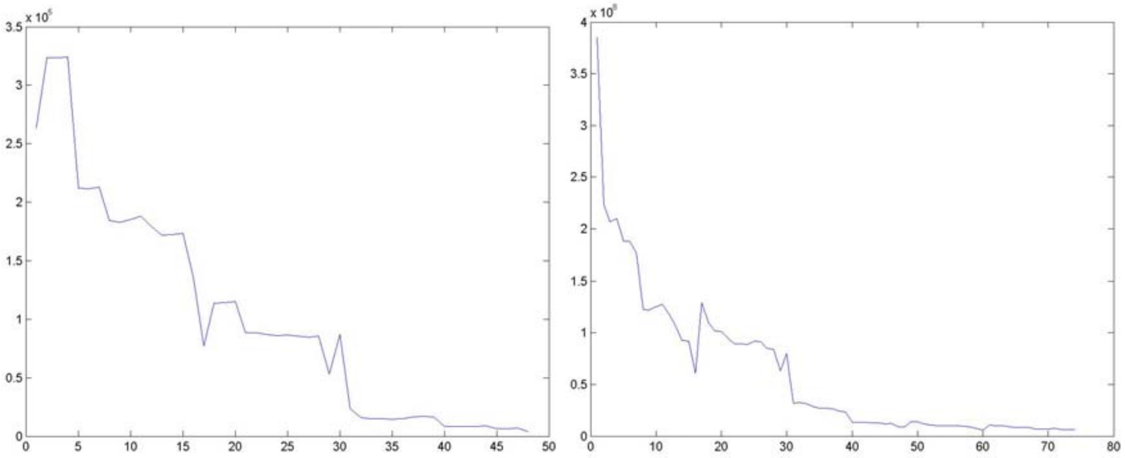


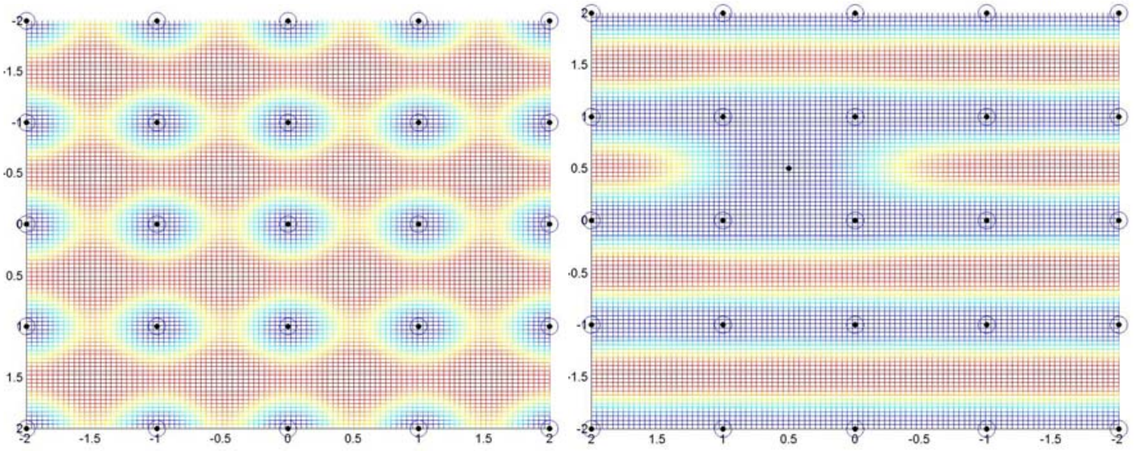
Fig 5.3 Percentage reduction in number of samples with MSE as ISC

- From these results it appears that for smoother functions about 30-40% of the total number of available samples should be used for initial sampling. The rest should be injected adaptively. For wavy functions, about 40-60% of the samples should be used for initial sampling, and the rest should be injected adaptively.
- Figure 5.4 shows the plot of NAD and IE with number of samples injected for the Goldstein price function. These plots were obtained for the Goldstein Price function with the use of adaptive sampling with 25 initial samples. Once again we see the step-like behavior in NAD. This kind of behavior often occurs with smoother functions when the number of initial samples is large.



(a) NAD with number of samples injected **(b) IE with number of samples injected**

Fig. 5.4 Effect of Adaptive sampling on the Goldstein Price function



(a) With 25 initial samples **(b) With 25 samples + 1 injected sample**

Fig 5.5 MSE for Goldstein Price function

- In order to understand this behavior we must have a look at the MSE that is formed with 25 uniformly distributed samples. Figure 5.5a shows the MSE after sampling 25 uniformly distributed points. It is seen that for a flat function like the Goldstein Price function, the MSE is also uniform. In every quadrilateral marked by 4 initial samples, there is a local maximum of the MSE. Figure 5.5b shows the MSE after injection of one more point. We see that after the injection of this point the MSE in the rectangular region surrounding that point subsides, while the MSE peaks in other regions remain. All these peaks are of similar height. The algorithm samples points one by one at successive peaks in the MSE. However, since the function is largely flat, not all of these points result in a reduction of the NAD. Not all of them result in significant reduction of IE too. Hence we get the step like behavior in NAD. More the number of initial samples, more the number of such peaks, and more and longer the "steps" in NAD.

As discussed with the observations about the peaks function, the step-like NAD is attributable to the fact the maxima of MSE do not coincide with the maxima of$|y(x)-y_\mathrm{p}(x)|$. This seen in Fig 5.6 which shows the plot of $|y(x)-y_\mathrm{p}(x)|$ for the Goldstein Price function with 25 uniformly

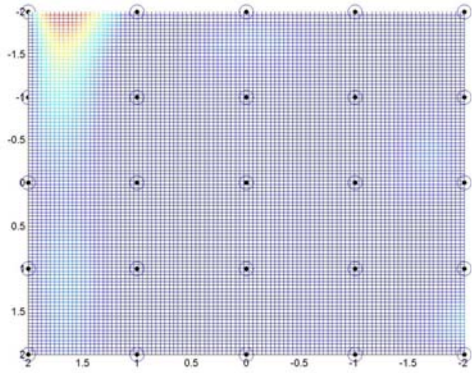


Fig 5.6 $|y(x)-y_\mathrm{p}(x)|$ **for Goldstein Price**

distributed samples. Comparing this with Fig 5.5a shows that there are gross differences between the two. We have hence developed a new ISC, which will try to reduce the number of redundant points sampled by the algorithm.

5.3 Dual criteria adaptive sampling

Before we present the new ISC, let us look at what information we have about the function as we try to approximate it. The information about the function is available from the initially sampled points, from the successively injected samples, successive DACE models, and the MSE. The algorithm described in Section 5.2 is weak since it uses only MSE and thus fails to use all the information available about the function. Our effort in this new "Dual Criteria ISC (DCISC)" is to use information from the previous DACE models to try and guess in which regions is the real error $|\, y(x) - y_{\mathrm{p}}(x) \,|$ is high.

When we make a series of DACE models in the adaptive sampling algorithm, each time we have a predicted value for the point where sample is to be injected. After the sample is injected (i.e. the real function is evaluated at that point), we know the real value of the function for that point. If the discrepancy between the predicted value and the real value is large as compared to $\sqrt{MSE}$, it means that in the vicinity of that point there is a need for more samples. The DCISC is based on this idea. It is defined as follows.

1. Consider the previous *k* models, and the previous *k* injected points.
2. Find the discrepancy between the predicted value and the real value at these *k* points.
3. Consider the point that shows the maximum discrepancy amongst these *k* points. Let it be x^* and let the maximum discrepancy be *d*.
4. Define Y_{step} as an *n*-D step function such that it has value = *d* in a δ box around x^* and is 0 everywhere.
5. The DCISC is the global maximum of $\alpha Y_{step} + \sqrt{MSE}$. The new sample is given by

$$x_{new} = \arg\max\left(\alpha Y_{step} + \sqrt{MSE}\right) \tag{5.4}$$

where $\alpha \in (0,1)$ is called *relative influence parameter.* It determines how much importance is to be given to the discrepancy relative to the $\sqrt{MSE}$. *k* is an integer greater than 1 and is called the *discrepancy memory parameter*. It determines the discrepancy from how many of the previous models influence the current one. δ is called the *spatial influence parameter*. It determines in how much vicinity of x^* does the influence of the discrepancy hold.

1. Peaks function

We shall now present results obtained with this new ISC on the peaks function defined in Section 5.2. These results have been obtained by taking α =0.5, *k*= *4,* and δ =1.5. The budget is the same in Section 5.2 and a similar set of simulations have been carried out.

Observations

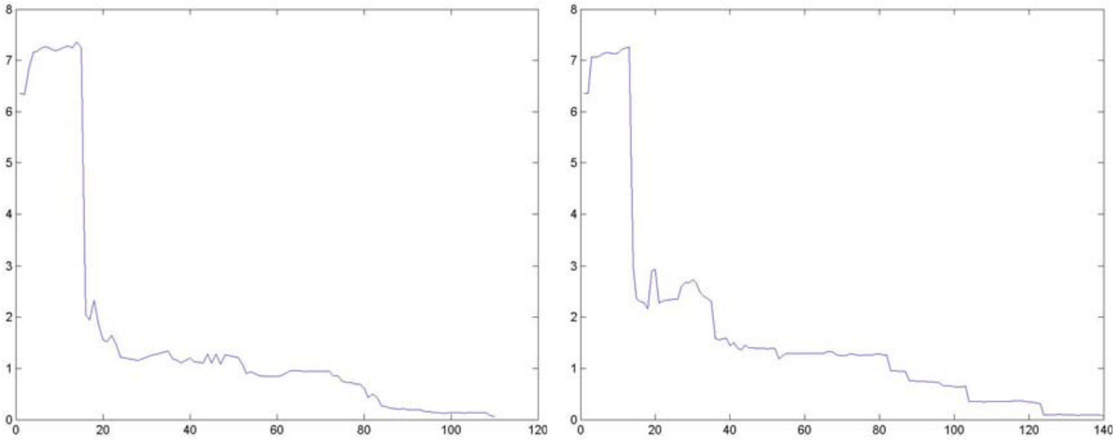
- The results presented in Table 5.2. As expected, the DCISC performs better than a proactive initial sampling. Just as with the MSE based sampling, here too the total number of samples needed varies with the number of points sampled initially.
- We see that this criterion performs better than the other criterion of using global maximum of MSE only. This is true for any number of initial samples. We see that the maximum reduction in the number of samples is obtained with 81 samples.

Number of initial samples	Percentage of maximum samples	Total number of samples required	Percentage reduction in number of samples (with dual criteria ISC)	Percentage reduction in number of samples (with only MSE)
9	5	164	16.33	14.80
16	8	150	23.47	14.80
25	13	135	31.12	15.82
36	18	172	12.24	7.14
49	25	160	18.37	9.69
64	33	143	27.04	16.33
81	41	133	32.14	10.71
100	51	173	11.73	10.2
121	62	162	17.35	17.86
144	73	164	16.33	16.33

Table 5.2 Dual criteria adaptive sampling on Peaks function

- When the number of initial samples is either too small or too large, the DCISC behaves similarly to the conventional criterion. However when the number of samples is 20-50% of the number of uniformly distributed samples required for $NAD \leq 0.01$, we see that there is a significant difference in performance between the DCISC and the conventional one. Again it is seen that the total number of samples needed does not vary uniformly with the number of initial samples.
- Let us compare the NAD for the peaks function with 81 initial samples with the two different criteria. Figure 5.7 shows the NAD plots for the 2 different approaches. We see that the NAD plots for the two different criteria are strikingly similar in some regions of the plot. Which means that in some phases of the sampling process DCISC

is degenerating to the conventional MSE based criterion, though the criterion overall performs better than the MSE based criterion. Such behavior is observed for any number of initial samples.



(a) With DCISC **(b) With only MSE**

Fig 5.7 NAD for the peaks function with 81 initial samples with different ISCs

2. **Goldstein Price, Branin's rcos, Rosenbrock's valley function**

Simulations carried out are just like in Section 5.2, with similar number of initial samples.

Observations

- We see that for the smoother variety of functions, like the Goldstein Price function, Branin's rcos function, Rosenbrock's valley functions the DCISC criterion performs significantly better than the conventional criterion. Table 5.3 shows the comparative results of the two different criteria on these functions.

- The improvement in performance is observed for any number of initial samples. In general we see that when the number of initial samples is around 10-35% of the number of uniformly sampled points needed to get $NAD \leq 0.01$, we get the greatest percentage reduction in the number of samples.

- As opposed to the case for wavy functions, this sampling criterion does not degenerate to the conventional criterion based on only MSE. But when number of initial samples is large the DCISC performs similar to the conventional criterion.

- As seen for the peaks function, here too the NAD plots resemble each other in some regions, while in other regions they are different - which again shows that there are phases when the DCISC degenerates to the conventional MSE based criterion.

Number of initial samples	Percentage of maximum samples	Total number of samples required	Percentage reduction in number of samples (with dual criteria ISC)	Percentage reduction in number of samples (with only MSE)
9	25	21	41.67	19.44
16	44	22	38.89	13.89
25	69	28	22.22	22.22

(a) Rosenbrock's valley function

Number of initial samples	Percentage of maximum samples	Total number of samples required	Percentage reduction in number of samples (with dual criteria ISC)	Percentage reduction in number of samples (with only MSE)
9	9	30	70.00	65
16	16	31	69.00	70
25	25	34	66.00	67
36	36	42	58.00	58
49	49	53	47.00	47

(b) Branin's rcos function

Number of initial samples	Percentage of maximum samples	Total number of samples required	Percentage reduction in number of samples (with dual criteria ISC)	Percentage reduction in number of samples (with only MSE)
9	9	69	31	18
16	16	75	25	25
25	25	71	29	28
36	36	69	31	23
49	49	76	24	28
64	64	70	30	30

(c) Goldstein Price function

Table 5.3 Comparative performance of DCISC and MSE based criterion

- Figure 5.8 also shows the percentage reduction in the number of samples for different number of initial samples, for various functions using DCISC.

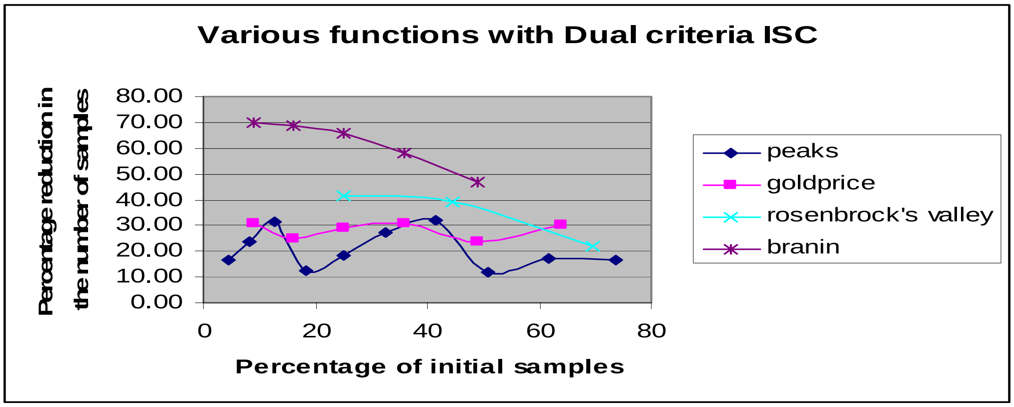


Fig 5.8 Percentage reduction in number of samples with DCISC

5.3.1 Some comments about Dual Criteria ISC

The performance of the DCISC based adaptive sampling algorithm is sensitive to the values of relative influence parameter, discrepancy memory parameter, and spatial influence parameter (α, k, δ). The values of α, k, δ we have used in our results have been arrived at by a process of trial and error, but their choice is critical to the success of this algorithm. Let us throw some light on the role of these parameters.

Relative influence parameter, α

In the DCISC, since Y_{step} is added to $\sqrt{MSE}$, α should be taken such that if *d* (as defined in Section 5.3) is comparable to $\sqrt{MSE}$, that region should get preference. It was found that α =0.5 yielded good results. If α is very small, the DCISC criterion reduces to using only MSE. If α is very large MSE loses its significance completely.

Discrepancy memory parameter, *k*

Often the largest discrepancies are found earliest stages of sampling. If *k* chosen to be very large, then we find that only one of the discrepancies dominate and a number of samples get accumulated in one region. If *k* is taken to be too small then DCISC degenerates to using only MSE. After some trial and error, a value of *k = 4* was found to appropriate.

Spatial influence parameter, δ

It was found that the number of initial samples and δ are closely related. When the domain is sampled uniformly, it is divided into rectangular subdomains of equal area. The results seem to be best when the area of each of these subdomains have is nearly equal to the area of the δ box. Again if δ is too small or too large, the algorithm degenerates to using only MSE.

5.4 Further research and exploration

From our results we find that the DCISC developed seems to have a lot of potential, particularly in the case of flat and smooth functions. It is able to capture several of the regions where samples are most needed and is able to also ignore the regions where samples are not required. It must be noted that in the worst case, DCISC performs as well as using only MSE as the ISC. The exact impact that the relative influence parameter, discrepancy memory parameter, and spatial influence parameter have on the algorithm is an interesting matter for further exploration.

Chapter 6

Conclusion

In this report we have surveyed the various existing methodologies for surrogate modeling and global optimization. We have dealt in depth with the theory associated with universal Kriging and DACE and chosen it as our surrogate model. We have developed and demonstrated the use of 2 methods to improve the approximation provided by the DACE predictor. The methods use a technique of adaptive sampling and successive improvements of the DACE model. Positioning the next sample requires us to solve a global optimization problem to satisfy some infill sampling criterion. Two sampling criteria were developed and compared. A new and original criterion, called the "dual criteria infill sampling criterion" was developed and was seen to produce better results than using only MSE.

The implementation of this "dual criteria infill sampling criterion" involves specifying values of 3 parameters - relative influence parameter, discrepancy memory parameter, and spatial influence parameter. Further research should be directed towards understanding the exact influence of these.

References

[1] Montgomery, D., *Design and Analysis of Experiments,* 3rd edition. New York, Wiley, 1991. pp 1-6.

[2] Sondergraad, J., *Optimization Using Surrogate Models – by the Space Mapping Technique,* Ph. D. Thesis, Technical University of Denmark, Kgs. Lyngby, Denmark, 2003.

[3] Schonlau, M., *Computer Experiments and Global Optimization*, Ph. D. Thesis, University of Waterloo, Waterloo, Canada, 1997.

[4] Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P., "Design and Analysis of Computer Experiments", *Statistical Science,* Vol 4, No 4, 1989, pp 409-423.

[5] Serafini, D.B., *A Framework for Managing Models in Nonlinear Optimization of Computationally Expensive Functions,* Ph. D. Thesis, Rice University, Houston, Texas, US, 1998.

[6] Nielsen H.B., *Surrogate models*, IMM Numerical Analysis Section, Technical University of Denmark, 2004.

[7] Krose, B., van der Smagt, P., *An Introduction to Neural Networks*, University of Amsterdam, 1996.

[8] Gutmann, H.M., "A Radial Basis Function Method for Global Optimization", *Journal of Global Optimizaton,* No 19, pp 201-227, 2001

[9] Myers, R., Montgomry D., Vining, G.G., *Generalized Linear Models*, John Wiley & Sons Inc., New York, US, 2002.

[10] Watson, G.S., "Smoothing and Interpolation by Kriging with Splines", *Mathematical Geology,* Vol 16, pp 231-257, 1995.

[11] Jones, D.R., Schonlau, M., Welch, W.J., "Efficient Global Optimization of Expensive Black- box Functions", *Journal of Global Optimization,* Vol 13, No 4, pp 455-492.

[12] Nielsen, H.B., Thuesen, K.F., "Kriging and Radial Basis Functions", IMM, Technical University of Denmark, Kongens Lyngby, Denmark.

[13] Lloyd, C.D., Atkinson, P.M., "Design Optimal Sampling Configurations with Ordinary and Indicator Kriging", *GeoComputation '99, http://www.geovista.psu.edu/sites/geocomp99/Gc99/065/gc_065.htm*

[14] Lophaven, S.N., Nielsen, H.B., Sondergraad, J., "Aspects of the DACE Matlab Toolbox", IMM, Technical University of Denmark, Kongens Lyngby, Denmark 2002, *www2.imm.dtu.dk/~hbn/dace/.*

[15] Lophaven, S.N., Nielsen, H.B., Sondergraad, J., "DACE, A Matlab Kriging Toolbox, Version 2.0", IMM, Technical University of Denmark, Kongens Lyngby, Denmark 2002, *www2.imm.dtu.dk/~hbn/dace/.*

[16] Simpson, T.W., Booker A.J., Ghosh, D., Giunta A.A., Koch, P.N., Yang, R., "Approximation Methods in Multidisciplinary Analysis and Optimization: A Panel Discussion", *Approximation Methods Panel, 9th AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Atlanta, GA*, 2002

[17] Moore A.W., "Learning with Maximum Likelihood", School of Computer Science, Carnegie Mellon University, *http://www.cs.cmu.edu/~awm*

[18] Levesque, L., "Revisiting the Nyquist Criterion and Aliasing in Data Analysis", *European Journal of Physics,* 22 (2001), pp 127-132.

[19] Sasena, M.J., Parkinson, M., Goovaerts, P., Papalamabros, P., Reed, M., "Adaptive Experimental Design Applied to an Ergonomics Testing Procedure", *in the Proceedings of DETC'02 ASME 2002 Design Engineering Technical Conferences and Computers and Information in Engineering Conference Montreal, Canada,* 2002.

[20] "Adaptive Sampling Designs",

http://www.eecs.umich.edu/~qstout/abs/Seattle97.html

[21] Atkinson, A., Pronzato, L. and Wynn, H.P. "MODA5: Advances in model-oriented data analysis and experimental design", in the *Proceedings of the 5th international workshop*, 1998.

[22] Atkinson, A., Hackl, P., and Muller, W., *Proceedings of MODA6: Model Oriented Data Analysis,* Physica Verlag, 2001.

[23] Flournoy, N., Rosenberger, W.F. and Wong, W.K., “New developments and applications in experimental design”, in the *Proceedings of the Joint AMS-IMS-SIAM Summer Research Conference*, 1997.

[24] Sasena, M.J., *Flexibility and Efficiency Enhancements for Constrained Global Design Optimization with Kriging Approximations*, Ph.D. Thesis, University of Michigan, Dept. of Mech. Engg., 2002

[25] Sasena, M.J., Papalambros, P.Y. and Goovaerts, P., “Exploration of Metamodeling Sampling Criteria for Constrained Global Optimization”, *Engineering Optimization*, 34(3):263-278

[26] Watson, A.G., and Barnes, R.J., “Infill sampling criteria to locate extremes” *Mathematical Geology*, 27(5), 1995, pp 589-608.

[27] Astolfi, A., *Optimization: An introduction,*

http://cap.ee.imperial.ac.uk/~astolfi/Courses/outs/Optim.pdf

[28] Panda, A., *Trust Region based Tunneling for Unconstrained Global Optimization,* M. Tech Thesis, Dept. of Aerospace Engg, IIT Bombay, India, 2005.

[29] Pinter, J., “Global Optimization”,

http://mathworld.wolfram.com/GlobalOptimization.html

[30] Levy A.V., Montalvo, A., “The Tunneling Algorithm for Global Minimization of Functions”, *SIAM J. of Sci. & Stat. Comput.*, Vol 6, No 15, 1985, pp 15-29.

Acknowledgements

I sincerely thank Prof Mujumdar for giving me a chance to work on such an interesting and enterprising topic. I especially thank him for having given me the freedom to take the topic wherever I wanted to. He has been extremely accommodating and understanding and has often gone out of his way to help out in my deficiencies. His kind and friendly mannerisms have made working with sheer pleasure and a memorable experience. I feel that I have grown up as an engineer during the course of this BTP.

I would like to specially thank Mr. Amitay Issacs for helping me throughout this semester with my understanding of the subject.

Ankur Kulkarni

Appendix 1

Maximum Likelihood Estimation

We shall first explain what the maximum likelihood estimation problem is. The maximum likelihood estimation (MLE) that we shall consider is a parametric form of density estimation problem. Suppose $X_1, X_2, ..., X_{\mathrm{m}}$ are independent and identically distributed (iid) with a common probability density function ψ. Suppose ψ is parameterized with respect to θ. The maximum likelihood estimation problem is to find θ such that the conditional probability of $X_1, X_2, ..., X_{\mathrm{m}}$ given θ, i.e. $P(X_1, X_2, ..., X_{\mathrm{m}} \mid \theta)$ is maximum [17]. Using that X_{i} 's are iid,

$$\theta_{MLE} = \arg\max(P(X_1, X_2, ..., X_{\mathrm{m}} \mid \theta))$$

Since X_{i} 's are iid,

$$\theta_{MLE} = \arg\max\left(\prod_1^{\mathrm{m}} P(X_{\mathrm{i}} \mid \theta)\right)$$

$$\theta_{MLE} = \arg\max\left(\prod_1^{\mathrm{m}} \psi(X_{\mathrm{i}}, \theta)\right)$$

In our case, we frame an MLE problem to find all of θ, σ, β such that the joint probability distribution given by $\prod_1^{\mathrm{m}} \psi(z(s_{\mathrm{i}}), \theta, \sigma, \beta)$ is maximized, where $z(s_{\mathrm{i}}) = y(s_{\mathrm{i}}) - f\left(s_{\mathrm{i}}\right)^{\mathrm{T}} \beta$ is the error at site s_{i} appearing out of a stochastic process. The Kriging model is a combination of multivariate normal model and a linear model. If the stochastic process is taken as Gaussian then the probability density function is given by [3]

$$\prod_1^{\mathrm{m}} \psi\left(z\left(s_{\mathrm{i}}\right), \theta, \sigma, \beta\right) = \frac{1}{\left(\left(2\pi\right)\sigma^2 \det(R)\right)^{\mathrm{m}/2}} \exp\left(\frac{-\left(Y - F\beta\right)^{\mathrm{T}} R^{-1}\left(Y - F\beta\right)}{2\sigma^2}\right)$$

where R, the correlation matrix is a function of θ, $Y = \left[y(s_1), y(s_2), ..., y(s_{\mathrm{m}})\right]^{\mathrm{T}}$ is the vector of outputs at the chosen sites, and $F = \left[f\left(s_1\right), f\left(s_2\right), ..., f\left(s_{\mathrm{m}}\right)\right]^{\mathrm{T}}$ is an $m \times p$ matrix holding the regression functions evaluated at the chosen sites and β is as defined in Eq. (3.9). Hence the MLE problem for our case is,

$$\left(\theta, \sigma, \beta\right)_{MLE} = \arg\max\left(\frac{1}{\left(\left(2\pi\right)\sigma^2 \det(R)\right)^{\mathrm{m}/2}} \exp\left(\frac{-\left(Y - F\beta\right)^{\mathrm{T}} R^{-1}\left(Y - F\beta\right)}{2\sigma^2}\right)\right)$$

Taking natural logarithms we get the *log likelihood problem*

$$(\theta,\sigma,\beta)_{MLE} = \arg\max\left(\frac{-1}{2}\left(m\ln\sigma^2 + \ln\det(R) + \frac{(Y-F\beta)^{\mathrm{T}} R^{-1}(Y-F\beta)}{\sigma^2}\right)\right)$$

To solve this we must differentiate the log likelihood with respect to $\theta,\sigma,\text{and}\beta$ and put the respective partial derivatives equal to 0. σ and R are independent of β, so the MLE of β, denoted by β^* is obtained as

$$\frac{\partial}{\partial\beta}\left((Y-F\beta)^{\mathrm{T}} R^{-1}(Y-F\beta)\right) = 0$$

$$\beta^* = \left(F^{\mathrm{T}}R^{-1}F\right)^{-1} F^{\mathrm{T}}R^{-1}Y$$

This MLE β^* is the same as the generalized least squares estimate of the regression problem. Similarly, the MLE of σ^2, denoted by σ^{*2} is obtained as

$$\frac{\partial}{\partial\sigma^2}\left(m\ln\sigma^2 + \frac{(Y-F\beta^*)^{\mathrm{T}} R^{-1}(Y-F\beta^*)}{\sigma^2}\right) = 0$$

$$\sigma^{*2} = \frac{1}{m}\left((Y-F\beta^*)^{\mathrm{T}} R^{-1}(Y-F\beta^*)\right)$$

We see that parameters β and σ^2 are decoupled. In fact it is clear that both σ^{*2} and β^* are both essentially functions of θ. Thus the MLE problem posed above is in fact a problem of finding the MLE of only θ. Although the least squares solution for β and the MLE β^* are identical, this fact could not have been established by finding least squares solution first and then imposing MLE separately on θ.

On solving the MLE for θ, σ^{*2} and β^* get fixed as a consequence. This means that by choosing n critical values in the vector θ, the model and as we shall see in the next section, the predictor can be determined. This remarkable simplicity is attributed to the choice of the stochastic process as Gaussian. MLE for θ is reposed as

$$\theta_{MLE} = \arg\max\left(\frac{-1}{2}\left(m\ln\sigma^{*2} + \ln\det(R)\right)\right)$$

This is an optimization problem that has to be solved numerically. References [14] and [15] provide algorithms for this. Having solved for θ and having found σ^{*2} and β^* we are in a position now to create a predictor for the function $y(x)$ to predict its value over the domain D.

Appendix 2

Deriving the BLUP using the method of Lagrange Multipliers

The BLUP (best linear unbiased predictor) requires that we solve:

1. Minimize $MSE = E\left(| y_p\left(x\right) - y\left(x\right)|^2\right)$ with respect to $c\left(x\right)$
2. subject to $E\left(y_p\left(x\right)\right) = E\left(y\left(x\right)\right)$

The mean square error at untried x is

$$\therefore MSE\left(x\right) = E\left\{| c\left(x\right)^{\mathrm{T}}\left(F\beta^* + Z\right) - f\left(x\right)^{\mathrm{T}}\beta^* - z\left(x\right)|^2\right\}$$

where $$Z = \left[z\left(s_1\right), z\left(s_2\right), ..., z\left(s_{\mathrm{m}}\right)\right]^{\mathrm{T}}$$

Since $E\left(z\right) = E\left(Z\right) = 0$ the unbiasedness condition becomes $F^{\mathrm{T}}c\left(x\right) = f\left(x\right)$

$$MSE\left(x\right) = \sigma^2\left(1 + c\left(x\right)^{\mathrm{T}} Rc\left(x\right) - 2c\left(x\right)^{\mathrm{T}} r\left(x\right)\right)$$

where $r\left(x\right)$ is a vector that holds the correlations between the untried x and sites in S.

$r\left(x\right) = \left[\rho\left(s_1, x\right), \rho\left(s_2, x\right), ..., \rho\left(s_{\mathrm{m}}, x\right)\right]^{\mathrm{T}}$

Thus the constrained optimization problem stated above is to minimize *MSE* (*x*) subject to $F^{\mathrm{T}}c\left(x\right) = f\left(x\right)$. This can be solved using the method of Lagrange multipliers. The Lagrange equation is with a vector Lagrange multipliers λ is

$$L\left(c, \lambda\right) = MSE\left(x\right) - \lambda^{\mathrm{T}}\left(F^{\mathrm{T}}c\left(x\right) - f\left(x\right)\right)$$

$$\therefore \frac{\partial L}{\partial \lambda} = 0 \Rightarrow F^{\mathrm{T}}c\left(x\right) - f\left(x\right) = 0 \quad \& \quad \frac{\partial L}{\partial c} = 0 \Rightarrow 2\sigma^2\left(Rc\left(x\right) - r\left(x\right)\right) - F\lambda = 0$$

$$\therefore\ c\left(x\right) = R^{-1}\left(r\left(x\right) + F\left(F^{\mathrm{T}}R^{-1}F\right)^{-1}\left(f\left(x\right) - F^{\mathrm{T}}R^{-1}r\left(x\right)\right)\right)$$

Thus the BLUP can now presented as below. Using that R is symmetric,

$$y_p\left(x\right) = r\left(x\right)^{\mathrm{T}} R^{-1}Y - \left(F^{\mathrm{T}}R^{-1}r\left(x\right) - f\left(x\right)\right)^{\mathrm{T}}\left(F^{\mathrm{T}}R^{-1}F\right)^{-1} F^{\mathrm{T}}R^{-1}Y$$

Appendix 3

Test functions for Adaptive Sampling

1. Peaks function

$$peaks(x_1, x_2) = 3(1-x_1)^2 \exp\left(-x^2-(y+1)^2\right)-10\left(\frac{x}{5}-x^3-y^5\right)\exp\left(-x^2-y^2\right)-\frac{1}{3}\exp\left(-(x+1)^2-y^2\right)$$

$$-3 \le x_1 \le 3, -3 \le x_2 \le 3$$

2. Ackley's path function

$$ackley(x_1, x_2) = -a * \exp\left(-b * \sqrt{(1/n)(x_1{}^2 + x_2{}^2)}\right) - \exp\left(\frac{1}{n}\left(\cos(cx_1) + \cos(cx_2)\right)\right) + a + e$$

$$a = 20, b = 0.2, c = 2\pi; x_1, x_2 \in [-30, 30]$$

3. Rosenbrock's valley function

$$rosen(x_1, x_2) = \sum_{i=1}^{2}\left[100\left(x_{i+1} - x_i{}^2\right)\right]^2 + (1 - x_i)$$

$$x_1, x_2 \in [-30, 30]$$

4. Branin's rcos function

$$branin(x_1, x_2) = a\left(x_2 - bx_1^2 + cx_1 - d\right)^2 + e(1-f)\cos x_1 + e$$

$$a = 1, b = 5.1/(4\pi^2), c = 5/\pi, d = 6, e = 10, f = 1/8\pi, \ -5 \le x_1 \le 10, 0 \le x_2 \le 15$$

5. Goldstein-Price function

$$goldsteinprice(x_1, x_2) = \left[1 + (x_1 + x_2 + 1)^2\left(19 - 14x_1 + 3x_1{}^2 - 14x_2 + 6x_1x_2 + 3x_2^2\right)\right]*$$

$$\left[30 + (2x_1 - 3x_2)^2\left(18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2\right)\right]$$

$$x_1, x_2 \in [-2, 2]$$