EXPERIMENTAL DESIGN METHODS FOR NANO-FABRICATION PROCESSES

A Thesis
Presented to
The Academic Faculty

by

Sungil Kim

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
December 2011

EXPERIMENTAL DESIGN METHODS FOR NANO-FABRICATION PROCESSES

Approved by:

Professor Jye-Chyi Lu, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Martha Grover, Co-advisor
School of Chemical and Biomolecular Engineering
Georgia Institute of Technology

Professor Jianjun Shi
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Yajun Mei
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Myong K. Jeong
Department of Industrial & Systems Engineering
Rutgers, the State University of New Jersey

Date Approved: 15 August 2011

To my wife,

Heeyoung Kim.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Jye-Chyi Lu, for his guidance and support during my doctoral study. Without his help and encouragement, the accomplishment of this dissertation would not be possible. I would also like to express my appreciation to my co-advisor, Dr. Martha Grover, for her constant assistance and deep discussion through my research. Every discussion with her inspired and helped me toward the achievement of this milestone.

I am also very grateful to Dr. Jianjun Shi, Dr. Yajun Mei, and Dr. Myong K. Jeong for serving on my committee and giving me insightful comments. Their valuable suggestions and comments make this dissertation more complete.

I would like to extend my gratitude to all my friends at the Georgia Institute of Technology for their continued care and help in my doctoral student life. Last but not least, I would like to express my deepest appreciation to my parents and wife, Heeyoung, for endless love and support. They give me strength and help me through difficult times in this challenging journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Most design of experiments assumes predetermined design regions. Design regions with uncertainty are of interest in the first chapter. When a design region is restricted by inequality engineering constraints like solubility constraints (Chrastil, 1982), there might exist uncertainty in a design region. This uncertainty comes from the fact that engineering models are based on simplified assumptions. If a design region has uncertainty on its boundary, any data collection plan (e.g., optimal designs) that depends on the location of the design region boundary is not appropriate. Optimal designs tend to place many design points at the extreme limits of boundary regions. However, the boundary of the region is not precisely known in many engineering experiments (e.g., mechanical and chemical experiments). This chapter proposes optimal designs under a two-part model to handle the uncertainty in the design region. In particular, the logit model in the two-part model is used to assess the uncertainty on the boundary of the design region. This chapter derives the information matrix of the two-part model and constructed optimal designs. Through several examples, we show how the two-part models explain uncertainty in design regions and can be used for inequality engineering constraint estimation.

The second chapter proposes an efficient and effective multi-layer data collection scheme (Layers of Experiments) for building accurate statistical models to meet tight tolerance requirement commonly encountered in nano-fabrication. In nano-fabrication processes, due to high material costs and processing time for physical experiments, number of experimental runs is very limited. However, the limited resources make it difficult to estimate statistical models that are required to be accurate

enough to meet a tight tolerance requirement. To overcome these difficulties, "Layers-of-Experiments" (LOE) obtain sub-regions of interest (layer) where the process optimum is expected to lie and collect more data in the sub-regions with concentrated focus. An evaluation metric is developed to measure the performance of statistical models for nano-fabrication quality prediction and the metric is used to decide whether further layers are needed. This chapter also discusses appropriate types of designs for each layer, e.g. space-filling designs or optimal designs.

The third chapter contributes a new design criterion combining model-based optimal design and model-free space-filling design in a constraint and compound manner. Optimal design criterion is for precise statistical inference, while the space-filling design criterion is for exploration over the design space. The weights between the two criteria in the combined design is controlled by an adaptive parameter (κ) depending on the available information provided for a specific application (see chapter 4 for more examples). The proposed design is useful when the fitted statistical model is required to have both characteristics: accuracy in statistical inference and design space exploration. We showed that combined designs have properties between optimal designs and space-filling designs and they are robust against model misspecification. Moreover, combined designs perform better than space-filling designs or optimal designs where partial information about underlying model is available.

The fourth chapter proposes a method to determine the adaptive parameter (κ) sequentially in the layers of experiments. The parameter reflects the uncertainty of each layer, e.g., less uncertainty on the design space, more weights on model-based optimal criteria. This chapter also develops methods to improve model quality by combining information from various layers and from engineering models. Combined designs are modified to improve its efficiency by incorporate collected field data from several layers of experiments. Updated engineering models are used to build more accurate statistical models.

CHAPTER I

INTRODUCTION

1.1 Model-free Design Criteria

Design of experiments is a method used to plan experiments to gain the most information possible from the experiments.

1.1.1 Geometrical Criteria

The idea of distance-based designs was first introduced by Johnson, Moore, and Ylvisaker (1990). These designs are selected based on how spread out the design points are according to a distance measure or metric. Let $\tau(\cdot, \cdot)$ be a distance measure. For example, a commonly used distance measure, the pth order distance between two points $\mathbf{x}_1$, $\mathbf{x}_2$ is given by

$$\tau(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{j=1}^{d} |x_{1j} - x_{2j}|^p \right]^{1/p}, \quad \text{where } p \geq 1.$$

The rectangular and Euclidean distances can be obtained as special cases of the pth order distance, by setting $p = 1$ and $p = 2$, respectively.

Let D denote an arbitrary design consisting of n distinct input sites $\{x_1, x_2, \ldots, x_n\}$, and $\mathscr{X}$ be the collection of all possible points in the experimental region. Then, $\xi \subset \mathscr{X}$.

The maximin criterion tries to spread out the points in $\mathscr{X}$ so that the minimum distance among the design points is maximized. Thus, the maximin design (ξ_{Mm}) can be defined as

$$\min_{x_i, x_j \in \xi_{Mm}} \tau_p(x_i, x_j) = \max_{\xi} \min_{x_i, x_j \in \xi} \tau_p(x_i, x_j).$$

In contrast, the minimax criterion tries to spread out the points in $\mathscr{X}$ so that the maximum distance from any point $x \in \mathscr{X}$ to the design is minimized. A design ξ_m

is said to be a minimax distance design if

$$\min_{\xi} \max_{\mathbf{x}\in\mathscr{X}} \tau_p(\mathbf{x}, \xi) = \max_{\mathbf{x}\in\mathscr{X}} \tau_p(\mathbf{x}, \xi_m).$$

Minimax distance designs ensure that every point in the experimental region is close to some point in the design. See Koehler and Owen (1996) for an interesting way to understand the concept behind maximin and minimax designs.

Other distance-based designs have been suggested that attempt to select design points such that the ‘average of some function of the distances between the points is minimized; see Santner et al. (2003).

1.1.2 Latin Hypercube Design

Latin hypercube sampling is an extension of the idea of stratified random sampling, in that Latin hypercube sampling ensures that all portions of the distribution of each input variable are represented in the sample. Designs generated by Latin hypercube sampling are called Latin hypercube designs. It is worth noting that Latin hypercube designs were first proposed by McKay, Beckman, and Conover (1979) for the purpose of numerical integration.

Let us consider the simple case of obtaining a design consisting of n points in a $[0,1]^2$ experimental region. First, each input dimension is divided into n partitions, that is, $[0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1]$, such that the experimental region is divided into a grid of n^2 cells, that is, $[0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1] \times [0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1]$. Next, n cells are chosen from this grid of n2 cells, such that each row and each column is represented by one cell only. This ensures that there is no replication, and that points are marginally spread (quite) evenly over the values of each input variable.

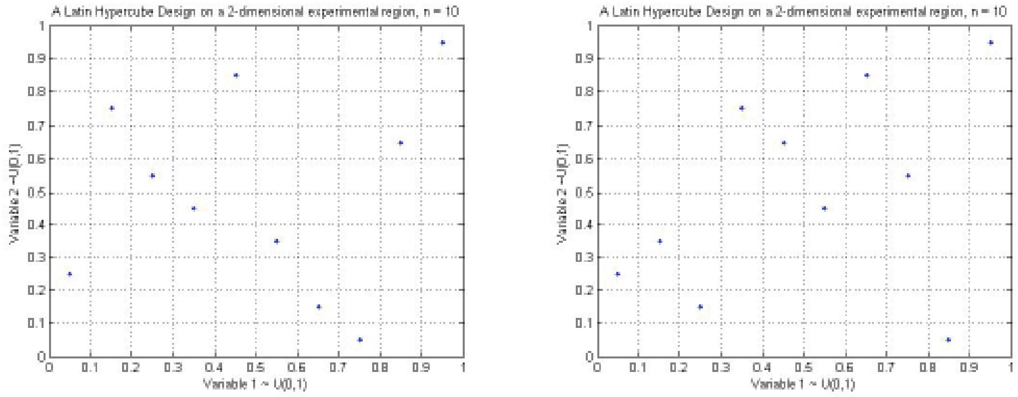


Figure 1: Two Latin hypercube designs on a 2-dimensional experimental region, $n = 10$

As can be seen in Figure 1, Latin hypercube designs are not unique, and there is a possibility that we might end up with a design that has all the design points lying along the diagonal; Figure 2 depicts such a design this design is not space-filling.

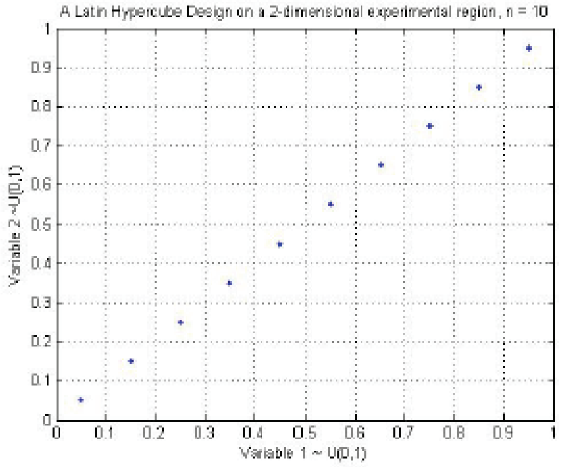


Figure 2: A Latin hypercube designs on a 2-dimensional experimental region, $n = 10$

Let us take a look at the details of constructing a Latin hypercube design consisting of n points, in a more general setting. Assume that we have d input variables, and

$G_j(.)$, $j = 1, \ldots, d$ is the marginal distribution of the jth input variable, X_j . Partition each axis into n segments, each with probability $1/n$, under $G_j(.)$. The division points for jth axis are then,

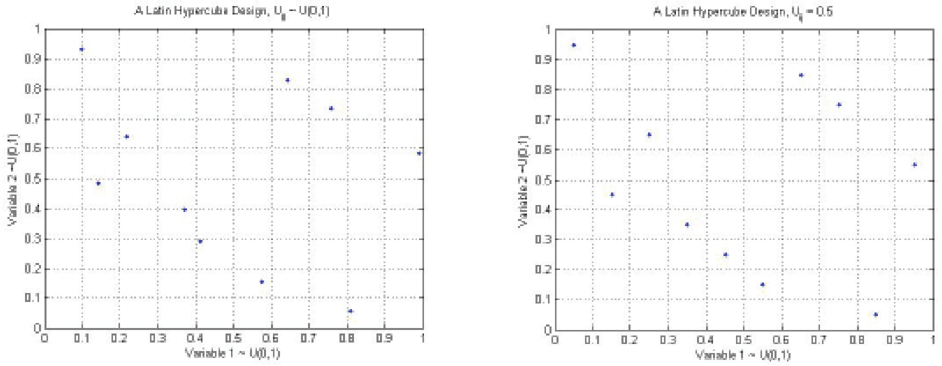
$$G_j^{-1}\left(\frac{1}{n}\right), G_j^{-1}\left(\frac{2}{n}\right), \ldots, G_j^{-1}\left(\frac{n-1}{n}\right).$$

Let, $\mathbf{\Pi} = (\Pi_{ij})$, $i = 1, \ldots, n$, and $j = 1, \ldots, d$ be an $n \times d$ matrix whose columns are d different randomly chosen permutations of $\{1, 2, \ldots, n\}$. Then the Latin hypercube design is given by

$$\mathbf{X}_{ij} = G_j^{-1}\left(\frac{1}{n}(\Pi_{ij} - 1 + U_{ij})\right),$$

where U_{ij} , $i = 1, \ldots, n$, and $j = 1, \ldots, d$ are independently and identically distributed $U(0, 1)$ random variables. Thus, the ith row of $\mathbf{\Pi}$ determines which cell the ith observation X_i, $i = 1, \ldots, n$ should be made in, and the corresponding uniform deviates U_{ij} determine the location of the ith observation in the chosen cell. Sometimes, U_{ij} can be set 0.5, for $i = 1, \ldots, n$, and $j = 1, \ldots, d$. Then,

$$\mathbf{X}_{ij} = G_j^{-1}\left(\frac{1}{n}(\Pi_{ij} - 1 + 0.5)\right).$$



(a) $U_{ij} \sim U(0, 1)$

(b) $U_{ij} = 0.5$

Figure 3: Two Latin hypercube designs on a 2-dimensional experimental region, $n = 10$: (a) uses $U_{ij} \sim U(0, 1)$ and (b) uses $U_{ij} = 0.5$

Figure 3 shows two Latin hypercube designs that use the same $\mathbf{\Pi}$ but the one on the left uses $U_{ij} \sim U(0,1)$ and the one on the right uses $U_{ij} = 0.5$, for $i = 1, \ldots, n$; $j = 1, \ldots, d$.

1.1.3 Other Space-filling Designs

There are many space-filling designs in the numerical integration literature that very easily lend themselves to designs for computer experiments. Uniform designs, good lattice points, and nets are examples of such designs. Niederreiter (1992) and Koehler et al. (1996) provide a thorough discussion of these.

Another type of space-filling design that is popular in numerical integration is that of Sobol' sequences. These are easy to generate and can be used to construct sequential designs for computer experiments. A very useful property of Sobol' sequences is that a longer sequence can be created by merely adding points to a shorter sequence. This is an advantage over Latin hypercube designs, where the designs have to be reconstructed if we want to increase the design size and maintain the Latin hypercube nature of the design. The thesis of Marin (2005) contains a very detailed discussion of Sobol' sequences.

1.2 Model-based Design Criteria

1.2.1 Traditional Design of Experiments

When constructing models, especially the structure zone model, an empirical model mentioned in section 1.5, the experimenters were required to do hundreds of experiments in order to validate their models. When one incorporates a design of experiment (DOE) approach, it is possible to quantify more characteristics of the model with fewer experiments, using a statistical approach. DOE is generally used for empirical models. One runs enough experiments to obtain statistically significant empirical parameters for a model. Common methods used are 2^k factorial approach (where k is the number of factors or process variables in the experiment) , as well as the fractional factorial DOE [84, 140]. Disadvantages of the fractional factorial are less precision in the model parameter estimates, and confounding or masking of main effects with interaction effects. Factorial experiments are commonly used in research [11, 14, 25, 73].

Response surface modeling (RSM) is a method to predict the local shape of the response surface of a system [84]. It is used mainly for optimizing system settings and to make a system more robust. RSM is most useful when the system does not have a linear response between the high and low levels of a factor, i.e. the center point result does not equal the average of the results with high and low settings. To this end there are several approaches to DOE. Most popular is the face-centered central composite design (CCF). A central composite design is a factorial or fractional factorial experiment with center points and a group of star points to allow for estimation of nonlinearity. Another design is the Box-Behnken design. Unlike the CCF, this design preserves rotability but the estimation of points on the corners of the box are poor. RSM is used in many current research projects for modeling of batch processes [57, 107].

Another popular experimental design technique is the Taguchi method. The Taguchi method is best when one is trying to find the most robust operating point for a process [99]. Again, the focus here is on the result (i.e. a more robust process) rather than the knowledge about the process gained from the experiments. This method has found use in batch process modeling as well [83].

Other experimental design techniques have been used to create a better sampling scheme. Defining a regular grid on the experimental space and randomly picking points from that grid is called Latin Hypercube sampling (LHS) [40]. Alternatively, one can space the grid points irregularly based on spatial variation of the function or adaptively based on previous samples and an experimental design objective [120]. All of these sampling methods are designed for better sampling of the entire experimental region, whereas here we are interested in designing our experiments for best prediction at the unknown optimal point of a process.

Work that combines experimental design with mechanistic and empirical models has been largely limited to studies for speeding up simulation times. Specifically, the concept of surrogate models has been introduced to replace complex mechanistic models with simpler empirical models [105, 143].

1.2.2 Alphabetic Optimality Experimental Design

The traditional motivation underlying the theory of optimal design is that experiments should be designed to achieve the most precise statistical inference possible. Kiefer (1981) stated that research work on optimal design arose in part as a reaction to earlier research on design, which emphasized attractive combinatoric properties rather than inferential properties. Design optimality was first considered by Smith (1918), and early work in the subject was done by Wald (1943), Hotelling (1944), and Elfving (1952). The major contributions to the area, however, were made by Kiefer (1958, 1959) and Kiefer and Wolfowitz (1959, 1960), who synthesized and greatly extended the previous work. Although the ideas of optimal design initially generated considerable controversy (see, for example, the discussion accompanying the paper by Kiefer 1959), they have since become well established in the statistical literature. In some areas, such as the design of block experiments, the use of optimal design theory is now accepted as a fundamental tool for comparing designs (see Section 9). In other areas, however, there is still disagreement over the applicability of optimal design theory (see, for example, the discussion in Section 6 on response surface designs).

Excellent reviews of research work on optimal design have appeared. For readers interested in the most recent developments in optimal design, we recommend the reviews by Atkinson (1982), Pazman (1980), and Ash and Hedayat (1978). The review by St. John and Draper (1975) provides a good introduction to the topic.

The recent book by Silvey (1980) presents a concise summary of the classical results in optimal design theory, and the book by Fedorov (1972) is a valuable compendium of results.

The influence of optimal design has extended to almost all areas of experimental design, and it will be useful to review some of the most basic definitions and results because they will be needed in subsequent sections. To apply optimal design theory in practice requires a criterion for comparing experiments and an algorithm for optimizing the criterion over the set of possible experimental designs. We will define the most commonly used criteria here but will defer the consideration of algorithms to Section 4. The classical criteria are derived within the context of linear model theory in which it is assumed that the experimental data can be represented by the equation

$$y_i = f^T(\mathbf{x}_i)\boldsymbol{\beta} + \varepsilon_i, \tag{1.2.1}$$

where y_i is the measured response from the ith experimental run, $\mathbf{x}_i$ is a vector of predictor variables for the ith run, f is a vector of p functions that model how the response depends on $\mathbf{x}_i$, $\boldsymbol{\beta}$ is a vector of p unknown parameters, and ε_i is the experimental error for the ith run.

A natural way to measure the quality of statistical inference with respect to a single parameter is in terms of the variance of the parameter estimate. If the errors are uncorrelated and have constant variance σ^2, the variance-covariance matrix of the least squares estimator $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \tag{1.2.2}$$

where $\mathbf{X}$ is the $n \times p$ matrix whose ith row is $f^T(\mathbf{x}_i)$. We will limit our discussion here to the case where $\mathbf{X}$ has full column rank. Another useful way to measure the quality of inference is in terms of the variance of the estimated response at $\mathbf{x}$, which, from Eq. (1.2.1), is given by

$$\sigma^2(\mathbf{x}) = \sigma^2 f^T(\mathbf{x}_i)(\mathbf{X}'\mathbf{X})^{-1}f(\mathbf{x}_i), \tag{1.2.3}$$

Both Eq. (1.2.2) and Eq. (1.2.3) depend on the experimental design only through the $p \times p$ matrix $(\mathbf{X}'\mathbf{X})^{-1}$, and suggest that a good experimental design will be one that makes this matrix small in some sense. Since there is no unique size ordering of the $p \times p$ matrices, various real-valued functionals have been suggested as measures of "smallness." The most popular of these optimality criteria are listed below:

1. D-Optimality: A design is said to be D-optimal if it minimizes $\det(\mathbf{X}'\mathbf{X})^{-1}$,where det denotes determinant.

2. A-Optimality: A design is said to be A-optimal if it minimizes $tr(\mathbf{X}'\mathbf{X})^{-1}$, where tr denotes trace.

3. E-Optimality: A design is said to be E-optimal if it minimizes the maximal eigenvalue of $(\mathbf{X}'\mathbf{X})^{-1}$.

4. G-Optimality: A design is said to be G-optimal if it minimizes $\max \sigma^2(\mathbf{x})$, where the maximum is taken over all possible vectors $\mathbf{x}$ of predictor variables.

5. I-Optimality: A design is said to be I-optimal if it minimizes $\int \sigma^2(\mathbf{x})\lambda(d\mathbf{x})$, where λ is a probability measure on the space of predictor variables. This criterion, which is sometimes called average integrated variance, also belongs to a more general class of L-optimality criteria discussed by Fedorov (1972).

CHAPTER II

DOE FOR PASS/FAIL AND LINEAR REGRESSION WITH INEQUALITY CONSTRAINTS INCLUDING UNCERTAINTY

2.1 Introduction

The assumption that underlies most research work in experimental design is that the experiment can be adequately described by an equation of the form:

$$\text{response} = \text{model}(x) + \text{error}, \quad x \in R, \tag{2.1.1}$$

where the model states the effect of the input variables x on the response variable and the error describes the general form of departures from the model. The input variables are restricted in the design region R.

As Eq. (2.1.1) states, generally, uncertainties can be classified into structural (or model) uncertainty, parameter uncertainty, and stochasticity uncertainty. Many studies have attempt to solve those uncertainty using appropriate designs. Model robust designs ([5, 35], and therein) seek designs that will yield reasonable results for the proposed model structure even though it is known to be inexact. Optimal designs such as D-,$A-$optimal designs, obtain designs that will reduce uncertainties in parameter estimation ([35]). Error-robust designs concern the implications for experimental design of inaccurate assumption about the error (see [35] for more detail).

Besides those kinds of uncertainties, we are interested in uncertainty in design region R. To the best of our knowledge, region uncertainty has not been reported in the literature. To help understand the concept of region uncertainty, let us explain different types of design regions.

Many different types of irregular-shaped experimental region R have been concerned in the experimental designs. Typically, input variable, x , vary between a minimum and a maximum value so that

$$x_{i,min} \leq x_i \leq x_{i,max} \quad (i = 1, \ldots, p). \tag{2.1.2}$$

The values of the upper and lower limits $x_{i,min}$and $x_{i,max}$ depend upon the range of the factors thought by the engineers to be interesting. For example, if pressure is one of the factors, the engineers have knowledge that experimental range will be bounded by the maximum safe working pressure of the equipment. However, $x_{i,max}$ may be less than this value if such high pressures are not of interest. If the limits (Eq.(2.1.2)) apply independently to each of the p factors, the experimental region will be an p-dimensional cube. For $p = 2$ this is the square as shown in Figure 4(a).

The cube design region is the most frequently encountered for quantitative variables. However, the nature of the experiment may sometimes cause more complicated specification of the factor intervals and of the design region. For example, the region will be spherical if it is defined by the equation

$$\sum_{i=1}^{p} x_i^2 \leq R^2$$

where the radius of the sphere is R(see Figure 4(b)).

Figure 4(c) shows another example, mixture experiments, in which the response depends only on the proportions of the components of a mixture and not at all on the total amount. An important feature of such experiments is that a change in the level of one of the factors leads to a change in the values of one, some, or all of the other factors. The constraints

$$\sum_{i=1}^{q} x_i = 1 \qquad x_i \geq 0$$

imposed on the q mixture components make the design region a $(q - 1)$ dimensional simplex.

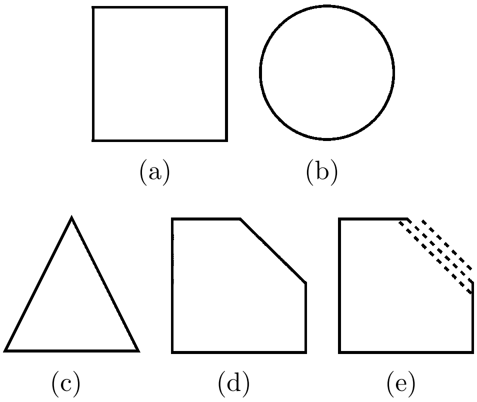


Figure 4: Some design regions of two factors x_1 and x_2 : (a) square(cubic for $p > 2$), (b) circular(spherical), (c) simplex for mixture experiments, (d) restricted to avoid simultaneous high values of x_1 and x_2(hard constraints), (e) restricted to avoid simultaneous high values of x_1 and x_2(soft constraints)

The experimental regions may be more irregular than previous examples, because of the imposition of extra constraints as shown in Figure 4(d). For such restricted experimental regions, standard designs may not be the best choice. Instead, optimal designs are appropriate for such situations.

Figure 4(a)-4(d) represent design regions which are assumed by traditional experimental design approaches such as factorial design, mixture design, and optimal design and so on. Figure 4(e) introduces a new kind of design region where an uncertain constraint is imposed on design space so that a part or all of the design region are not known with certainty. We call this type of constraint, a *soft constraint*, to distinguish it from usual ones. The soft constraint does not divide inside region and outside region clearly.

In Section 2, we explain the concept of soft constraints in detail and formulate research problems. We briefly review two-part models and derive optimal designs under the two-part model in Section 3. In Section 4 we consider several examples to verify our approaches. Finally, we conclude this chapter and discuss future work in Sections 5.

2.2 Focused Research Problem Formulation

To help understand the concept of soft constraints, Table 1 presents the average aspect ratio of of ZnO Nanowires recorded in [41]. “no growth” indicates failure in our discussion. Each run represents different experimental conditions. As Table 1 shows, the failure region is not clearly distinguishable from the success region, because there exists partially success and partially failure regions like run 2 and 8. Also, failure occurs more frequently in some design regions than the others. Soft constraints are defined to explain this kind of uncertain boundary between success regions and failure regions. This uncertainty is due to some uncontrollable or unknown noise factors during the growth process.

Table 1: The average aspect ratio of of ZnO Nanowires

run	trial 1	trial 2	trial 3
1	17.2	14.7	13.6
2	no growth	6.9	9.9
3	8.3	10.1	17.7
4	18.9	9.6	28.4
5	10.8	14.1	15.0
6	7.8	8.4	11.2
7	14.1	17.9	18.8
8	no growth	9.1	no growth

Here is another example of a soft constraint. A solubility model constraints the precursor amount. If the precursor amount does not exceed the maximum soluble amount (S), then all of the platinum precursor can be solubilized in the fluid phase; otherwise, the experimental setting is not appropriate for running an experiment. Precursor remaining in the solid phase will be wasted and more importantly will be detrimental to catalyst activity. The solubility of the precursor in sc-CO_2 has been measured and then modeled using the Chrastil model [6].

$$\ln(S) = k\ln(\rho(T,P)) + \frac{a}{T+b} \tag{2.2.1}$$

where S is the maximum soluble amount of precursor, T is the temperature, P is the pressure, and ρ is the density of the sc-CO$_2$. k, a, b are the adjustment parameters. Therefore, the precursor amount should not exceed S and this inequality constraint restricts the design space of T and P. However, the model (2.2.1) contains uncertainty due to model bias, parameter estimation, some uncontrollable or unknown noise factors, and so on. Because of those uncertainty, the soft constraint (2.2.1) is likely to produce similar results as shown in Table 1.

This chapter proposes optimal designs under a two-part model to handle the uncertainty in the design region. This approach may provide a solution for resolving a long-standing issue in optimal design. The use of optimal design theory in response surface studies has been criticized [4], because the optimal designs tend to place many design points at the extreme limits of the region. However, the boundary of the region is not precisely known in many engineering experiments (e.g., mechanical and chemical experiments). The proposed method uses the logit model in the two-part model to assess the uncertainty in engineering models. There has been an attempt to incorporate engineering models into statistical models as constraints [42]. However, our approach is different in that we applies a mixture model for pass/fail data in nanofabrication processes and focus on data collection scheme under uncertainty. Also, we showed how proposed method can be used to estimate inequality engineering constraint (unknown).

2.3 Methodology

2.3.1 *D*-optimal Design under Two-part Model

One considers optimum designs to achieve the most precise statistical inference possible for an underlying model. We assume that an observation $y_i(x_i)$ may be written

$$y_i = \mathbf{f}^T(x_i)\boldsymbol{\theta} + \epsilon_i, \quad i = 1, 2, \ldots, N, \tag{2.3.1}$$

where the x_i's are elements of a compact design space, $\mathscr{X}$, $\mathbf{f}$ is continuous on $\mathscr{X}$, and the ϵ_i's are uncorrelated random variables with mean zero and variance σ^2. Exact and approximate D-optimal designs maximize the determinant of the information matrix,

$$M(\xi) = \int_{\mathscr{X}} \mathbf{f}(x)\mathbf{f}^T(x)d\xi(x). \tag{2.3.2}$$

Suppose that a soft constraint $g_s(x) \leq \varepsilon$ is imposed on the design space $\mathscr{X}$. This constraint divides the design space into two regions: success region and failure region. Failure region is the design space where there is no yield in that region. To construct D-optimal design under the soft constraint is of interest. In order to model the two regions divided by the soft constraint, we use a two-part model. Two-part model have appeared in econometric analysis for nearly two decades [26, 30]. However, optimal design for a two-part model has not been studied much. Han [16] finds D-optimal designs under a two-part model analytically in a simple setting: one variable having two design points. We consider more a practical situation: several variables having multiple design points. Furthermore, we are more interested in the meaning of a two-part model in the experimental design.

A response variable might be a mixture of two or several random variables. y_i can be recoded as a mixture of two random variables, U and V.

$$U_i = \begin{cases} 1 & \text{if } y_i \neq 0 \\ 0 & \text{if } y_i = 0 \end{cases}$$

and

$$V_i = \begin{cases} g(y_i) & \text{if } y_i \neq 0 \\ \text{irrelevant} & \text{if } y_i = 0, \end{cases}$$

where g is a monotone increasing function (e.g., log) that will make V_i approximately Gaussian. This regards a response variable as the result of two processes, one determining whether the response is zero and the other determining the actual level if it is non-zero.

The observations from physical experiments, y_i are modeled by the two separate models: one for the logit of $P(U_i = 1)$ and one for the mean conditional response

$E(V_i|U_i = 1)$. The logit, η_i, can be modeled by

$$\eta(\mathbf{x}_i) = \mathbf{f}_1^T(\mathbf{x}_i)\boldsymbol{\theta}_1, \qquad (2.3.3)$$

where $\eta_i = \log \frac{P(U_i=1)}{1-P(U_i=1)}$, $\mathbf{x}_i$ is a vector of design points for the ith run, $\mathbf{f}_1$ is vector of p functions that model how η_i depends on $\mathbf{x}_i$, and $\boldsymbol{\theta}_1$ is a vector of p unknown logit model parameter.

The linear regression model for the continuous response is

$$V_i = \mathbf{f}_2^T(\mathbf{x}_i)\boldsymbol{\theta}_2 + \epsilon_i,$$

where $\mathbf{x}_i$ is a vector of design points for the ith run, $\mathbf{f}_2$ is a vector of q functions that model how the V_i depends on $\mathbf{x}_i$, $\boldsymbol{\theta}_2$ is a $q \times 1$ vector of linear regression model parameters that need to be estimated, and ϵ_i is the experimental error for the ith run ($\epsilon_i \sim N(0, \sigma^2)$).

Proposition 1. *The information matrix for a two-part model with logit model,* $\eta_i = \mathbf{f}_1^T(\mathbf{x}_i)\boldsymbol{\theta}_1$*, and linear regression model,* $V_i = \mathbf{f}_2^T(\mathbf{x}_i)\boldsymbol{\theta}_2 + \epsilon_i$*, is*

$$\mathbf{M}(\xi, \boldsymbol{\theta}_1) = \begin{pmatrix} \sum_i \frac{\exp(\eta_i)}{(1+\exp(\eta_i))^2}\mathbf{f}_1(\mathbf{x}_i)\mathbf{f}_1(\mathbf{x}_i)^T & \mathbf{0} \\ \mathbf{0} & \sum_i \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\mathbf{f}_2(\mathbf{x}_i)\mathbf{f}_2(\mathbf{x}_i)^T \end{pmatrix}, \qquad (2.3.4)$$

where ξ *is a design* $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$.

Proof. Appendix □

The information matrix of the two-part model(2.3.4) consider model structure of both logit model and linear regression model. Note that the information matrix (Eq. (2.3.4)) depends on η, which contains unknown parameters to be estimated. It means two-part model is a non-linear model which requires initial guess to construct optimal designs.

2.3.2 Property Investigation of the Logit Model

The logit model (2.3.3) plays a role to separate success region where $P(U = 1) = 1$ and failure region where $P(U = 1) = 0$ through $\eta_i = \log \frac{P(U_i=1)}{1-P(U_i=1)}$.

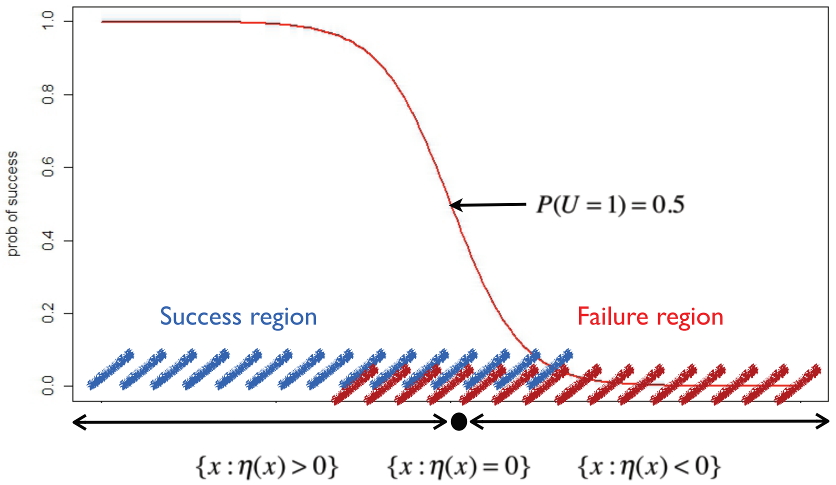


Figure 5: The logit model and success & failure regions

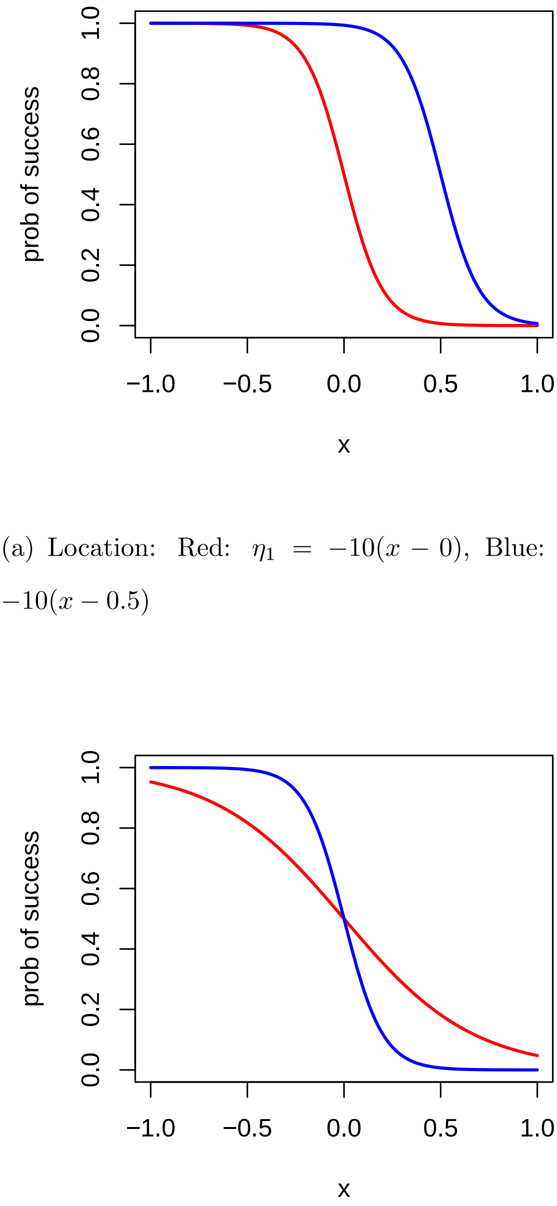
As Figure 5 illustrates, design space can be separated by the logit model into three spaces: $\{x : \eta(x) = 0\}$, $\{x : \eta(x) > 0\}$, and $\{x : \eta(x) < 0\}$. Let us call those three spaces as follows.

Definition 1. *Soft Boundary (SB) is* $SB = \{x : \eta(x) = 0\}$.

Definition 2. *Success Region (SR) is* $SR = \{x : \eta(x) > 0\}$.

Definition 3. *Failure Region (FR) is* $FR = \{x : \eta(x) < 0\}$.

Note that there are a region near SB, whose experimental results are sometimes success and sometimes fail. Let us call the region a mixed region and the probability of success in the mixed region is $(0 < P(U = 1) < 1)$. The mixed region is the consequence of soft constraints. We attempt to explain the mixed region using logit models. Here is an important insight on logit models.



(a) Location: Red: $\eta_1 = -10(x - 0)$, Blue: $\eta_2 = -10(x - 0.5)$

(b) Dispersion: Red: $\eta_3 = -3(x - 0)$, Blue: $\eta_1 = -10(x - 0)$

Figure 6: Location and Dispersion of the Logit model

As Figure 6 depicts, the mixed region can be defined by its location and dispersion and the estimate of logit models provides the location and dispersion information of the mixed region.

With respect to the location, SB can be used as the location indicator of the mixed region, because SB is particularly the region such that $P(U = 1) = 0.5$ as proved in Proposition 2. On the SB, the chance to be success is equal to the chance to be failure and so SB can be used as the location of the mixed region, in other words, the location of the soft constraint. For example, $\eta_1 = -10x$ and $\eta_2 = -10(x - 0.5)$ have 0.5 difference in SB as shown in Figure 6(a).

Proposition 2. *For $x_i \in SB$, $P(U_i = 1) = 0.5$.*

Proof. Since $x_i \in SB$,

$$\eta(x_i) = \mathbf{f}_1^T(x_i)\boldsymbol{\theta}_1 = \log \frac{P(U_i = 1)}{1 - P(U_i = 1)} = 0,$$

Thus, $P(U_i = 1) = 0.5$. □

With respect to dispersion, the dispersion of the mixed region can be measured by the steep of the $P(U = 1)$ line at SB. That is, $|\eta'(x_i)|$ for $x_i \in SB$. Larger value of $|\eta'(x_i)|$ indicates steeper line of $P(U = 1)$ on the points of SB, which means small dispersion. As shown in Figure 6(b), blue line is steeper than red line at $x = 0$ and the steep can be confirmed by $|\eta'(x_i)|$ by

$$|\eta_3'(1)| < |\eta_1'(1)|.$$

Then, we state that η_3 is more disperse than η_1.

Therefore, to estimate $\boldsymbol{\theta}_1$ in the logit model is equivalent to find soft constraints. Once the logit model is estimated by $\hat{\eta}(x_i) = \mathbf{f}_1^T(x_i)\hat{\boldsymbol{\theta}}_1$, the location of the mixed region is $SB = \{x : \hat{\eta}(x) = 0\}$ and its dispersion is $|\hat{\eta}'(x_i)|$ for $x_i \in SB$. That is, we can model location and dispersion of a soft constraint through the logit model in a two-part model.

2.3.3 Soft Constraint Estimation

By definition, D-optimal design, ξ^*, satisfies

$$|\mathbf{M}(\xi^*, \boldsymbol{\theta}_1)| = \max_{\xi} |\mathbf{M}(\xi, \boldsymbol{\theta}_1)|.$$

Since $\boldsymbol{\theta}_1$ is unknown in practice, initial guess of $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_1^{(0)}$, is required to find D-optimal design under a two-part model. Denote the D-optimal design with $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^{(0)}$ by $\xi^{*(0)}$. At design points $(\xi^{*(0)})$, we collect the observations from physical experiments $y(\xi^{*(0)})$ and recode it into U based on the rule,

$$U_i^{(0)} = \begin{cases} 1 & \text{if } y(\xi_i^{*(0)}) \neq 0 \\ 0 & \text{if } y(\xi_i^{*(0)}) = 0 \end{cases}.$$

Using $(\xi^{*(0)}, \mathbf{U}^{(0)})$ in the logit model, we can estimate $\boldsymbol{\theta}_1$. This sequential experiment keeps going as follows until sample size reaches the allowed size.

$$\boldsymbol{\theta}_1^{(0)} \to (\xi^{*(0)}, \mathbf{U}^{(0)}) \to \hat{\boldsymbol{\theta}}_1^{(1)} \to (\xi^{*(1)}, \mathbf{U}^{(1)}) \to \hat{\boldsymbol{\theta}}_1^{(2)} \to \cdots.$$

2.4 Illustrative Examples

The purpose of the illustrative examples is to justify the use of optimal designs with a two-part model when design regions include uncertainty, and to show some of characteristics of the design. Two illustrative examples show optimal designs under the two-part model and some preliminary results of the impact of the two-part model. Example 1 demonstrates the result of proposed designs for various design spaces restricted by soft constraints. In Example 2, the proposed method of soft constraint estimation is evaluated. To construct D-optimal design under a two-part model, we use modified Fedorov exchange algorithm [8], which is most commonly used in literature for optimal design construction.

2.4.1 Example 1

The purpose of Example 1 is to show that how proposed design is different from the usual D-optimal design. Let us consider two variables, x_1 and x_2, which define the design space as $[-1, 1] \times [-1, 1]$. D-optimal design under 2nd order polynomial regression model with $N = 8$ is shown in Figure 7. All design points are at the extreme limits of the region. This design is only valid under the assumption that the design space is surely defined $[-1, 1] \times [-1, 1]$.

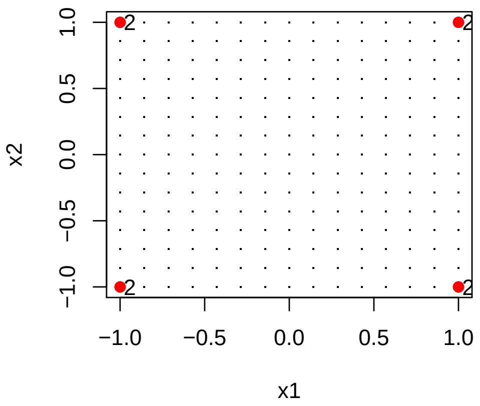


Figure 7: D-optimal design on under 2nd order polynomial model with $N = 8$. The number next to design points indicates the number of replications.

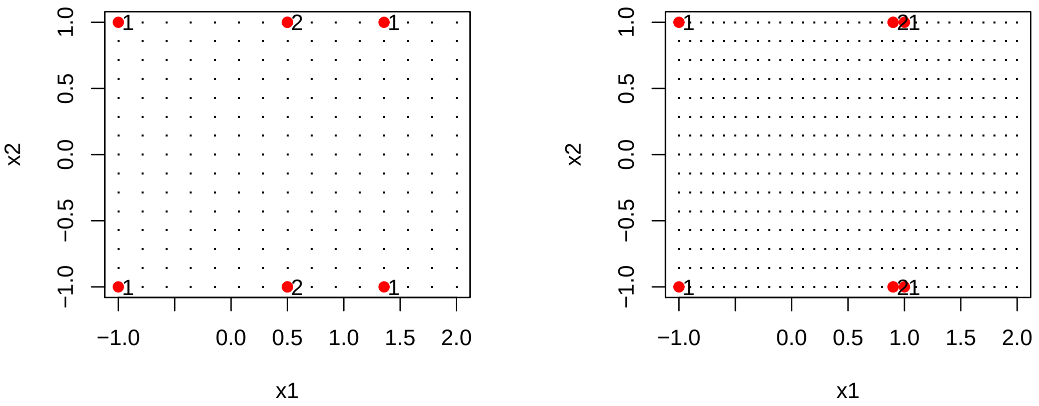
As mentioned before, to construct D-optimal design under a two-part model, an initial guess on the parameter of the logit model part is required. The initial guess $\boldsymbol{\theta}_1^{(0)}$ is assumed to be given by engineers. Let us compare two designs corresponding to two initial guesses,

$$(1) \qquad \eta = -3(x_1 - 1) \tag{2.4.1}$$

$$(2) \qquad \eta = -100(x_1 - 1). \tag{2.4.2}$$

Note that both guesses have the same SB at $x_1 = 1$, but the dispersions are different.

Eq. (2.4.1) is more disperse than Eq. (2.4.2), because the dispersion of Eq. (2.4.1) is $\frac{\partial}{\partial x_1}\eta(x)\big|_{x_1=1} = 3$, while the dispersion of Eq. (2.4.2) is 100.



(a) $\eta = -3(x_1 - 1)$

(b) $\eta = -100(x_1 - 1)$

Figure 8: D-optimal designs under a two-part model with different initial guess on $\boldsymbol{\theta}_1$

Figure 8 shows the results of D-optimal designs under a two-part model with different initial guesses. The points at $(-1, 1)$ and $(-1, 1)$ are same as the D-optimal design in Figure 7, but the positions of other points are quite different. First, the proposed designs reflect the existence of soft constraint in the design space. Secondly, two proposed designs in Figure 8(a) and 8(b) are different depending on the initial guesses. Since Eq. (2.4.1) requested more dispersion than Eq. (2.4.2), the design in Figure 8(a) spreads out its design points near SB, $x_1 = 1$, more than design in Figure 8(b).

The proposed method, D-optimal design under a two-part model, works well for any kind of initial guess other than lines. Engineers may guess more complicated soft constraints like

$$(3) \quad \eta = -10x_1 - 5x_1^2 - 5x_2^2. \tag{2.4.3}$$

The success region corresponding to Eq (2.4.3) is

$$\{(x_1, x_2) : 5(x_1 + 1)^2 + 5x_2^2 < 5^2\},$$

and Figure 9 shows the D-optimal design based on the initial guess (2.4.3).

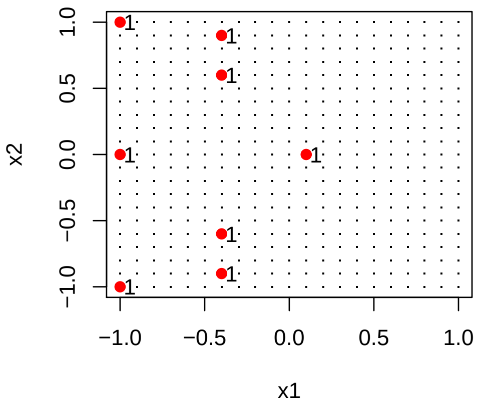


Figure 9: D-optimal design based on the initial guess $\eta = -10x_1 - 5x_1^2 - 5x_2^2$ with $N = 8$

2.4.2 Example 2

In this example, we demonstrate the proposed method of soft constraint estimation. One of the main advantages of D-optimal design under a two-part model is that the design is able to find the soft constraint sequentially.

Let us consider two variables, x_1 and x_2, which are defined in the design space $[-1, 2] \times [-1, 1]$. Suppose that a true logit model is

$$\eta(x_1, x_2) = 50 - 50x_1 - 25x_2, \tag{2.4.4}$$

which is unknown. That is, the true parameter of the logit model is

$$\boldsymbol{\theta}_1^* = (50, -50, -25).$$

In other words, there is a soft constraint whose SB is $\{(x_1, x_2) : 50 - 50x_1 - 25x_2 = 0\}$ in the design space. Blue dash lines in Figure 10 represent SB from the true logit model (2.4.4). For linear regression model part in the two-part model, 2nd order polynomial regression model is assumed.

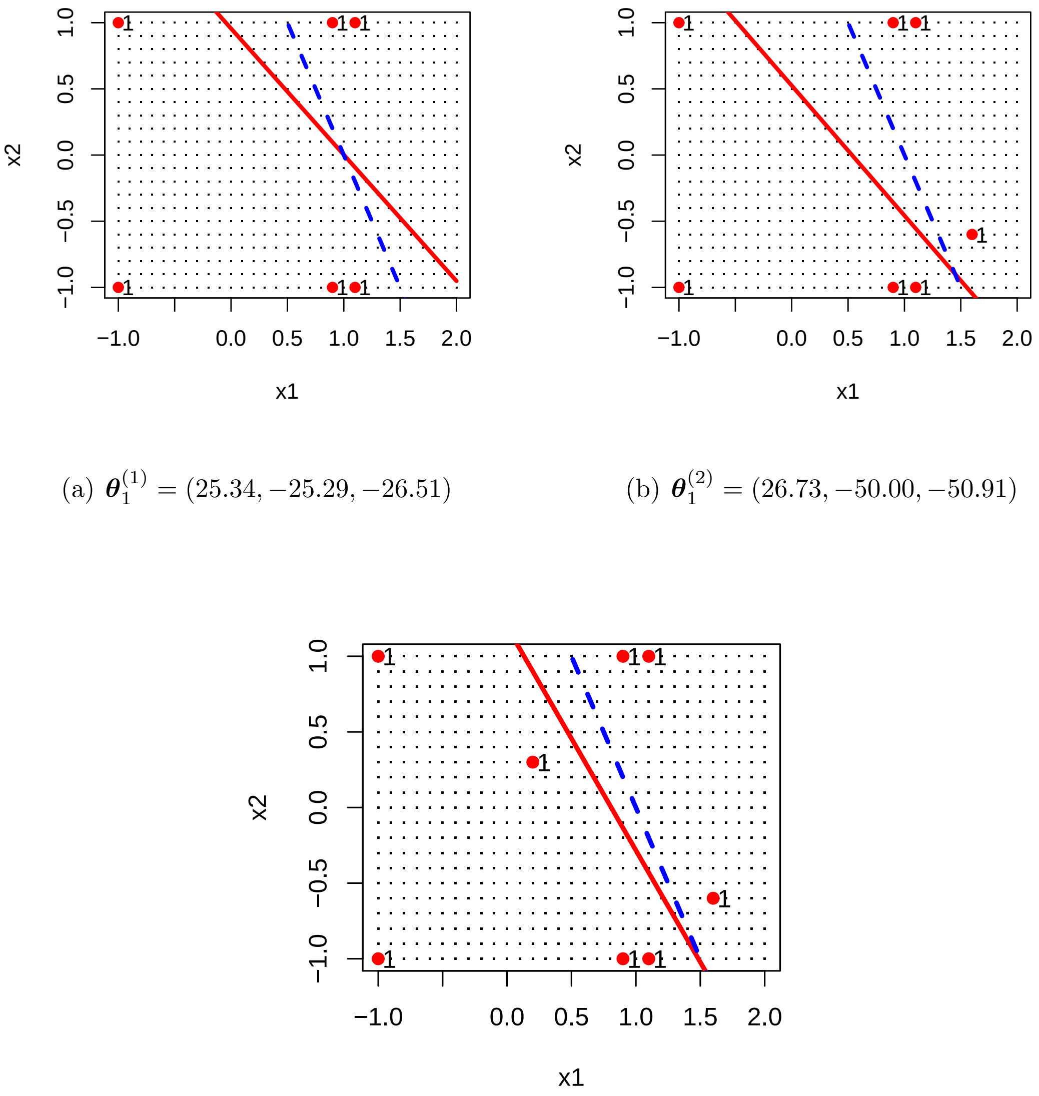
Now, we want to construct initial D-optimal design under a two-part model with six initial sample size. As mentioned before, design construction requires initial guess of $\boldsymbol{\theta}_1$. Initial guess of the logit model part is given by engineers as

$$\boldsymbol{\theta}_1^{(0)} = (20, -20, 0)$$

Figure 10(a) shows the constructed design ($\xi^{*(0)}$) using given $\boldsymbol{\theta}_1^{(0)}$. Using the true relationship 2.4.4, we obtain $\eta(\xi^{*(0)})$ and corresponding $U^{(0)}$. Using $(\xi^{*(0)}, U^{(0)})$ in the logit model, we can estimate $\boldsymbol{\theta}_1$, which is depicted by red line in Figure 10(a). This sequential experiments keep going as follows until sample size reaches the allowed size $N = 8$.

$$\begin{aligned} \boldsymbol{\theta}_1^{(0)} &= (20, -20, 0) \\ &\downarrow \\ \hat{\boldsymbol{\theta}}_1^{(1)} &= (25.34, -25.29, -26.51) \\ &\downarrow \\ \hat{\boldsymbol{\theta}}_1^{(2)} &= (26.73, -50.00, -50.91) \\ &\downarrow \\ \hat{\boldsymbol{\theta}}_1^{(3)} &= (49.63, -61.36, -41.51) \end{aligned}$$

As shown in Figure 10, the proposed method finds design points sequentially so as to estimate the parameter in the logit model. The SB lines corresponding to $\hat{\boldsymbol{\theta}}_1$ get close to the true line.



(a) $\boldsymbol{\theta}_1^{(1)} = (25.34, -25.29, -26.51)$

(b) $\boldsymbol{\theta}_1^{(2)} = (26.73, -50.00, -50.91)$

(c) $\boldsymbol{\theta}_1^{(3)} = (49.63, -61.36, -41.51)$

Figure 10: Sequential experiments for soft constraint estimation. Blue dash lines represent the true line and red solid liens are estimated soft constraints

2.5 Conclusion

Design of experiments assumes predetermined design regions. Design regions with uncertainty are of interest in this chapter. If a design region has uncertainty on its boundary, any data collection plan depends on the location of the design region boundary is not appropriate, for example, optimal designs. We observed uncertainty in a design region when a design region is restricted by inequality engineering constraints like solubility constraints. This results from the fact that engineering models based on simplified assumptions are subject to be biased and ignore noise variables, measurement errors, and so on. Thus, we propose optimal designs under a two-part model to tackle the uncertainty in the design region. We derived the information matrix of the two-part model and constructed optimal designs. Through several examples, we showed how two-part models explain uncertainty in design regions and can be used for soft constraint estimation.

Proof of Proposition 1 The information matrix for a two-part model with logit model, $\eta_i = \mathbf{f}_1^T(\mathbf{x}_i)\boldsymbol{\theta}_1$, and linear regression model, $V_i = \mathbf{f}_2^T(\mathbf{x}_i)\boldsymbol{\theta}_2 + \epsilon_i$, is

$$\mathbf{M}(\xi, \boldsymbol{\theta}_1) = \begin{pmatrix} \sum_d \frac{\exp(\eta_i)}{(1+\exp(\eta_i))^2}\mathbf{f}_1(\mathbf{x}_i)\mathbf{f}_1(\mathbf{x}_i)^T & \mathbf{0} \\ \mathbf{0} & \sum_d \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\mathbf{f}_2(\mathbf{x}_i)\mathbf{f}_2(\mathbf{x}_i)^T \end{pmatrix},$$

where ξ is a design $\{\mathbf{x_1}, \mathbf{x_2}, \ldots\}$.

Proof. Let $d_0 = \{i : y_i = 0\}$ and $d_1 = \{i : y_i \neq 0\}$, n_0 and n_1 be the number of elements in d_0 and d_1, respectively, and $d = d_0 \cup d_1$. The likelihood function L for the two-part model is

$$L \propto \prod_{d_1}\left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}L_{V_i}\right)\prod_{d_0}\left(\frac{1}{1+\exp(\eta_i)}\right),$$

where

$$L_{V_i} \propto \exp\left(-\frac{(V - \mathbf{f}_2^T(\mathbf{x}_i)\boldsymbol{\theta}_2)^2}{2}\right)$$

comes from the linear regression.

Then, the log-likelihood function, l, is

$$\begin{aligned} l &\propto \sum_{d_1}\log\left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right) + \sum_{d_1}\log L_{V_i} - \sum_{d_0}\log(1+\exp(\eta_i)) \\ &\propto \sum_{d_1}\eta_i + \sum_{d_1}\log L_{V_i} - \sum_{d}\log(1+\exp(\eta_i)) \end{aligned}$$

To calculate the information matrix,

$$\begin{aligned} \frac{\partial l}{\partial \theta_1} &= \frac{\partial}{\partial \theta_1}\left(\sum_{d_1}\eta_i - \sum_d \log(1+\exp(\eta_i))\right) \\ &= \sum_{d_1}\mathbf{f}_1(\mathbf{x}_i) - \sum_d \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\mathbf{f}_1(\mathbf{x}_i) \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \theta_1 \partial \theta_1} &= -\sum_d \mathbf{f}_1(\mathbf{x}_i)\frac{\partial}{\partial \theta_1}\left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right) \\ &= -\sum_d \mathbf{f}_1(\mathbf{x}_i)\left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)} - \frac{\exp(\eta_i)\exp(\eta_i)}{(1+\exp(\eta_i))^2}\right)\mathbf{f}_1(\mathbf{x}_i)^T \\ &= -\sum_d \left(\frac{\exp(\eta_i)}{(1+\exp(\eta_i))^2}\right)\mathbf{f}_1(\mathbf{x}_i)\mathbf{f}_1(\mathbf{x}_i)^T \end{aligned}$$

$$\begin{aligned}\frac{\partial l}{\partial \theta_2} &= \frac{\partial}{\partial \theta_2}\sum_{d_1}\left(\frac{(V_i - \mathbf{f}_2(\mathbf{x}_i)^T\boldsymbol{\theta}_2)^2}{2}\right) \\ &= \sum_{d_1}\left(V_i - \mathbf{f}_2(\mathbf{x}_i)^T\boldsymbol{\theta}_2\right)\mathbf{f}_2(\mathbf{x}_i)\end{aligned}$$

$$\frac{\partial^2 l}{\partial \theta_2 \partial \theta_2} = -\sum_{d_1}\mathbf{f}_2(\mathbf{x}_i)\mathbf{f}_2(\mathbf{x}_i)^T$$

Since

$$\frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} = \frac{\partial^2 l}{\partial \theta_2 \partial \theta_1} = 0$$

and

$$\begin{aligned}E\left[\sum_{d_1}\mathbf{f}_2(\mathbf{x}_i)\mathbf{f}_2(\mathbf{x}_i)^T\right] &= E\left[\sum_{d}\mathbf{f}_2(\mathbf{x}_i)\mathbf{f}_2(\mathbf{x}_i)^T\cdot I(i\in d_1)\right] \\ &= \sum_{d}\mathbf{f}_2(\mathbf{x}_i)\mathbf{f}_2(\mathbf{x}_i)^T\cdot E\left[I(i\in d_1)\right] \\ &= \sum_{d}\mathbf{f}_2(\mathbf{x}_i)\mathbf{f}_2(\mathbf{x}_i)^T\cdot P(i\in d_1) \\ &= \sum_{d}\mathbf{f}_2(\mathbf{x}_i)\mathbf{f}_2(\mathbf{x}_i)^T\cdot\frac{\exp\eta_i}{1+\exp(\eta_i)}.\end{aligned}$$

$$\mathbf{M}(\xi,\boldsymbol{\theta}_1) = \begin{pmatrix}\sum_d \frac{\exp(\eta_i)}{(1+\exp(\eta_i))^2}\mathbf{f}_1(\mathbf{x}_i)\mathbf{f}_1(\mathbf{x}_i)^T & \mathbf{0} \\ \mathbf{0} & \sum_d \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\mathbf{f}_2(\mathbf{x}_i)\mathbf{f}_2(\mathbf{x}_i)^T\end{pmatrix}$$

□

CHAPTER III

LAYERS OF EXPERIMENTS FOR BUILDING STATISTICAL MODELS

3.1 Introduction

Statistical modeling for nano-fabrication process to achieve consistently good performance has become increasingly important [25]. Although statistical quality control and engineering-driven statistical analysis of traditional manufacturing processes have achieved great success in yield and productivity improvement [40, 33, 28], the nanomanufacturing dealing with nanoparticles in nanometer-scale results in new challenges for quality control and statistical analysis.

In most existing literature, the synthesis of nanomaterials are lack of theoretical guidance for achieving high quality and reproducible nanomaterials [9]. Yet, statistical modeling is required to use a few experiment runs due to high material costs and processing time for physical experiments in nanomanufacturing [1]. At the same time, the model is required to meet a tight tolerance requirement so that the model should achieve high level of prediction accuracy. Unfortunately, these two requirement, few experiment runs and tight tolerance requirement, conflict with each other.

We overcome theses difficulties by Layers of Experiments. We build a statistical model on the region of interest and obtain subregion where we predict the optimum lies. On the smaller region we build a more accurate model and continue the search. The new process improvement methodology for nanomanufacturing is referred to as Layers of Experiments. A layer denotes a region of interest and the next layer is the design regions that one needs to conduct the next batch of experiments to improve model quality and to access the process optimum.

Sequential experiments on the smaller design space have been studied by some researchers. Bernardo et al (1992) [3] proposed multistage experiments, a sequential strategy for optimizing integrated circuit design in computer experiments manner. However, this approach is for computer experiments rather than physical experiments. So, the motivation of multistage experiments is similar to layers of experiments in that both consider reduced subregions, but strategies for design and modeling are different. Wissmann and Grover (2009) [39] developed grid algorithm to reduce the design region of interest using confidence intervals which contain the true optimum of the model. They attempt to improve model prediction around the optimal points. The grid algorithm selects one next design point for each iteration using the most probable model. The most probable model is decided by mean squared error among many candidate models at each iteration. However, the grid algorithm is lack of statistical analysis as well as it's focus is to find optimal condition rather than build a statistical model with good statistical inference.

This chapter proposes an efficient and effective multi-layer data collection scheme (Layers of Experiments) for building accurate statistical models to meet tight tolerance requirement commonly encountered in nano-fabrication. We developed a method to decide the location and size of sub-regions (layers) using resampling techniques. An evaluation metric is introduced to measure the performance of statistical models for nano-fabrication quality prediction and the metric is used to decide whether further layers are needed. Moreover, this chapter also discusses appropriate types of designs for each layer, e.g. space-filling designs or optimal designs.

In Section 2, we formulate the problem and overview the layers of experiments and then we propose new methodology for the focused problems in Section 3. Results are presented for illustrative examples in Section 4.

3.2 Brief Overview of Layer-of-Experiments and Focused Research Problem Formulation

Briefly, layers of experiments have the following six steps. Step 1) Postulate a tentative statistical model. Step 2) Plan an experiment and collect the data. Step 3) Use the data to fit the models. Step 4) Check the accuracy of prediction. Step 5) If the model are insufficiently accurate, choose a subregion for the next experiment and return to step 1. Step 6) When the models are sufficiently accurate, optimize the objective (loss, yield, etc) using the fitted model in place of the performance functions. Figure 11 illustrates each procedure.

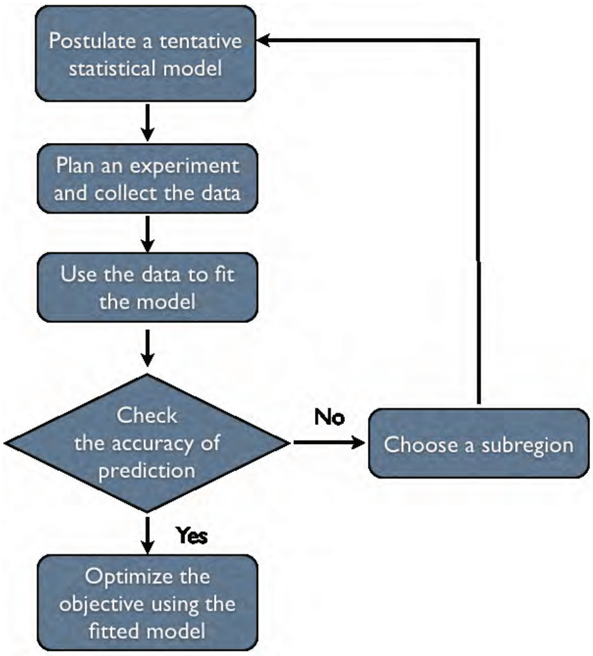

Postulate a tentative statistical model
Plan an experiment and collect the data
Use the data to fit the model
Check the accuracy of prediction
No
Choose a subregion
Yes
Optimize the objective using the fitted model

Figure 11: Flowchart for Layers of Experiments

In this chapter, based on the flowchart, we develop methodologies for each steps to solve our own problem for nano-fabrication process: building accurate statistical models to meet tight tolerance requirement with limited run sizes.

The key part of these approaches is step 5 that choose a subregion for the next experiment, which is also main difference from sequential experiments. Focused research problems in this chapter are shown in Figure 12.

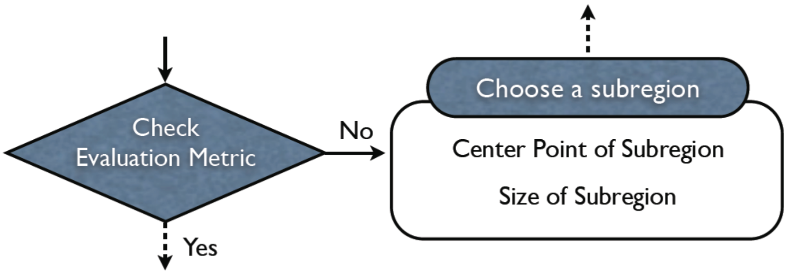


Figure 12: Focused research problems

First, we develop an evaluation metric to check whether estimated statistical model meets tolerance requirement near the process optimum. If the model does not meet the requirement, we choose a subregion for further experiments. Second focused problem is to decide the subregion (window) for the next experiment. To maximize efficiency, next experiments should be carried out in the refined subregion. Refining input design space requires decision procedures whether a certain region is important with respect to the purpose of experiment or not. There exists the risk of wrong decision that may cause severe inefficiency. Thus, the location and size of subregions are very important.

3.3 Methodology

3.3.1 Uncertainty Measurement (Evaluation Metric)

Researchers often categorize uncertainty in experiments into three components: structural, parameter, and stochasticity [34]. Structural uncertainty refers to uncertainty due to lack of knowledge about the correct model. Parameter uncertainty is associated with the uncertainty introduced by having to use values of model parameters that are not known with certainty. Finally stochasticity occurs when parameters or other quantities are not fixed but may vary. Stochasticity is generally viewed as uncertainty that is not reducible, while structural and parameter uncertainty are viewed as reducible (at least in principle) as more information is gathered. For example, in multiple regression analysis structural uncertainty is associated with model misspecification. Model evaluation assumes a certain general structure (e.g. multiple linear) and the model is built through adding terms which are significant or which aid in prediction. Parameter uncertainty is typically discussed as a problem of estimation. Stochasticity is commonly dealt with as a measurement problem, under the assumption of an additive error.

The problem of structural uncertainty has been focused less than parameter and stochastic uncertainty. One of the most common uncertainty measurement in statistical modeling is confidence interval. However, it only measure parameter and stochastic uncertainty, in other words, precision rather than accuracy (see Figure 13). However, once a model misspecified, model accuracy could be very poor due to structural uncertainty from model misspecification. Hence, we propose a new uncertainty measurement, which consider structural, parameter, and stochasticity uncertainty.

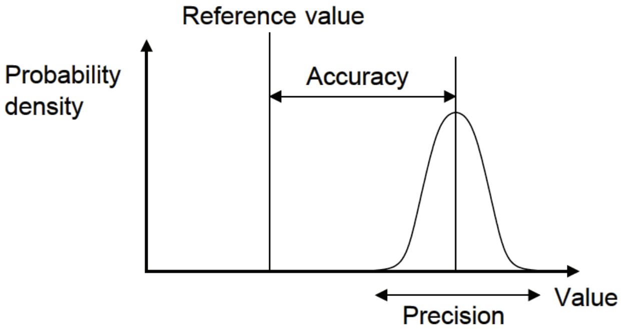


Figure 13: The concept of accuracy and precision (figure from Wikipedia)

We assume that an observation $y_i(x_i)$ may be written

$$y_i = \mathbf{f}^T(x_i)\boldsymbol{\theta} + \varepsilon_i, \quad i = 1, 2, \ldots, N, \tag{3.3.1}$$

where the x_i's are elements of a compact design space, $\mathscr{X}$, $\mathbf{f}$ is continuous on $\mathscr{X}$, and the ε_i's are uncorrelated random variables with mean zero and variance σ^2. Letting

$\mathbf{f}^T(x_i)$ denote the $(p \times 1)$ ith row of a matrix $\mathbf{X}$, Eq. (3.3.1) may be rewritten

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}.$$

The prediction variance $\sigma^2(x)$ of $\hat{y}(x)$ is

$$\sigma^2(x) = \hat{\sigma}^2 \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{x} \tag{3.3.2}$$

where $\mathbf{x}' = [f_1, \ldots, f_p]$ is a vector of p real valued functions of x based on the model terms. The model variance $\hat{\sigma}^2$ used in Eq. (3.3.2) can be estimated as follows.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(y(x_i) - \hat{y}(x_i))^2}{N-p}.$$

The model variance $\sigma^2(x)$ is used to calculate the confidence interval on $\hat{y}(x)$

$$CI(x) = \pm t_{\alpha/2,n-p}\sqrt{\sigma^2(x)}, \tag{3.3.3}$$

where α is the level of confidence desired.

Then, given the target value (T) (or reference value) of the output yield and tolerance requirement (d), we want to minimize the fluctuation around T caused by model uncertainty. This suggests maximum distance of $(1-\alpha)\%$ confidence interval of $\hat{y}$ from T

$$P\left\{|\hat{y}(x) - T| \geq d\right\} \leq \alpha, \tag{3.3.4}$$

where $\hat{y}(x)$ is a fitted model. Based on Eq. (3.3.4), the evaluation metric is defined by

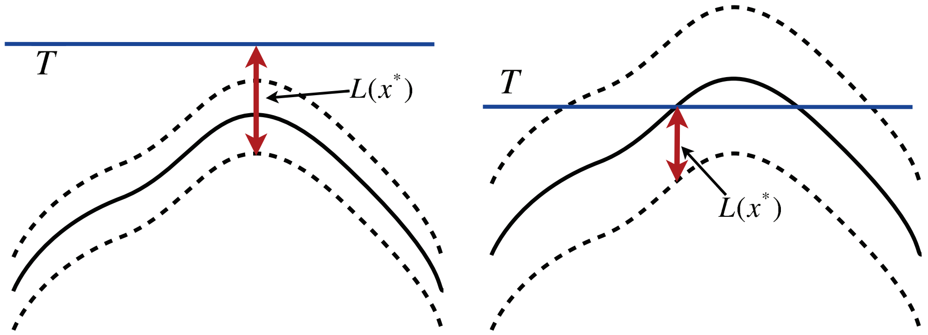
$$L(x) = \max\left\{|T - (\hat{y}(x) + CI(x))|, |T - (\hat{y}(x) - CI(x))|\right\}, \tag{3.3.5}$$

Then, we can find x^* such that

$$x^* = \arg\min_{x} L(x)$$

and let L^* denotes $L(x^*)$.

Figure 14 illustrates two different cases of evaluation metric at x^*: 1) when mean response of fitted model does not reach the target (T) and 2) when it reaches the target.



(a) when mean response of fitted model does not reach the target

(b) when mean response of fitted model reaches the target

Figure 14: Evaluation Metric $L(x)$: dash lines are $(1-\alpha)\%$ confidence intervals of $\hat{y}(x)$

Formally, the problem is to build a statistical model so that

$$L^* \leq d. \tag{3.3.6}$$

3.3.2 Subregion Decision for the Next Experiment

Choosing subregion for the next experiment is the key part in layers of experiments. Let L_U be a upper layer and L_L be a lower layer (subregion). There are two steps in subregion decision: 1) find the center of the new subregion (location), and 2) choosing new limits of the subregion (size).

3.3.2.1 The center of new subregion

One naive way to decide center point(c) is

$$c = \arg\min_{x} |T - \hat{y}(x)|.$$

However, this approach may not appropriate under the lack of knowledge about the correct model. Also, large stochasticity uncertainty is problematic.

Thus, we resample residuals using bootstrapping to generate the distribution of the center point (c). The method proceeds as follows.

1. Fit a model and retain the fitted values $\hat{y}_i$ and the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$ for $i = 1, \dots, n$.

2. For each pair, (x_i, y_i), in which x_i is the (possibly multivariate) explanatory variable, add a randomly resampled residual, $\hat{\varepsilon}_j$, to the response variable y_i. In other words create synthetic response variables $y_i^* = \hat{y}_i + \hat{\varepsilon}_j$ where j is selected randomly from the list $(1, \dots, n)$ for every i.

3. Fit candidate models using the fictitious response variables y_i^*, and retain the quantities of interest: the center point ($\tilde{c}$). Suppose that there are several candidate models M_k, $k = 1, \dots, m$, and denote $\tilde{c}^{(k)}$ by a center point using (M_k). A priori estimates $P(M_k)$ are available for each model,
$$\sum_{k=1}^{m} P(M_k) = 1.$$

4. Repeat steps 2 and 3 for a statistically significant number of times (N). Let n_k denote the number of samples from a candidate model $P(M_k)$. Then, for total N samples,
$$\begin{aligned} \sum_{k=1}^{m} n_k &= N, \\ n_k &\simeq N \cdot P(M_k). \end{aligned}$$
The center of new subregion (c^*) should be the mean of most frequent bin in histogram of $\tilde{c}$ s.

Suppose that we obtain $\tilde{c}_{l_k}^{(k)}$, for $l = 1, \dots, n_k$ $k = 1, \dots, m$, under m candidate models and the histogram of $\tilde{c}_{l_k}^{(k)}$ is shown in Figure 15. The m candidate models $(M_1, M_2, \dots, M_m)$ are different polynomial regression models where

$$P(M_1) + P(M_2) + \cdots + P(M_m) = 1. \tag{3.3.7}$$

The center of new subregion (c^*) is the mean of most frequent bin in histogram of $\tilde{c}_{l_k}^{(k)}$.

For example, Figure 15 illustrates the histogram of $\tilde{c}_{l_k}^{(k)}$ in a design space $[-3, 8]$ using three candidate models: 2nd, 3rd, 4th order polynomial regression models with $n_1 = n_2 = n_3 = 100$. From the histogram, we can compute $c^* = 2.3$. With a few data points, the proposed method finds the center of subregion efficiently. This approach reflects model uncertainty by considering multiple candidate models and also measurement error by considering resampling techniques.

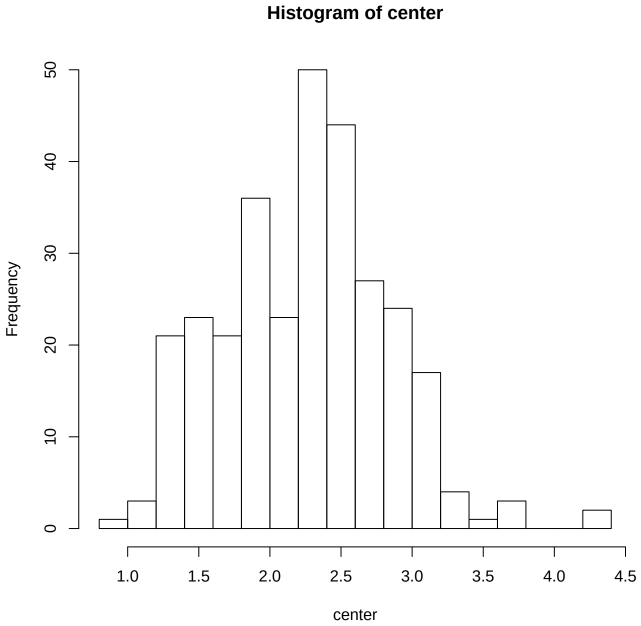


Figure 15: Histogram of $\tilde{c}_{l_k}^{(k)}$ in the design space $[-3, 8]$, where $k = 1, 2, 3$ and $n_1 = n_2 = n_3 = 100$. $c^* = 2.3$

3.3.2.2 The size of new subregion

Next, we should decide the size of the a subregion. If the size of a subregion is too small, it may miss true optimum, while too large size of subregion may cause inefficiency. Thus, we want to select the optimal size in some sense, rather than arbitrary.

Figure 15 illustrates a histogram of all possible center points in the subregion using proposed resampling technique. This is an important indicator for subregion size decision. Given a center point c^*, larger size of a subregion can cover more possible centers. The probability of success for subregion to include true optimum with respect to size r is defined by

$$\psi(r) = \frac{1}{N}\sum_{k=1}^{m}\sum_{l=1}^{n_k} I(c^* - r \leq \tilde{c}_{l_k}^{(k)} \leq c^* + r). \tag{3.3.8}$$

Then, we compute the size of new subregion (r^*) by

$$r^* = \min\{r : \psi(r) \geq 0.95\}.$$

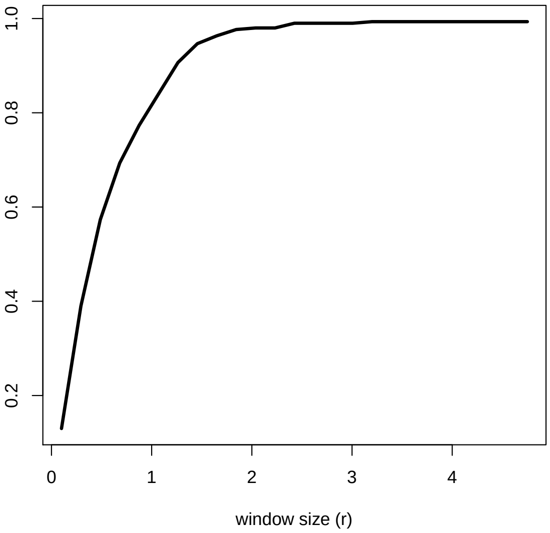


Figure 16: Plot of r against $\psi(r)$

For example, Figure 16 illustrates $\psi(r)$ with respect to subregion size r. Note that $\psi(r)$ is a increasing function of r and $0 \leq \psi(r) \leq 1$, $\psi(r) \geq 0.95$ for $r \geq 1.9$. Therefore, the size of subregion is $r^* = 1.9$ and subregion $[2.3 - 1.9, 2.3 + 1.9]$ covers 95% possible center points under three candidate models, even though the size of design space reduced from 11.0 to 3.8.

3.3.3 Zoom-out procedure

Note that the proposed subregion decision method does not always reduce the size of design space. That is, upper layer is not necessarily smaller than lower layer in its size, nor is lower layer a subset of upper layer. If many $\tilde{c}$ locate outside of upper layer, then we should zoom-out the design space to include possible true optimums.

3.4 Property Investigation

The goal of the experiment is to build a statistical model such that

$$L^* \leq d.$$

To achieve this goal, we take three strategies: 1) sample size increasing, 2) optimal design, and 3) design space zoom-in. First two strategies are commonly used in physical experiments. Third one is related to layers of experiments, which is the method we propose. In order to see the impact of those strategies, we decompose L^* into two part: confidence interval part and mean part.

With respect to the confidence interval, we focus the term inside of square root of confidence interval (3.3.3),

$$\mathbf{x}'(\mathbf{X}'\mathbf{X})^{\mathbf{-1}}\mathbf{x}. \tag{3.4.1}$$

Since $(\mathbf{X}'\mathbf{X})^{\mathbf{-1}}$ is determined by design points, Eq. (3.4.1) is rewritten using information matrix $\mathbf{M}$ with a design ξ as follows.

$$\mathbf{x}'\mathbf{M}^{\mathbf{-1}}(\xi)\mathbf{x}. \tag{3.4.2}$$

Here, $\mathbf{x}' = [f_1, \ldots, f_p]$ is a vector of p real valued functions of a specific x of interest based on the model terms. For us, x is process optimum that exists somewhere in the current design space R. However, process optimum is unknown, so we modify Eq. (3.4.2) to

$$\gamma(\xi) = \frac{1}{n_g}\sum_{l=1}^{n_g} \mathbf{x}_l'\mathbf{M}^{\mathbf{-1}}(\xi)\mathbf{x}_l, \tag{3.4.3}$$

where $\xi \in R$ and $\mathbf{x}_1, \ldots, \mathbf{x}_{n_g}$ represent vectors at n_g evenly spaced grid points over the design space R.

First, it is obvious that $\gamma(\xi)$ decreases as sample size increases. Second, under the assumption that the model structure $(f_1, \ldots, f_p)$ is known, we can find a design that minimize $\gamma(\xi)$. Among many alphabetical optimal designs, I-optimal design is the one. As Figure 17 shows, the kind of design used in the experiments affects $\gamma(\xi)$.

However, the result in Figure 17 does not mean that the optimal design is always better than others because we do not know true underlying model structure in practice. Optimal designs are known that they are dependent on the model assumption much.

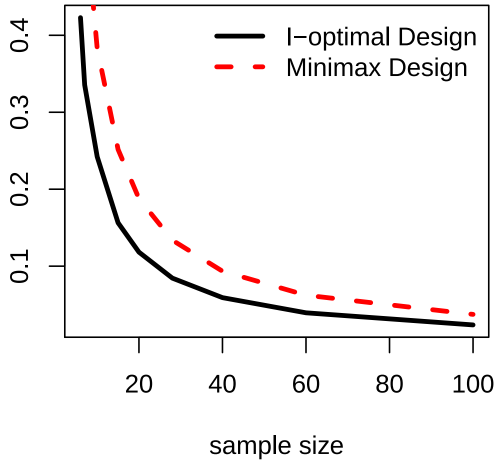


Figure 17: Performance of $\gamma(\xi)$ as sample size increases. one dimensional second-order polynomial model is considered in $[0, 1]$

Proposition 3 says the third strategy does not help on confidence interval reduction. That is, same sample size and same design type obtain same value in $\gamma(\xi)$ regardless the size of design space R. This implies layers of experiments have impact more on mean part.

Proposition 3. *If a design $\xi_B \in R_B$ can be expressed by another design $\xi_A \in R_A$ by $\xi_B = a\xi_A + b$ for $a, b \in \mathbb{R}$, then*

$$\gamma(\xi_A) = \gamma(\xi_B)$$

Proof. Since $\tilde{\xi}_B = \xi_B - (a\bar{\xi}_A + b) = a(\xi_A - \bar{\xi}_A)$, it is enough to consider the case $b = 0$. Let $\mathbf{x}' = [f_1, \ldots, f_p]$ be a vector of p real valued functions of x based on the model terms. For a design $\xi = \{x_1, x_2, \dots, x_n\}$, a $n \times p$ matrix $\mathbf{X}_\xi$ is defined by

$$\mathbf{X}_\xi = \begin{pmatrix} \mathbf{x'_1} \\ \mathbf{x'_2} \\ \vdots \\ \mathbf{x'_n} \end{pmatrix}.$$

For $\xi_A = \{x_{1A}, x_{2A}, \dots, x_{nA}\}$ and $\xi_B = \{x_{1B}, x_{2B}, \dots, x_{nB}\}$

$$\begin{aligned} \mathbf{X}_{\xi_\mathbf{B}} &= \begin{pmatrix} \mathbf{x'_{1B}} \\ \mathbf{x'_{2B}} \\ \vdots \\ \mathbf{x'_{nB}} \end{pmatrix} = \begin{pmatrix} \mathbf{x'_{1A}} \\ \mathbf{x'_{2A}} \\ \vdots \\ \mathbf{x'_{nA}} \end{pmatrix} \mathbf{diag}\left(f_1(a), f_2(a), \cdots, f_p(a)\right) \\ &= \mathbf{X}_{\xi_\mathbf{A}} \mathbf{diag}\left(f_1(a), f_2(a), \cdots, f_p(a)\right) \end{aligned}$$

and

$$\mathbf{M^{-1}}(\xi_\mathbf{B}) = \left(\mathbf{X}_{\xi_\mathbf{B}}{}'\mathbf{X}_{\xi_\mathbf{B}}\right)^{-1} = \mathbf{M^{-1}}(\xi_\mathbf{A})\mathbf{diag}\left(\frac{1}{f_1^2(a)}, \frac{1}{f_2^2(a)}, \cdots, \frac{1}{f_p^2(a)}\right),$$

because $f_i(x_B) = f_i(ax_A) = f_i(a)f_i(x_B)$ for $i = 1, \dots, p$.

Similarly, for $l = 1, \dots, 10$,

$$\mathbf{x}'_{l,B} = \mathbf{diag}(f_0(a), f_1(a), f_2(a), \cdots)\mathbf{x}'_{l,A}.$$

Therefore,

$$\gamma(\xi_B) = \frac{1}{n_g}\sum_{l=1}^{n_g} \mathbf{x}'_{l,B}\mathbf{M^{-1}}(\xi_\mathbf{B})\mathbf{x}_{l,B} = \frac{1}{n_g}\sum_{l=1}^{n_g} \mathbf{x}'_{l,A}\mathbf{M^{-1}}(\xi_\mathbf{A})\mathbf{x}_{l,A} = \gamma(\xi_A).$$

□

With respect to mean part, it is required to develop very accurate statistical model $\hat{y}(x)$. Since the final goal of the experiment is to find optimal conditions whose response is near the target T, the response from fitted model should be close to the target. In the process optimization point of view, it is crucial to select design points near the optimal regions. If one fails to have design points around optimal regions, there is no hope to find appropriate optimal conditions. Many space-filling designs are developed for this purpose, while optimal designs are not appropriate when design space exploration is needed because optimal designs tend to place many design points at the extreme limits of the region.

Two school of designs: optimal designs and space-filling designs are both advantages and disadvantages. Then, a question is what kind of design is appropriate for upper layer and what for lower layer. Space-filling designs are thought to be particularly appropriate for upper layer, because in general they spread the design points out nearly evenly or uniformly throughout the region of experimentation. This is desirable feature if the experimenter does not know the form of the model that is required, and believes that interesting phenomena are likely to be found in different regions of the experimental space. On the other hand, optimal designs are more appropriate for lower layer, because upper layer experiments can provide information about the form of the model and the location of interesting regions. Furthermore, optimal designs contain replicate runs, which is a desirable feature if the experimenter want to estimate a model as accurately as possible under large measurement error.

3.5 Layers of Experiments

Layers of Experiments is multi-layer experiment strategy for building a model under tight tolerance requirement. We now detail the six-step scheme, outlined in previous section.

Step 1) Postulate a Tentative Statistical Model: Low-order polynomial models are used.

$$y(x) = \sum_{i=1}^{p} \beta_i f_i(x) + \varepsilon, \tag{3.5.1}$$

where $\varepsilon \sim N(0, \sigma^2)$.

Whether to include stochastic process terms in the model (3.5.1) can be dealt with in individual problems. Model (3.5.1) with stochastic process terms is called the Gaussian process model. The Gaussian process model is essentially a spatial correlation model, where the correlation of the response between two observations decreases as the values of the design factors become further apart. However, when design points are close together this causes ill-conditioning in the data for the Gaussian process model, much like multicollinearity resulting from predictors that are nearly linearly dependent in linear regression models. Thus, Gaussian process models may be not appropriate in the physical experiment considering replication.

Step 2) Plan an Experiment and Collect the Data: In two-layers of experiments, space-filling designs are used for upper layer and optimal designs are used for lower layer. Note that optimal designs are constructed under the assumption that the model structure $(f_1, \ldots, f_p)$ in (3.5.1) is known. The question of sample size is a difficult one. Criteria for selecting sample sizes are the subject of ongoing research, but here is one guideline.

Step 3) Use the Data to Fit the Model: We estimate the parameters in the model (3.5.1) and obtain $\hat{y}(x)$. Note that we do not consider multiple models, because optimal designs depend on underlying model. Designs based on multiple candidate models require more physical experiments, but we cannot afford it due to limited resources. Instead, multiple candidate models are considered in step 5 to simulate possible center points.

Step 4) Check the Accuracy of Prediction and Plot the Parameter Effects: To check whether the fitted model accuracy meet tight tolerance requirement, we compute the evaluation metric. If the evaluation metric is sufficiently accurate for the required tolerance, go to step 6; otherwise proceed to step 5.

Step 5) Choose a Subregion for the Next Experiment: An optimization routine can be used to find the center of the new subregion, while the sensitivity analysis can be used for choosing new limits for the new subregion. We will discuss this step in detain in the following section. Then repeat steps 1 to 4, with data drawn from the new subregion.

Step 6) Optimize the Evaluation Metric: The evaluation metric depends on the goal of the experiments. We replace y by $\hat{y}$ and seek to optimize the resulting predicted evaluation. Then, given the target value of the output yield (T) and tolerance requirement (d), we want to minimize the fluctuation around T caused by model uncertainty. This suggests maximum distance of 95% confidence interval of $\hat{y}$ from T

To satisfy Eq. (3.3.6), we should construct very accurate statistical model near T. After finding an estimate of the optimum we do a confirmatory run. If the confirmatory run is unsatisfactory, we take steps to improve the models. A new stage with further data might be necessary if we cannot improve the fit of the models.

The six steps just described clearly can accommodate other classes of models in step 1 and other optimizing object function in step 6. We have found that our particular choices make the sequential process efficient.

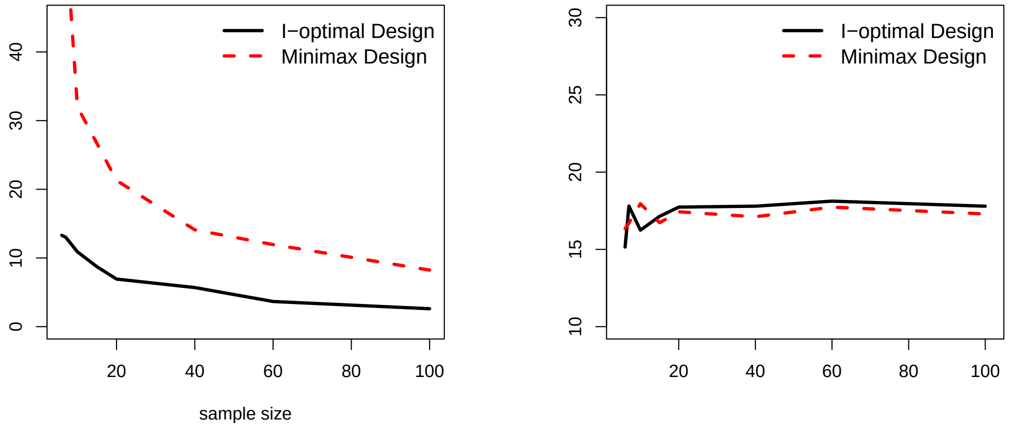
3.6 Illustrative Example

The goal of this example is to show how sample size, design types, and design space refinement affect the evaluation metric L^*. Given tight tolerance requirement, $L^* \leq d$, we show layers of experiments are an effective and efficient approach for building a statistical model. For comparison, we present the performance of single layer approach first.

3.6.1 Single layer experiments

A single variable cubic function, $y(x) = f(x)+\varepsilon = 2x^3-32x+1+\varepsilon$, is considered where $\varepsilon \sim N(0, 10)$. $f(x)$ is used to represent a computation-intensive design function. In the designated design space $[-3, 5]$, L^* is computed based upon approximation by 2nd order polynomial regression model. Note that there is difference between fitting model and true model. This represents model misspecification that occasionally occurs in physical experiments.

For comparison, we use two different types of designs: I-optimal design and minimax design with various sample sizes and investigate $\gamma(\xi)$ and $|T - \hat{y}(x)|$ separately as shown in Figure 18(a) and 18(b).



(a) The value of $\gamma(\xi)$ according to different sample size

(b) The value of $\min_x |T - \hat{y}(x)|$ according to different sample size

Figure 18: The impact of sample size on evaluation metric

Figure 18(a) depicts the impact of sample size and design types on $\gamma(\xi)$. Generally, $\gamma(\xi)$ decreases as sample size increases and an optimal design, as it stands, provides accurate inference performance. However, there is no difference in $\min_x |T - \hat{y}(x)|$ between two different types of designs. Although the minimax design explores design space more evenly, it does not help on $\min_x |T - \hat{y}(x)|$ under the environment of model misspecification. In reality, model misspecification is inevitable, because the true model is unknown and sometimes too complex. Once model is misspecified, large sample size and design strategy hardly make fitted model to reach the target (T).

3.6.2 Two Layers of Experiments

The concept of the layer of experiments is illustrated with the same function from previous example. The difference from previous example is that lower layer experiments are conducted sequentially after upper layer experiment.

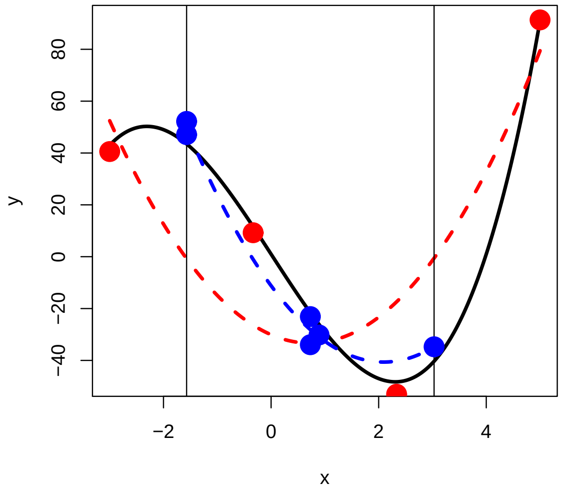


Figure 19: Two-layer experiments: Four samples (red dots) are collected by space-filling design at the upper layer. Six samples (blue dots) are collected by optimal design at the lower layer.

Based upon a minimax design, four experimental points are obtained in the upper layer $[-3, 5]$ to approximate the "unknown curve, as illustrated in Figure 19. If one applies a second-order model to carry out the approximation, the first fitted function using the least square method will be $\hat{y}_u(x) = -20.42 - 10.11x + 6.06x^2$.

To improve the approximation accuracy, we apply the proposed method to conduct an experiment in the lower layer. The lower layer, new design space, is determined at $[-1.8, 2.5]$. At the lower layer I-optimal design with $n_{lower} = 6$ is use to fitting 2nd-order polynomial model. (see blue dash line in Figure 19) The second fitted model, $f(x) = -8.91 - 22.80x + 4.98x^2$, performs much better in terms of evaluation metric

(L^*).

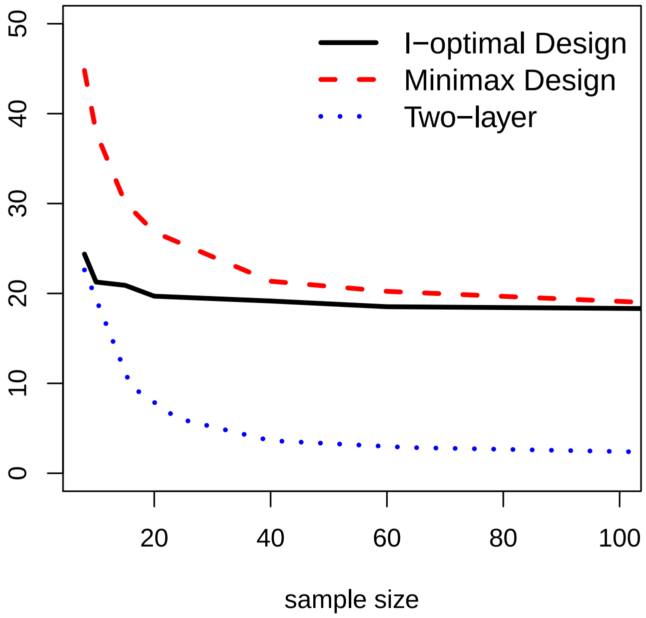


Figure 20: The performance of two-layers of experiments in L^*

Now, we compare a single layer of experiment and two layers of experiments in terms of evaluation metric (L^*). For a single layer experiment, we use the same setting as illustrated in Figure 18. For two layers of experiments, two different types of designs are used: minimax design for upper layer and I-optimal design for lower layer. For simplicity, half of samples are collected in the upper layer and the other half samples are used in the lower layer. In both layers, approximation models are 2nd order polynomial regression model and I-optimal designs are constructed based on the model structure. Evaluation metric L^* is computed by the average of ten experiments for each experimental setting.

Figure 20 shows the results of comparison. From sample size 10 to 100, a single layer experiment and two layers of experiments are compared in terms of evaluation metric. A single layer experiment cannot reduce model accuracy more even with a large sample size after it reaches a certain level of accuracy. On the other hand, layers of experiments provide a way to reduce evaluation metric dramatically with same sample size. The experiment in the lower layer alleviates the negative effect of model misspecification. For example, with sample size $n = 16$, two layers of experiments results in $L^* = 9.67$, while a single layer experiment gives 21.16 for I-optimal design and 30.45 for minimax design, respectively.

Next, we carry out same experiments except the sequence of design type to see the performance of recommended sequence: space-filling design for upper layer and optimal design for lower layer. Table 2 shows the L^* values for two-layers of experiments in four different sequence of design type: 1)space-filling / optimal, 2)space-filling / space-filling, 3)optimal / optimal, and 4)optimal / space-filling.

Table 2: The performance of two layer experiments in different design strategy

Sample size (upper/lower)	Design (upper/lower)	L^*
n=16 (8/8)	space-filling / optimal	9.67
n=24 (12/12)	space-filling / optimal	6.09
n=40 (20/20)	space-filling / optimal	3.62
n=16 (8/8)	space-filling / space-filling	11.96
n=24 (12/12)	space-filling / space-filling	6.79
n=40 (20/20)	space-filling / space-filling	4.95
n=16 (8/8)	optimal / optimal	50.30
n=24 (12/12)	optimal / optimal	33.82
n=40 (20/20)	optimal / optimal	32.19
n=16 (8/8)	optimal / space-filling	43.83
n=24 (12/12)	optimal /space-filling	31.89
n=40 (20/20)	optimal / space-filling	28.61

The results in Table 2 confirm that space-filling / optimal combination is better than the others. Upper layer requires a design having space-filling properties to find the design space of lower layer accurately. The inefficiency from optimal designs in the upper layer is mainly caused by wrong subregion decision. However, how to allocate sample size into upper layer and lower layer is remained issue. Proper design sequence in more than two layers of experiments will be covered in the next chapter.

3.7 Conclusion

To find optimal process conditions of complicated response with large uncertainty, it is crucial to select design points near the optimal regions. If one fails to have design points around optimal regions, there is no hope to find appropriate optimal conditions. However, the given resources are limited and so one should allocate enough resources to important regions. We proposed a systematic procedure to give more weight of using given resources on the optimal regions. We called it 'Layers of Experiments'.

'Layers of Experiments' is motivated from nano-manufacturing where the yield function of nano-fabrication is very complex and requires very tight tolerance requirement. So, commonly used experiment schemes do not work for this situation. Layers of Experiments zoom into subregions and conduct sequential experiments. By redefine design space, we can relax the model complexity in the subregions and put more resource in the focused subregions. However, this approach contains a risk of missing process optima.

Illustrative examples show layers of experiments are able to satisfy tight tolerance requirement, which is hardly made by a single layer experiment. Also, we recommend space-filling design for upper layer and optimal design for lower layer in two layers of experiments. A design strategy for multi-layer experiments is developed in the next chapter.

CHAPTER IV

COMBINED OPTIMAL AND SPACE-FILLING DESIGNS

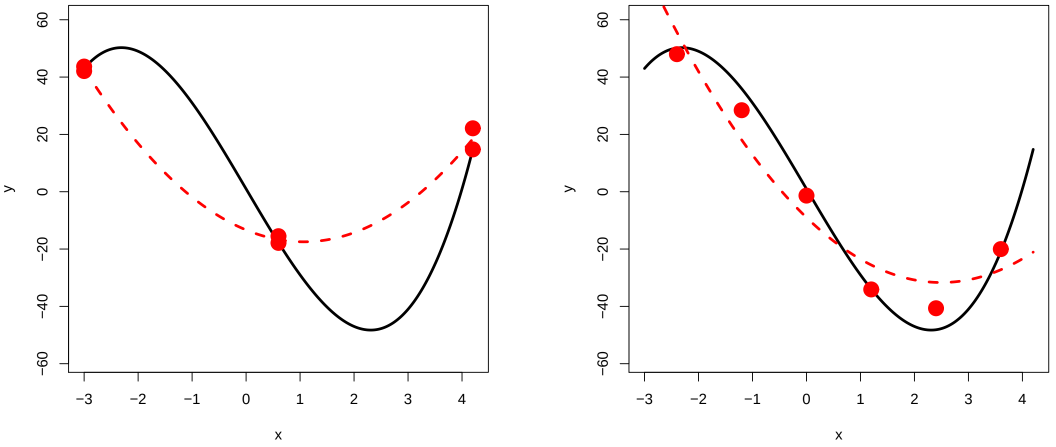
4.1 Introduction

Statistical modeling for nano-fabrication process to achieve consistently good performance has become increasingly important [25]. So, we need to build a statistical model that meets tight tolerance requirement near the process optimum (or a given target (T)). To check whether the requirement is met or not, we developed an evaluation metric in the previous chapter as

$$L(x) = \max\left\{\left|T - (\hat{y}(x) + CI(x))\right|, \left|T - (\hat{y}(x) - CI(x))\right|\right\}.$$

In this chapter, a new design criterion for the evaluation metric is of interest. To meet the evaluation metric, the new design criterion should satisfy two characteristics: accuracy in statistical inference and design space exploration. The design should be able to construct statistical models as accurately as possible. The component of confidence interval in the evaluation metric quantifies the accuracy of the fitted model. On the other hand, the design should be spread out as evenly as possible over the design space to explore the process optimum. The difference between T and $\hat{y}(x)$ can be reduced by design space exploration. However, those two characteristics conflict with each other.

To help understand the conflict between these two characteristics, Figure 21 shows the two different types of designs and their results in data collection and modeling. Both designs are used to build 2nd order polynomial regression models to fit unknown true physical process. We assume that there exists a mismatch between the true process and the statistical models. Figure 21(a) illustrates six D-optimal design points under the 2nd order models. Red dots on the plot represent collected field



(a) Model-based optimal design (D-optimal design)

(b) Model-free space-filling design (minimax design)

Figure 21: Accuracy in statistical inference vs. design space exploration. Black solid lines are true physical process, while red dash lines are fitted statistical models

data. Optimal designs usually allow replication to improve statistical inference as shown in Figure 21(a). On the other hand, Figure 21(b) shows a minimax design and corresponding collected field data. Contrasting with Figure 21(a), the design points are spread out evenly over the design space to reduce the chance of missing the process optimum. For the given target $T = -48.27$, the values of $\min_x |T - \hat{y}(x)|$ are 30.79 for the D-optimal design and 16.60 for the minimax design. The values of the average confidence interval (CI) over the design space are 6.60 for the D-optimal design and 31.20 for the minimax design. We see that there is a trade-off between the two characteristics. Here, the D-optimal design and the minimax design are representative model-based optimal design and model-free space-filling design, respectively.

Much of the statistical work on process optimization has concerned the use of optimal design under the underlying models. With limited resources, optimal designs provide optimal accuracy in statistical inference (e.g. parameter estimation of the models). Model-based optimal design approaches are generally used in physical experiments where model validation is needed due to random errors and where a high level of model accuracy is required. However, optimal designs have been criticized because they need to assume underlying models and so they are quite sensitive to their assumptions [4, 5].

On the other hand, space-filling designs such as minimax designs [18], Latin hypercube designs [27] and uniform designs [12] do not need underlying model assumptions. They are adequate for exploring complex response surfaces with a minimum number of runs. However, the space-filling designs have limitations. They are hardly used for model validation because they do not allow replications and so the fitted model may be more erroneous and less accurate than optimal designs. Since each of them has its own advantages and limitations, it is natural consider a combination of those design criteria.

Table 3: Combined design criteria

	Model-based	Model-free	Example
One after the other		Maximin distance, Latin Hypercube	Morris and Mitchell (1995) [29]
		Latin Hypercube, pairwise correlations	Owen (1994) [31]
	Orthogonal	Latin Hypercube	Tang (1993) [36]
	Adjusted Optimal	Uniform	Fang and Wang (1994) [13]
Combinations	D-optimality	Latin Hypercube	Goel et al. (2008) [15]
Constrained or Compound	D , T-optimality		Atkinson (2008) [2]
		Latin Hypercube, pairwise correlations	Joseph and Hung (2008) [20]

We propose a combined design criterion between model-based optimal design and model-free space-filling design. Combined criteria are discussed in several papers (see Table 3). There are three types of procedures for combining design criteria. The first type is that the two design criteria are applied one after the other. That is, designs are first restricted by the first criterion and then a second criterion is applied to this in order to achieve additional desirable properties. For example, maximin distance designs may be applied within the class of Latin Hypercube Designs (LHDs) [29] (see more examples in Table 3).

The second type is that a part of the total runs are collected by one design criterion and the rest of runs by another design criterion. This type of combined design is called a combination design [15]. For example, a total of 100 runs consists of 50 runs by model-based D-optimal design and the remaining 50 runs by geometry-based LHS criterion.

The third type is to combine two design criteria using constrained optimal designs [7] or compound optimal designs [22, 23]. For example, the DT-optimality criterion [2] combines the T-optimality criterion and D-optimality criterion, where the interest lies in model discriminating and estimating the parameters. This type of combined design creates a new design criterion and the criterion has both characteristics of original criteria partially. Generally, this type of combined design introduces a tuning parameter to control these characteristics.

We combine design criteria using constraint or compound optimization technique. Recall that our purpose is to build accurate statistical models for approximation of a complicated true function with a few runs. For this purpose, the third type of combined design is more appropriate because each design point in this type of designs has both characteristics, thereby requiring smaller run size than the other types. The contribution of this chapter is to propose a new design criterion combining model-based design criteria and model-free design criteria in a constraint or compound manner. To the best of our knowledge, we could not find an attempt to this in the literature. The proposed design is useful when the fitted statistical model is required to have both characteristics: accuracy in statistical inference and design space exploration. The weight of these two characteristics can be easily controlled compared to existing combined designs.

This chapter is organized as follows. In Section 2, performance measures for evaluating the goodness of optimal design and a minimax criterion are described. In Section 3, we propose a multi-objective criterion combining the two performance measures. In Section 4, we introduce the algorithm for generating combined designs. Some property investigations and several examples are presented in Section 5 and 6.

4.2 Performance Measures

One considers optimum designs to achieve the most precise statistical inference possible for an underlying model. We assume that an observation $y_i(x_i)$ may be written

$$y_i = \mathbf{f}^T(x_i)\boldsymbol{\theta} + \varepsilon_i, \quad i = 1, 2, \ldots, N, \tag{4.2.1}$$

where the x_i's are elements of a compact design space, $\mathscr{X}$, $\mathbf{f}$ is continuous on $\mathscr{X}$, and the ε_i's are uncorrelated random variables with mean zero and variance σ^2. Letting $\mathbf{f}^T(x_i)$ denote the $(q \times 1)$ ith row of a matrix $\mathbf{X}$, Eq. (4.2.1) may be rewritten

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}.$$

A design ξ_N is an N-point exact design if ξ_N is a probability measure on $\mathscr{X}$ and $N\xi_N(x)$ is a nonnegative integer for every $x \in \mathscr{X}$. Denote the space of N-point exact designs on $\mathscr{X}$ by $\Omega_{\mathscr{X}}^N$ For approximate designs, the restriction that $\xi(x)$ be a multiple of $1/N$ is relaxed. Thus, an approximated design is simply an element of the space $\Omega_{\mathscr{X}}$, of probability measures on $\mathscr{X}$.

Exact and approximate D-optimal designs maximize the determinant of the information matrix,

$$M(\xi) = \int_{\mathscr{X}} \mathbf{f}(x)\mathbf{f}^T(x) d\xi(x). \tag{4.2.2}$$

If $|M(\xi)| \neq 0$, the dispersion matrix, $M^{-1}(\xi)$. Note that for an N-point exact design ξ_N,

$$M(\xi_N) = \mathbf{X}^T\mathbf{X}/N.$$

Hence, if we set $N\xi_N = \xi[N]$, we have

$$M(\xi[N]) = \mathbf{X}^T\mathbf{X}.$$

The variance of the least squares predictor $\hat{y}(x)$ is given by

$$\sigma^2(x) = \sigma^2\mathbf{x}'\mathbf{M}^{-1}(\xi[N])\mathbf{x} = \sigma^2 v(x, \xi[N]),$$

where $\mathbf{x}' = [f_1, \dots, f_p]$ is a vector of p real valued functions of x based on the model terms and $v(x, \xi)$ is defined to be the variance function of the design ξ.

Now we will discuss a performance measure based on the geometric distance. Let ξ be the design of N points, $x \in \mathbb{R}^k$ represent an arbitrary point in the feasible region, and $\rho(x, \xi)$ the distance between x and its closest design point, i.e.

$$\rho(x, \xi) = \min_{x_i \in \xi} \tau(x, x_i).$$

A minimax design ξ^* of N points then has a distance

$$\rho^* = \min_{\xi \subset \Omega} \max_{x \in \Omega} \rho(x, \xi), \tag{4.2.3}$$

or

$$\xi^* = \min_{\xi \subset \Omega} \rho(\xi),$$

where $\rho(\xi)$ is defined by $\max_{x \in \Omega} \rho(x, \xi)$ and Ω denote a set of sites.

The distance ρ^* is referred to as the minimal covering radius of the design. For example, with ordinary distance operating on $[0, 1]$ minimax design places N points at elements of

$$\{(2i - 1)/2N, \quad i = 1, \dots, N\}$$

while $\rho^* = 1/2N$. In case of 7 congruent l^2-circles the minimal radius needed to cover the unit square is $\rho^* \approx 0.2743$ (see Figure 22). In this figure the diamonds($\diamond$) depict remote sites, i.e. points in the square that are at distance ρ from the design.

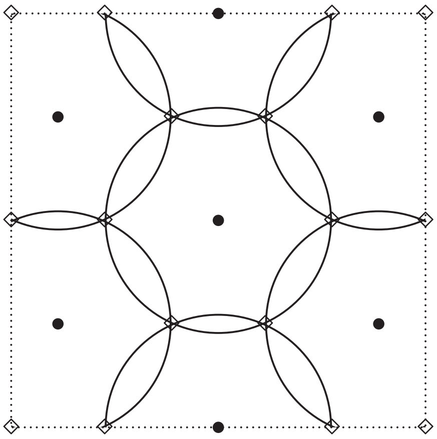


Figure 22: Two-dimensional l^2-minimax design of 7 points on the design space $[0,1]\times[0,1]$; $\rho^* \approx 0.2743$. (from [18])

The design that minimizes ρ will be a minimax design. In the next section we propose a new criterion which combines the performance measures in (4.2.2) and (4.2.3).

4.3 Combined Design

Our objective is to find a combined design that minimizes both $|M^{-1}(\xi)|$ and $\rho(\xi)$. A common approach in multi-objective optimization is to optimize a weighted average of all the objective functions [20]. The objective function of a weighted average method is

$$\kappa|M^{-1}(\xi)| + (1-\kappa)\rho(\xi), \tag{4.3.1}$$

where κ is pre-specified positive weights. In the weighted-sum method, all the objectives are aggregated into a single objective by using a weight vector. Although the weighted-sum method is simple and easy to use, there are three major problems. Firstly, it is not easy to choose appropriate values of κ, because the objectives have different scale. Secondly, the performance of the method is heavily dependent on the shape of the feasible region so that it cannot find all the optimal solutions for problems that have a non-convex feasible region. Thirdly, κ values may not linear in the value of $\rho(\xi)$ or $M(\xi)$. For example, even though $\kappa = 0$ means $|M^{-1}(\xi)| = 1$, $\kappa = 0.8$ does not necessarily mean $|M^{-1}(\xi)| = 0.8$ in the weighted sum method. $\rho(\xi)$ could be 0.7, 0.6, or any other value depending on the relationship between κ and the objectives. Without exact knowledge of the relationship, it is difficult to control the characteristic of the combined design through κ.

A procedure that overcomes some of the problems of the weighted sum technique is the ϵ-constraint method. This involves maximizing a primary objective, $|M(\xi)|$, and expressing the other objective in the form of inequality constraints as

$$\begin{aligned} \max_{\xi} \quad & |M(\xi)| \\ s.t. \quad & \rho(\xi)_s \leq \kappa, \end{aligned} \tag{4.3.2}$$

where $\rho(\xi)_s$ is the performance measure of $\rho(\xi)$ that scaled into $[0, 1]$. We choose $|M(\xi)|$ as a primary objective, because it facilitates to construct the combined design using existing optimal design construction algorithm. In other words, we find D-optimal design under the restricted design space by inequality constraints. We will call a optimal design from the objective function (4.3.2) with pre-specified κ κ-CbindD.

To compute $\rho(\xi)_s$, we need lower bound (ρ_L) and upper bound (ρ_U) of $\rho(\xi)$. ρ_L is known by definition of minimax design and ρ_U is $\rho(\xi^\dagger)$ where $\xi^\dagger$ is the D-optimal design as the results of Proposition 4. Denote $\rho(\xi)_s$ by $(\rho(\xi) - \rho_L)/(\rho_U - \rho_L)$. Then, $\rho(\xi)_s \in [0, 1]$ has the same range as $\kappa \in [0, 1]$.

Proposition 4. *Suppose that $\xi^\dagger$ is a D-optimal design and $\Omega = \{\xi : |M(\xi)| \leq |M(\xi^\dagger)|\}$. Then, the upper bound of $\rho(\xi)$ is $\rho(\xi^\dagger)$. That is,*

$$\min_{\xi \in \Omega} \rho(\xi) \leq \rho(\xi^\dagger).$$

Proof. The objective function (4.3.1) can be rewritten by

$$\begin{aligned} \min_{\xi} \quad & \rho(\xi) \\ s.t. \quad & |M(\xi)| \geq c, \end{aligned}$$

Consider two designs: ξ_A and ξ_B, which $|M(\xi_A)| \leq |M(\xi_B)|$. Let $\Omega_A = \{\xi : |M(\xi)| \geq |M(\xi_A)|\}$ and $\Omega_B = \{\xi : |M(\xi)| \geq |M(\xi_B)|\}$. Then

$$\min_{\xi \in \Omega_A} \rho(\xi) \leq \min_{\xi \in \Omega_B} \rho(\xi),$$

because $\Omega_A \subset \Omega_B$. Hence, for $|M(\xi)| \leq |M(\xi^\dagger)|$,

$$\min_{\xi \in \Omega} \rho(\xi) \leq \rho(\xi^\dagger),$$

where $\xi^\dagger$ is a D-optimal design and $\Omega = \{\xi : |M(\xi)| \leq |M(\xi^\dagger)|\}$. □

4.4 Algorithm

An exchange algorithm is used to find designs from a discrete candidate list of possible design points. Exchange algorithms are the most common techniques to construct optimal designs. All exchange algorithms share the same basic operations. Points in the current design are exchanged with those in a candidate list of possible design points. Exchanges are accepted when they improve objective function; otherwise the exchange is rejected. If constraints are exist in the objective function, exchanges are made only for the candidates that satisfy the constraints. According to Fedorov algorithm [14], during the ith iteration, a point x_j is deleted and another point x in a design space $\mathscr{X}$ $(x \in \mathscr{X})$ is added in such a way that the resulting increase in the determinant is maximal.

$$\frac{|M(\xi_{i+1}[N])|}{|M(\xi_i[N])|} = 1 + \Delta_i(x_j, x)$$

where

$$\Delta_i(x_j, x) = [1 + v(x, \xi_i[N])] \times [v(x, \xi_i^x[N+1]) - v(x_j, \xi_i^x[N+1])].$$

$\xi_i^x[N+1]$ is used to denote the design $\xi_i[N]$ augmented by the point x.

Cook and Nachtsheim modified Fedorov exchange algorithm [8]. Each iteration corresponds to an iteration of the Fedorov algorithm. However, in the modified one, an iteration s is broken down into N stages, one for each support point in the design at the start of the iteration. At stage i, the first argument of the delta function is fixed at the ith support point, the point $x^* \in \mathscr{X}$ is found such that

$$\max_{x \in \mathscr{X}} \Delta_s(x_i, x) = \Delta_s(x_i, x^*),$$

and then x_i is exchanged for x^* (i.e., the design is updated). The support points may be randomly ordered at the start of an iteration. A single iteration consists of a number of exchanges equal to the number of support points at the start of the iteration. Since

$$\max_{x \in \mathscr{X}} \Delta_s(x_i, x) \geq \Delta_s(x_i, x_i) = 0,$$

an exchange will never result in a decrease in the determinant of the information matrix.

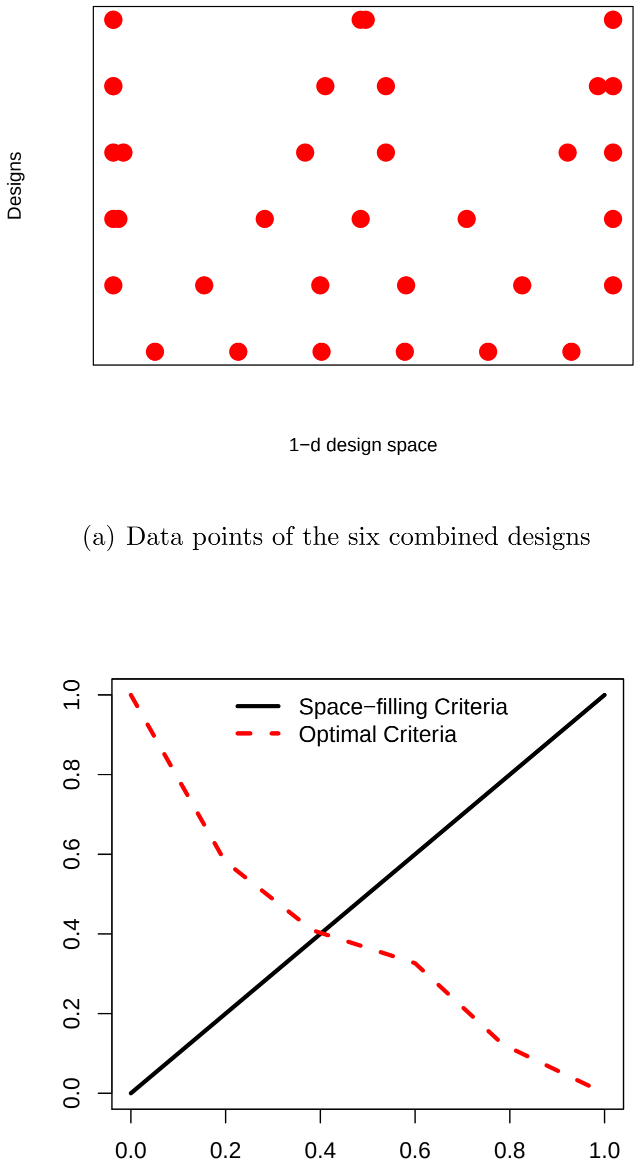
4.5 Property Investigation

Table 4: Six-points combined designs under 2nd order polynomial model structure with different values of κ

κ	0.0	0.2	0.4	0.6	0.8	1.0
	0.08	0.00	0.01	0.02	0.00	0.00
	0.25	0.18	0.30	0.38	0.42	0.49
	0.42	0.41	0.49	1.00	0.00	0.00
	0.58	0.59	1.00	0.00	1.00	1.00
	0.75	0.82	0.71	0.91	0.97	0.51
	0.92	1.00	0.00	0.55	0.55	1.00

We construct combined designs using the proposed new criterion and the exchange algorithm. The exchange algorithm is used to find designs from a discrete candidate list of 100 grid points over a design space $[0, 1]$. Table 4 shows six-points combined designs on $[0, 1]$ with different values of κ. The first column is a minimax design and the last column is a D-optimal design under the 2nd order polynomial model.

From 2nd column to 5th column represent combined designs between minimax design criteria and D-optimal design criteria under the 2nd order polynomial models.



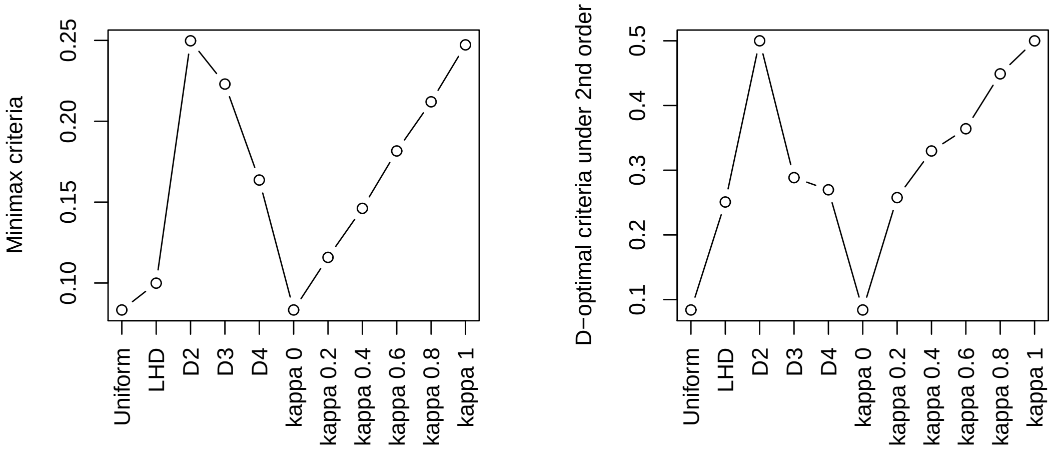
(a) Data points of the six combined designs

(b) The values of $\rho(\xi)_s$ (Black solid line) and $M(\xi)_s$ (red dash line) in the six combined designs.

Figure 23: Combined designs

Figure 23(a) depicts six combined designs in Table 4. Six data points at the bottom line are the minimax design, while six data points at the upper line are the D-optimal design. The proposed combined design criterion and the algorithm find designs that have both properties of minimax and D-optimal designs and the weights of two properties changes depending on the κ value. Figure 23(b) shows the values of $\rho(\xi)_s$ and scaled $M(\xi)$ ($M(\xi)_s$) of the six combined designs. Note that $\rho(\xi)_s$ has linear relationship with κ. This is because $\rho(\xi)_s$ is restricted by κ value in the secondary objective of (4.2.3) when the design is constructed. This is one of aforementioned advantages of ϵ-constraint method over weighted sum method.

Now, the properties of combined designs with several different values of κ are investigated and they are compared with the popularly used space-filling designs and optimal designs. Combined designs of various κ are compared with uniform design, LHD, and D-optimal designs in terms of $\rho(\xi)$ and $M(\xi)$.



(a) Minimax criteria(smaller-the-better)

(b) D-optimal criteria under 2nd order(larger-the-better)

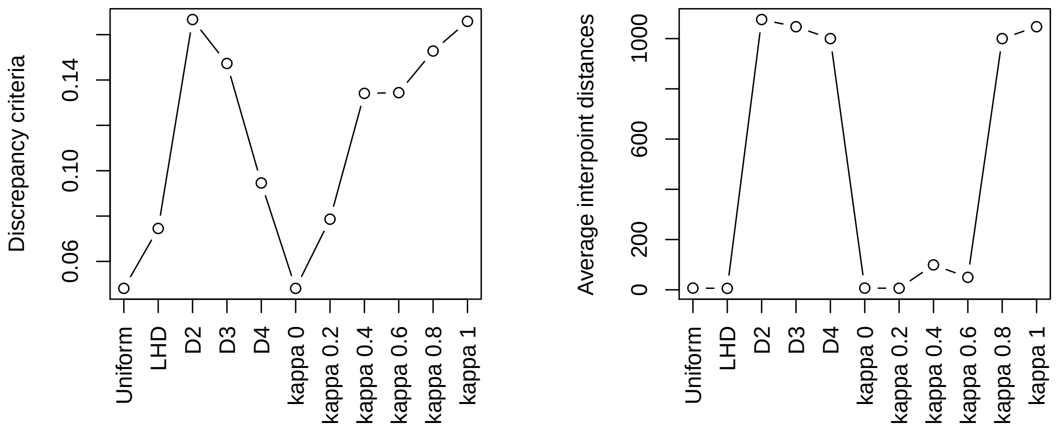
Figure 24: Comparison among combined designs and commonly used experimental designs in minimax criteria and D-optimal criteria. $D2$, $D3$, and $D4$ represents D-optimal criteria under 2nd, 3rd, 4th polynomial regression model, respectively.

As shown in Figure 24(a) and Figure 24(b), space-filling designs such as uniform design and LHD show better results than optimal designs in minimax criteria($\rho(\xi)$). On the other hand, $D2$(D-optimal design under 2nd order polynomial regression model) is the best in terms of D-optimal criteria under 2nd order model structure. Note that $D2$ is actually equivalent to combined design with $\kappa = 1$ by definition of combined design. However, $D3$ and $D4$ performs similar but both poorer than $D2$. This is because optimal designs are sensitive underlying model misspecification. That is, wrong model assumption makes optimal designs ineffective.

Let us compare those designs with other criteria: discrepancy and average interpoint distance. The centered $L2$-discrepancy criterion (see [12]) is a design criterion to construct the uniform design, whereas the average interpoint distance can be used to construct LHD design (see [29]) by maximizing

$$\phi_p = \left(\frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \frac{1}{\tau(x_i, x_j)^p} \right)^{\frac{1}{p}}, \tag{4.5.1}$$

where $\tau(x_i, x_j)$ denote the Euclidean distance between two points x_i and x_j of a design $\xi = \{x_1, x_2, \ldots, x_n\}$. When p is sufficiently large, it can be shown that the r criterion is just equivalent to the maximin criterion [18]. In our study, we choose a value $p = 15$ and use ϕ_{15}.



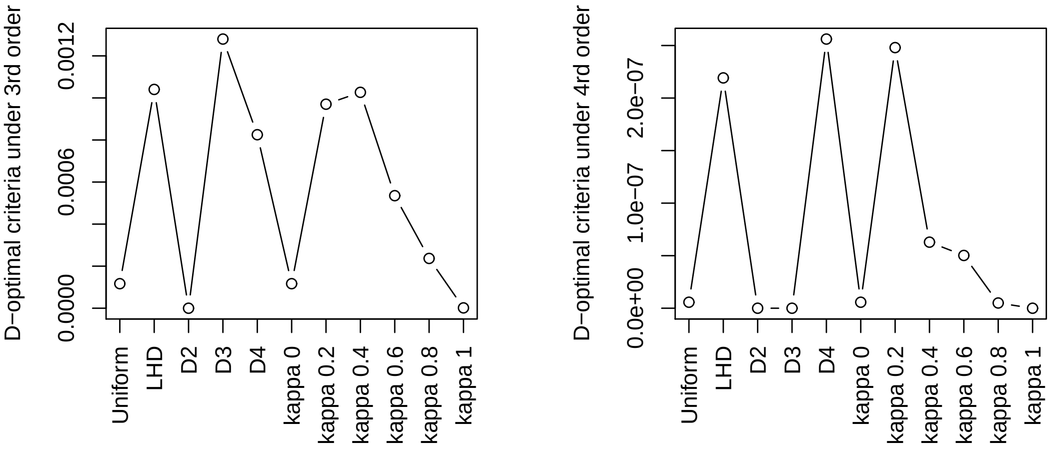
(a) Discrepancy criteria(smaller-the-better)

(b) Average interpoint distances(smaller-the-better)

Figure 25: Comparison among combined designs and commonly used experimental designs in discrepancy criteria and average interpoint distances

With respect to discrepancy criteria as shown in Figure 25(a), it is similar to the comparison using minimax criteria. As expected, uniform design indicates the smallest value in discrepancy criteria. 0.4-CbindD and 0.6-CbindD show similar value in discrepancy criteria.

Figure 25(b) indicates that the average interpoint distance is not appropriate criterion for comparison optimal designs because the optimal designs allow replication of design points. As you see in Eq. (4.5.1), replication makes the criteria infinity.



(a) D-optimal criteria under 3rd order(larger-the-better)

(b) D-optimal criteria under 4th order(larger-the-better)

Figure 26: Comparison among combined designs and commonly used experimental designs in D-optimal criteria

The weakness of combined designs is that they are also dependent to underlying model assumption, because optimal design criteria part in combined design criteria requires the information of model structure. So, the performance of combined designs are poor in terms of D-optimal criteria with different order from the one used in design construction (see Figure 26). If correct model is 3rd order (Figure 26(a)) or 4th order (Figure 26(b)) polynomial model, the performance of designs assuming incorrect model may be poor. $D2$ and $D4$ in Figure 26(a) and $D2$ and $D3$ in Figure 26(b) are the case.

However, combined designs are less sensitive to model misspecification than pure optimal designs, because of the space-filling criteria part in combined design criteria. 0.2-CbindD and 0.4-CbindD outperform both D-optimal design ($\kappa = 1$) and minimax design ($\kappa = 0$), which means combined designs may be robust against model misspecification because they have both space-filling properties and optimal design properties. Also, strength of the proposed combined designs is the two properties can be controlled by κ value.

4.6 Conclusion

We employed a combined design criterion: one from optimal design criteria and one from space-filling design criteria. In another words, it is combination between model-based design and model-free design. One for precise statistical inference, while the other one is for exploration over the design space. A threshold (κ) in the combined design criteria controls the weight between the two criteria.

We showed that combined designs have properties between optimal designs and space-filling designs and they are robust against model misspecification. Moreover, combined designs perform better than space-filling designs or optimal designs where partial information about underlying model is available.

The combined design can be used in the layers of experiments. That is, as layers go further, the combined design criterion moves from space-filling criterion to optimal criterion as more information accumulates. The threshold of the combined design should adaptive to model uncertainty of each layer. Thus, as layers go further, the combined design criterion moves from space-filling criterion to optimal criterion. We will study the adaptiveness of combined design criteria in the next chapter.

CHAPTER V

ADAPTIVE COMBINED DESIGNS WITH ENGINEERING MODELS

5.1 Introduction

Statistical models are commonly used in quality improvement studies. Since such models are basically data-driven models, they tend to perform poorly when predictions are made far away from the observed data points. Moreover, the experimental data required for estimating the statistical models can be expensive.

On the other hand, engineering models are developed based on the engineering/physical laws governing the process, which include analytical models and finite element models. However, engineering models have limitations in that predictions derived from engineering models are often not accurate. This is because engineering models are developed based on several simplifying assumptions, which may not hold true in practice.

As reviewed in Section 2, various formulations are available for constructing an update model based on the original engineering model $y_m(x)$. Without loss of generality, we use one typical engineering model updating formulation

$$y_{m'}(\mathbf{x}) = y_m(\mathbf{x}) + \omega(\mathbf{x}) + e, \tag{5.1.1}$$

where, $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ are n controllable input variables, e is an unobservable output variable, also assumed random, to capture the experimental uncertainty associated with a model output. $\omega(\mathbf{x})$ represent the engineering model bias, which is unknown.

Let $y(\mathbf{x})$ be the observations (field data) from physical experiments in the layer R, $\mathbf{x} \in R$, and $y_s(\mathbf{x})$ be the output of a statistical model. The proposed approach models the relationships among the observation from physical experiments, $y(\mathbf{x})$, the engineering model output $y_m(\mathbf{x})$, and the statistical model output $y_s(\mathbf{x})$ via

$$y(\mathbf{x}) = y_s(\mathbf{x}) + \epsilon = y_m(\mathbf{x}) + \omega(\mathbf{x}) + e, \tag{5.1.2}$$

where ϵ is the experimental error of the observation from physical experiments in the layer. Engineering models derived using the underlying physics of the process do not always match satisfactorily with reality.

The bias function $\omega(\mathbf{x})$ is used to capture the model systematic bias, but not intended to account for the experimental uncertainty. $\omega(\mathbf{x})$ could be parameterized in various ways, for example, with a regression model $\omega(\mathbf{x}; \boldsymbol{\beta})$ parameterized by $\beta_{\omega 0}, \beta_{\omega 1}, \ldots, \beta_{\omega p}$. Here the bias function $\omega(\mathbf{x}; \boldsymbol{\beta})$ is treated to be a deterministic function that does not contribute to the model output uncertainty.

$$\text{Find } \boldsymbol{\beta} \text{ Minimizing SSE } = \sum_{i=1}^{N} w_i e_i^2 = \sum_{i=1}^{N} w_i [y(\mathbf{x}_i) - y_m(\mathbf{x}_i) - \omega(\mathbf{x}_i; \boldsymbol{\beta})]^2, \tag{5.1.3}$$

where $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]^T$ $(i = 1, \ldots, N)$ are sample points. w_i $(i = 1, \ldots, N)$ are the weights for different experimental observations reflecting the quality of experimental data, $\boldsymbol{\beta} = [\beta_{\omega 0}, \beta_{\omega 1}, \ldots, \beta_{\omega p}]^T$ are unknown parameters to be estimated.

Once $\omega(\mathbf{x}_i; \boldsymbol{\beta})$ is estimated using field data by (5.1.3), we can update engineering model by (5.1.1). Let us call $y_{m'}(\mathbf{x})$ *statistically updated engineering model* or *adjusted engineering model.*

Statistical models are useful when engineering models can be quite complex and expensive to compute. Furthermore, statistical models can be underlying models of most of model-based designs and more appropriate prediction, control, and optimization in many cases. Therefore, we want to build an accurate statistical model using field data as well as *statistically updated engineering model.* Here, we call this kind of statistical models *engineering adjusted statistical model.*

The contribution of this chapter is to present an adaptive combined design to build the *engineering adjusted statistical model.* In engineering model updating methods, finding appropriate designs have not been main issue because updating methods focus bias-correction. So, space-filling designs such as Latin hypercube design or uniform designs are used, because design points should be spread out over the design space as evenly as possible to reduce $\omega(\mathbf{x})$ efficiently. However, to build the *engineering adjusted statistical model*, the field data update engineering model as well as estimate the statistical model $y_s(\mathbf{x})$ due to limited resources, as shown in Eq. (5.3.4). Those two characteristics: design space exploration and accuracy in statistical inference are conflict with each other.

Another contribution is to combine information from various layers in Layers of Experiments. With limited resource, efficient design collection scheme is important. We modify the design criteria of combined designs in order to utilize all information from various layers.

Following the introduction, in Section 2, we review literature in engineering model updating methods. In Section 3, we propose two methodologies: determining adaptive parameter (κ) for adaptive combined designs and combining information from various layers for layers of experiments. In Section 4, the proposed methodology is evaluated using illustrative examples.

5.2 Literature Review: Engineering model updating methods

There are two main approaches to update engineering models. Model bias-correction approaches and model calibration approaches.

5.2.1 Model bias-correction approaches

Bias-correction is useful when accuracy improvement cannot be accomplished by calibrating model parameters [11, 17]. Bias-correction approach captures the potential model error due to model misspecification [21]. There are various formulations of bias-correction in the literature. In the Bayesian bias-correction model proposed by

Wang et al. [38], a plain additive bias-correction model is formulated as

$$y(\mathbf{x}) = y_m(\mathbf{x}) + \omega(\mathbf{x}) + \varepsilon, \tag{5.2.1}$$

where the bias function $\omega(\mathbf{x})$ is a direct measure of the difference between the engineering model $y_m(\mathbf{x})$ and the physical process $y(\mathbf{x})$. The bias function $\omega(\mathbf{x})$ is assumed to be a Gaussian Process model.

In addition to Eq. (5.2.1), a bias correction approach may employ a combination of multiplicative bias and additive bias, as shown in the following formulation [32],

$$y(\mathbf{x}) = \nu(\mathbf{x})y_m(\mathbf{x}) + \omega(\mathbf{x}) + \varepsilon, \tag{5.2.2}$$

where $\nu(\mathbf{x})$ is modeled as a simple linear regression model w.r.t. $\mathbf{x}$, ε is assumed to be a zero-mean Gaussian random variable. The scaling function $\nu(\mathbf{x})$ in Eq. (5.2.2) brings more flexibility to the constant adjustment parameter ν used in Kennedy and OHagan [21]. The regression coefficients of $\nu(\mathbf{x})$ can be estimated by the Maximum Likelihood Estimation (MLE) method [32].

One inherent limitation of the bias-correction method is that it assumes all inputs $(\mathbf{x})$ of both the engineering model $(y_m(\mathbf{x}))$ and the physical process $(y(\mathbf{x}))$ are observable and controllable. In practice, some of the model input parameters cannot be directly observed and measured in the physical experiments. This limitation can be addressed using the model calibration approach.

5.2.2 Model calibration approaches

With a typical model calibration approach, the inputs of a computer model are divided into controllable inputs $(\mathbf{x})$ and uncontrollable parameters $(\boldsymbol{\theta})$ that are assumed to be fixed over the experiment. Note that it is $\boldsymbol{\theta}$ that are to be calibrated. A engineering model for the given input vector $(\mathbf{x}, \boldsymbol{\theta})$ is denoted as $y_m(\mathbf{x}, \boldsymbol{\theta})$, while the physical process is denoted to be $y(\mathbf{x})$ as a function of controllable inputs x only.

5.2.2.1 Deterministic calibration approach

A conventional way to carry out a deterministic parameter calibration is to formulate the problem through the following equation

$$y(\mathbf{x}) = y_m(\mathbf{x}, \boldsymbol{\theta}) + e,$$

where e is the residual between the prediction from the calibrated engineering model $y_m(\mathbf{x}, \boldsymbol{\theta})$ and the experimental observation $y(\mathbf{x})$. The optimal values of the calibration parameters $\boldsymbol{\theta}$ are found by minimizing the (weighted) sum of the squared error (SSE) between the model predictions and the physical experiments [24], i.e.,

$$\text{Find } \boldsymbol{\beta} \text{ Minimizing SSE } = \sum_{i=1}^{N} w_i e_i^2 = \sum_{i=1}^{N} w_i [y(\mathbf{x}) - y_m(\mathbf{x}, \boldsymbol{\theta})]^2$$

where $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{ik}]^T$ $(i = 1, 2, \ldots, N)$ are sample points, w_i $(i = 1, 2, \ldots, N)$ are the weights for different experimental observations reflecting the quality of experimental data, $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_m]^T$ are unknown physical constants, and k is the number of input variables. Deterministic calibration approaches are generally plausible and easy to apply, but they cannot account for uncertainties in both engineering model simulation and physical experimentation.

5.2.2.2 Non-deterministic Bayesian calibration approach

Non-deterministic parameter calibration is also called calibration under uncertainty (CUU) [37]. Kennedy and OHagan [21] first developed a Bayesian approach to simultaneously calibrate a engineering model and characterize the potential bias (discrepancy) between the model output and the physical experiments. Their method is based on the following relation,

$$y(x) = \nu \cdot y_m(\mathbf{x}, \boldsymbol{\theta}) + \omega(\mathbf{x}) + \varepsilon, \tag{5.2.3}$$

where ν is an unknown regression parameter (an adjustment parameter), $\omega(\mathbf{x})$ is a bias (discrepancy) function assumed to be the realization of a Gaussian Process, ε

is the experimental error assumed to be a zero-mean Gaussian random variable. In essence, the formulation shown in Eq. (5.2.3) is a combination of both parameter calibration and bias-correction. Several variants and applications of Kennedy and OHagans approach [21] exist in the literature.

5.3 Methodology

5.3.1 Determining adaptive parameter (κ)

Layers of Experiments (LoE) are multi-stage experiments, each stage of whose has different size of design space called a layer. Generally, the size of layer gets smaller sequentially to carry out experiments in more interested region in terms of experimental goal.

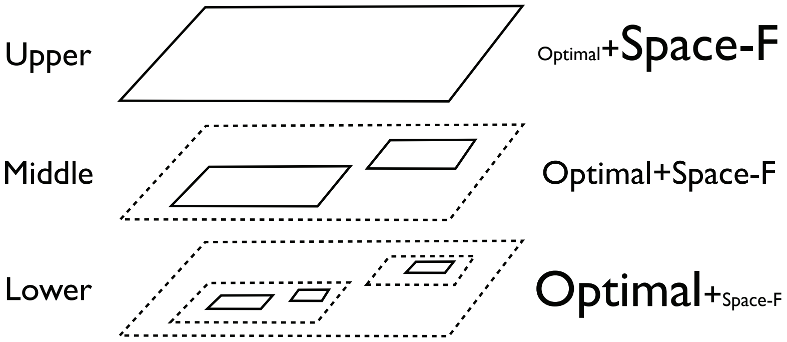


Figure 27: The concept of adaptive design in three layers of experiments

LoE employs combined designs for efficient data collection. That is, as layers move to more focused local region, combined designs have the properties of optimal designs. Figure 27 illustrates the concept of combined design in the LoE. The weight between optimal design criteria and space-filling design criteria changes over layers. That is, it is adaptive to data collected in the previous layer.

Adaptive design has been studied much mainly in clinical research. There are many different types of adaptiveness (see [10], and literature cited therein), but we restrict our attention about adaptiveness to the modification in design features in combined designs.

The value of κ plays a role to control weight between optimal design and space-filling design in combined design criteria. The combined design is developed to be adaptive to model uncertainty through the value of κ. Thus, the value of κ depends on the model uncertainty in a certain design space. Since uncertainty measurement is proposed in Eq. (3.3.5),

$$L(x) = \max \left\{ \left| T - (\hat{y}(x) + CI(x)) \right|, \left| T - (\hat{y}(x) - CI(x)) \right| \right\},$$

the problem is now to link it to κ value.

There are two conditions for κ. First, the value of κ should be between $[0, 1]$. Second, as defined in a combined design criteria, $\kappa = 0$ makes it a pure space-filling criterion while it becomes a pure optimal criterion when $\kappa = 1$. Hence, the combined design has more space-filling property as κ is close to zero and more optimal property as κ is close to one.

For each layer, we can compute L_k^*, $k = 1, \ldots, n_l$. L_k^* is defined by

$$L_k^* \equiv L_k(x^*),$$

where $x^* = \arg\min_x \hat{y}_k(x)$ and $\hat{y}_k(x)$ is a fitted model in the kth layer.

We conduct experiments in the sequentially zoomed-in design space until the evaluation metric meets tight tolerance requirement. Then, evaluation metric, L_k^*, in the kth layer should be less than the one of previous layer and greater than the one of next layer.

$$L_{k-1}^* < L_k^* < L_{k+1}^* \tag{5.3.1}$$

As more information is gathered, the amount of uncertainty should decrease. If not, there is no reason to collect more data to conduct additional experiments.

To re-scale the evaluation metric into $[0, 1]$, we need upper bound and lower bound of L^*. According to Eq. (5.3.1), the evaluation metric in the initial layer is the largest one and the smallest one should be less than or equal to d. For simplicity, assume that $L^*_{n_l} - d = 0$ in the lowest (last) layer. Then κ_k in the kth layer can be computed as

$$\kappa_k = 1 - \frac{\max\{0, L^*_{k-1} - d\}}{L^*_0 - d}, \tag{5.3.2}$$

where L^*_0 is the evaluation metric with simple mean model $y(x) = \mu + \varepsilon$. In this way, the value of κ reflects the uncertainty in a certain layer, and also satisfies two conditions above.

5.3.2 Combining information from various layers

Combined design developed in the previous chapter is also appropriate to utilize information from various layers. Once subsequent design space (R_B) is decided, data collected in the previous layers (R_A) should be used to construct the design in the subsequent layer (R_B). How to reflect given information into a subsequent design is an important issue for efficiency in design of experiments.

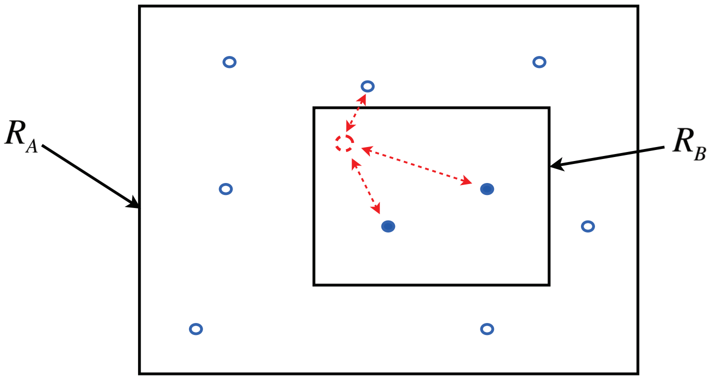


Figure 28: The concept of combining information from various layers. Open blue circles composite a design ξ_A in the R_A. Closed blue circles are design points of ξ_A in the R_B.

Suppose that a design ξ_A is given in the previous layer (R_A), which means $\xi_A \subset R_A$. Data are collected on ξ_A and R_B is determined based on the collected data. Assume that a set S_{AB} is not an empty set, where

$$S_{AB} = \{x : (x \in \xi_A) \cap (x \subset R_B)\} \neq \emptyset.$$

closed blue circles indicate the elements of S_{AB} in Figure 28. Now, we present the way to combine information for D-optimal design criteria and minimax design criteria, respectively.

Given S_{AB}, D-optimal designs in the R_B maximizes the determinant of the modified information matrix,

$$M(\xi) = \int_{R_B} \mathbf{f}(x)\mathbf{f}^T(x)d\xi(x) + \int_{x \in S_{AB}} \mathbf{f}(x)\mathbf{f}^T(x)d\xi(x). \tag{5.3.3}$$

Similarly, a minimax design given S_{AB} in the R_B is

$$\xi^* = \min_{\xi \subset R_B} \max_{x \in R_B} \rho(x, \xi \cup \xi_A), \tag{5.3.4}$$

where

$$\rho(x, \xi \cup \xi_A) = \min_{x_i \in \{\xi \cup \xi_A\}} \tau(x, x_i)$$

Eq. (5.3.4) is modified from Eq. (4.2.3) in that ξ is replaced by $\xi \cup \xi_A$. Given information ξ_A is added in minimax design criteria. Note that we use $\xi \cup \xi_A$ instead of $\xi \cup S_{AB}$. This is because some data points just outside of R_B are also useful for space-filling design. For example, in Figure 28, the left-upper corner of R_B has been already explored much by a data point located in the just outside of left-upper corner of R_B. So, in space-filling point of view, left-upper corner of R_B may be less attractive as the location of new design point.

Thus, we can easily modify optimal criteria and space-filling criteria as Eq. (5.3.3) and (5.3.4) and construct combined criteria as we proposed in the previous chapter. The proposed modified design criteria present a flexible way to combine information from various layers regardless from upper layer (zoom-in procedure) or lower layer (zoom-out procedure). Also, this method is applicable in the irregular design space.

5.4 Layers of Experiments with Engineering Models

Step 1) Postulate a Tentative Statistical Model: Low-order polynomial models are used.

$$y(x) = \sum_{i=1}^{p} \gamma_i f_i(x) + \varepsilon, \tag{5.4.1}$$

where $\varepsilon \sim N(0, \sigma^2)$.

Step 2) Plan an Experiment and Collect the Data: With computed κ from upper layer, find design ξ using proposed combined design criteria in Eq. (4.3.2). Collect $y(\xi)$ from physical experiments at the design points of ξ. An updated engineering model is given from upper layer and a simulator can compute $y_m(\xi)$ at ξ where physical experiments are conducted.

Step 3) Update Engineering Model: Estimate the bias function $\omega(\xi; \boldsymbol{\beta})$ by

$$\text{Find } \boldsymbol{\beta} \text{ Minimizing SSE } = \sum_i \left[y(\xi_i) - y_m(\xi_i) - \omega(\xi_i; \boldsymbol{\beta})\right]^2,$$

Once $\omega(\mathbf{x}_i; \boldsymbol{\beta})$ is estimated, update engineering model by (5.1.1).

Step 4) Use the Data and the Updated Engineering Model to build the Engineering Adjusted Statistical Model: We estimate the parameters in the model (5.4.1) and obtain $\hat{y}$. Both the observations (field data) from physical experiments and the adjusted engineering model outputs are obtained and they are valuable information to build an accurate statistical model. In addition to the field data $\mathbf{x}_f$, collect grid points $\mathbf{x}_m$ from updated $\hat{y}_m$ to build an *engineering adjusted statistical model* ($\hat{y}_s$). For ease to apply design criteria, a statistical model is restricted to polynomial regression models. We recommend the order of polynomial regression model does not exceed the order which the size of $\mathbf{x}_f$ allows.

Step 5) Check the Accuracy of Prediction: The accuracy of the fitted model $\hat{y}$ can be measured by evaluation metric (Eq. (3.3.5)). If prediction is sufficiently accurate for the required tolerance, go to step 7; otherwise proceed to step 6. At this step, the threshold, κ, in combined criteria (Eq.(4.3.2)) is updated for lower layer.

Step 6) Choose a Subregion for the Next Experiment: An optimization routine can be used to find the center of the new subregion, while the sensitivity analysis can be used for choosing new limits for the new subregion. Then repeat steps 1 to 5, with data drawn from the new subregion.

Step 7) Find Optimal Process Conditions: Statistical model in the lowest layer is accurate enough to meet tight tolerance requirement. Also, the lowest layer is expected to include process optimum. Using fitted statistical mode, we find optimal process conditions.

Figure 29 summarizes the layers of experiments procedures.

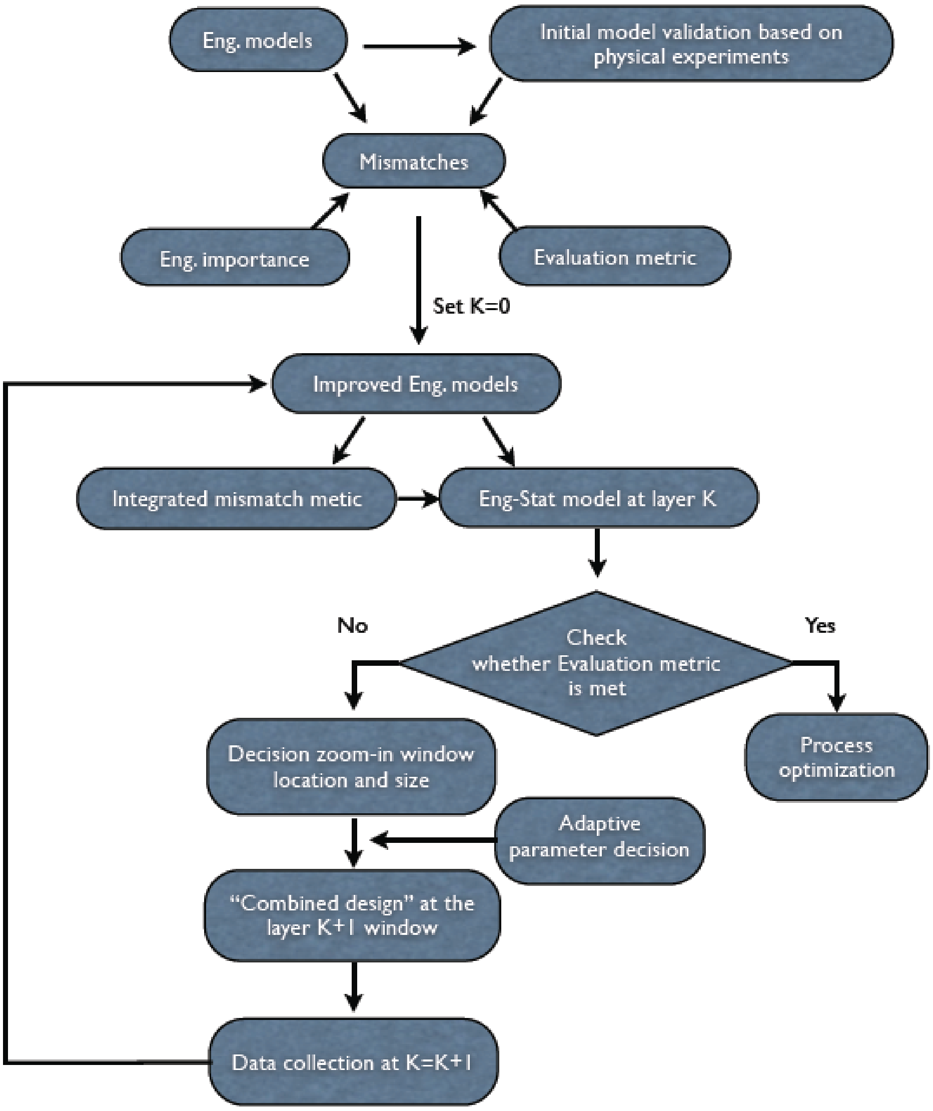

Eng. models
Initial model validation based on physical experiments
Mismatches
Eng. importance
Evaluation metric
Set K=0
Improved Eng. models
Integrated mismatch metic
Eng-Stat model at layer K
No
Check whether Evaluation metric is met
Yes
Decision zoom-in window location and size
Process optimization
Adaptive parameter decision
"Combined design" at the layer K+1 window
Data collection at K=K+1

Figure 29: Layers of Experiments

5.5 Illustrative Example

The purpose of illustrative examples is to justify the proposed methodology, Layers of Experiments (LoE), by showing its performance when the true response is complex and engineering models are biased. Example 1 is about combined design criteria. We show that the threshold, κ, varies depending on engineering model bias.

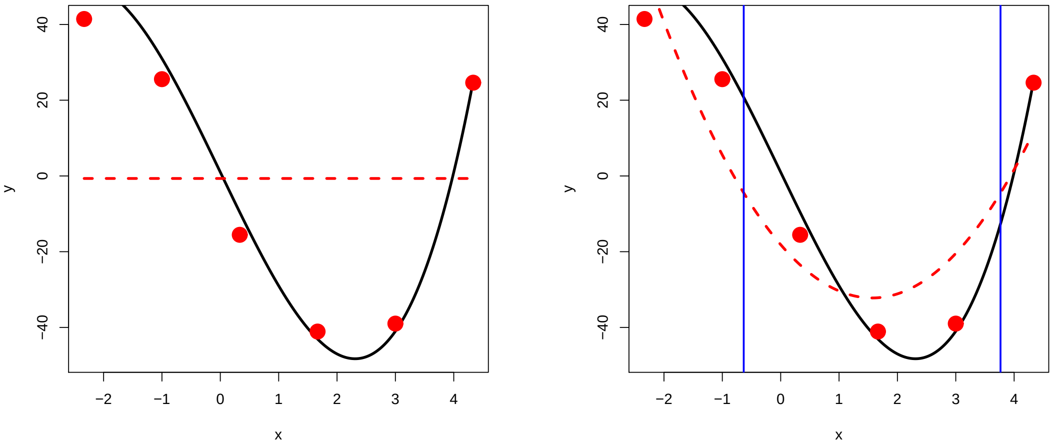
5.5.1 Example 1

The concept of the layer of experiments can be illustrated through a single variable cubic function, $f(x) = 2x^3 - 32x + 1 + \sigma$, where $\sigma \sim N(0, 5)$. $f(x)$ is used to represent a computation-intensive design function. In the fist layer design space $[-3, 5]$, six experimental points (a combined design with $\kappa = 0$) are obtained to approximate the unknown curve with 2nd order polynomial regression model, as illustrated in Figure 30(b). However, its evaluation metric ($L_1^* = 69.01$) is much larger than tolerance requirement ($d = 10$). Thus, we decided to conduct six additional experiments in the second layer. The zoom-in procedure is applied, which is introduced in the previous chapter. The new design space is now between the points $[-0.63, 3.76]$.

Using Eq. (5.3.2) we can find $\kappa_2 = 0.215$ for the new layer,

$$\kappa_k = 1 - \frac{\max\{0, L_{k-1}^* - d\}}{L_0^* - d},$$

where L_0^* is the evaluation metric with simple mean model $y(x) = \mu + \varepsilon$ as illustrated in Figure 30(a).

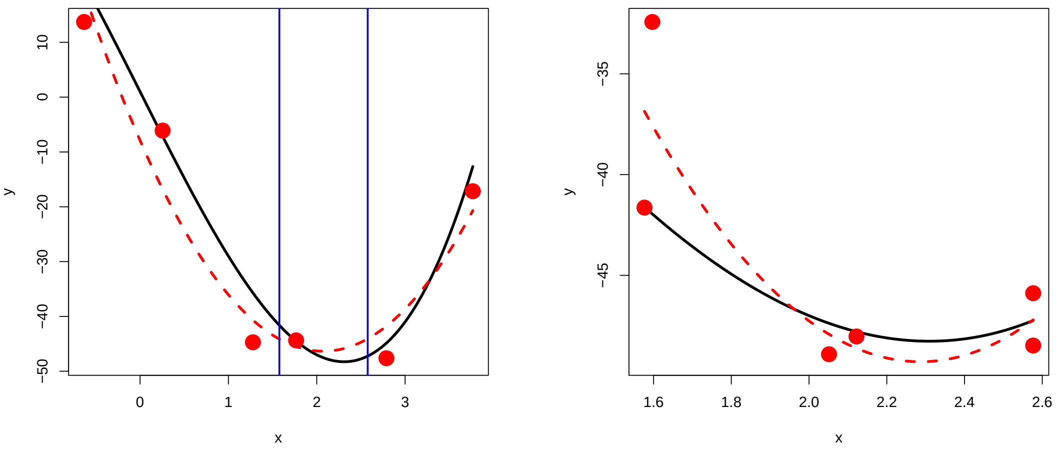


(a) L_0^* is the evaluation metric with simple mean model $y(x) = \mu + \varepsilon$

(b) First layer: A combined design with $\kappa = 0$ and a new design space $(-0.63, 3.76)$ for second layer

Figure 30: Six points of combined design with $\kappa = 0$ (a minimax design) in the first layer $[-3, 5]$.

Then, we applies the combined design with $\kappa_2 = 0.215$ again over the reduced design space $[-0.63, 3.76]$, and produces a second fitted model shown in Figure 31(a). This second fitted model yields a much better evaluation metric ($L_2^* = 19.65$), but it is not enough close to the tight tolerance requirement ($d = 10$). By continuing this process, we zoom-in again for third layer $[1.58, 2.58]$ and compute $\kappa_3 = 0.872$.

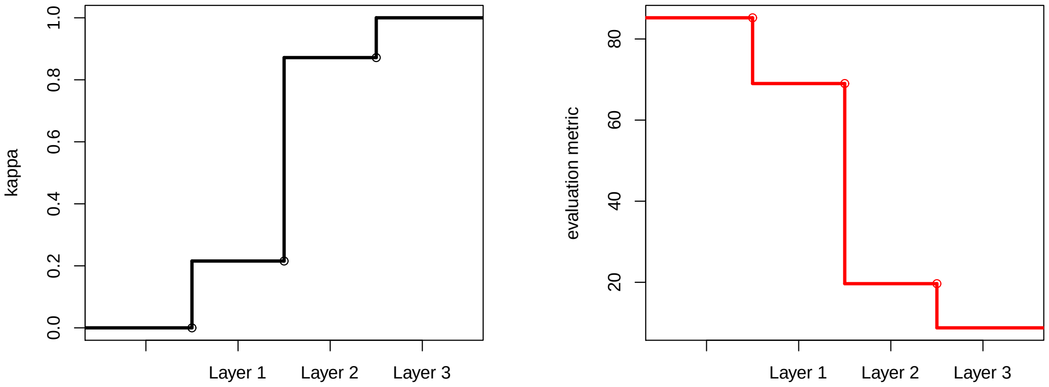


(a) Second layer: A combined design with $\kappa = 0.215$ and a new design space $(1.58, 2.58)$ for third layer

(b) Third layer: A combined design with $\kappa = 0.872$ and finally meet the tolerance requirement

Figure 31: Six points of combined designs in the second layer and third layer.

In the third layer, the evaluation metric ($L_3^* = 8.77 < 10$) finally meets tolerance requirement. So, we stop further experiments and find the optimal condition from the final accurate statistical model.



(a) The value of κ for each layer

(b) The value of evaluation metric L^* for each layer

Figure 32: κ value decision and corresponding evaluation metric in the three layers of experiments

Figure 32(a) illustrates how the values of κ changes over three layers. As we defined before, κ value measures the amount of uncertainty existing in a certain layer. As κ value moves from 0 to 1, corresponding combined designs changes from space-filling to optimal designs, and evaluation metric gets improved as shown in Figure 32(b).

5.5.2 Example 2

Suppose that a true response function $f(x)$ is a nonlinear complicated function of x. Observations (y) of physical experiments from the true response function may be modeled by

$$y = f(x) + \sigma,$$

where $\sigma \sim N(0, 5)$. In the first layer design space $[-18, 28]$, six experimental points (a combined design with $\kappa = 0$) are obtained as a space-filling design (minimax design in this example).

$$\xi_1 = \{-14.17, -6.50, 1.17, 8.83, 16.50, 24.17\}$$

The design is used to approximate the unknown true function

$$f(x) = 10^{-4}\left(-x(x-1)(x-5)(x-7)(x-10) + 200(x-5)^3 + 10^{-4}\exp(x)\right)$$

with 2nd order polynomial regression model, as illustrated in Figure 33. However, its evaluation metric ($L_1^* = 82.53$) is much larger than tolerance requirement ($d = 10$). Thus, we decided to conduct six additional experiments in the zoomed-in layer. The proposed zoom-in procedure is applied and the it yields $L_2 = [-11.86, 15.94]$. Since the set S_{12} defined in the previous section is not an empty set,

$$S_{12} = \{x : (x \in \xi_1) \cap (x \subset L_2)\} = \{-6.50, 1.17, 8.83\} \neq \emptyset,$$

we are able to utilize the collected data in previous layer to construct a combined design in a next layer.

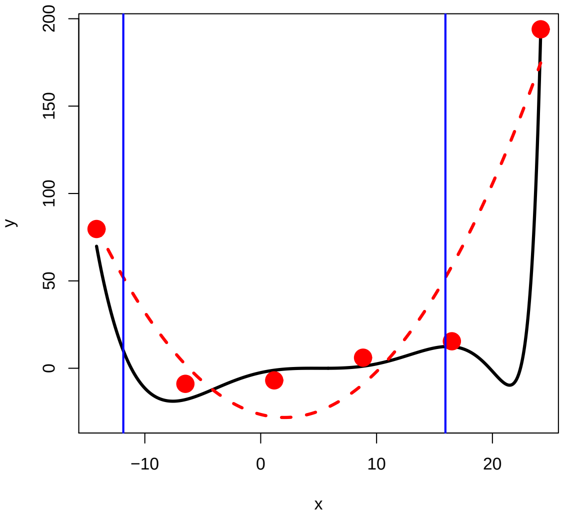


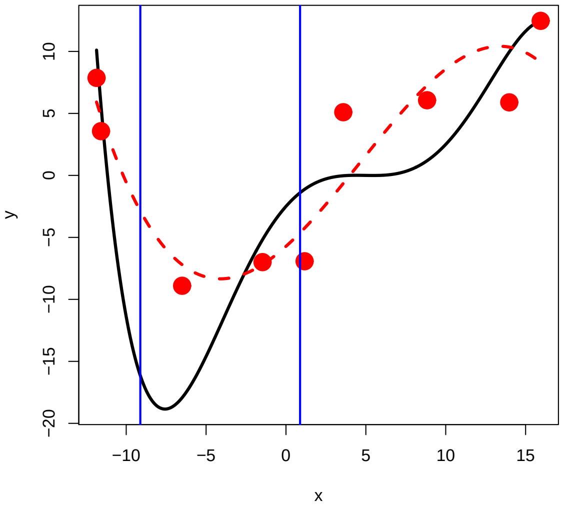
Figure 33: First layer: A combined design with $\kappa = 0$ and a new design space $(-11.86, 15.94)$ for the second layer. Red dashed line indicates statistical model approximation using 2nd order polynomial regression model. Black solid line represents the true model. Three points ($(-6.50, -8.90), (1.17, -6.92), (8.83, 6.06)$) are useful in the second layer

Then, with S_{12} and $\kappa_2 = 0.476$ obtained from Eq. (5.3.2), we apply the combined design again over the zoomed-in design space $L_2 = [-11.86, 15.94]$, and approximate the unknown $f(x)$ with 3rd order polynomial regression model as shown in Figure 34(a). Note that the three points in S_{12} affect the combined design in the L_2. The combined design with κ_2,

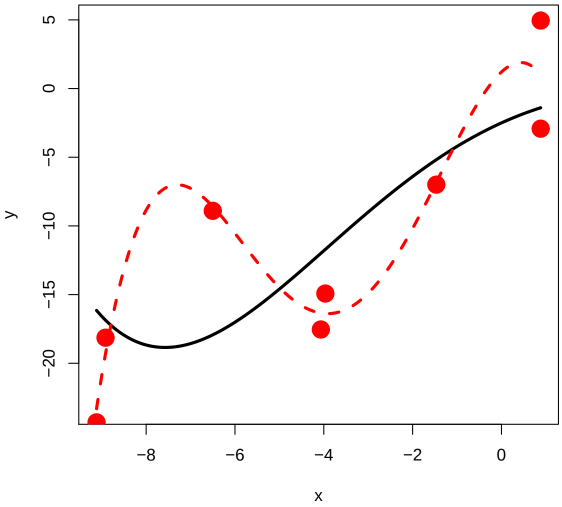
$$\xi_2 = \{-11.86, -1.47, -11.58, 3.59, 13.98, 15.943\},$$

is designed by the method of combining information from various layers explained in the previous section.

This fitted model yields a much better evaluation metric ($L_2^* = 17.40$), but it is not enough close to the tight tolerance requirement ($d = 10$). So, we decide to conduct next layer experiments.



(a) Second layer: A combined design with $\kappa = 0.476$ and a new design space $(-9.12, 0.88)$ for third layer. Red dashed line indicates statistical model approximation using 3rd order polynomial regression model.



(b) Third layer: A combined design with $\kappa = 0.91$ and finally meet the tolerance requirement. Red dashed line indicates statistical model approximation using 4th order polynomial regression model.

Figure 34: Six points of combined designs in the second layer and third layer.

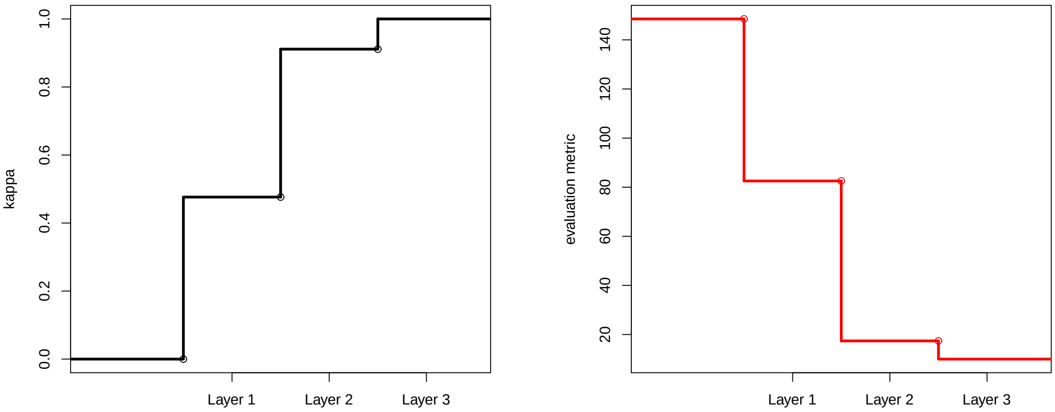
By continuing this process, we zoom-in again for third layer $L_3 = [-9.12, 0.88]$ and compute $\kappa_3 = 0.91$. With $S_{23} = \{-1.47, -6.50\}$ and κ_3, the combined design in the L_3 is

$$\xi_3 = \{-8.91, -4.067, -9.12, -3.97, 0.88, 0.88\}.$$

In the third layer, the evaluation metric from approximated 4th order polynomial regression model (see Figure 34(b)) finally meets tolerance requirement,

$$L_3^* = 9.99 < 10$$

Therefore, we stop further experiments and find the optimal condition from the final accurate statistical model.



(a) The value of κ for each layer

(b) The value of evaluation metric L^* for each layer

Figure 35: κ value decision and corresponding evaluation metric in the three layers of experiments

Figure 35(a) illustrates how the values of κ changes over three layers. As we defined before, κ value measures the amount of uncertainty existing in a certain layer. As κ value moves from 0 to 1, corresponding combined designs changes from space-filling to optimal designs, and evaluation metric gets improved as shown in Figure 35(b).

5.5.3 Example 3

In this example, we show the procedure to build engineering adjusted statistical model. Six field data are obtained from physical experiments in the design space $[-15, 24]$ and engineering model $(y_m(x))$ is given as shown in Figure 36. We follow the procedure proposed in [19] to check whether the given engineering model is adequate or not. After testing the adequacy of the engineering model, it confirms that the engineering model is not good. So, we proceed to adjust engineering model with field data.

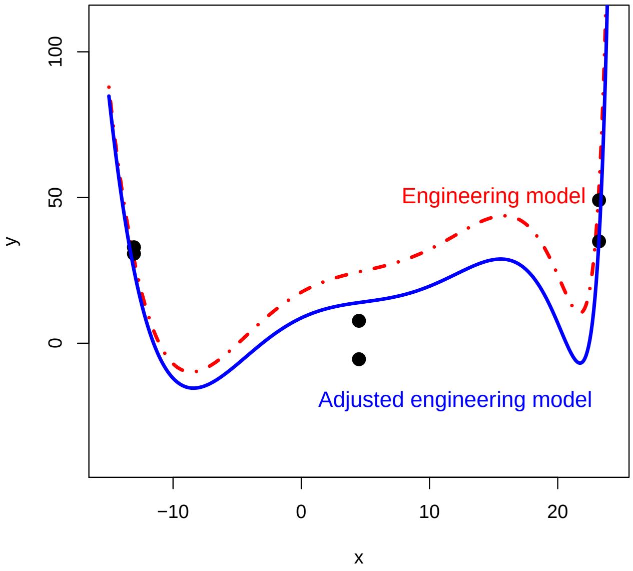
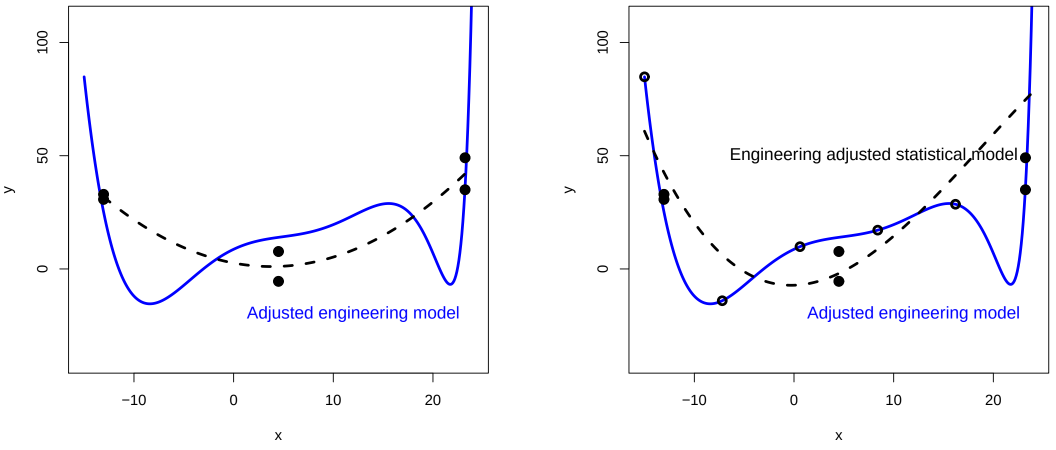


Figure 36: Adjusted engineering model

Using the constant adjustment model (5.1.1) for $y(x) - y_m(x)$ on $y_m(x) - y_m$, we obtain $\hat{\beta}_0 = -10.693$ and $\hat{\beta}_1 = 0.076$. Hereafter, the adjusted engineering model is

$$\hat{y}_m = y_m(x) - 10.693 + 0.076(y_m(x) - \bar{y}_m), \qquad (5.5.1)$$

A plot of the adjusted engineering model in shown in Figure 36, which clearly shows great improvement compare to the original engineering model. That is, original engineering model is adjusted by field data to reduce mismatch between the engineering model and field data. To quantify the performance, we can carry out the model inadequacy test as before.



(a) Statistical model relies only on field data for estimation

(b) Engineering adjusted statistical model utilizes field data as well as simulated data from adjusted engineering model

Figure 37: Statistical model vs. Engineering adjusted statistical model

Without the engineering model, statistical model estimation relies only on collected field data. The quality of the statistical model depends on the number of data and its quality. If physical experiments for the field data are time-consuming or expensive, we cannot expect qualified statistical model (see Figure 37(a)).

Now, we combine information from adjusted engineering model to estimate statistical model. In addition to six field data, six different type of data are collected from the adjusted engineering model as shown in Figure 37(b). Dash line in Figure 37(b) depicts statistical model estimated by both field data and the adjusted engineering model. We call it the engineering adjusted statistical model. Compared to the dash line in Figure 37(a), the engineering adjusted statistical model is clearly more accurate than the one without using information from engineering model.

5.6 Conclusion

In the process optimization point of view, it is crucial to select design points near the optimal regions. If one fails to have design points around optimal regions, there is no hope to find appropriate optimal conditions. However, the given resources are limited and so one should allocate enough resources to important regions. We proposed a systematic procedure to give more weight of using given resources on the optimal regions. We called it 'Layers of Experiments'. As layers go further, the uncertainty of underlying model gets decreased and the region of interest gets restricted. We employed combined design criteria: one from optimal design criteria and one from minimum energy criteria in the Layers of Experiments.

The *engineering adjusted statistical model* is a statistical model based on both field data and updated engineering model in layers of experiments. To build an accurate *engineering adjusted statistical model*, both characteristics (design space exploration and accuracy in statistical inference) are required. So, combined designs are appropriate for the *engineering adjusted statistical model.* The adaptive parameter (κ) in the combined design criteria controls the weight between the two criteria. The value of κ is adaptive to model uncertainty of each layer. Thus, as layers go further, the combined design criterion moves from space-filling criterion to optimal criterion.

We proposed the method to determine adaptive parameters sequentially based on uncertainty at each layer. However, the proposed method to determine κ is an practical guideline, rather than rigorous way from a statistical perspective. Future study on the property of adaptive parameter is needed. Also modified combined design criteria are presented to improve its efficiency by combining information from various layers.

REFERENCES

[1] *2005 NCMS Survey of Nanotechnology in the U.S.* National Center for Manufacturing Science, Manufacturing Industry, National Center for Manufacturing Science, Ann Arbor, MI., 2006.

[2] Atkinson, A. C., "Dt-optimum designs for model discrimination and parameter estimation," *Journal for Statistical Planning and Inference*, vol. 138, pp. 56–64, 2008.

[3] Bernardo, M. C., Buck, R., Liu, L., Nazaret, W. A., Sacks, J., and Welch, W. J., "Integrated circuit design optimization using a sequential strategy," *IEEE Transactions on Computer-aided Design*, vol. 11, no. 3, 1992.

[4] Box, G., "Measures of lack of fit for response surface designs and predictor variable transformations," *Technometrics*, vol. 23, pp. 1–8, 1982.

[5] Box, G. and Draper, N., "A basis for the selection of response surface design," *Journal of the American Statistical Association*, vol. 54, pp. 622–653, 1959.

[6] Chrastil, J., "Solubility of solids and liquids in supercritical gases," *Journal of Phys. Chem.*, vol. 86, pp. 3016–3021, 1982.

[7] Cook, D. and Fedorov, V., "Constrained optimization of experimental-design," *Statistics*, vol. 26, no. 2, pp. 129–178, 1995.

[8] Cook, R. D. and Nachtsheim, C. J., "A comparison of algorithms for constructing exact *d*-optimal design," *Technometrics*, vol. 22, pp. 315–324, 1980.

[9] Dasgupta, T., C. M. e. a., "Statistical modeling and analysis for robust synthesis of nanostructures," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 594–603, 2008.

[10] Dragalin, V. and Fedorov, V., "Adaptive designs for dose-finding based on efficacy- toxicity response," *Journal of Statistical Planning and Inference*, vol. 136, pp. 1800–1823, 2006.

[11] Easterling, R. and Berger, J., "Statistical foundations for the validation of computer models," *Presented at Computer Model Verification and Validation in the 21st Century Workshop*, p. Johns Hopkins University, 2002.

[12] Fang, K., Lin, D., Winker, P., and Zhang, Y., "Uniform design: Theory and application," *TECHNOMETRICS*, vol. 42, no. 3, pp. 237–248, 2000.

[13] Fang, K. and Wang, Y., *Number-theoretical Methods in Statistics.* Chapman & Hall, London, 1994.

[14] Fedorov, V. . V., *Theory of optimal experiments.* Preprint No. 7 LSM, Izd-vo Moscow State University, Moscow, USSR, 1969.

[15] Goel, T., Haftka, R. T., Shyy, W., and Watson, L. T., "Pitfalls of using a single criterion for selecting experimental designs," *INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN ENGINEERING*, vol. 75, pp. 127–155, 2008.

[16] Han, C., "A note on optimal designs for a two-part model," *Statistics & Probability Letters*, vol. 65, pp. 343–351, 2003.

[17] Hasselman, T., Yap, K., Lin, C., and Cafeo, J., "A case study in model improvement for vehicle crashworthiness simulation," *23rd International Modal Analysis Conference, Orlando, Florida*, p. January 31February 3, 2005.

[18] Johnson, M. E., Moore, L. M., and Ylvisaker, D., "Minimax and maxmin distance design," *Journal for Statistical Planning and Inference*, vol. 26, pp. 131–148, 1990.

[19] Joseph, V. R. and Melkote, S. N., "Statistical adjustments to engineering models," *Journal of Quality Technology*, vol. 41, no. 4, pp. 362–375, 2009.

[20] Joseph, V. R. and Hung, Y., "Orthogonal-maximin latin hypercube designs," *Statistica Sinica*, vol. 18, pp. 171–186, 2008.

[21] Kennedy, M. and O'Hagan, A., "Bayesian calibration of computer models(with discussion)," *Journal of Royal Statistical Society-Series B*, vol. 63, pp. 425–264, 2001.

[22] Läuter, E., "Experimental planning in a class of models," *Mathematische Operationsforschung und Statistik*, vol. 5, pp. 673–708, 1974.

[23] Läuter, E., "Optimal multipurpose designs for regression models," *Mathematische Operationsforschung und Statistik*, vol. 7, pp. 51–68, 1976.

[24] Lindgren, L.-E., Alberg, H., and Domkin, K., "Constitutive modelling and parameter optimization," *7th International Conference on Computational Plasticity, Barcelona, Spain*, p. April 710, 2003.

[25] Lu, J.-C., Jeng, S.-L., and Wang, K., "A review of statistical methods for quality improvement and control in nanotechnology," *Journal of Quality Technology*, vol. 41, no. 2, 2009.

[26] Manning, W., Morris, C., Newhouse, J., Orr, L., Duan, N., Keeler, E., Leibowitz, A., Marquis, K., Marquis, M., and Phelps, C., *A Two-Part Model of the Demand for Medical Care: Preliminary Results From the Health Insurance Experiments.* J. van der Gaag and M. Perlman, Editors, Health, Economics and Health Economics, North Holland, Amsterdam, 1981.

[27] McKay, M.D., R. J. B. and Conover, W. J., "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, no. 2, pp. 239–245, 1979.

[28] Montgomery, D., *Design and Analysis of Experiments, seventh edition.* John Wiley & Sons, New York, 2009.

[29] Morris, M. D. and Mitchell, T. J., "Exploratory designs for computational experiments," *Journal of Statistical Planning and Inference*, vol. 43, no. 3, pp. 381–402, 1995.

[30] Olsen, M. K. and Schafer, J. L., "A two-part random-effects model for semi-continuous longitudinal data," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 730–745, 2001.

[31] Owen, A., "Controlling correlations in latin hypercube samples," *Journal of the American Statistical Association*, vol. 89, pp. 1517–1522, 1994.

[32] Qian, Z. and Wu, C., "Bayesian hierarchical modeling for integrating low-accuracy and high accuracy experiments, in: Twelfth annual international conference on statistics," *Combinatorics, Mathematics and Applications, Auburn, AL, December 24*, 2005.

[33] Shi, J., *Stream of Variation Modeling and Analysis for Multistage Manufacturing Processes.* CRC Press, Baca Raton, FL., 2006.

[34] Smith, E., *Uncertainty analysis: Encyclopedia of Environmetrics.* John Wiley & Sons, Ltd, Chichester, 2002.

[35] Steinberg, D. and Hunter, W., "Experimental design: Review and commnet," *Technometrics*, vol. 26, no. 2, 1984.

[36] Tang, B., "Orthogonal array-based latin hypercubes," *Journal of the American Statistical Association*, vol. 88, pp. 1392–1397, 1993.

[37] Trucano, T., Swilera, L., Igusab, T., Oberkampf, W., and Pilch, M., "Calibration, validation, and sensitivity analysis: Whats what," *Reliab. Engrg. Syst. Safe.*, vol. 91, p. 13311357, 2006.

[38] Wang, S., Chen, W., and Tsui, K., "Bayesian validation of computer models," *Technometrics*, vol. 51, no. 4, 2009.

[39] Wissmann, P. J. and Grover, M. A., "A new approach to batch process optimization using experimental design," *AIChE J.*, vol. 55, no. 2, 2009.

[40] Wu, C. and Hamada, M., *Experiments: Planning, Analysis, and Parameter Design Optimization.* John Wiley & Sons, New York, 2000.

[41] Xu, S., Adiga, N., Ba, S., Dasgupta, T., Wu, C. F. J., and Wang, Z. L., "Optimizing and improving the growth quality of zno nanowire arrays guided by statistical design of experiments," *American Chemical Society, Nano*, vol. 3, no. 7, pp. 1803–1812, 2009.

[42] Zhao, H., Jin, R., Wu, S., and Shi, J., "Pde-constrained gaussian process model on material removal rate of wire saw slicing process," *Journal of Manufacturing Science and Engineering*, vol. 133, no. 2, 2011.