INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN, ICED'07
28 - 31 AUGUST 2007, CITE DES SCIENCES ET DE L'INDUSTRIE, PARIS, FRANCE

Q2S2: QUALITATIVE AND QUANTITATIVE SEQUENTIAL SAMPLING - A NOVEL APPROACH TO EXPLOIT QUALITATIVE DESIGN INFORMATION

Rahul Rai[1], and Matthew I. Campbell[2]

[1] NSF Center for e-Design, Virginia Tech, Blacksburg, VA-24061, USA
[2] Department of Mechanical Engineering, University of Texas at Austin, TX-78712, USA

ABSTRACT

Sequential sampling refers to a set of design of experiment (DOE) methods where the next sample point is determined by information from previous experiments. This paper introduces a qualitative and quantitative sequential sampling (Q2S2) technique, in which optimization and user knowledge is used to guide the efficient choice of sample points. This method combines information from multiple fidelity sources including computer simulation models of the product, first principals involved in design, and designer's qualitative intuitions about the design. Both *quantitative* and *qualitative* information from different sources are merged together to arrive at a new sampling strategy. This is accomplished by introducing the concept of a confidence function, C, which is represented as a field that is a function of the decision variables, **x**, and the performance parameter, f. We compare the sampling plans generated by Q2S2 to previously known sample plans on two test functions using various metrics. In each case, the performance of Q2S2 is highly encouraging.

Keywords: Design of Experiments (DOE), Metamodeling, Qualitative Information

1 INTRODUCTION

The field of study, sometimes referred to "Design of Experiments" (or DOE) deals with what efficient set of experiments should be run to best understand a given phenomena [1-9]. DOE is a structured, organized approach for effectively and efficiently exploring the cause-and-effect relationship between numerous impact variables (the **x**'s) and the resulting performance variable (the f). DOE and in particular Metamodeling [10-16] comprise a group of statistical techniques for empirical model building and analysis. Design of experiments research is pertinent to not only engineering sciences, but also to any physical or natural sciences seeking to better understand a specific phenomenon [17, 18]. Several surveys of DOE and metamodeling approaches have been conducted [11, 19, 20] with respect to their suitability for various applications.

In recent years, there has been a major advance in DoE in the form of sequential sampling. The basic gist of sequential sampling is to determine experimental points based on previous experiments as opposed to traditional approaches which determine all experiments prior to any known results [21-27]. In this paper, we improve sequential sampling by drawing not only on the results of previous experiments but also on the qualitative knowledge that the user knows about the phenomenon under study. Furthermore, the concepts introduced in the paper, can be used to merge data returned from results of varying fidelity, for example, results from detailed experiments, results from computer simulations, or results from approximations specified by the user.

Prior information from earlier related experiments, observational studies, or subjective beliefs from personal observations, can be valuable in deciding how to allocate experiments efficiently, leading to more informative experiments. Because information is usually available prior to experimentation and, indeed, often motivates doing the experiment, Bayesian methods are suited to contribute to experimental design. The Bayesian approach to experimental design provides a formal way to incorporate such prior information into the design process. Lindley [28] presented a two-part decision theoretic approach to experimental design, which provides a unifying theory for most work in Bayesian experimental design today. Lindley's approach involves specification of a suitable utility function reflecting the purpose and costs of the experiment; the best design is selected to maximize expected utility. The reader is refered to a detailed review of Bayesian experimental design by

Chaloner and Verdinelli [29] for additional details on these criteria, references, and relationships of other Bayesian utility functions to standard optimality criteria. Bayesian experimental design has gained popularity in the past two decades. But also like many areas of Bayesian statistics, application to actual experiments still lag behind the theory [29]. Chaloner and Verdinelli [29] report that apart from Flournoy [30] there are no true "case studies" where Bayesian ideas have been formerly applied to the design of an actual scientific experiment before it is done. This paper presents a deterministic approach to integrate *apriori* knowledge into DoE.

Since information in engineering problems can be both quantitative (QT) and qualitative (QL) in nature, combining both types of information can be more beneficial in solving real world design problems. Even though QT models are useful in providing detailed information about a design problem, they can be ineffective in situations where the mathematical formulation of design is not available or is partially defined. In such cases QL information can provide a valuable access to the design problem by taking advantages of human approximate reasoning to improve the complex design problem representation. However, all the sequential sampling approaches reported in the literature are purely quantitative in nature and thus not conducive to process information in such a mixed environment. In much of the research on sequential sampling, QT measures (such as Mean Squared Error or Maximum Likelihood Estimation) are used as criterion for selecting the next sample point[24, 25]. These sequential sampling methods have following drawbacks:

- Existing techniques are unable to merge information from different sources such as actual physical experiment (Non-deterministic and QT), and computer simulations (Deterministic and QT). In fact, most techniques proposed work well only for design and analysis of computer experiment (deterministic experiments).
- Sampling strategies are dependent upon the properties of the metamodel which is used to fit the model for the data.
- For most design problems, there exists qualitative information in the form of user guesses, first principles, monotonicity information and sensitivity information. No technique in the literature exploits this readily available and useful source of information about the design problem.

These drawbacks can be attributed to the fact that there exists no method to assimilate QT and QL information coming from various sources. This necessitates the development of a method that can easily handle, merge and manipulate both QT and QL information. Secondly, the method developed should be independent or decoupled of the properties of the metamodel. This paper proposes a mathematically rigorous methodology for handling integrated QT and QL information to suitably represent and explore search space in real world design problems. To the best of our knowledge this is the first attempt to merge qualitative information via deterministic approach in DOE process. An important distinction is often made between DOE procedures that seek to maximize/minimize response value, and those geared toward gaining information about the problem. This paper is motivated toward information gathering process.

The remainder of the paper is organized as follows. Section 2 briefly presents an illustrative example to elucidate the scenario where our methodology could be useful. Section 2 also presents basic definitions pertaining to this paper. We then describe the concept of confidence and mathematical models that captures both quantitative and qualitative data in section 3. Section 4 describes the qualitative and quantitative sequential sampling (Q2S2) methodology in which optimization and user knowledge is used to guide the efficient choice of sample points. The measures of the efficiency are presented in section 5. The applicability of Q2S2 to real problems is also demonstrated in Section 5 for four example problems where the structure of response model is not known *apriori* and is constructed by the methodology. Section 6 concludes the paper.

2 ILLUSTRATIVE EXAMPLE

The illustrative example is founded on a simple projectile problem available in standard undergraduate physics book. This 2-D test problem models the distance traversed by a projectile. Consider a projectile thrown at some elevation angle φ and initial velocity v_0. The goal of the problem is to determine the flight distance d of the projectile in terms of these inputs. Ignoring air resistance and assuming the start and the landing position of the projectile are at the same ground level, the flight distance is given by $d = 2v_0^2 \sin\varphi \cos\varphi/g$. Qualitatively, a greater initial velocity of the projectile

results in a longer flight distance. However, flight distance also depends on the elevation angle. The maximal flight distance is achieved by throwing the projectile at elevation angle $\varphi = \pi/4$ (45degrees). Qualitatively, flight distance monotonically increases with elevation angle if $\varphi < \pi/4$, and decreases with elevation angle if $\varphi > \pi/4$. In both cases, flight distance monotonically increases with initial velocity. Thus, we have two piecewise monotonicities with respect to variable φ and one full range monotonicity with respect to the variable v_0 (for φ the range considered is only between 0 and 90 degrees). It is "easy" to guess that with initial velocity equal to 0, the distance traversed by the projectile will be 0. Furthermore, it is also "easy" to guess that with elevation φ being equal to 0 or $\pi/2$ (90 degrees) i.e., if you throw a projectile directly upward or downward, the distance traversed by the projectile will be 0. As one can notice that this information about the problem is qualitative in nature and is readily available even before any black box or actual experimentation based approach is applied to fit a metamodel for the problem. Generally, for any problem in hand, and especially design problems we have qualitative information of this nature at our disposal. We believe that using these qualitative information in conjunction with quantitative information available from existing sequential sampling techniques will lead to better and effective sequential sampling techniques.
In the next section we present the concept of confidence that captures both quantitative and qualitative data. Section 3 also presents mathematical models and their properties which help us to capture information from different sources such as actual physical experiment, computer simulation and user guesses.

3 CONFIDENCE AS A FUNCTION OF X AND F

The basis for this method is a qualitative measure called confidence that ranges from -1 and 1. As an illustrative example, we have run a single experiment in the testing of a heat sink design (Figure 1) where a design variable is set to a length of 0.4 in. The result of the experiment shows a resulting surface temperature of 600°F. Given this information, one has a better sense of how other values of length will perform. Qualitatively, our confidence about values near 0.4 in. is likely to be higher than values far from 0.4 in. As a method to capture this notion, a **confidence** function is developed that starts at *one* at *l*=0.4 in. and diminishes exponentially to *zero* as we move away from that value (in the a–a direction of Figure 1). Furthermore, if we consider moving vertically along the temperature direction (in the b–b direction of Figure 1), we would be surprised if the true value deviated significantly from the experimental value of 600°F. This is captured in the confidence function as an exponential decay to negative one. It is not that we lose confidence in the prediction but rather that we

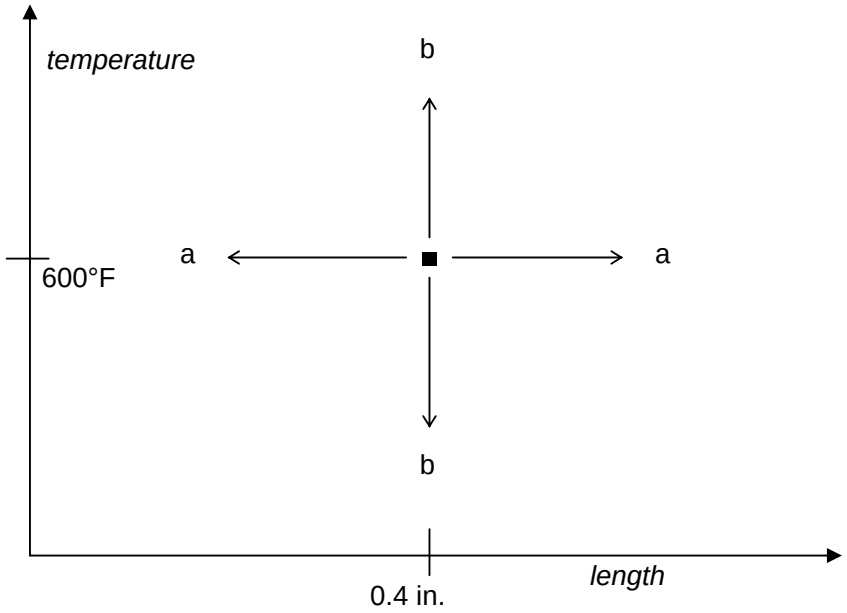

Figure 1: A single experiment in the design of a heat sink shows an experiment run at l = 4 in. yields a surface temperature of 600°F

Table 1. Different functions for representing C fields.

Name	*R(d)*
Exponential	$\exp(-\lvert d\rvert)$
General Exponential	$\exp(-\lvert d\rvert^{\theta_{n+1}}),\quad 0<\theta_{n+1}\le 2$
Gauss	$\exp(-d^2)$
Linear	$\max\{0,1-\lvert d\rvert\}$
Spherical	$1-1.5\xi+0.5\xi^2 \qquad \xi=\min\{1,\lvert d\rvert\}$
Cubic	$1-3\xi^2+2\xi^3 \qquad \xi=\min\{1,\lvert d\rvert\}$
Spline	$\begin{cases} 1-15\xi^2+30\xi^3 \; for 0\le\xi\le 0.2 \\ 1.25(1-\xi)^3 \; for 0.2\le\xi\le 1 \\ 0 \; for \xi>1 \end{cases}$ where $\xi=\lvert d\rvert$

become confident that such a value is incorrect (unless, of course, the experiment was greatly flawed). For example, running a replicate experiment at *l*=0.4in. would likely produce temperatures around 600°F, and unlikely to produce values at, say 10°F or 1000°F. The newly defined parameter of Confidence, or *C*, is a function of both the decision variables, **x**, and the phenomenon in question *f*. The values of *C* range from *-1* to *1*, where:

$C[f_0 : \mathbf{x_0}] = 1$ corresponds to absolute certainty that the value of $\mathbf{x}_0$ does produce the value of f_0,
$C[f_0 : \mathbf{x_0}] = -1$ corresponds to absolute certainty that the value of $\mathbf{x}_0$ does **not** produce the value of f_0,
$C[f_0 : \mathbf{x_0}] = 0$ corresponds to absolute **un**certainty in whether the value of $\mathbf{x}_0$ can produce the value of f_0.

Adding this Confidence field affords more benefits than the harm introduced by adding this new dimension. For example, one can now include information of varying degrees of fidelity. We can update the C-field to include information about detailed experiments (Figure 2a) or monotonicity information (Figure 2b). The equation for the C-field of actual experiment for a one variable design problem presented in Figure 2a is:

$$C(f : \vec{x}) = (2e^{-\left(\frac{f-f_0}{f_{tol}}\right)^2} - 1)(e^{-\left(\frac{\vec{x}-\vec{x}_0}{\vec{x}_{\inf}}\right)^2}) \qquad (1)$$

where, xinf or xinfluence is similar to the □ in Kriging, and the ftol or ftolerance is the expected error or standard deviation in the experiment due to the fidelty aspect of the information source. By varying the values of ftol value we can integrate the information from different sources while taking into account the fidelty issue. Typically, the ftol value will be higher for a user guess, than a computer simulation followed by an actual physical experiment in that order (i.e, ftol_userguess>ftol_simulation>ftol_actualexperiment). Since, confidence is a newly defined qualitative value; other formulations may be pursued (for example, algebraic or trigonometric expressions as opposed to this exponential one). For actual experiments, simulation or user guesses the equation for the C-field as a function of both $\vec{x}$ and *f* and can be written in the convolution form as follows:

$$C_e(f : \vec{x}) = \left(f_f(u_f) * x_1(u_1) * x_2(u_2) * \ldots x_N(u_N)\right) \qquad (2)$$

$$\text{Or } C_e(f : \vec{x}) = f_f(u_f) * \prod_{i=1}^{I} x_i(u_i) \qquad (3)$$

where the subscript *e* denotes that the c-field is for actual experiment or user guess. The function can be written as follows:

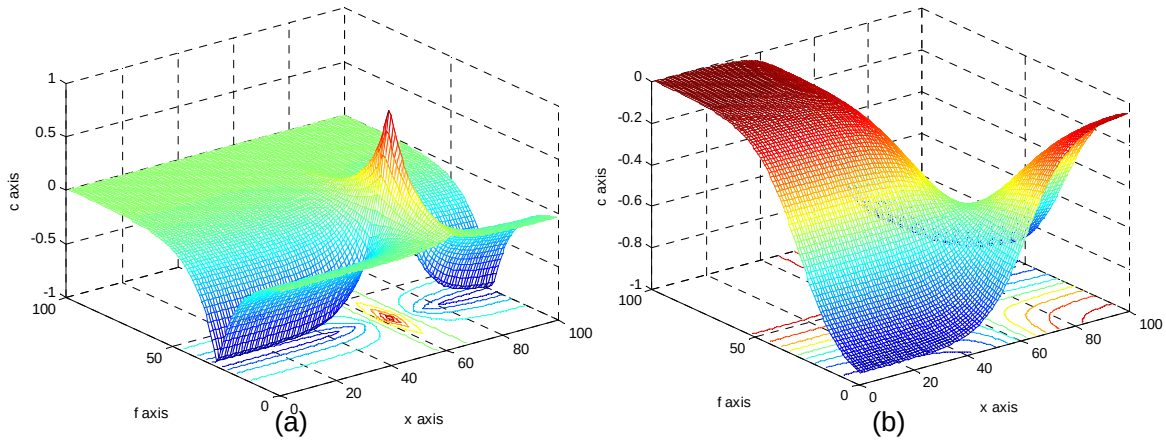

Figure 2: a) the C-field created by introducing the results of a single experiment; b) the C-field that results from including monotonicity information

$$f_f(u_f) = (2R_f(d_f) - 1). \tag{4}$$

The subscript *f* here denotes the dependence of C field on the objective function, and $x_i(u_i) = R_i(d_i)$.

Where, $d_f = \left(\frac{f - f_0}{f_{tol}}\right)$ and $d_i = \left(\frac{x_i - f_{i0}}{x_{itol}}\right)$. (5)

More specifically, there could be 8 different functional choices for *R(d)* as presented in Table 1. For one objective function and 3 variables if *R(d)* is same for all and is of Gaussian form then the C-field equation for a given experiment or user guess can be written as follows;

$$C(f : \vec{x}) = (2 * \exp(-d_f^{\,2}) - 1) * \prod_{i=1}^{3} \exp(-d_i^{\,2}) \tag{6}$$

where d_f and d_i are given by equation 5.

The actual C function could be formulated by using any of the 8 mentioned form of *d* for *f* and any $\vec{x}_i$. Different forms of R(d) can be used in the same problem as well. For example, for 2 variables and 1 objective function, the 1st variable can have a Gaussian form whereas the second variable can have spline form for *d*. The *f* can also have any of the eight forms.

In Figure 2b, a plot is shown for monotonicity information. The equation for the C-field shown in the Figure 2b is inspired by fuzzy-set theory [31], where a 1-dimensional sigmoid function is used to "black out" regions that are impossible to reach.

This monotonicity information can be provided through user queries. For example in the heat sink problem, a user may claim that the temperature decreases with longer lengths. Given a single experiment at length equal to 0.4 in. ($\mathbf{x_0}$ = .4) and temperature at 600^0C (*f* = 600), the C-field can be adjusted at that point to eliminate the areas that are impossible due to this monotonicity. The equation for the C-field for one variable monotonicity information is shown in Figure 2b and is represented by the following equation:

$$C(f : \vec{x}) = -\left\{\left(\frac{1}{1 + e^{-\left|\frac{f - f_0}{f_{tol}}\right|}}\right)\left(\frac{1}{1 + e^{-\left|\frac{\vec{x} - \vec{x}_0}{\vec{x}_{\inf}}\right|}}\right) + \left(\frac{1}{1 + e^{\left|\frac{f - f_0}{f_{tol}}\right|}}\right)\left(\frac{1}{1 + e^{\left|\frac{\vec{x} - \vec{x}_0}{\vec{x}_{\inf}}\right|}}\right)\right\}. \tag{7}$$

We consider two types of monotonicity information i.e., positive monotonicity and negative monotonicity. The C-field value for monotonicity information ranges from 0 to -1. We use sigmoid function from the mathematical literature to represent the monotonicity information. For the convenience purpose let us define following:

$$sigmoid(y) = \phi^+(y) = \frac{1}{1 + e^{-y}} \tag{8}$$

$$sigmoid(-y) = \phi^{-}(y) = \frac{1}{1+e^{y}} \tag{9}$$

Then equation for representing C-fields of negative monotonicity in terms of sigmoid function can be written as follows:

$$C_{-mon}(f:\vec{x}) = -\left(\phi^{+}(d_f) * \prod_{i=1}^{I}\phi^{+}(d_i) + \phi^{-}(d_f) * \prod_{i=1}^{I}\phi^{-}(d_i)\right) \tag{10}$$

And the equation for representing C-fields of positive monotonicity in terms of sigmoid function can be written as follows:

$$C_{+mon}(f:\vec{x}) = -\left(\phi^{-}(d_f) * \prod_{i=1}^{I}\phi^{+}(d_i) + \phi^{+}(d_f) * \prod_{i=1}^{I}\phi^{-}(d_i)\right) \tag{11}$$

Where d_f and d_i are defined by equation 5.

The C-field values of monotonicity function always have a range [-1 0]. This is because the monotonicity function just gives us the appropriate information where the solution can not lie (or the space that need not be searched) but it does not provide any information about the space where the solution lies.

In the next section we present our (Q2S2) methodology which uses the models of C fields information presented in this section for sequential sampling purposes. Different facets of the methodology are elaborated and present with a mathematical rigor where necessary.

4 QUALITATIVE AND QUANTITATIVE SEQUENTIAL SAMPLING

The methodology begins by taking inputs from the user about number of variables and objective functions (Step 1 Figure 3). Then it interacts with user by querying for specific qualitative information (Step 2 Figure 3). Our current implementation includes the following interactions with the user:

Interval Information: The program begins by asking about the ranges of ***x*** and *f* it needs to consider. Once the user provides the ranges of ***x*** and *f* (or the ranges they think are relevant to the problem at hand) the program proceeds to ask further questions.

Monotonicity Information: Once the interval information is inputted, the program asks the user whether he/she has monotonicity information. If the user has any monotonicity information the program represents this information as a C-field described in equation 2. The monotonicity information handled by the program can be for any number of decision variables. For example, if a user knows that the problem at hand is monotonic with respect to only one *x* in a three variable problem but not with respect to the other two variables, the program can still capture and use this

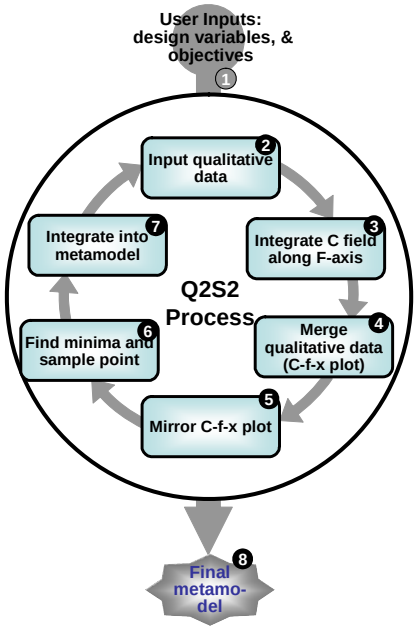

Figure 3: The Q2S2 process is cycle of six main steps.

information. The process asks user to explicitly input the type of monotonicity (positive or negative). The code written can also capture piecewise monotonicity (if any) for each *x*.

Actual experiment or simulation data: The program also asks about the values of **x** and *f* for which an actual experiment or simulation was performed. The distinction between the simulation and actual experiment is captured by the f_{tol} values of equation 1. The tolerance value for computational simulations may be kept relatively high, whereas the tolerance value for actual experiment is kept small to account for experimental variations. The program asks for the tolerance values if the data is from actual experiment.

Approximation or user guess: The program then proceeds to ask user whether they know of any approximate values at the current stage. Any approximations, no matter how imprecise can improve the model in location where no information currently exists.

It is important to note that experiments exist only at values of ***x***. and not at values of both ***x*** and *f*. Therefore, in Step 3 (of Figure 3), the methodology integrates C-field along the *f* dimension for each ***x*** to produce a new value of confidence as a simpler function of just ***x***:

$$C'(\vec{x}) = \int_{f_{min}}^{f_{max}} C(f : \vec{x}_0)df \ . \qquad (12)$$

Then the methodology progresses to step 4 (details in Subsection 4.1), where it merges all the information at hand to create a single overall *C'-f-**x*** function for the problem. In Step 5, the methodology mirrors the value of the constructed *C'-f-**x*** function. Since *C'* ranges from -1 to 1, the mirroring functions reflects the negative C' values to positive values, thus making it now range from 0 to 1. This is done to identify where values are closest to zero since zero corresponds with no knowledge about the model. This mirroring, represents one of the main advantage in creating confidence as a field ranging from -1 to 1. Not only can this range represent whether a relationship exists or not, but it can also be normalized to identify where the least (and most) information is known (details in Subsection 4.2). Based on this new refined C′, Step 6 performs an optimization using simple gradient based search techniques to identify the minimizer of C′ as the subsequent sample point. This sampling point is then integrated into the previous metamodel (which may be Kriging or B-splines). The Steps 2-8 are repeated till we exhaust the number of sampling points we can expend or when there is no more information available. The output of the process is both the resulting C-field and the associated metamodel.

The Q2S2 was implemented using MATLAB® R12. The optimization toolbox in MATLAB® was used for the search process. In next subsections we discuss mathematical details pertaining to various steps of Q2S2 methodology. The next section presents example problems. The optimization space appears to become multimodal relatively quickly thus complicating the search for the global optimum. It is important to note that the minimization of the co-variance in Kriging requires a similar optimization process, and to solve the values of co-variance for each point sampled requires the set up and solving of an *n*-by-*n* matrix (where n is the number of existing sample points). Similarly in our approach, the combined C-field expression is indeed composed of n terms; however, by using the exponential equation we can create an analytical expression of not only our C-field but of it gradient and Hessian as well. In doing so, gradient based optimization methods are used quite efficiently since there is no need to find such information with finite difference approaches. The multimodality of the space is alleviated by initiating parallel optimization at the known local maxima. These maxima are known since they correspond to the peaks caused by the actual experiments.

Merging Functions

Step 4 of the flowchart (Figure 3) requires that different C-field information gathered from different sources or during different iterations to be merged to create a unified C-field. The combination function should be such that it does not violate the definition of C-field, i.e., for two given point a, $b \in [-1\ +1]$ the combination function $f(a, b) \rightarrow [-1\ \ +1]$. Two different merging functions that exhibit the property mentioned above can be used to merge different C-fields:

Normalizing merging function:

$$f(a,b) = (a^p + b^p)^{1/p} \qquad (13)$$

Where, *p* is odd integer. The restriction of *p* being an odd number is due to fact that range is [-1 1]. Merging a negative number with positive number can be preserved only when *p* is odd.

Averaging function:

(14)

$$f(a_1, a_2, \ldots a_n) = \frac{f(a_1) + f(a_2) + \ldots f(a_n)}{N}$$

Where, N is total number of C-fields data.

For analytical brevity and due to savings in computational efforts normalizing function was selected to proceed further.

Mirroring Function of C-f-x plots

There are two mirroring function that can be used.

Absolute Function: $\left| C(f : \vec{x}) \right|$ (15)

Square Function: $\left(C(f : \vec{x}) \right)^2$ (16)

Out of these two, square function is preferred over the absolute function due to following reasons.

Using Absolute function has the pitfall of creating potential cusps in the overall C'-f-x plot.

The Square function is more amenable for calculating the derivate of analytical expressions.

Analytical expression Derivation

For the purpose of analytical expression derivation, the Gaussian form was selected for actual experiment/user guess information. The sigmoid function was selected for representing both the positive and the negative monotonicity information. The primary reason behind these selections was the ease of differentiability of these functions. These functions can be easily differentiated easily to provide Gradient and Hessian information which can be used by the gradient based optimization technique. In case we want to use algorithms which do not require gradient information any of the above mentioned function can be used (Table 1).

Let us assume the actual experiment or user guess information is modeled through Guassian form in Table 1. Then for I variable problem the actual experiment/ user guess information can be modeled by.

$$C_e(f : \vec{x}) = (2e^{-d_f^2} - 1) * \prod_{i=1}^{I} e^{-d_i^2}, \quad (17)$$

Where d_f is given by equation 5.

$$\text{Let } \prod_{i=1}^{I} e^{-d_i^2} = A, \text{ Then, } C_e(f : \vec{x}) = A(2e^{-d_f^2} - 1) \quad (18)$$

Let us assume the negative monotonicity information is modeled through sigmoid form in equation 10. . Then for I variable problem the negative monotonicity information can be modeled by

$$C_{-mon}(f : \vec{x}) = -\left((1+e^{-d_f})^{-1} \prod_{i=1}^{I} (1+e^{-d_i})^{-1} + (1+e^{d_f})^{-1} * \prod_{i=1}^{I} (1+e^{d_i})^{-1} \right) = -\left(B(1+e^{-d_f})^{-1} + C(1+e^{d_f})^{-1} \right)$$

$$\text{Where } B = \prod_{i=1}^{I} (1+e^{-d_i})^{-1} \text{ and } C = \prod_{i=1}^{I} (1+e^{d_i})^{-1} \quad (19)$$

Let us assume the positive monotonicity information is modeled through sigmoid form in equation 11. Then for I variable problem the positive monotonicity information can be modeled by

$$C_{+mon}(f : \vec{x}) = -\left((1+e^{d_f})^{-1} \prod_{i=1}^{I} (1+e^{-d_i})^{-1} + (1+e^{-d_f})^{-1} * \prod_{i=1}^{I} (1+e^{d_i})^{-1} \right) = -\left(B(1+e^{d_f})^{-1} + C(1+e^{-d_f})^{-1} \right)$$

$$\text{Where } B = \prod_{i=1}^{I} (1+e^{-d_i})^{-1} \text{ and } C = \prod_{i=1}^{I} (1+e^{d_i})^{-1} \quad (20)$$

In step 3 (equation 12), the process integrates C-field along the *f* dimension for each ***x*** to produce a new value of confidence as a simpler function of just ***x***:

$$C'(\vec{x}) = \int_{f_{\min}}^{f_{\max}} C(f, \vec{x}_0) df .$$

Which can be rewritten in terms of d_f by simple substitution

$$C'(\vec{x}) = f_{tol} \int_{f_{\min}}^{f_{\max}} C(f, \vec{x}_0) dd_f \quad (21)$$

This integration when performed on experiment/ user guess information gives

(22)

$$C'_e(\vec{x}) = \int C_e(f : \vec{x})dd_f = \int A(2e^{-d_f^2} - 1)dd_f = A(\sqrt{\pi}erf(d_f) - d_f) + C$$

Here A is constant with respect to d_f as it is just a function of the input variable ***x***'s and not f. Similarly, this integration when performed on negative monotonicity and positive monotonicity equations we have,

$$C'_{-mon}(\vec{x}) = \int C_{-mon}(f : \vec{x})dd_f = -\left(B\ln\left(\frac{e^{-d_f}}{1+e^{-d_f}}\right) + C\ln\left(\frac{e^{d_f}}{1+e^{d_f}}\right)\right) + C \quad (23)$$

$$C'_{+mon}(\vec{x}) = \int C_{+mon}(f : \vec{x})dd_f = -\left(B\ln\left(\frac{e^{d_f}}{1+e^{d_f}}\right) + C\ln\left(\frac{e^{-d_f}}{1+e^{-d_f}}\right)\right) + C \quad (24)$$

Let us assume that when we reach step 4, we have L user guesses or actual experiment data, M positive monotonicity data, and N negative monotonicity data, and assuming normalized merging function being used for merging purpose, then mathematically step 4 can be written as follows:

$$C_{merge}(\vec{x}) = \left\{\sum_{l=1}^{L}\left(a_l\prod_{i=1}^{I}A_{li}\right)^p + \sum_{m=1}^{M}(-1)^p\left(b_m\prod_{i=1}^{I}B_{mi} + c_m\prod_{i=1}^{I}C_{mi}\right)^p + \sum_{n=1}^{N}(-1)^p\left(b_n\prod_{i=1}^{I}B_{mi} + c_n\prod_{i=1}^{I}C_{mi}\right)^p\right\}^{1/p} \quad (25)$$

Where, different coefficient a_l, b_m, c_m, b_n, and c_n are found from evaluation of integrations in step 3. The final equation at step 5 is just the square (assuming square function being used for mirroring) of the previous equation. The equation used during the optimization step can be written as:

$$C_{merge}(\vec{x}) = \left\{\sum_{l=1}^{L}\left(a_l\prod_{i=1}^{I}A_{li}\right)^p + \sum_{m=1}^{M}(-1)^p\left(b_m\prod_{i=1}^{I}B_{mi} + c_m\prod_{i=1}^{I}C_{mi}\right)^p + \sum_{n=1}^{N}(-1)^p\left(b_n\prod_{i=1}^{I}B_{mi} + c_n\prod_{i=1}^{I}C_{mi}\right)^p\right\}^{2/p} \quad (26)$$

This equation is in analytical form can still be differentiated to provide slope and Hessian information which can be easily used by gradient based search techniques.

5 PERFORMANCE COMPARISON

In this section, we investigate the performance of the Q2S2 technique by comparing its performance to that of several other classes of sampling plans found in the literature. A sampling technique results in a specified number of data points over the design domain where the actual function is evaluated. In the present paper, we use three representative sampling techniques for comparison purposes. These sampling techniques are Full Factorial, Latin Hybercube Design and Hammersley Sequential Sampling [18, 19, 28, 29]. The overall performance of the metamodels is evaluated using six standard accuracy measures from literature [11, 32]. These measures are *RSME, NRSME, RSQUARE, RAAE, RMAE,* and *MAXAB*. Only higher values of R Square are desirable, while smaller values of RSME, NRSME, RMAE, RAAE and MAXAB are sought. The effect of sample size on metamodel accuracy is also an important criterion for selecting an appropriate sampling approach. 16 sample points for 2 variable problems and 128 sample points for 7 variable problems were selected for comparison purposes.

Table 2. Test functions and monotonicity information used

Name	Function	# of variables	Design Domain
Test Function1 (Projectile)	$2v_0^2 \sin\varphi \cos\varphi/g$	2	$0 \le v_0 \le 18$ $0 \le \varphi \le 90$
Monotonicity Information used in Q2S2: $\varphi[0\ 30]^+$ $\varphi[60\ 90]^-$			
Test Function 2	$(x_1-10)^2+5(x_2-12)^2+x_3^4+3(x_4-11)^2+10x_5^6+7x_6^2+x_7^4-4x_6x_7-10x_6-8x_7$	7	$0 \le x_i \le 90$
Monotonicity Information used in Q2S2: $x_3[0\ 3]^+$ $x_3[9\ 10]^+$ $x_5[0\ 3]^+$ $x_5[9\ 10]^+$			

The error measure RSME is defined as:

$$RSME = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(f_i - \tilde{f}_i \right)^2 \right\}^{1/2} \tag{27}$$

Where, $\tilde{f}_i$ is the corresponding predicted value for the observed value f_i. The Normalized Root Mean Squared Error (NRMSE) as follows:

$$NRSME = \left\{ \sum_{i=1}^{n} \left(f_i - \tilde{f}_i \right)^2 / \sum_{i=1}^{n} \left(f_i \right)^2 \right\}^{1/2} \times 100 \tag{28}$$

Third measure is R square, we calculate the R square as follows:

$$RSQUARE = 1 - \sum_{i=1}^{n} \left(f_i - \tilde{f}_i \right)^2 / \sum_{i=1}^{n} \left(f_i - \bar{f}_i \right)^2 = 1 - \frac{MSE}{VARIANCE} \tag{29}$$

where, $\bar{f}_i$ is the mean of the observed values. While MSE (Mean Square Error) represents the departure of the metamodel from the real simulation model, the variance captures how irregular the problem is. The fourth measure is Relative Average Absolute Error (RAAE):

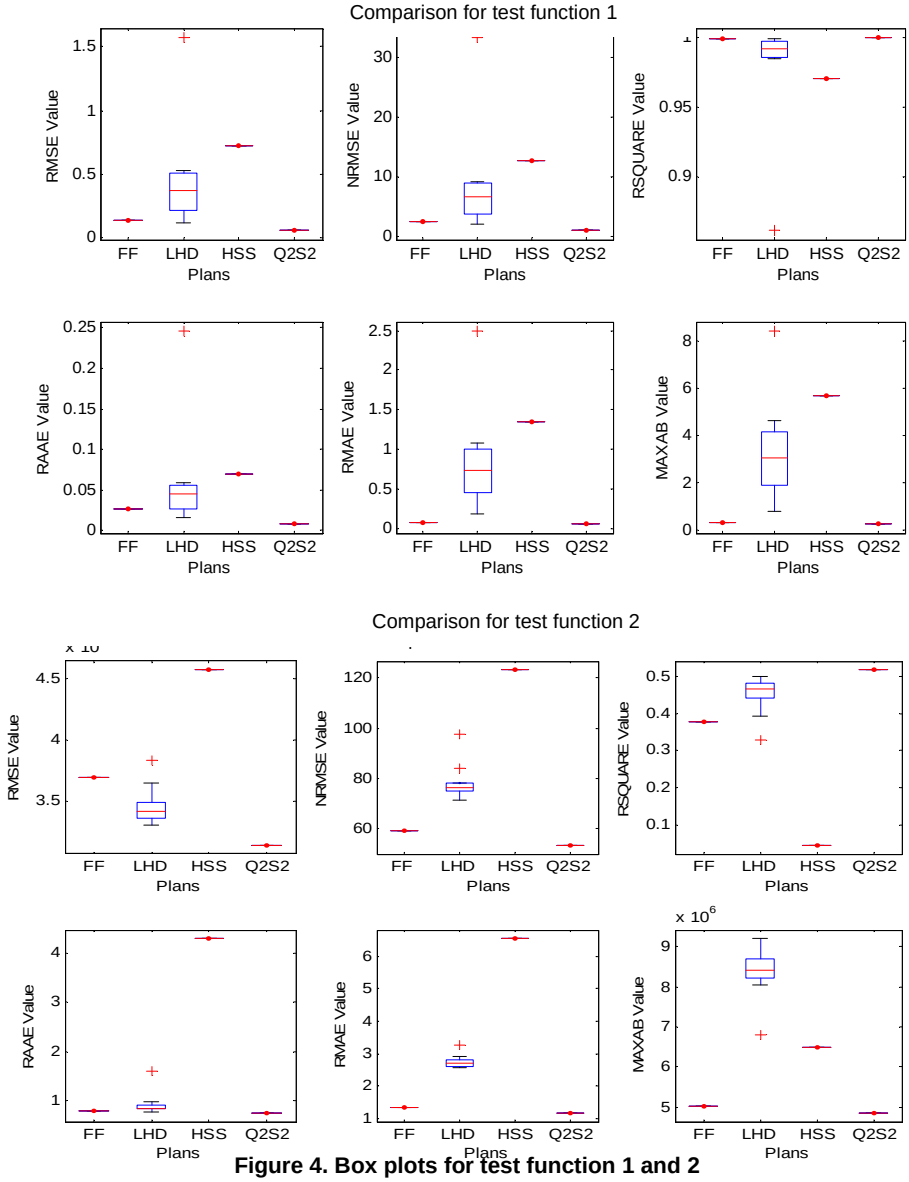

Figure 4. Box plots for test function 1 and 2

$$RAAE = \frac{\sum_{i=1}^{n} \left| f_i - \tilde{f}_i \right|}{n \times STD} \tag{30}$$

where STD stands for standard deviation. Fifth measure is Relative Maximum Absolute Error (RMAE):

$$RMAE = \frac{\max\left(\left| f_1 - \tilde{f}_1 \right|, \quad \left| f_2 - \tilde{f}_2 \right|, \ldots, \quad \left| f_n - \tilde{f}_n \right| \right)}{STD} \tag{31}$$

A *small* RMAE is preferred. Sixth measure is Maximum Absolute Bias (MAXAB):

$$MAXAB = \max\left(\left| f_1 - \tilde{f}_1 \right|, \quad \left| f_2 - \tilde{f}_2 \right|, \ldots, \quad \left| f_n - \tilde{f}_n \right| \right) \tag{32}$$

Based on the proposed scheme for comparative study, Kriging metamodels are created for the 2 test problems [4, 32] (from literature), using different sampling techniques (see Table 2). Table 2 also presents assumed qualitative information about the problems that has been used for Q2S2 method. Different techniques are compared based on the results from different metamodel accuracy measures. Boxplot graphs are used for comparisons (Figure 4).

Once the qualitative information about the problems are supplied to Q2S2, the efficiency of the sampling increases significantly. For each of the test function shown the Q2S2 produces the best sample points. As the number of variables increases Q2S2 is clearly better than LHD and HSS sampling techniques in every measure.

6 CONCLUSIONS AND FUTURE WORK

In this paper, a new approach that integrates both qualitative and quantitative information into a sequential sampling process is presented. Five example problems (four problems existing in current literature) were emulated and compared to show the effectiveness of the methodology. To the best of our knowledge this is the first attempt to include qualitative knowledge into sequential sampling techniques. The Q2S2 allows us to incorporate information of varying degrees of precision including detailed experiments, computational simulations, back-of-the-envelop calculations, and human intuition. The Q2S2 relying on human qualitative knowledge in conjunction with already existing quantitative techniques focuses efforts on important regions of the design space, thus eliminating extra sampling points and reducing the number of actual experiments performed.

The current implementation of Q2S2 does not rely upon existing metamodeling techniques. However, it will be beneficial to develop a metamodeling technique that is customized to incorporate this new qualitative information. For example, if we know that in certain regions of $\boldsymbol{x}$ there exists piecewise monotonicity (linear, quadratic so on) and directly map that information into the metamodel, it will be able to model the true nature of the problem more accurately. Current research is aimed at suitably modifying existing Kriging metamodels that would serve this purpose.

REFERENCES

[1] Box, G.E.P. and J.S. Hunter., The 2k-p fractional factorial designs. Technometrics, 1961. **3**: p. 311-351.

[2] McKay, M.D., Sensitivity and uncertainty analysis using a statistical sample of input values in Uncertainty Analysis, Y. In Ronen, editor,, Editor. 1988, CRC Press: Boca Raton, FL. p. 145–186.

[3] Montgomery, D.C., *Design and Analysis of Experiments*. 5th ed ed. 2001, New York: Wiley.

[4] Palmer, K. and K.-L. Tsui, A Minimum Bias Latin Hypercube Design. IIE Transactions, 2001. **33**(9): p. 793-808.

[5] Pacheco, J.E., C.H. Amon, and S. Finger, Bayesian Surrogates Applied to Conceptual Stages of the Engineering Design Process Journal of Mechanical Design, 2003. **125**(4): p. 664–672.

[6] Hilla, G. and B.-G. Irad, Designing experiments for robust-optimization problems: the V_s-optimality criterion. IIE Transactions (after Jan 1, 2003), 2006. **38**(6): p. 445-461.

[7] Chang, S., An algorithm to generate near D-optimal designs for multiple response surface models. IIE Transactions, 1997. **29**(12): p. 1073-1081.

[8] Myers, H.R. and D.C. Montgomery, *Response surface methodology* 1995, New York, NY: Wiley.
[9] Tsui, K.-L., Modeling and analysis of dynamic robust design experiments. IIE Transactions, 1999. **31**(12): p. 1113-1122.
[10] Wang, N. and P. Ge, Study of Metamodeling Techniques and Their Applications in Engineering Design ASME-MED Manufacturing Science and Engineering, 1999. **10**: p. 89-95.
[11] Simpson, T.W., Peplinski, J.D., Koch, P.N., and Allen, J.K., *Metamodels for Computer-based Engineering Design: Survey and Recommendations.* Engineering with Computers, 2001. **17**: p. 129-150.
[12] Friedman, J.H., Multivariate Adaptive Regression Splines The Annals of Statistics, 1991. **19**(1): p. 1-141.
[13] Barton, R.R. Simulation Metamodels. in ACM Proc. of the 1998 Winter Simulation Conference. 1998.
[14] Barton, R.R. Metamodeling: A State-of-the-Art Review. in ACM Proc. of the 1994 Winter Simulation Conference. 1994.
[15] Wheyming, S., S. Chien, and C. Wei-Chin, A pseudo-random number assignment strategy for the estimation of quadratic metamodels. IIE Transactions (after Jan 1, 2003), 2005. **37**(7): p. 657-666.
[16] Feng, C.-X., Z.-G. Yu, and A. Kusiak, *Selection and validation of predictive regression and neural network models based on designed experiments.* IIE Transactions, 2006. **38**(1): p. 13-23.
[17] Rikards, R., Abramovich, H., Kalnins, K., and Auzins, J., Surrogate modeling in design optimization of stiffened composite shells. Composite Structures, 2006. **73**(2): p. 244-251.
[18] Goh, T.N., A pragmatic approach to experimental design in industry. Journal of Applied Statistics, 2001. **28**(3): p. 391-398.
[19] Victoria, C., et al., A review on design, modeling and applications of computer experiments. IIE Transactions (after Jan 1, 2003), 2006. **38**(4): p. 273-291.
[20] Jeff Wu, C., M. Hamada, and V. Joseph, A Review of: â€œExperiments: Planning, Analysis, and Parameter Design Optimizationâ€□. IIE Transactions, 2006. **38**(6): p. 521-522.
[21] Sasena, M., Parkinson, M. B., Goovaerts, P., Papalambros P., and Reed, M. P.,. *Adaptive Experimental Design Applied to an Ergonomics Testing Procedure DAC-34091*. in *Proc. of the2002 ASME DETC Conferences*. 2002. Montreal, Quebec, Canada.
[22] Sasena, M., Flexibility and Efficiency Enhancements for Constrained Global Design Optimization with Kriging Approximations 2002, The University of Michigan: Ann Arbor, MI.
[23] Turner, C.J., M.I. Campbell, and R.H. Crawford. Generic sequential sampling for metamodel approximations CIE-48230 in Proc. of the2003 ASME DETC Conferences. 2003. Chicago, Illinois, USA.
[24] Sasena, M., Parkinson, M. B., Goovaerts, P., Papalambros P., and Reed, M. P.,, *Improving an Ergonomics Testing Procedure via Approximation-based Adaptive Experimental Design.* Journal of Mechanical Design, 2005. **127**: p. 1006-1013.
[25] Jin, R., W. Chen, and A. Sudjianto. On Sequential Sampling for Global Metamodeling in Engineering Design. in ASME Design Automation Conference Paper No. DETC-DAC34092. 2002: ASME.
[26] Pielaat, A., Lewis, M. A., Lele, S., and de-Camino-Beck, T., Sequential sampling designs for catching the tail of dispersal kernels. Ecological Modelling, 2006. **190**(1-2): p. 205-222.
[27] Ben-Gal, I. and M. Caramanis, *Sequential DOE via dynamic programming.* IIE Transactions, 2002. **34**(12): p. 1087-1100.
[28] Lindley, D.V., *Bayesian Statistics: A Review*, in *SIAM*. 1972: Philadelphia, PA, US.
[29] Chaloner, K., and Verdinelli, I., Bayesian Experimental Design: A Review Statistical Science, 1995. **10**: p. 273-304.
[30] Flournoy, N., ed. A clinical experiment in bone marrow transplantation: Estimating a percentage point of a quantal response curve Case Studies in Bayesian Statistics I, ed. J.H. C. Gatsonis, R. Kass, and N. Singpurwalla. 1993, Springer-Verlag. 324-336.
[31] Zimmermann, H.J., *Fuzzy Set Theory and its Applications*. 4th edition ed. 2001: Kluwer Academic Publishers.
[32] Mullur, A.A. and A. Messac, Metamodeling using extended radial basis functions: a comparative approach. Engineering with Computers 2005 **21**(3): p. 203 - 217.