Sequential Optimization through Adaptive Design of Experiments

by

Hungjen Wang

Submitted to the Engineering Systems Division
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Science in Engineering Systems
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
March 2007

Author .....
Engineering Systems Division
Mar 05, 2007

Certified by .....
Daniel D. Frey
Robert N. Noyce Career Development Professor
Thesis Supervisor

Certified by .....
Gordon M. Kaufman
Morris A. Adelman Professor of Management
Thesis Advisor

Certified by .....
Roy E. Welsch
Professor of Statistics and Management Science
Thesis Advisor

Accepted by .....
Richard de Neufville
Professor of Engineering Systems,
Chair, ESD Education Committee

Sequential Optimization through Adaptive Design of Experiments

by

Hungjen Wang

Submitted to the Engineering Systems Division
on Mar 05, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Science in Engineering Systems

Abstract

This thesis considers the problem of achieving better system performance through adaptive experiments. For the case of discrete design space, I propose an adaptive One-Factor-at-A-Time (OFAT) experimental design, study its properties and compare its performance to saturated fractional factorial designs. The rationale for adopting the adaptive OFAT design scheme become clear if it is imbedded in a Bayesian framework: it becomes clear that OFAT is an efficient response to step by step accrual of sample information. The Bayesian predictive distribution for the outcome by implementing OFAT and the corresponding principal moments when a natural conjugate prior is assigned to parameters that are not known with certainty are also derived.

For the case of compact design space, I expand the treatment of OFAT by the removal of two restrictions imposed on the discrete design space. The first is that the selection of input level at each iteration depends only on observed best response and does not depend on other prior information. In most real cases, domain experts possess knowledge about the process being modeled that, ideally, should be treated as sample information in its own right-and not simply ignored. Treating the design problem Bayesianly provides a logical scheme for incorporation of expert information. The second removed restriction is that the model is restricted to be linear with pairwise interactions – implying that the model considers a relatively small design space. I extend the Bayesian analysis to the case of generalized normal linear regression model within the compact design space. With the concepts of **c**-optimum experimental design and Bayesian estimations, I propose an algorithm for the purpose of achieving optimum through a sequence of experiments. I prove that the proposed algorithm would generate a consistent Bayesian estimator in its limiting behavior. Moreover, I also derive the expected step-wise improvement achieved by this algorithm for the analysis of its intermediate behavior, a critical criterion for determining whether to continue the experiments.

Thesis Supervisor: Daniel D. Frey

Title: Robert N. Noyce Career Development Professor

Thesis Supervisor: Gordon M. Kaufman
Title: Morris A. Adelman Professor of Management

Thesis Supervisor: Roy E. Welsch
Title: Professor of Statistics and Management Science

Acknowledgments

I would like to thank my committee for staying with me for the entire course of my PhD study. Professor Frey has been the most supportive thesis supervisor, who constantly reminds me what my thesis is about and who always listens and helps. Professor Kaufman, my research mentor, guided me through this onerous process with good research discipline, insightful ideas, and great patience. I would like to thank Professor Welsch for his thorough thinking that helps me stay focused and keeps me in check with the details. I am extremely fortunate to have worked with such a great committee.

I am grateful for the company and friendship of my ESD and Robust Lab fellows. Many moments of laughers and disappointments I have shared with you.

Finally, I would like to thank every one of my family, who had made me who I am today and gives me unconditional love and endless understanding and support.

Contents

1 Introduction **15**

1.1 Motivation 15

1.2 Literature Review 16

1.2.1 Stochastic Approximation Problem 16

1.2.2 Multi-Armed Bandit Problem 18

1.2.3 Bayesian Experimental Design 20

1.2.4 Response Surface Methodology 23

1.3 Structure of the Thesis 24

2 The Discrete Space and the One-Factor-at-A-Time Experimental Scheme **27**

2.1 The Fundamental Model 27

2.2 A Bayesian Analysis for Adaptive Experiments 28

2.2.1 After the first experiment 30

2.2.2 After the second experiment 31

2.2.3 After k experiments with $3 \leq k \leq q$ 35

2.3 Adaptive One-Factor-at-A-Time Experimental Scheme 35

2.3.1 The Processes of Adaptive OFAT Scheme 36

2.4 Preposterior Analysis of OFAT Scheme 40

2.4.1 The First Step in Adaptive OFAT 40

2.4.2 The Second Step in Adaptive OFAT 42

2.4.3 Subsequent Steps in Adaptive OFAT 43

2.5 Comparison with Fractional Factorial Experiments 44

2.6 The Posterior Analysis 47
2.6.1 Posterior Analysis for OFAT Scheme with Unknown σ_ϵ^2 47
2.6.2 Posterior Analysis for Saturated Fractional Factorial Experimental Scheme with Unknown σ_ϵ^2 48
2.6.3 A Brief Discussion 51
2.7 Proofs 51

3 The Optimization Scheme Through Sequential Experiments in Compact Design Space 59
3.1 The Fundamental Model 60
3.2 The Proposed Algorithm 60
3.2.1 A Brief Discussion On The Problem of Optimization Through Sequential Experiments 62
3.3 Bayesian Analysis with Known σ_ϵ^2 63
3.3.1 Consistency of the Bayesian Estimator $\mathbf{m}_k$ 63
3.3.2 Discussion for Selection of Sequence $\{\alpha_k\}$ 65
3.3.3 Expected Improvement in Each Iteration 67
3.3.4 An Illustrative Example 68
3.4 Bayesian Analysis with Unknown σ_ϵ^2 69

4 A Case Study for Adaptive OFAT Scheme and Sequential Response Optimization Scheme 75
4.1 Electric Powered Aircraft Flight Duration 75
4.1.1 Background for the case study 75
4.1.2 Adaptive OFAT Scheme 77
4.1.3 Sequential Response Optimization Scheme 81
4.2 Uniformity in the Manufacture of Integrated Circuits 89
4.2.1 Background for the case study 89
4.2.2 Adaptive OFAT Scheme 90
4.2.3 Sequential Response Optimization Scheme 94

5 Concluding Remarks **101**

A A Brief Summary of Bayesian Analysis for Normal Linear Regression Models **105**

A.1 Known Error Variance σ_ϵ^2 105

A.2 Unknown Error Variance σ_ϵ^2 108

List of Figures

2-1 The change of covariance with the increase of different numbers in input elements 30

2-2 Illustration of the adaptive OFAT applied to a system with three two-level factors (A, B, and C) 37

2-3 Comparison of $\mathbb{E}[y(\mathbf{x}_{\text{OFAT}})]$ and $\mathbb{E}[y(\mathbf{x}_{III})]$ 46

3-1 The Performance of Proposed Algorithm with $\sigma_\epsilon^2 = 1$ 72

3-2 The Performance of Proposed Algorithm with $\sigma_\epsilon^2 = 1$ 72

3-3 The Performance of Proposed Algorithm with $\sigma_\epsilon^2 = 0.5$ 73

3-4 The Performance of Algorithm Proposed byPronzato (2000) with $\sigma_\epsilon^2 =$ 1: Bad performance 73

3-5 The Performance of Algorithm Proposed byPronzato (2000) with $\sigma_\epsilon^2 =$ 1: Good performance 74

3-6 The Performance of Algorithm Proposed byPronzato (2000) with $\sigma_\epsilon^2 = 1$ 74

4-1 Flight Duration from OFAT under Various Experimental Error 79

4-2 Flight Duration from Fractional Factorial Experiments under Various Experimental Error 79

4-3 The Maximum Eigenvalue of Covariance Matrix $\mathbf{V}_k$ at Each Stage . . 84

4-4 Norm of Differences between Estimators of Two Consecutive Iterations 84

4-5 Estimated Optimal Response at each Iteration 85

4-6 True Response of the Estimated Optimum Levels of Factors 85

4-7 True Response of the Estimated Optimum Levels of Factors 86

4-8 Uniformity Capability from OFAT under Various Experimental Error 92

4-9 Uniformity Capability from Fractional Factorial Experiments under Various Experimental Error 92
4-10 Performance Comparison between OFAT and Fractional Factorial Schemes 93
4-11 The Maximum Eigenvalue of Matrix **V** for Every Iteration 98
4-12 Norm of Differences between Estimators of Two Consecutive Iterations 98
4-13 Norm of Differences between Estimated and True Parameters for Each Iteration 99
4-14 True Response of the Estimated Optimum Levels of Factors 99

List of Tables

4.1 The factors and levels in the electric powered aircraft experiment . . 76

4.2 The largest 12 effects in magnitude 77

4.3 The Response of the 2^7 Full Factorial Experiments 88

4.4 The factors and levels in the uniformity experiment 89

4.5 Analysis of Buckner et al. (1997) 2^{7-3}_{IV} Experiment source: (Mee and Peralta, 2000) 90

4.6 True and Estimated Parameters under Different Experiment Error Variance σ^2_ϵ 100

Chapter 1

Introduction

1.1 Motivation

Inspired by Box (1999) who pointed out that "there should be more studies of statistics from the dynamic point of view" and that "it is the investigation itself that must be regarded as the unit, and the success of the investigation that must be regarded as the objective," we consider the problem of developing mathematical results regarding adaptive experiments and the relationship between expected improvement and information gathered from adaptive experiments. More specifically, we would like to investigate whether such an adaptive experiment approach could attain better result and how much the improvement one could expect from one stage of experiment to the next. In fact, conducting experiments adaptively to achieve better system performance is in general discouraged by many statistical studies, especially in the field of Design of Experiments (see Wu and Hamada, 2000; Logothetis and Wynn, 1995). One key reason is that, in terms of regression, conducting experiments adaptively could not make unbiased estimates while keeping the variance of the estimates minimum. However, this goal could be easily achieved by many other experimental designs such as fractional factorial design (Wu and Hamada, 2000). Therefore, since the estimates from the adaptive experiments are less accurate, the expected value of the estimated best performance would not be the real optimum. This argument assumes that estimating unknown parameters with minimum variance would lead to

exploiting the best expected value of the performance. However, the study of (Frey et al., 2002; Frey, 2003), which indicate that conducting experiments adaptively could achieve better expected performance improvement through hundreds of case studies, renders a strong counterexample. Moreover, as discussed in (Lai, 2001) and the corresponding discussion by Wei, the roles of mean and variance in this kind of adaptive procedures still requires satisfactory theories.

1.2 Literature Review

The research topics related to the problem considered in this thesis are very rich. However, most of them could be regarded as a special case within the following four categories: Stochastic Approximation, Multi-Armed Bandit Problem, Bayesian Experimental Design, and Response Surface Methodology. The reason why we review these four topics is not only because they are closed related to the adaptive experiments but also because the mathematical techniques and the concepts developed in these topics are inspiring to us.

1.2.1 Stochastic Approximation Problem

The study of sequential design of experimentations could be traced back to (Robbins, 1952), in which the author proposed the idea of *stochastic approximation* and *multi-armed bandits*. The typical stochastic approximation (SA) problem could be described as follows. Consider a regression model

$$y_i = F(x_i, \theta) + \epsilon_i, \qquad i = 1, 2, \ldots$$

where y_i is an observed response contaminated with unobservable random error/noise ϵ_i, and $F(x_i, \theta) : \mathbb{R}^2 \to \mathbb{R}$ is a smooth regression function with unknown parameters θ and input x_i. The goal of the SA problem is to choose the root of the partial derivative of $F(x, \theta)$ with respect to the input x through a sequence of observation of response y_i's. Note that to find a maximum, minimum or a saddle point of a smooth

function $F : \mathbb{R}^2 \to \mathbb{R}$ is equivalent to find the x^* which satisfies $\frac{\partial F(x)}{\partial x}(x^*) = 0$. Hence, the SA problem sometime is termed as sequential optimization problem (Lai, 2001). To solve this problem, Robbins and Monro (1951) proposed the following recursive scheme

$$x_{n+1} = x_n + \alpha_n y_n$$

which has been generalized to (see Kushner and Yin, 2003)

$$x_{n+1} = x_n + \alpha_n \frac{\partial \hat{F}(x, \theta)}{\partial x}(x_n, y_n)$$

where $\frac{\partial \hat{F}(x,\theta)}{\partial x}(x_n, y_n)$ is the estimate of the gradient of the function $F(\cdot, \cdot)$ at x_n and the estimate is made according to the observation y_n. Moreover, the coefficients α_n have the property

$$\sum_{n=1}^{\infty} \alpha_n = \infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} \alpha_n^2 < \infty.$$

Later Blum (1954) proved that the recursive scheme in (Robbins and Monro, 1951) would converge to the root of the real partial function $\frac{\partial F(\cdot,\cdot)}{\partial x}$ with a strict assumption on error ϵ. Kiefer and Wolfowitz (1952) proposed the following recursive algorithm to deal with the SA problem

$$x_{n+1} = x_n + \alpha \frac{\partial \hat{F}(x, \theta)}{\partial x}(x_n, y_n)$$

where

$$y_n^{(i)} = F(x_n^{(i)}, \theta) + \epsilon_n^{(i)}, \quad i = 1, 2$$

are the two samples at the stage n experiment with inputs

$$x_n^{(1)} = x_n - \beta_n \qquad \text{and} \qquad x_n^{(2)} = x_n + \beta_n,$$

where the positive coefficients α_n and β_n satisfy

$$\sum_{n=1}^{\infty} \alpha_n = \infty, \quad \sum_{n=1}^{\infty} \left(\frac{\alpha_n}{\beta_n}\right)^2 < \infty, \quad \text{and} \quad \beta_n \to \infty,$$

and the estimate of the derivative of the smooth function $F(x, \theta)$ is defined as

$$\frac{\partial \hat{F}(x,\theta)}{\partial x}(x_n, y_n) = \frac{y_n^{(2)} - y_n^{(1)}}{2\beta_n}.$$

Blum (1954) also proved that the above recursive algorithm would converge to the maximum of $F : \mathbb{R}^2 \to \mathbb{R}$ under certain conditions on function F and error ϵ. After the schemes proposed by Robbins and Monro (1951) and Kiefer and Wolfowitz (1952), there is a vast literature of variants of the two recursive schemes, and the problem has soon been extended to multivariate case. That is, $F(\mathbf{x}, \theta) : \mathbb{R}^p \to \mathbb{R}$ with $p \geq 2$. For example, (Spall, 1992) and (Spall and Cristion, 1994) have proposed a perturbation gradient estimation scheme to estimate $\frac{\nabla F(\mathbf{x},\theta)}{\nabla \mathbf{x}}$ at each stage of experiments. See (Ruppert, 1991) for a detailed survey of algorithms related to the SA problems. Although the stochastic approximation problem is to find the maximizer of a smooth regression function through a sequential observations of responses contaminated with errors, a problem which is very close to the one we consider in this thesis, yet the convergence conditions is usually difficult to justify in advance, and the SA problem in general does not consider the case in which the design points lie in a constraint set. Most importantly, as indicated in (Lai, 2001), the interaction between the estimates of mean and variance through the sequential experiments remains unclear in the literature concerning the SA problem.

1.2.2 Multi-Armed Bandit Problem

The pioneering work of Robbins (1952) also introduces a new problem called the *multi-armed bandit* problem, which could be interpreted as follows. Imagine a slot machine with $q \geq 2$ slots. When one pulls a slot, say slot i, one would obtain a reward y_i which is a random variable with mean $\mu(\theta_i)$ and with a density function $f(\theta_i)$. In the setup of the multi-armed bandit problem, the density function $f(\theta_i)$ is usually known; however, the parameter θ_i remains unknown but belongs to some set Θ. The key difference between the multi-armed bandit problem and the stochastic approximation problem is that the latter deals with a continuous case where the design points could be

chosen from $\mathbb{R}$, while the former chooses the design points from a discrete set with $q \geq 2$ elements. The goal of the multi-armed bandit problem is to maximize the *total* expected rewards from the sequence of N trials. That is, to maximize $\mathbb{E}\left[\sum_{i=1}^{N} y_i\right]$. There is a fundamental dilemma between *exploration*, such as choosing different slots to estimate the parameter θ, and *exploitation*, such as determining which slot with the highest mean $\mu(\theta)$ so as to achieve highest $\mathbb{E}\left[\sum_{i=1}^{N} y_i\right]$.

After the multi-armed bandit problem has been proposed, this topic has been extensively studied in the fields of engineering, economics and statistics. In the engineering field, the extensive study of the bandit problem was motivated by the adaptive control of finite state Markov chains with a finite control set (Bertsekas, 2001). See (Kumar, 1985) for a detail survey for the development of the adaptive control. Lai and Robbins (1985) proposed a sampling rule called uniformly good rule, and developed an asymptotic regret-type performance lower bound for this uniform good sampling rule. Graves and Lai (1997) extended this result to the case in which the rewards behave as a Markov chain random variable. Moreover, they used the sequential testing theory to demonstrate that the derived lower bound is achievable. Auer et al. (1995) revised the problem and assumed that an adversary, instead of the conventionally assumed a well-defined density function, will completely govern the rewards. They proposed an algorithm and showed the algorithm converging to the best arm at the rate of $O(N^{-\frac{1}{2}})$. This line of research has become a branch of game theory in the Economics field, (see Foster and Vohra, 1997; Vovk, 1999; Auer, 2000; Piccolboni and Schindelhauer, 2001; Blum et al., 2003). The multi-armed bandit problem also has been studied in pricing under demand uncertainty, portfolio selection, and resource allocations in economics, (see Rothschild, 1974; Parkes and Huberman, 2001; Brezzi and Lai, 2000). In the economic literature, the multi-armed bandit problem would incorporate the discount factor into its objective function, and thus the future rewards would become the present values. Gittins (1979) showed the optimal solution of this *discounted* multi-armed bandit problem under certain conditions.

The multi-armed bandit problem as well as many variants of this problem has been extensively studied and many fruitful results have been discovered by researchers.

However, the problem has several differences from the problem we are interested in this thesis. The most important is that the objectives are different. The objective for the multi-armed bandit problem is to maximize the total rewards which one could receive along the way he explores the slots and exploits what he learns in the exploring stage. This implicitly implies that, by sequentially trying different arms, i.e. conducting sequentially experiments, he would learn about the slots and then identify the arm with highest expected reward *asymptotically.* On the other hand, the goal of our problem is not to pursue the *sum* of the rewards along the sequential experiments but to apply the information gathered from the limited finite number of experiments to identify the best possible arm.

1.2.3 Bayesian Experimental Design

The basic idea for Design of Experiments (DOE) is that the statistical inferences about the quantitative objectives of interests could be improved by appropriately selecting the levels of control variables (see Wu and Hamada, 2000). It is no wonder why the DOE is usually perceived as a technique for extracting the most useful information from the data to be collected. Moreover, DOE could also serve different purposes. For instance, it serves as a suitable vehicle for constructing the link between estimation, prediction, control, and optimization. The Bayesian Experimental Design is the DOE under the Bayesian framework, in which one would assume that the unknown parameters are random variables and the associated prior distributions are known. When the prior distributions contain no information, for instance, the corresponding variances are infinite, and then the Bayesian experimental design could be dealt with the ordinary DOE.

Lindley and Smith (1972) applied the decision theory to lay out the fundamental structure for the Bayesian experimental design. His argument is as follows. Consider a design η (see notiations in Fedorov, 1972), chosen from a design space $\mathcal{X}$, and let $\mathbf{y}$ denote the observed data. According to the data $\mathbf{y}$, an experimenter needs to make a decision d from the decision space $\mathcal{D}$. The experimenter encounters a problem on two folds: (1) how to determine the design η from $\mathcal{X}$ and (2) how to select d from

$\mathcal{D}$. Let the prior information be a density function $f(\theta)$ with unknown parameters $\theta \in \Theta$. Then the objective that the experimenter needs to maximize is

$$U(\eta) = \int \max_{d \in \mathcal{D}} \int_{\Theta} U(d, \eta, \mathbf{y}; \theta) f(\theta \,|\, \mathbf{y}, \eta) f(\mathbf{y}|\eta) \, d\theta d\mathbf{y}$$

where $U(d, \eta, \mathbf{y}; \theta)$ could be regarded as a utility function of the experimenter and $f(\cdot)$ denotes a probabilistic measure. Note that the structure of this objective is in the Bayesian format. To find the best experimental design $\eta^\star$, we need to maximize the utility backwards, i.e., we need to find $\eta^\star$ such that

$$U(\eta^\star) = \max_{\eta \in \mathcal{X}} U(\eta) = \max_{\eta \in \mathcal{X}} \int \max_{d \in \mathcal{D}} \int_{\Theta} U(d, \eta, \mathbf{y}; \theta) f(\theta \,|\, \mathbf{y}, \eta) f(\mathbf{y}|\eta) \, d\theta d\mathbf{y}$$

It is obvious that the selection of the utility function $U(d, \eta, \mathbf{y}; \theta)$ is crucial in this framework. DeGroot (1962) chooses the mutual information as the utility function for the purpose of inference, and this choice leads to the Bayesian $D-$optimality (Fedorov, 1972; Silvey, 1980) in the linear regression model (see Bernardo, 1979, for a detailed discussion). It is clear that the estimation is not necessarily equivalent to prediction. Suppose now the goal of the experimenter is to predict a linear combination of the unknown parameters θ, then the utility function the experimenter encounters could be

$$U(\theta) = -\int \int (\theta - \hat{\theta}) A (\theta - \hat{\theta}) f(\mathbf{y}, \theta \,|\, \eta) \, d\theta \, d\mathbf{y}$$

where the utility function represents a quadratic loss function, $\hat{\theta}$ denotes the estimate of θ, and matrix A is a symmetric positive definite matrix with a proper dimension. Chaloner (1984) solved the above problem with the assumption of Gaussian linear regression model, and concluded that the optimal design $\eta^\star$ would maximize the criterion $-tr\left\{A(nX^TX + R)^{-1}\right\}$, where n is the number of experiments, X is the design matrix according to an experimental design η, and R denotes the variance matrix of the prior distribution. Note that this is equivalent to the A-optimum design (see Silvey, 1980). A special selection of matrix A is that $A = \mathbf{c}\mathbf{c}^T$. Since matrix A is symmetric positive definite, there must exist a column vector $\mathbf{c}$ satisfying

the selection. This responds to the Bayesian c-optimum design (see Silvey, 1980). Chaloner (1984) gave a geometric argument for this case, which is extended by Dette (1996). There are other types of optimum designs closely related to the Bayesian experimental design; see (Chaloner and Verdinelli, 1995) for a detailed review.

From the above discussion, we know that the Bayesian D-optimum design could lead to better estimation while the Bayesian c-optimum design corresponds to better prediction. Suppose now the goal of the experimenter is to maximize the response while decreasing the variance of the estimate, which is very close to the problem we consider in this thesis. A combination of the D- and c- optimum design could be expected. Verdinelli and Kadane (1992) proposed the following expected utility objective function

$$U(\eta) = \int \int \left(\sum_i y_i + \alpha \log f(\theta \,|\, \mathbf{y}, \eta) \right) f(\theta \,|\, \mathbf{y}, \eta)\, d\theta\, d\mathbf{y}$$

where α is the weight of the importance in exploration and exploitation. Note that this utility function corresponds to the problem of multi-armed bandit problem discussed earlier. Verdinelli (1992) extended this utility to

$$U(\eta) = \int \int \log f\,(y_{n+1} \,|\, \mathbf{y},\, \eta\mathbf{y})\, f(\mathbf{y},\, y_{n+1} \,|\, \eta)\, d\mathbf{y}\, d\, y_{n+1} + \alpha \int \int \log f\,(\theta \,|\, \mathbf{y}, \eta)\, f(\theta \,|\, \mathbf{y}, \eta)\, d\theta\, d\mathbf{y}$$

where y_{n+1} could be regarded as the outcome of the experimental design and the coefficient α denotes the weights between the prediction and estimation. A similar problem could be found in (Pronzato, 2000; Pronzato and Thierry, 2003).

In certain linear regression problems, both the Bayesian and non-Bayesian experiments are independent of the observations (Lindley and Smith, 1972), implying that the sequential experiments could not gain any more information. However, for most problems, including the linear ones, it is still not clear whether the sequential designs could be better, especially when the goal is to maximize the response as we discussed earlier. Moreover, for nonlinear regression models, the posterior analysis clearly depends on the observation $\mathbf{y}$ (Silvey, 1980). We expect that there is a gain

from choosing design inputs sequentially.

1.2.4 Response Surface Methodology

Response Surface Methodology (RSM) is a collection of mathematical and statistical techniques developed for modeling and analyzing problems of determining optimum operating points through a sequence of experiments. Particularly, RSM is a useful technique in cases where system optimization is the goal and the system is largely unknown. Pioneered by Box and Wilson (1951) in the field of experimental design and analysis, the RSM has been vastly studied and implemented in a wide range of fields. Box and Draper (1975) was the first to point out the need to deal with the model uncertainty in design, and applied the idea of RSM to establish the robust design concept. Taguchi (1987) extended this frontier and gave the routine use of fractional factorial designs a central role in the design of products and process, a simple practical approach which allows engineers to design, perform and analyze their experiments within a context of a unified scheme called robust parameter design (see Nair, 1992, for the discussion of Taguchi's method and its drawbacks). In the last two decades, many researchers have encouraged the use of RSM in solving parameter design problems; for example, (Myers et al., 1992; Lucas, 1994; Lin and Tu, 1995; Engel and Huele, 1996; V. N et al., 2002). The increase has broaden the use of RSM and has shifted the focus of RSM. The computer-aided design has made much emphasis on RSM focused on finding regions where there is demonstrated improvements in response over what has been achieved by current operating points instead of finding optimum response (Myers, 1999). A detailed review concerning the response surface methodology could be found in (Hill and Hunter, 1966; Myers et al., 1989; Myers, 1999).

Even the RSM has been demonstrated in numerous cases, there still lacks a theoretical foundation for RSM to justify its achievements in these cases. Moreover, from the claim made by Wu and Hamada (2000): "if the input factors are quantitative and there are only a few of them, response surface methodology is an effective tool for studying this relationship," we know that the application of RSM is strongly influ-

enced by the scaling properties of the techniques comprising the methodology. This is opposite to the problem we consider in this thesis when the design point is finite, since what we would like to investigate is how to make improvements within very few experiments. Additionally, note that the RSM is for one interested in designing experiments on a set of design points and analyzing the experimental data with the goal of determining variables that provide improvements (Box et al., 1978). Although the goal of RSM is very close to ours in this thesis, we would like to put more emphasis on the roles of estimating mean and variances as the experiments proceeds, which is ignored in the response surface methodology.

1.3 Structure of the Thesis

We consider the cases when the design spaces discrete and compact respectively because of their unique properties, which will be explored in the following chapters. In Chapter 2, we will first consider the case of discrete design space. In this case, we assume that the resource for the experiments is very limited, and the goal is to make improvements in response as large as possible. I propose an adaptive One-Factor-at-A-Time (OFAT) experimental design, study its properties and compare its performance to saturated fractional factorial designs. The rationale for adopting the adaptive OFAT design scheme become clear if it is imbedded in a Bayesian framework: it becomes clear that OFAT is an efficient response to step by step accrual of sample information. The Bayesian predictive distribution for the outcome by implementing OFAT and the corresponding principal moments when a natural conjugate prior is assigned to parameters that are not known with certainty are also derived. Chapter 3 will consider the case in which the design space is continuous. In this case, I will extend the treatment of OFAT by the removal of two restrictions imposed on the discrete design space. The first is that the selection of input level at each iteration depends only on observed best response and does not depend on other prior information. In most real cases, domain experts possess knowledge about the process being modeled that, ideally, should be treated as sample information in its own right-and

not simply ignored. Treating the design problem Bayesianly provides a logical scheme for incorporation of expert information. The second removed restriction is that the model is restricted to be linear with pairwise interactions – implying that the model considers a relatively small design space. I extend the Bayesian analysis to the case of generalized normal linear regression model within the compact design space. Much of the literature on adaptive optimization of experimental designs focuses on **D**-optimum schemes, minimizing the uncertainty over the entire design space. I instead focus on **c**-optimality, in which the objective is to minimize uncertainty in the neighborhood of design points of interests, and propose an adaptive experiment algorithm. I will prove that this proposed algorithm would generate a consistent Bayesian estimator almost surely in its limiting behavior. Moreover, I also derive the expected step-wise improvement achieved by this algorithm for the analysis of its intermediate behavior, a critical criteria for determining whether to continue the experiments. Moreover, this improvement scheme would be shown to be applicable when neither regression parameters nor residual error variance are known with certainty. In Chapter 4, we conduct a case study to illustrate the practical value of the proposed algorithms in the cases of discrete and compact design space. We conclude and discuss future research directions in Chapter 5. Since we will use the Bayesian analysis technique frequently, a succinct survey of the Bayesian techniques is presented in Appendix A.

Chapter 2

The Discrete Space and the One-Factor-at-A-Time Experimental Scheme

We begin with presenting the regression model for the case in which the design space is discrete. We will propose an adaptive experimental scheme, the One-Factor-at-A-Time scheme, and demonstrate its ability to achieve a better response improvement under the restriction of scare resource for experiments. We also provide relational for the construction of the adaptive One-Factor-at-A-Time scheme with a Bayesian analysis. Finally, we will explore the properties of this adaptive experimental scheme and compare its performance to that of conventional fractional factorial schemes.

2.1 The Fundamental Model

We consider the following linear Gaussian regression model

$$y(\mathbf{x}) = \beta_0 + \beta^T f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \beta_{ij}\, x_i x_j + \epsilon,$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_q)^T$ denotes a q-element input vector, and

$$f(\mathbf{x}) = (x_1, \ldots, x_q, x_1x_2, \ldots, x_{q-1}x_q)^T$$

includes the main as well as interaction effects. Moreover, $q \geq 2$ denotes the number of factors, β_0 is a constant, and ϵ denotes the experimental error. Furthermore, in this chapter, we consider a discrete 2^q experimental design, and thus the design space is

$$\mathbf{x} \in \mathcal{X} = \left\{(\pm 1, \pm 1, \ldots, \pm 1)^T \in \mathbb{N}^q\right\}.$$

Since we consider the engineering applications, we assume that β_0 is sufficiently large such that $y(\mathbf{x})$ is positive almost surely. Finally, the prior of the main and interaction coefficients $\beta = (\beta_1, \ldots, \beta_q, \beta_{12}, \ldots, \beta_{(q-1)q})$ is

$$\begin{aligned} \beta \,|\, \sigma_\epsilon^2 &\sim \mathcal{N}(\mathbf{0}, \mathbf{V}) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2) \\ \mathbf{V} &= \mathrm{Diag}(\underbrace{\sigma_{\mathrm{ME}}^2, \ldots, \sigma_{\mathrm{ME}}^2}_{q \text{ main coefficients}}, \underbrace{\sigma_{\mathrm{INT}}^2, \ldots, \sigma_{\mathrm{INT}}^2}_{\frac{q(q-1)}{2} \text{ interactions}}) \end{aligned} \tag{2.1}$$

where $\mathrm{Diag}(\mathbf{v})$ denotes a diagonal matrix with its diagonal vector $\mathbf{v}$, and ϵ's are all independently, identically distributed throughout the experiment. We also assume that the linear model follows the hierarchical structure (Wu and Hamada, 2000), and therefore $\sigma_{\mathrm{ME}}^2 \geq \sigma_{\mathrm{INT}}^2$. Notice that, this prior assumes that all the coefficients in β are mutually independent, which is a strong assumption.

2.2 A Bayesian Analysis for Adaptive Experiments

The analysis approach we take in this section is to consider all the possible observations. We first construct the design matrix X for all possible inputs and then numerate the inputs. That is, we set

$$X = [f(\mathbf{x}_1),\, f(\mathbf{x}_2), \ldots, f(\mathbf{x}_{2^q})]$$

in which we always have $\mathbf{x}_1 = (1, 1, \ldots, 1)^T$ and $\mathbf{x}_{2^q} = (-1, -1, \ldots, -1)^T$. For example, when $q = 3$, we have

$$X = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_8)] = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix}$$

a 6×8 matrix. Then, the vector

$$\begin{aligned} Y(X) &= \begin{bmatrix} y(\mathbf{x}_1) \\ y(\mathbf{x}_2) \\ \vdots \\ y(\mathbf{x}_{2^q}) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, W) \\ W &= X^T V X + \sigma_\epsilon^2 I_{2^q} \end{aligned}$$

Let $W(i, j)$ denote the element of W in the ith row and jth column. Then, by construction, we have the diagonal terms

$$W(i, i) = q\,\sigma_{\text{ME}}^2 + \frac{q(q-1)}{2}\,\sigma_{\text{INT}}^2 + \sigma_\epsilon^2, \quad \forall\, i = 1, \ldots, 2^q.$$

Furthermore, suppose, for some $j \neq k$, we observe

$$\sum_{i=1}^{q} |X(i, j) - X(i, k)| = 2r$$

for some integer $1 \leq r \leq q$. Then we have

$$W(j, k) = (q - 2r)\,\sigma_{\text{ME}}^2 + \frac{q^2 - (4r+1)q + 4r^2}{2}\,\sigma_{\text{INT}}^2. \tag{2.2}$$

Note that r denotes the number of different elements in input vectors $\mathbf{x}_j$ and $\mathbf{x}_k$, and

$W(j,k)$ denotes the covariance of $y(\mathbf{x}_j)$ and $y(\mathbf{x}_k)$. It is clear that the covariance function (2.2) is a convex function in r. Moreover, by a straightforward analysis, $W(i,k)$ has the following property:

$$\text{If } r \begin{cases} < \frac{q}{2} + \frac{\sigma_{\text{ME}}^2}{2\sigma_{\text{INT}}^2}, \text{ then } W(j,k) \text{ is decreasing} \\ = \frac{q}{2} + \frac{\sigma_{\text{ME}}^2}{2\sigma_{\text{INT}}^2}, \text{ then } W(j,k) \text{ attains minimum} \\ > \frac{q}{2} + \frac{\sigma_{\text{ME}}^2}{2\sigma_{\text{INT}}^2}, \text{ then } W(j,k) \text{ is increasing} \end{cases}$$

Figure 2-1 illustrates this fact. From Fig. 2-1, we observe that the value would achieve its highest when $r = 1$ and behaves as a U-shape curve, which is convex. Last, we note that $W(i,i) \gg W(i,j)$ for all possible i and $j \neq i$. This property will reveal its importance in its later usage.

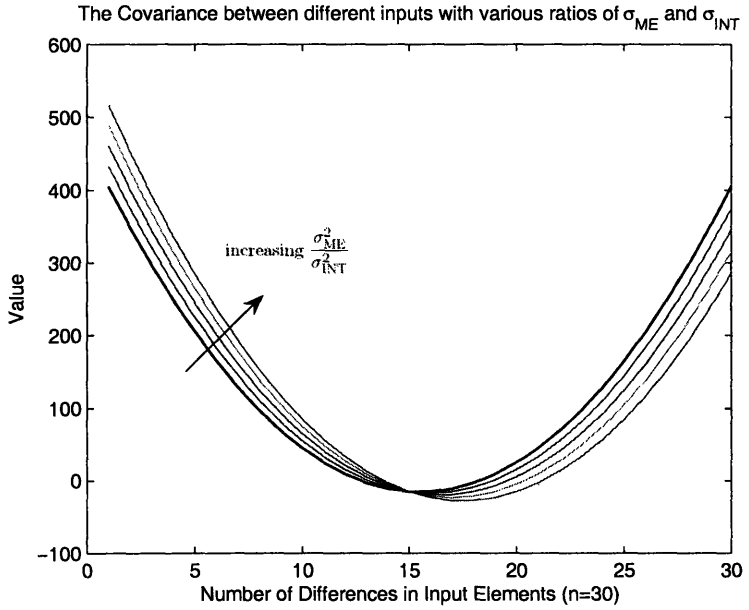

Figure 2-1: The change of covariance with the increase of different numbers in input elements

2.2.1 After the first experiment

Without loss of generality, we pick the input vector of the first experiment as $\mathbf{x}_{2^q}$, and obtain an observation y_{2^q}. As indicated before, $y_{2^q} \geq 0$ almost surely. With the goal

of searching the input $\mathbf{x}$ which achieves the highest response $\beta_0 + \beta^T f(\mathbf{x})$, we select the highest posterior estimation of response other than the $\mathbf{x}_{2^q}$ as the next sample input.

Then the posterior estimation of response $y_i = y(\mathbf{x}_i)$ for all $i = 1, 2, \ldots, 2^q - 1$ becomes

$$\hat{y}_i = \frac{W(i, 2^q)}{W(2^q, 2^q)} y_{2^q}$$

To select the highest $\hat{y}_i$ is equivalent to search for the input $\mathbf{x}_i$ with the highest $W(i, 2^q)$. Hence, we obtain the following observation.

Proposition 1

$$W(2^q - 1, 2^q) \geq W(i, 2^q), \quad \forall i = 1, 2, \ldots, 2^q - 1.$$

Proof: Choose any j with $\sum_{i=1}^{q} |X(i,j) - X(i, 2^q)| = 2r \geq 2$. Then we have

$$\begin{aligned} W(2^q - 1, 2^q) &- W(j, 2^q) \\ &= (q-2)\,\sigma^2_{\text{ME}} + \frac{q^2 - 5q + 4}{2}\,\sigma^2_{\text{INT}} (q - 2r)\,\sigma^2_{\text{ME}} + \frac{q^2 - (4r+1)q + 4r^2}{2}\,\sigma^2_{\text{INT}} \\ &= [(q-2) - (q-r)]\sigma^2_{\text{ME}} + \left[\frac{q^2 - 5q + 4}{2} - \frac{q^2 - (4r+1)q + 4r^2}{2}\right]\sigma^2_{\text{INT}} \\ &= 2(r-1)\left[\sigma^2_{\text{ME}} + (q - (r+1))\sigma^2_{\text{INT}}\right] \\ &\geq 0 \end{aligned}$$

□

Hence, after the first experiment, the input with one factor different from the first input vector is the optimal one to choose.

2.2.2 After the second experiment

After the second experiment is conducted, we obtain two observations y_{2^q-1} and y_{2^q}, and, without loss of generality, we assume that $y_{2^q} \geq y_{2^q-1}$. Then the posterior

estimate of the other response becomes

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_{2^q-2} \end{bmatrix} = \frac{1}{Var^2(y_{2^q}) - Cov^2(y_{2^q-1}, y_{2^q})} \begin{bmatrix} Cov(y_1, y_{2^q-1}) & Cov(y_1, y_{2^q}) \\ Cov(y_2, y_{2^q-1}) & Cov(y_2, y_{2^q}) \\ \vdots & \vdots \\ Cov(y_{2^q-2}, y_{2^q-1}) & Cov(y_{2^q-2}, y_{2^q-1}) \end{bmatrix} \times \begin{bmatrix} Var(y_{2^q}) & -Cov(y_{2^q-1}, y_{2^q}) \\ -Cov(y_{2^q-1}, y_{2^q}) & Var(y_{2^q}) \end{bmatrix} \begin{bmatrix} \hat{y}_{2^q-1} \\ \hat{y}_{2^q} \end{bmatrix}$$

Hence, for some $i \in \{1, 2, \ldots, 2^q - 2\}$, we have

$$\begin{aligned} \hat{y}_i &= \frac{1}{\text{Var}^2(y_{2^q}) - \text{Cov}^2(y_{2^q-1}, y_{2^q})} \left[\text{Var}(y_{2^q}) \text{Cov}(y_i, y_{2^q-1}) - \text{Cov}(y_i, y_{2^q}) \text{Cov}(y_{2^q-1}, y_{2^q}) \right] y_{2^q-1} \\ &+ \frac{1}{\text{Var}^2(y_{2^q}) - \text{Cov}^2(y_{2^q-1}, y_{2^q})} \left[\text{Var}(y_{2^q}) \text{Cov}(y_i, y_{2^q}) - \text{Cov}(y_i, y_{2^q-1}) \text{Cov}(y_{2^q-1}, y_{2^q}) \right] y_{2^q} \end{aligned}$$

Since $\text{Var}^2(y_{2^q}) - \text{Cov}^2(y_{2^q-1}, y_{2^q}) > 0$, we need to find

$$\begin{aligned} i = \arg \max_{i \in \{1,2,\ldots,2^q-2\}} & \left[\text{Var}(y_{2^q}) \text{Cov}(y_i, y_{2^q-1}) - \text{Cov}(y_i, y_{2^q}) \text{Cov}(y_{2^q-1}, y_{2^q}) \right] y_{2^q-1} \\ & + \left[\text{Var}(y_{2^q}) \text{Cov}(y_i, y_{2^q}) - \text{Cov}(y_i, y_{2^q-1}) \text{Cov}(y_{2^q-1}, y_{2^q}) \right] y_{2^q} \end{aligned} \quad (2.3)$$

To facilitate our discussion, we define the function

$$\psi(r) = \text{Cov}(y_i, y_j)$$

for all corresponding $\mathbf{x}_i$ and $\mathbf{x}_j$ such that

$$\sum_{k=1}^{q} |\mathbf{x}_i(k) - \mathbf{x}_j(k)| = 2r, \quad \forall\, r = 0, 1, \ldots, q.$$

We will show that the best input regarding the information collected so far is $\mathbf{x}_{2^q-2}$ which is only one element deviate from $\mathbf{x}_{2^q}$. Recall that $\mathbf{x}_{2^q}$ generates the highest

response in assumption.

Proposition 2 *Given $y_{2^q} \geq y_{2^q-1} > 0$, we have*

$$\hat{y}_{2^q-2} - \hat{y}_i \geq 0, \quad \forall\, i = 1, 2, \ldots, 2^q - 2.$$

Proof: Since $\mathbf{x}_{2^q}$ and $\mathbf{x}_{2^q} - 1$ have only one different elements, all other elements could have either r elements different from $\mathbf{x}_{2^q}$ and $r+1$ elements different from $\mathbf{x}_{2^q-1}$ with $r \geq 1$, or r elements deviate from $\mathbf{x}_{2^q}$ and $r-1$ elements from $\mathbf{x}_{2^q-1}$ with $r \geq 2$.

We first consider $\mathbf{x}_i$ with r different elements from $\mathbf{x}_{2^q}$ and $r+1$ elements from $\mathbf{x}_{2^q-1}$.

$$\begin{aligned}
\hat{y}_{2^q-2} - \hat{y}_i &= \frac{1}{\psi(0)^2 - \psi(1)^2}\Bigg(\Big[\psi(0)\psi(2) - \psi(1)\psi(1)\Big] y_{2^q-1} + \Big[\psi(0)\psi(1) - \psi(2)\psi(1)\Big] y_{2^q} \\
&\qquad - \Big[\psi(0)\psi(r+1) - \psi(r)\psi(1)\Big] y_{2^q-1} - \Big[\psi(0)\psi(r) - \psi(r+1)\psi(1)\Big] y_{2^q}\Bigg) \\
&= \frac{1}{\psi(0)^2 - \psi(1)^2}\Bigg(\Big[\psi(0)\psi(2) - \psi(1)^2 - \psi(0)\psi(r+1) + \psi(r)\psi(1)\Big] y_{2^q-1} \\
&\qquad + \Big[\psi(0)\psi(1) - \psi(2)\psi(1) - \psi(0)\psi(r) + \psi(r+1)\psi(1)\Big] y_{2^q}\Bigg) \\
&\geq \frac{1}{\psi(0)^2 - \psi(1)^2}\Big(\psi(0)\Big[\psi(2) - \psi(r+1) + \psi(1) - \psi(r)\Big] \\
&\qquad + \psi(1)\Big[\psi(r) - \psi(1) + \psi(r+1) - \psi(2)\Big]\Big) y_{2^q-1} \\
&= \frac{1}{\psi(0)^2 - \psi(1)^2}\Big(\Big[\psi(0) - \psi(1)\Big]\,[\psi(2) - \psi(r+1) + \psi(1) - \psi(r)]\Big) y_{2^q-1}
\end{aligned}$$

From Eq.(2.2), we know that $\psi(1) > \psi(2)$ and it remains to show that $\psi(2) \geq \psi(q)$. Hence,

$$\begin{aligned}
\psi(2) - \psi(q) &= (q-2)\sigma^2_{\text{ME}} + \frac{q^2 - 5q + 4}{2}\sigma^2_{\text{INT}} + q\sigma^2_{\text{ME}} - \frac{q^2 - (4q+1)q + 4q^2}{2}\sigma^2_{\text{INT}} \\
&= 2(q-1)\sigma^2_{\text{ME}} - 2\sigma^2_{\text{INT}} \\
&\geq 2(q-1)\sigma^2_{\text{ME}} - 2\sigma^2_{\text{ME}} \\
&\geq 2(q-2)\sigma^2_{\text{ME}} \geq 0
\end{aligned}$$

The other case is for $\mathbf{x}_i$ with r different elements from $\mathbf{x}_{2^q}$ and $r-1$ elements from $\mathbf{x}_{2^q-1}$ with $r \geq 2$. Following the identical argument from above, we have

$$\begin{aligned}
\hat{y}_{2^q-2} - \hat{y}_i &= \frac{1}{\psi(0)^2 - \psi(1)^2}\Bigg(\Big[\psi(0)\psi(2) - \psi(1)\psi(1)\Big]y_{2^q-1} + \Big[\psi(0)\psi(1) - \psi(2)\psi(1)\Big]y_{2^q} \\
&\qquad - \Big[\psi(0)\psi(r-1) - \psi(r)\psi(1)\Big]y_{2^q-1} - \Big[\psi(0)\psi(r) - \psi(r-1)\psi(1)\Big]y_{2^q}\Bigg) \\
&\geq \frac{1}{\psi(0)^2 - \psi(1)^2}\Big(\psi(0)\Big[\psi(2) - \psi(r-1) + \psi(1) - \psi(r)\Big] \\
&\qquad + \psi(1)\Big[\psi(r) - \psi(1) + \psi(r-1) - \psi(2)\Big]\Big)\, y_{2^q-1} \\
&= \frac{1}{\psi(0)^2 - \psi(1)^2}\Big(\Big[\psi(0) - \psi(1)\Big]\,[\psi(2) - \psi(r-1) + \psi(1) - \psi(r)]\Big)\, y_{2^q-1} \\
&\geq 0
\end{aligned}$$

□

Hence, from Prop. 2, the best thing one could do for the next experiment after the second is to choose the input which is one element different from the input with higher response.

An alternative, heuristic way to obtain the desired next input $\mathbf{x}$ is to rearrange elements in Eq.(2.3) and find the difference between $\mathbf{x}_i$ and $\mathbf{x}_j$ as follows

$$\begin{aligned}
&\mathrm{Var}(y_{2^q})\Big(\mathrm{Cov}(y_i, y_{2^q-1}) - \mathrm{Cov}(y_j, y_{2^q-1})\Big)y_{2^q-1} \\
+\;&\mathrm{Var}(y_{2^q})\Big(\mathrm{Cov}(y_i, y_{2^q}) - \mathrm{Cov}(y_j, y_{2^q})\Big)y_{2^q} \qquad (2.4) \\
+\;&\mathrm{Cov}(y_{2^q-1}, y_{2^q})\Big(\mathrm{Cov}(y_j, y_{2^q}) - \mathrm{Cov}(y_i, y_{2^q})\Big)y_{2^q-1} \\
+\;&\mathrm{Cov}(y_{2^q-1}, y_{2^q})\Big(\mathrm{Cov}(y_j, y_{2^q-1}) - \mathrm{Cov}(y_i, y_{2^n-1})\Big)y_{2^n}
\end{aligned}$$

Since $\mathrm{Var}(y_{2^q})$ dominates all other covariance and $y_{2^q} > y_{2^q-1}$, we consider only Eq.(2.4) and choose the next input vector accordingly. By our discussion above, we should choose the input with only one element different from the input $\mathbf{x}_{2^q}$.

2.2.3 After k experiments with $3 \leq k \leq q$

Suppose that we have conducted k experiments with $3 \leq k \leq q$, and that y_s, $1 \leq s \leq k$, dominates all other observed response. According to (Horn and Johnson, 1990), we know that the inverse of a nonsingular matrix A could be represented as

$$A^{-1} = \frac{1}{\det(A)} \operatorname{adj}(A)$$

where $\det(A)$ denotes the determinant of matrix A, and $\operatorname{adj}(A)$ denotes the adjugate of matrix A. Hence, by the definition of $\operatorname{adj}(A)$, the prediction of y_i would involve the term

$$\operatorname{Var}^{k-1}(y_s)\operatorname{Cov}(y_i, y_s)\, y_s.$$

Recall that $\operatorname{Var}(y_{2^q})$ dominates all other covariance terms. Thus, given $y_s > 0$, to achieve the highest response is to select the appropriate $\mathbf{x}_i$ such that $\operatorname{Cov}(y_i, y_s)$ is maximum. Therefore, the input vector which differs only one elements from $\mathbf{x}_s$ is the optimal choice. Suppose now that we impose one more restriction that the experimenter will not toggle the factors that has been investigated before, then we will have the following adaptive one-factor-at-a-time (OFAT) experiment scheme.

2.3 Adaptive One-Factor-at-A-Time Experimental Scheme

In this section, we will propose an adaptive experimental scheme, the adaptive One-Factor-at-A-Time (OFAT) Scheme, according to the analysis in the above section. In the adaptive OFAT scheme, we consider the q factor normal linear model

$$y(\mathbf{x}) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \beta_{ij} x_i x_j + \epsilon \tag{2.5}$$

where $y(\mathbf{x})$ denotes the response at the input level $\mathbf{x}^T = (x_1, \ldots, x_q)$, β_i denotes the main effects associated with the ith factor, β_{ij} represents the interaction between

factors x_i and x_j, and, finally, ϵ is the experimental error. We omit the term β_0 here because we could not control the constant term and thus β_0 becomes redundant in our analysis. Notice that, in this model, we have $p = \frac{q(q+1)}{2}$ elements in the coefficient vector β. For simplicity, we assume that the coefficients and the errors are mutually independent and the distributions are

$$\begin{aligned} \beta_i \,|\, \sigma_\epsilon^2 &\sim \mathcal{N}(0, \sigma_{\text{ME}}^2), \quad \forall i = 1, \ldots, q \\ \beta_{ij} \,|\, \sigma_\epsilon^2 &\sim \mathcal{N}(0, \sigma_{\text{INT}}^2), \quad \forall i = 1, \ldots, (q-1),\ j = i+1, \ldots, q \\ \epsilon \,|\, \sigma_\epsilon^2 &\sim \mathcal{N}(0, \sigma_\epsilon^2) \end{aligned} \tag{2.6}$$

This prior implies that we have no specific information about the magnitude and sign of the coefficients, and thus the variances σ_{ME}^2 and σ_{INT}^2 is expected to be large. That is, the prior would be weak. The analysis of the adaptive OFAT scheme is very sensitive to the assumption of prior, and we will discuss the prior in details later.

2.3.1 The Processes of Adaptive OFAT Scheme

We illustrate the process of the adaptive OFAT scheme with the case of three two-level factors labeled here as A, B, and C, which is shown in Figure 2-2. First, an experiment is conducted at some baseline point in the design space. In Figure 2-2 this baseline point is A=-1, B=+1, C=+1, and any of the eight points in the space have the same probability to be chosen. Next a factor is varied and another experiment is run. In Figure 2-2, factor A is varied first. If the experimental results suggest there was an improvement in the latter response, then the change in factor A will be retained. Thus, all future experiments are affected by the results of the first two experiments. Next, another factor is changed. In Figure 2-2, B is toggled from +1 to -1. Another experiment is conducted and compared to the best result observed so far. If the most recent change does not seem favorable, it is reversed before proceeding. The process ends when all the factors have been changed. The final settings are determined by the best observation in the sequence.

Note that this adaptive OFAT scheme has several features. First, it requires

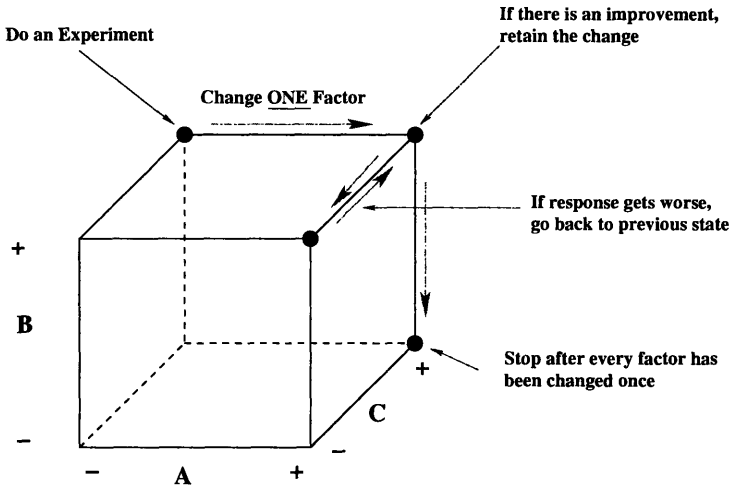

Figure 2-2: Illustration of the adaptive OFAT applied to a system with three two-level factors (A, B, and C)

$q + 1$ experiments, which is significantly less than almost all traditional design of experiment schemes. In fact, only the saturated fractional factorial designs could require the same experimental runs. Thus, we choose the fractional factorial scheme to be the benchmark for our comparison later. Further, the factors flip the sign of its corresponding inputs one at a time, regardless of its history. This is one of the key reasons why this analysis of its performance is tractable. It also implies that the adaptive OFAT scheme explores each dimension one by one while keeping all other dimensions fixed. However, it does *not* behave like the gradient search as we usually see in the optimization algorithms, because the OFAT scheme considers only the sign of difference but not the direction of most improvement. Moreover, due to the symmetry of the prior distribution of β, it would not make any stochastic difference in selecting the initial inputs and therefore the order of the sign change of inputs could be arbitrary. We will discuss the symmetry property of the prior distribution later. Finally, the determination of the inputs depends on the observational difference between the corresponding two consecutive experiments. It implies that the OFAT scheme implicitly has a short memory; it only considers the current and the previous experiment results and ignores all other experiments.

In the mathematical framework, the input levels of the adaptive OFAT scheme

could be described as follows:

$$\begin{aligned}
\mathbf{x}_{(0)} &= \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_q^{(0)}\right)^T \\
\mathbf{x}_{(1)} &= \left(-x_1^{(0)}, x_2^{(0)}, \ldots, x_q^{(0)}\right)^T \\
\mathbf{x}_{(2)} &= \left(x_1^{(0)} \operatorname{sign}\left(y(\mathbf{x}_{(0)}) - y(\mathbf{x}_{(1)})\right), -x_2^{(0)}, x_3^{(0)}, \ldots, x_q^{(0)}\right)^T \\
&\vdots \\
\mathbf{x}_{(k+1)} &= \left(x_1^{(k)}, \ldots, x_{k-1}^{(k)}, x_k^{(0)} \operatorname{sign}\left(y(\mathbf{x}_{(k-1)}) - y(\mathbf{x}_{(k)})\right), -x_{k+1}^{(0)}, x_{k+2}^{(0)}, \ldots, x_q^{(0)}\right)^T \quad (2.7) \\
&\vdots \\
\mathbf{x}_{(q)} &= \left(x_1^{(q-1)}, \ldots, x_{q-2}^{(q-1)}, x_{q-1}^{(0)} \operatorname{sign}\left(y(\mathbf{x}_{(q-2)}) - y(\mathbf{x}_{(q-1)})\right), -x_q^{(0)}\right)^T
\end{aligned}$$

where the $\mathbf{x}_{(0)}$ denotes the initial input level of the OFAT scheme, the superscript k of scalar $x_i^{(k)}$ denotes the kth experimental run, and the associated subscript i denotes the ith element of the input $\mathbf{x}_{(k)}$, the kth input level. The final decision of the input level for the adaptive OFAT is

$$\mathbf{x}_{\text{OFAT}} = \left(x_1^{(q)}, \ldots, x_{q-1}^{(q)}, \ldots, x_q^{(0)} \operatorname{sign}\left(y(\mathbf{x}_{(q-1)}) - y(\mathbf{x}_{(q)})\right)\right)$$

We would like to address several remarks for the process of the OFAT scheme in the 2^q experimental environment.

- all elements of input level $\mathbf{x}_{(k)}$ $(k > 1)$ are identical to those of $\mathbf{x}_{(k-1)}$ except the $(k-1)$th and the kth elements, and the change of the kth element is static
- the $(k-1)$th element of $\mathbf{x}_{(k)}$ $(k \geq 2)$ depends on the values of $y(\mathbf{x}_{(k-1)})$ and $y(\mathbf{x}_{(k-2)})$; if $y(\mathbf{x}_{(k-1)})$ wins, i.e., $y(\mathbf{x}_{(k-1)}) \geq y(\mathbf{x}_{(k-2)})$, then we keep the $(k-1)$th element of $y(\mathbf{x}_{(k-1)})$ in $\mathbf{x}_{(k)}$; otherwise, we keep that of $y(\mathbf{x}_{(k-2)})$
- although the selection of the input levels does follow the posterior Bayesian analysis after each experiment runs as we did earlier, it is more convenient to regard it depending on the observations of previous two experiments

Before we proceed to analyze the OFAT scheme, we would like to review several

historical remarks in the DOE literature regarding the scheme. As pointed out by Frey and Wang (2006), OFAT experimental scheme is generally discouraged by experts in experimental design and quality improvement (see Box et al., 1978; Logothetis and Wynn, 1995; Czitrom, 1999; Wu and Hamada, 2000). Reasons cited include

1. It requires more runs for the same precision in effect estimation;
2. It cannot estimate some interactions;
3. The conclusions from its analysis are not general;
4. It can miss optimal settings of factors;
5. OFAT can be susceptible to bias due to time trends because OFAT cannot be randomized in the same sense that fractional factorial designs frequently are.

While these cautions are valid and should be taken into account when considering use of OFAT, some researchers have articulated a role for OFAT and demonstrated that it has some advantages under some conditions. Friedman and Savage (1947) suggested that a one-factor-at-a-time approach might be used in preference to balanced factorial plans when the experimenter seeks an optimum within a system likely to contain interactions. They suggested that OFAT might offer advantages since it concentrates observations in regions that are likely to contain the optimum. Daniel (1973) suggested that OFAT may be preferred when an experimenter wishes to react more quickly to data and can safely be used in those cases in which factor effects are three or four times the standard deviation due to pure experimental error. Koita (1994) showed that a one-factor-at-a-time method was effective for identifying selected interactions after running fractional factorial designs as part of an overall approach to sequential experimentation. McDaniel and Ankenman (2000) provided empirical evidence that, for "small factor change problems," a strategy including one-factor-at-a-time and Box-Behnken designs often worked better than a comparable strategy employing fractional factorial designs when there is no error in the response. Qu and Wu (2005) used one-factor-at-a-time techniques to construct resolution V designs within an economical run size.

2.4 Preposterior Analysis of OFAT Scheme

In this preposterior analysis, we will assume that the parameter $\sigma_\epsilon > 0$ is known, which is a common assumption in many DOE literatures. An analysis for unknown σ_ϵ will be presented later. In this section, we will focus on the performance of the OFAT scheme from one experiment to the next. There are two performances we will analyze in this section: one is the expected response value after experiments and the other is the exploitation probability of a coefficient in our model Eq.(2.5). Before starting the analysis, we define the exploitation in the following.

Definition 1 *An effect β_i or β_{ij} is said to be* ***exploited*** *if the product of the suggested x_i and/or x_j has the same sign as that of the corresponding effect.*

2.4.1 The First Step in Adaptive OFAT

Theorem 1 *If adaptive OFAT scheme is applied to a response model (2.5), then the expected improvement after the first step (the second experiment) is*

$$\mathbb{E}\left[y(\mathbf{x}_{(2)})\right] = \mathbb{E}\left[\beta_1 x_1^{(2)}\right] + (q-1)\mathbb{E}\left[\beta_{1j} x_1^{(2)} x_j^{(0)}\right]$$

where

$$\mathbb{E}\left[\beta_1 x_2^{(1)}\right] = \sqrt{\frac{2}{\pi}} \frac{\sigma_{ME}^2}{\sqrt{\sigma_{ME}^2 + (q-1)\sigma_{INT}^2 + \frac{\sigma_\epsilon^2}{2}}}$$

and

$$\mathbb{E}\left[\beta_{1j} x_1^{(2)} x_j^{(0)}\right] = \sqrt{\frac{2}{\pi}} \frac{\sigma_{INT}^2}{\sqrt{\sigma_{ME}^2 + (q-1)\sigma_{INT}^2 + \frac{\sigma_\epsilon^2}{2}}}$$

Recall that, in Eq.(2.7), $x_i^{(j)}$ denotes the ith element of the jth experiment in the adaptive OFAT scheme. For the ease of reading, all the proofs in this section will be delivered later.

As Theorem 1 indicates, the expected response in the second experiment following the adaptive OFAT arises due to the main effect of the first main factor and all the $q-1$ interaction factors in which the first factor participates. This combination of a

main effect and related interactions is often called the *conditional main effect* since it represents the effect of the factor conditioned on the current initial settings, i.e. $y(\mathbf{x}_0)$ and $y(\mathbf{x}_1)$. Moreover, Theorem 1 also suggests that, when $\sigma_{\text{ME}} > \sigma_{\text{INT}}$, most of the expected improvement is due to the main effect; otherwise, most of the improvement would be from the interactions.

Although there is an improvement in response realized due to interactions after the first step in OFAT, note that none of interaction factors has been exploited (see Def. 2.4). All of the factors except for $x_1^{(\cdot)}$ will be toggled in subsequent steps of adaptive OFAT. Depending on what is observed, their final state may be different from the state after the first experiment. In that case, the contributions due to interaction factors β_{1j} may potentially be reversed as the process continues. By contrast, the first main effect β_1 has been exploited and its contribution to the expected value is permanent. No subsequent steps in the adaptive OFAT process will affect the contribution of β_1 to expected response in Eq.(2.5). Thus, the probability of β_1 being exploited can be fully determined by analyzing the behavior of the first step. This probability is given in the following theorem.

Theorem 2 *If adaptive OFAT scheme is applied to a response model (2.5), then the probability of the main effect β_1 being exploited would be*

$$P\left(\beta_1 x_1^{(2)} > 0\right) = \frac{1}{2} + \frac{1}{\pi} \arcsin \frac{\sigma_{ME}}{\sqrt{\sigma_{ME}^2 + (q-1)\sigma_{INT}^2 + \frac{\sigma_\epsilon^2}{2}}}$$

Theorem 2 shows that the probability of exploiting the main effect approaches 100% as the main effect strength is much larger than interactions and experimental error, i.e., $\sigma_{\text{ME}} \gg \max\{\sigma_{\text{INT}}, \sigma_\epsilon\}$. It also shows that the probability will drop to 50% as the experimental error σ_ϵ increases.

2.4.2 The Second Step in Adaptive OFAT

Theorem 3 *If adaptive OFAT scheme is applied to a response model (2.5), then the expected improvement after the second step (the third experiment) is*

$$\mathbb{E}\left[y(\mathbf{x}_{(3)})\right] = 2\mathbb{E}\left[\beta_1 x_1^{(3)}\right] + 2(q-1)\mathbb{E}\left[\beta_{1j} x_1^{(3)} x_j^{(0)}\right] + \mathbb{E}\left[\beta_{12} x_1^{(3)} x_2^{(3)}\right]$$

where $\mathbb{E}\left[\beta_1 x_1^{(3)}\right]$ *and* $\mathbb{E}\left[\beta_{1j} x_1^{(2)} x_j^{(0)}\right]$ *are identical to those in Thoerem 1, and*

$$\mathbb{E}\left[\beta_{12} x_1^{(3)} x_2^{(3)}\right] = \sqrt{\frac{2}{\pi}} \frac{\sigma_{INT}^2}{\sqrt{\sigma_{ME}^2 + (q-1)\sigma_{INT}^2 + \frac{\sigma_\epsilon^2}{2}}}$$

Theorem 3 reveals that, after the second step in the OFAT process, the response has an additional increase due to main effects and also added contributions due to interactions. The improvements at this stage arise due to three different contributors: 1) two main effects that might be exploited, 2) a two-factor interaction β_{12} which may have been exploited, and 3) a set of $2(q-2)$ interactions which involve exactly one of the two main effects that have been toggled so far.

After the second variable is set by adaptive OFAT, the interaction β_{12} will not be affected in any way by subsequent experiments. The probability of exploiting the interaction factor β_{12} can therefore be determined by analyzing the process at the second step of adaptive OFAT. The following theorem arises from such an analysis.

Theorem 4 *If adaptive OFAT scheme is applied to a response model (2.5), then the probability of the interaction effect* β_{12} *being exploited would be*

$$P\left(\beta_{12} x_1^{(3)} x_2^{(3)} > 0\right) = \frac{1}{2} + \frac{1}{\pi} \arctan \frac{\sigma_{INT}}{\sqrt{\sigma_{ME}^2 + (q-2)\sigma_{INT}^2 + \frac{\sigma_\epsilon^2}{2}}}$$

Note that the probability given in Theorem 4 is greater than 50% for all systems with nonzero interactions. This presents a paradox. The adaptive OFAT process confounds the two-factor interaction β_{12} with the main effects β_1 and β_2. Nevertheless, the probability of exploiting the interaction is better than that provided by random

chance and the experimenter gains some improvement from the interaction on average. Thus, the experimenter benefits from an effect which he cannot resolve.

2.4.3 Subsequent Steps in Adaptive OFAT

As the adaptive process proceeds through subsequent steps, the mathematical results become increasingly complex. Exact closed form solutions to the probabilities and expected values become cumbersome. However, simple bounds and approximations can still be derived and provide useful insights.

Theorem 5 *If adaptive OFAT scheme is applied to a response model (2.5), then the probability of exploiting interaction β_{ij} depends only on j for all $1 \leq i < j \leq q$.*

Theorem 5 argues that any interaction factor β_{ij} would not be exploited until the input $x_j^{(q)}$ is determined, which matches our earlier observations. Therefore, whether an interaction β_{ij} will contribute to the response improvement will not be clear until the $j+1$ experiment. More importantly, the result of Theorem 5 would lead to the following observation.

Theorem 6 *If adaptive OFAT scheme is applied to a response model (2.5), then the probability of exploiting a two-factor interaction β_{ij} is greater than or equal to the probability of exploiting the two-factor interaction β_{12}, i.e.,*

$$P\left(\beta_{ij}x_i^{(q)}x_j^{(q)}\right) \geq P\left(\beta_{12}x_1^{(q)}x_2^{(q)}\right)$$

Note that $x_i^{(q)}$ denote the final decision of ith input of OFAT scheme. Theorem 6 shows that β_{12} is the least likely exploited interaction factor. One conjecture is that the exploiting probability would increases as the OFAT scheme proceeds alone.

Finally, we derive a lower bound for the output of the adaptive OFAT scheme. Since it is extremely difficult to obtain a theoretical closed form for the final expected response of the adaptive OFAT scheme, Frey and Wang (2006) proposed a closed form from a practical observation. Fortunately, Theorem 6 renders us a nice way to derive the following performance lower bound, which would be clear in the proof.

Theorem 7 *Given the number of control variable $q \geq 3$, the expected value of the outcome of the OFAT scheme, denoted by $y(\mathbf{x}_{OFAT})$, obeys the following*

$$\begin{aligned}\mathbb{E}\left[y\left(\mathbf{x}_{OFAT}\right)\right] \;\geq\; & q \times \left(\frac{2}{\pi}\right)^{\frac{3}{2}} \times \arctan\left(\frac{\sigma_{ME}}{\sqrt{(q-1)\sigma_{INT}^2 + \frac{\sigma_\epsilon^2}{2}}}\right) \times \sigma_{ME} \\ & + \frac{q(q-1)}{2} \times \left(\frac{2}{\pi}\right)^{\frac{3}{2}} \times \arctan\left(\frac{\sigma_{INT}}{\sqrt{\sigma_{ME}^2 + (q-2)\sigma_{INT}^2 + \frac{1}{2}\sigma_\epsilon^2}}\right) \times \sigma_{INT}\end{aligned} \tag{2.8}$$

Note that the equality would hold when $\sigma_{\text{INT}} = \sigma_\epsilon = 0$. Theorem 7 provides a lower bound for the final performance on the expected response. The lower bound indicates that the final performance of adaptive OFAT scheme is sustained by strength of both main factor, i.e. σ_{ME}, and the interaction factor, i.e. σ_{INT}. When the strength of main factors is much larger than that of the interaction factors, i.e. $\sigma_{\text{ME}} > \sigma_{\text{INT}}$, the first term of Eq.(2.8) would be large although the otehr term would be small. On the other hand, when $\sigma_{\text{INT}} > \sigma_{\text{ME}}$, the lower bound would be sustained by the second term in Eq.(2.8). This lower bound will be used to compare the performance of adaptive OFAT to that of conventional fractional factorial experimental scheme.

2.5 Comparison with Fractional Factorial Experiments

The previous section enables estimates of the expected response provided by adaptive OFAT and reveals some of the mechanisms providing the improvements. To understand whether those improvements are large or small, it is worthwhile to have a basis for comparison. Saturated resolution III fractional factorial experiments are useful in this regard. Resolution III designs are used frequently for screening and also sometimes used in improvement processes, especially in robust parameter design. A saturated resolution III design, if it exists for a given q, can be carried out using *the same number of experiments as adaptive OFAT*; that is, $q + 1$ experiments. Thus,

with resolution III designs and adaptive OFAT, we are comparing alternatives with similar resource demands. To make a quantitative comparison, the following result is useful.

Theorem 8 *If a saturated resolution III two-level fractional factorial experiment is applied to the response model Eqs.(2.5) and the factor levels are chosen among their discrete levels to achieve the maximum response based on the main effect estimates, then the expected performance is*

$$\mathbb{E}\left[y(\mathbf{x}_{III})\right] = \sqrt{\frac{2}{\pi}} \frac{q\,\sigma_{ME}^2}{\sqrt{\sigma_{ME}^2 + \frac{q}{2(q+1)}\sigma_{INT}^2 + \frac{1}{q+1}\sigma_\epsilon^2}}$$

where $\mathbf{x}_{III}$ denotes the optimal input chosen by the saturated resolution III fractional factorial experiment scheme.

Theorem 8 shows that the final performance of the saturated resolution III experiment is proportional to the strength of main effect, i.e. σ_{ME}^2, and is contaminated by the strength of interaction factors and the errors, i.e. σ_{INT}^2 and σ_ϵ^2. Moreover, the influence of the interaction factors on the performance is far more than that of error. In other words, the expected improvement provided by the resolution III design is more sensitive to interactions than to errors. This observation makes perfect sense because resolution III experiments are devised so as to provide information about main effects with the greatest possible robustness to experimental error. These designs are not well suited for providing information about two-factor interactions. Resolution III designs, unlike resolution IV designs, are negatively affected by two-factor interactions. Although Resolution III designs cannot exploit interactions, they are optimally robust to experimental error and therefore should provide better results than adaptive OFAT under some conditions.

Figure 2.3 illustrate the performance between OFAT and saturated resolution III factorial fractional experiment by the lower bound derived in Theorem 7 and the result obtain in Theorem 8 with three error strength, $\sigma_\epsilon^2 = 0$, $\sigma_\epsilon^2 = 1$, and $\sigma_\epsilon^2 = 10$. From Fig. 2.3, one could conclude that the OFAT scheme would outperform the

saturated fractional factorial experimental scheme when the strength of the interaction factors is large, at least no less than 45% of the main factor strength σ_{ME} when the error strength is worst. As the error strength σ_ϵ increases, the performances for both schemes become weak. However, the saturated resolution III experimental scheme is more inferiorly influenced by the adaptive OFAT scheme because the OFAT would still make observable improvements while the other scheme make less and less improvements when the error increases.

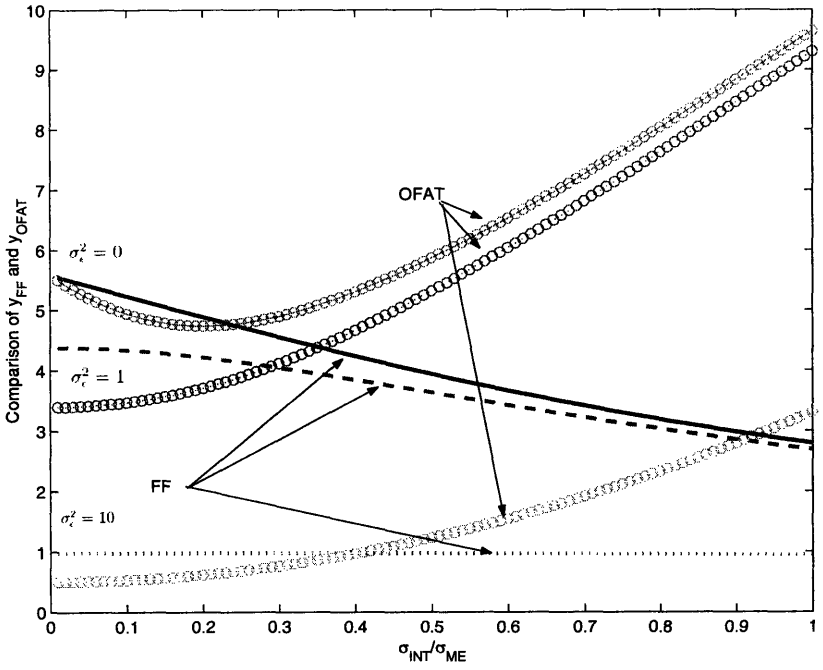

Figure 2-3: Comparison of $\mathbb{E}[y(\mathbf{x}_{\text{OFAT}})]$ and $\mathbb{E}[y(\mathbf{x}_{III})]$

For completeness, we discuss the exploiting probability of the saturated fractional factorial experimental scheme.

Theorem 9 *If a saturated two-level fractional factorial experiment scheme is applied to Eqs.(2.5), the main effect β_i will be exploited with probability*

$$\frac{1}{2} + \frac{1}{\pi} \arctan \frac{\sigma_{ME}}{\sqrt{\frac{(q-1)(q-2)}{2(q+1)}\sigma_{INT}^2 + \frac{\sigma_\epsilon^2}{q+1}}}$$

Finally, in a saturated two-level fractional factorial experiment, the interaction effect β_{ij} will be exploited with probability $\frac{1}{2}$. It is because whether the choice of the resolution III experiment for the ith input element $x_i^{(III)}$ as well as the jth element $x_j^{(III)}$ equal to +1 or -1 are independent of β_{ij} for all possible i and j, the probability of exploiting the interaction effect is $\frac{1}{2}$.

2.6 The Posterior Analysis

In this section, we will conduct the posterior analysis for the adaptive OFAT scheme and the fractional factorial scheme with an emphasis on the case where the error strength σ_ϵ is *unknown*.

2.6.1 Posterior Analysis for OFAT Scheme with Unknown σ_ϵ^2

When σ_ϵ^2 is unknown, we assume that σ_ϵ^2 has the prior

$$f(\sigma_\epsilon^2) = \frac{\left(\frac{a}{2}\right)^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \left(\sigma_\epsilon^2\right)^{-\frac{d+2}{2}} \exp\left(-\frac{a}{2\sigma_\epsilon^2}\right)$$

with parameter a, $d > 0$. Then, recall the Eqs.(2.7) and apply results in Appendix A, the posterior distribution of the response at $y(\mathbf{x}_{\text{OFAT}})$ would be

$$\begin{aligned}\mathbf{x}_{\text{OFAT}} \;=\; \Big(&x_1^{(0)} \operatorname{sign}\left(y(\mathbf{x}_{(0)}) - y(\mathbf{x}_{(1)})\right),\; x_2^{(0)} \operatorname{sign}\left(y(\mathbf{x}_{(1)}) - y(\mathbf{x}_{(2)})\right), \ldots \\ &\ldots,\; x_q^{(0)} \operatorname{sign}\left(y(\mathbf{x}_{(q-1)}) - y(\mathbf{x}_{(q)})\right)\Big)^T\end{aligned}$$

$$\begin{aligned}f\Big(y(\mathbf{x}^\star) \,|\, \mathbf{y}_{\text{OFAT}}\Big) \;=\; &\frac{\left(a^\star_{\text{OFAT}}\right)^{\frac{q+d}{2}} \Gamma\left(\frac{q+d+1}{2}\right)}{\left(\pi\left(\mathbf{x}^{\star T}\mathbf{V}^\star_{\text{OFAT}}\mathbf{x}^\star\right)\right)^{\frac{1}{2}} \Gamma\left(\frac{q+d}{2}\right)} \\ &\times \left(a^\star_{\text{OFAT}} + \frac{\left(y(\mathbf{x}^\star) - \mathbf{x}^{\star T}\mathbf{m}^\star_{\text{OFAT}}\right)^T \left(y(\mathbf{x}^\star) - \mathbf{x}^{\star T}\mathbf{m}^\star_{\text{OFAT}}\right)}{1 + \mathbf{x}^{\star T}\mathbf{V}^\star_{\text{OFAT}}\mathbf{x}^\star}\right)\end{aligned}$$

where

$$
\begin{aligned}
\mathbf{X}_{\text{OFAT}} &= \sum_{i=0}^{q} \mathbf{x}_{(i)}\mathbf{x}_{(i)}^{T} \\
\mathbf{y}_{\text{OFAT}} &= \Big(y(\mathbf{x}_{(0)}),\, y(\mathbf{x}_{(1)}), \ldots, y(\mathbf{x}_{(q)})\Big)^{T} \\
\mathbf{m}^{\star}_{\text{OFAT}} &= \left(\text{Diag}\left(\sigma^{-2}_{\text{ME}}\, I_q,\, \sigma^{-2}_{\text{INT}}\, I_{\frac{q(q-1)}{2}}\right) + (\mathbf{X}_{\text{OFAT}})^{T}(\mathbf{X}_{\text{OFAT}})\right)^{-1} (\mathbf{X}_{\text{OFAT}})^{T}\mathbf{y}_{\text{OFAT}} \\
\mathbf{V}^{\star}_{\text{OFAT}} &= \left(\text{Diag}\left(\sigma^{-2}_{\text{ME}}\, I_q,\, \sigma^{-2}_{\text{INT}}\, I_{\frac{q(q-1)}{2}}\right) + (\mathbf{X}_{\text{OFAT}})^{T}(\mathbf{X}_{\text{OFAT}})\right)^{-1} \\
a^{\star}_{\text{OFAT}} &= a + \mathbf{y}^{T}_{\text{OFAT}}\left(I_{q+1} - \mathbf{X}_{\text{OFAT}}\left(\text{Diag}\left(\sigma^{-2}_{\text{ME}}\, I_q, \sigma^{-2}_{\text{INT}}\, I_{\frac{q(q-1)}{2}}\right) + (\mathbf{X}_{\text{OFAT}})^{T}(\mathbf{X}_{\text{OFAT}})\right)^{-1}(\mathbf{X}_{\text{OFAT}})^{T}\right)\mathbf{y}_{\text{OFAT}}
\end{aligned}
$$

and Diag($\mathbf{A}$, $\mathbf{B}$) denotes a matrix consists of matrix ($\mathbf{A}$, $\mathbf{B}$) on its diagonal and keeps all other elements zero. We have the expectation and variance of the response at $\mathbf{x}^{\star}_{\text{OFAT}}$

$$
\begin{aligned}
\mathbb{E}\left[y\left(\mathbf{x}_{\text{OFAT}}\right)\right] &= \mathbf{f}\left(\mathbf{x}_{\text{OFAT}}\right)^{T}\mathbf{m}^{\star}_{\text{OFAT}} \\
\text{Var}\Big(y(\mathbf{x}_{\text{OFAT}})\Big) &= \frac{a^{\star}_{\text{OFAT}}\Big(1 + (\mathbf{x}_{\text{OFAT}})^{T}\mathbf{V}^{\star}_{\text{OFAT}}\mathbf{x}_{\text{OFAT}}\Big)}{q + d - 2}
\end{aligned}
$$

2.6.2 Posterior Analysis for Saturated Fractional Factorial Experimental Scheme with Unknown σ^2_{ϵ}

In a general case with q input factors, each main effect would have $\left(\frac{q-1}{2}\right)$ aliased interactions. Hence, the model Eqs.(2.5) is now modified to

$$\mathbf{y} = \tilde{\mathbf{X}}\tilde{\beta} + \epsilon$$

where $\mathbf{y}$ denotes the observations through the entire saturated fractional factorial experiments, $\tilde{\mathbf{X}}$ denotes the design matrix in this modified model, and $\tilde{\beta}$ is an $(q \times 1)$ vector with prior information

$$\tilde{\beta}\,|\,\sigma^2_{\epsilon} \sim \mathcal{N}\left(\mathbf{0},\, \left(\sigma^2_{\text{ME}} + \frac{q-1}{2}\sigma^2_{\text{INT}}\right) I_q\right) = \mathcal{N}\left(\mathbf{0},\, \sigma^2_{\epsilon}\left(\frac{2\sigma^2_{\text{ME}} + (q-1)\sigma^2_{\text{INT}}}{2\sigma^2_{\epsilon}}\right) I_q\right) \tag{2.9}$$

Furthermore, from (Wu and Hamada, 2000), we know that the design matrix $\tilde{\mathbf{X}}$ would have the following property

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = (q+1)I_q \tag{2.10}$$

From Eq.(2.9), and Eq.(2.10) and results in Appendix A, we obtain the following posterior distribution of $\tilde{\beta}$

$$\begin{aligned}
\tilde{\beta}\,|\,\mathbf{y}, \sigma_\epsilon^2 &\sim \mathcal{N}\left(\mathbf{m}_{\text{FF}},\, \mathbf{V}_{\text{FF}}\right), \quad \text{where} \\
\mathbf{m}_{\text{FF}} &= \left(\frac{(q+1)(2\sigma^2_{\text{ME}} + (q-1)\sigma^2_{\text{INT}})}{(q+1)(2\sigma^2_{\text{ME}} + (q-1)\sigma^2_{\text{INT}}) + 2\sigma^2_\epsilon}\right)\tilde{\mathbf{X}}^T\mathbf{y} \\
\mathbf{V}_{\text{FF}} &= \sigma^2_\epsilon\left(\frac{(q+1)(2\sigma^2_{\text{ME}} + (q-1)\sigma^2_{\text{INT}})}{(q+1)(2\sigma^2_{\text{ME}} + (q-1)\sigma^2_{\text{INT}}) + 2\sigma^2_\epsilon}\right) I_q
\end{aligned}$$

We now turn to the case in which σ_ϵ is unknown with the prior distribution

$$f(\sigma^2_\epsilon) = \frac{\left(\frac{a}{2}\right)^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}\left(\sigma^2_\epsilon\right)^{-\frac{d+2}{2}} \exp\left(-\frac{a}{2\sigma^2_\epsilon}\right)$$

where $a,\, d > 0$. Then the prior distribution of joint $(\tilde{\beta},\, \sigma^2_\epsilon)$ becomes

$$f(\tilde{\beta},\, \sigma^2_\epsilon) = \frac{\left(\frac{a}{2}\right)^{\frac{d}{2}}}{\left(2\pi\,\left(\sigma^2_{\text{ME}} + \frac{q-1}{2}\sigma^2_{\text{INT}}\right)\right)^{\frac{q}{2}}\Gamma\left(\frac{d}{2}\right)}\left(\sigma^2_\epsilon\right)^{-\frac{d+q+2}{2}} \exp\left(\frac{\left(\sigma^2_{\text{ME}} + \frac{q-1}{2}\sigma^2_{\text{INT}}\right)^{-1}\tilde{\beta}^T\tilde{\beta} + a}{2\sigma^2_\epsilon}\right)$$

and therefore

$$f(\tilde{\beta}) = \frac{(a)^{\frac{d}{2}}\,\Gamma\left(\frac{d+q}{2}\right)}{\left(\pi\left(\sigma^2_{\text{ME}} + \frac{q-1}{2}\sigma^2_{\text{INT}}\right)\right)^{\frac{q}{2}}\,\Gamma\left(\frac{d}{2}\right)}\left(a + \left(\sigma^2_{\text{ME}} + \frac{q-1}{2}\sigma^2_{\text{INT}}\right)^{-1}\tilde{\beta}^T\tilde{\beta}\right)^{-\frac{d+q}{2}}$$

Finally, from Appendix A, the posterior distribution of $(\tilde{\beta},\, \sigma^2_\epsilon)$ is

$$f(\tilde{\beta},\, \sigma^2_\epsilon|\,\mathbf{y}) = \frac{\left(\frac{a^*}{2}\right)^{\frac{d+q}{2}}}{(2\pi)^{\frac{q}{2}}|\mathbf{V}_{\text{FF}}|^{\frac{1}{2}}\Gamma(\frac{d+q}{2})}\left(\sigma^2_\epsilon\right)^{-\frac{d+2q+2}{2}} \exp\left(-\frac{\left(\tilde{\beta} - \mathbf{m}_{\text{FF}}\right)^T(\mathbf{V}_{\text{FF}})^{-1}\left(\tilde{\beta} - \mathbf{m}_{\text{FF}}\right) + a_{\text{FF}}}{2\sigma^2_\epsilon}\right)$$

where

$$\begin{aligned}
\mathbf{m}_{\text{FF}} &= \frac{\sigma^2_{\text{ME}} + \frac{q-1}{2}\sigma^2_{\text{INT}}}{1+(q+1)\left(\sigma^2_{\text{ME}} + \frac{q-1}{2}\sigma^2_{\text{INT}}\right)}\tilde{\mathbf{X}}^T\mathbf{y} \\
\mathbf{V}_{\text{FF}} &= \frac{\sigma^2_{\text{ME}} + \frac{q-1}{2}\sigma^2_{\text{INT}}}{1+(q+1)\left(\sigma^2_{\text{ME}} + \frac{q-1}{2}\sigma^2_{\text{INT}}\right)} I_q \\
a_{\text{FF}} &= a + \frac{1}{1+(q+1)\left(\sigma^2_{\text{ME}} + \frac{q-1}{2}\sigma^2_{\text{INT}}\right)}\mathbf{y}^T\mathbf{y}
\end{aligned} \tag{2.11}$$

By Eq.(2.11), we choose the optimum input level as

$$x_i^\star = \text{sign}\left(\sum_{j=1}^{q}\tilde{X}_{ij}y_j\right), \quad \forall\, i = 1,\ldots,q$$

Then, the posterior distribution of the response at $y(\mathbf{x}^\star)$ would be

$$\begin{aligned}
\mathbf{x}^\star &= (x_1^\star,\ldots,x_q^\star,\, x_1^\star x_2^\star,\ldots,x_{q-1}^\star x_q^\star)^T \\
f(y(\mathbf{x}^\star)|\,\mathbf{y}) &= \frac{\left(a^\star_{\text{FF}}\right)^{\frac{q+d}{2}}\Gamma\left(\frac{q+d+1}{2}\right)}{\left(\pi\left(\mathbf{x}^{\star T}\mathbf{V}^\star_{\text{FF}}\mathbf{x}^\star\right)\right)^{\frac{1}{2}}\Gamma\left(\frac{q+d}{2}\right)}\left(a^\star_{\text{FF}} + \frac{\left(y(\mathbf{x}^\star) - \mathbf{x}^{\star T}\mathbf{m}^\star_{\text{FF}}\right)^T\left(y(\mathbf{x}^\star) - \mathbf{x}^{\star T}\mathbf{m}^\star_{\text{FF}}\right)}{1+\mathbf{x}^{\star T}\mathbf{V}^\star_{\text{FF}}\mathbf{x}^\star}\right)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{m}^\star_{\text{FF}} &= \left(\text{Diag}\left(\sigma^{-2}_{\text{ME}}\, I_q,\, \sigma^{-2}_{\text{INT}}\, I_{\frac{q(q-1)}{2}}\right) + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
\mathbf{V}^\star_{\text{FF}} &= \left(\text{Diag}\left(\sigma^{-2}_{\text{ME}}\, I_q,\, \sigma^{-2}_{\text{INT}}\, I_{\frac{q(q-1)}{2}}\right) + \mathbf{X}^T\mathbf{X}\right)^{-1} \\
a^\star_{\text{FF}} &= a + \mathbf{y}^T\left(I_{q+1} - \mathbf{X}\left(\text{Diag}\left(\sigma^{-2}_{\text{ME}}\, I_q,\, \sigma^{-2}_{\text{INT}}\, I_{\frac{q(q-1)}{2}}\right) + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y}
\end{aligned}$$

Notice that $f(y(\mathbf{x}^\star)|\mathbf{y})$ is a standard t-student distribution. Therefore, we have the expectation and variance of the response at $\mathbf{x}^\star$

$$\begin{aligned}
\mathbb{E}[y(\mathbf{x}^\star)] &= (\mathbf{x}^\star)^T\,\mathbf{m}^\star_{\text{FF}} \\
\text{Var}\left(y(\mathbf{x}^\star)\right) &= \frac{a^\star_{\text{FF}}\left(1+(\mathbf{x}^\star)^T\mathbf{V}^\star_{\text{FF}}\mathbf{x}^\star\right)}{q+d-2}
\end{aligned}$$

2.6.3 A Brief Discussion

It is very difficult to determine whether the matrix $(\mathbf{V}^{\star}_{\text{OFAT}} - \mathbf{V}^{\star}_{\text{FF}})$ is positive definite or not. Observe

$$\mathbf{V}^{\star}_{\text{OFAT}} - \mathbf{V}^{\star}_{\text{FF}} = \text{Diag}\left(\sigma^2_{\text{ME}}\, I_q,\, \sigma^2_{\text{INT}}\, I_{\frac{q(q-1)}{2}}\right)\Bigg\{\mathbf{X}^T_{\text{FF}}\left(I_p + \mathbf{X}_{\text{FF}}\text{Diag}\left(\sigma^{-2}_{\text{ME}}\, I_q, \sigma^{-2}_{\text{INT}}\, I_{\frac{q(q-1)}{2}}\right)\mathbf{X}^T_{\text{FF}}\right)\mathbf{X}_{\text{FF}}$$

$$-\mathbf{X}^T_{\text{OFAT}}\left(I_p + \mathbf{X}_{\text{OFAT}}\text{Diag}\left(\sigma^{-2}_{\text{ME}}\, I_q,\, \sigma^{-2}_{\text{INT}}\, I_{\frac{q(q-1)}{2}}\right)\mathbf{X}^T_{\text{OFAT}}\right)\mathbf{X}_{\text{OFAT}}\Bigg\}\text{Diag}\left(\sigma^2_{\text{ME}}\, I_q,\, \sigma^2_{\text{INT}}\, I_{\frac{q(q-1)}{2}}\right)$$

implying that it is equivalent to determine whether the matrix

$$\mathbf{X}^T_{\text{FF}}\left(I_p + \mathbf{X}_{\text{FF}}\text{Diag}\left(\sigma^{-2}_{\text{ME}}\, I_q, \sigma^{-2}_{\text{INT}}\, I_{\frac{q(q-1)}{2}}\right)\mathbf{X}^T_{\text{FF}}\right)\mathbf{X}_{\text{FF}} - \mathbf{X}^T_{\text{OFAT}}\left(I_p + \mathbf{X}_{\text{OFAT}}\text{Diag}\left(\sigma^{-2}_{\text{ME}}\, I_q, \sigma^{-2}_{\text{INT}}\, I_{\frac{q(q-1)}{2}}\right)\mathbf{X}^T_{\text{OFAT}}\right)\mathbf{X}_{\text{OFAT}}$$

is positive definite. However, since we do not have sufficient information for $\mathbf{X}_{\text{OFAT}}$, the analysis is difficult to proceed. Although the case study indicates that the matrix is positive definite most of the time, it still remains to find the possibility that the matrix is positive definite.

Another comparison is

$$\mathbb{E}\left[y\left(\mathbf{x}^{\star}_{\text{OFAT}}\right)\right] - \mathbb{E}\left[y\left(\mathbf{x}^{\star}_{\text{FF}}\right)\right] = \left(\mathbf{x}^{\star}_{\text{OFAT}}\right)^T \mathbf{m}^{\star}_{\text{OFAT}} - \left(\mathbf{x}^{\star}_{\text{FF}}\right)^T \mathbf{m}^{\star}_{\text{FF}}$$

which is also hard to compare directly.

2.7 Proofs

We need to following lemma to prove Theorem 1.

Lemma 1 *Let X and Y denote two independent Gaussian random variables with distribution $\mathcal{N}(0, \sigma^2_X)$ and $\mathcal{N}(0, \sigma^2_Y)$ respectively. Then, we have*

$$\mathbb{E}\left[X \cdot sgn(X+Y)\right] = \sqrt{\frac{2}{\pi}}\frac{\sigma^2_X}{\sigma^2_X + \sigma^2_Y}$$

where the sign function $sgn(Z) = 1$ if $Z \geq 0$ and $sgn(Z) = -1$ otherwise.

Proof: We integrate by two parts: $X > -Y$ and $X \leq -Y$ as follows

$$\begin{aligned}\mathbb{E}\left[X \cdot sgn(X+Y)\right] &= \int_{-\infty}^{\infty}\left(\int_{-y}^{\infty} x\phi(x,\,\sigma_X^2)\phi\phi(y,\,\sigma_Y^2)dx \right.\\ &\qquad \left. + \int_{-\infty}^{-y}(-x)\phi(x,\,\sigma_X^2)\phi\phi(y,\,\sigma_Y^2)dx\right)dy \\ &= \sqrt{\frac{2}{\pi}}\frac{\sigma_X^2}{\sigma_X^2+\sigma_Y^2}\end{aligned}$$

where $\phi(x,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}$ denotes a normal density function with mean 0 and variance σ^2 at value x. □

Proof: (Theorem 1) Without loss of generality, we choose $\mathbf{x}_0 = (+1, +1, \ldots, +1)$ in this proof. By the OFAT scheme indicated in Eq.(2.7), we have

$$x_1^{(2)} = sgn\left(\beta_1 + \sum_{j=2}^{q}\beta_{1j} + \frac{\epsilon_0 - \epsilon_1}{2}\right)$$

which leads to the expected response

$$\mathbb{E}\left[y(\mathbf{x}_2)\right] = \mathbb{E}\left[\beta_1 x_1^{(2)}\right] + \mathbb{E}\left[\sum_{j=2}^{q}\beta_{1j}x_1^{(2)}\right] + \mathbb{E}\left[\sum_{j=2}^{q}\beta_j + \sum_{i=2}^{q-1}\sum_{j=i+1}^{q}\beta_{ij}\right]$$

From Eq.(2.6), the last term reduces to 0. Moreover, the terms $\mathbb{E}\left[\beta_{1j}x_1^{(2)}\right]$ are identical. Therefore, we have

$$\mathbb{E}\left[y(\mathbf{x}_2)\right] = \mathbb{E}\left[\beta_1 x_1^{(2)}\right] + (q-1)\mathbb{E}\left[\beta_{1j}x_1^{(2)}x_j^{(0)}\right]$$

Now applying Lemma 1 by taking $X = \beta_1$ and $Y = \sum_{j=2}^{q}\beta_{1j} + \frac{\epsilon_0-\epsilon_1}{2}$, in which X and Y are independent by assumption, we obtain

$$\mathbb{E}\left[\beta_1 x_2^{(1)}\right] = \sqrt{\frac{2}{\pi}}\frac{\sigma_{\text{ME}}^2}{\sqrt{\sigma_{\text{ME}}^2 + (q-1)\sigma_{\text{INT}}^2 + \frac{\sigma_\epsilon^2}{2}}}$$

Repeating to apply Lemma 1 by having $X = \beta_{1j}$ and $Y = \beta_1 + \sum_{k\neq j,k=2}^{q} \beta_{1k} + \frac{\epsilon_0+\epsilon_1}{2}$, we have

$$\mathbb{E}\left[\beta_{1j} x_1^{(2)} x_j^{(0)}\right] = \sqrt{\frac{2}{\pi}} \frac{\sigma^2_{\text{INT}}}{\sqrt{\sigma^2_{\text{ME}} + (q-1)\sigma^2_{\text{INT}} + \frac{\sigma^2_\epsilon}{2}}}$$

□

We need the following lemma to prove the Theorem 2.

Lemma 2 *Let X and Y denote two independent Gaussian random variables with distribution $\mathcal{N}(0, \sigma_X^2)$ and $\mathcal{N}(0, \sigma_Y^2)$ respectively. Then, we have*

$$P(X \geq 0,\, X + Y \geq 0) = \frac{1}{4} + \frac{1}{2\pi} \arcsin\left(\frac{\sigma_X}{\sigma_X^2 + \sigma_Y^2}\right)$$

Proof: By definition,

$$\begin{aligned} P(X \geq 0,\, X + Y \geq 0) &= \int_0^\infty \int_{-x}^\infty \phi(x, \sigma_X^2)\phi(y,\, \sigma_Y^2) dy\, dx \\ &= \frac{1}{4} + \frac{1}{2\pi} \arcsin\left(\frac{\sigma_X}{\sigma_X^2 + \sigma_Y^2}\right) \end{aligned}$$

□

Proof: (Theorem 2) Follow the notations in the proof of Theorem 1, we have

$$\begin{aligned} P\left(\beta_1 x_1^{(2)} \geq 0\right) &= P\left(\beta_1 \geq 0,\, x_1^{(2)} = 1\right) + P\left(\beta_1 \leq 0,\, x_1^{(2)} = -1\right) \\ &= P\left(\beta_1 \geq 0,\, \beta_1 + \sum_{j=2}^{q} \beta_{1j} + \frac{\epsilon_0 - \epsilon_1}{2} \geq 0\right) \\ &\quad + \left(\beta_1 \leq 0,\, \beta_1 + \sum_{j=2}^{q} \beta_{1j} + \frac{\epsilon_0 - \epsilon_1}{2} \leq 0\right) \\ &= 2P\left(\beta_1 \geq 0,\, \beta_1 + \sum_{j=2}^{q} \beta_{1j} + \frac{\epsilon_0 - \epsilon_1}{2} \geq 0\right) \end{aligned}$$

where the last equality holds because the symmetric prior assumption Eqs.(2.6). Now we apply Lemma 2 by setting $X = \beta_1$ and $Y = \sum_{j=2}^{q} \beta_{1j} + \frac{\epsilon_0 - \epsilon_1}{2}$, we have

$$P\left(\beta_1 x_1^{(2)} > 0\right) = \frac{1}{2} + \frac{1}{\pi} \arcsin \frac{\sigma_{\text{ME}}}{\sqrt{\sigma_{\text{ME}}^2 + (q-1)\sigma_{\text{INT}}^2 + \frac{\sigma_\epsilon^2}{2}}}.$$

□

We will apply the following lemma to prove Theorem 3.

Lemma 3 *Let* $X \sim \mathcal{N}(0, \sigma_X^2)$, $Y_1, Y_2 \sim \mathcal{N}(0, \sigma_Y^2)$, *and* $Z \sim \mathcal{N}(0, \sigma_Z^2)$ *denote mutually independent random variables. Then we have*

$$\mathbb{E}\left[X \cdot sgn\left(X + Y_1 - Y_2 + (Z + Y_1 + Y_2)\,\alpha\right)\right] = \sqrt{\frac{2}{\pi}} \left(\frac{\sigma_X^2}{\sigma_X^2 + 4\sigma_Y^2 + \sigma_Z^2}\right)$$

where $\alpha \in \{+1, -1\}$.

Proof: This is an extension to Lemma 1. Let $Y = Y_1 - Y_2 + (Z + Y_1 + Y_2)\,\alpha$, and we have $Y \sim \mathcal{N}(0, 4\sigma_Y^2 + \sigma_Z^2)$ for either $\alpha = +1$ or $\alpha = -1$. Moreover, X and Y are independent by assumption. Then the proof is complete by applying Lemma 1. □

Proof: (Theorem 3) By Eqs.(2.7), we have

$$x_2^{(3)} = sgn\left(\beta_2 + \beta_{12} x_1^{(3)} + \sum_{j=3}^{q} \beta_{2j} + \frac{\frac{1+x_1^{(3)}}{2}\epsilon_0 + \frac{1-x_1^{(3)}}{2}\epsilon_1}{2} - \epsilon_2\right)$$

Hence,

$$\begin{aligned}\mathbb{E}\left[y(\mathbf{x}_3)\right] \;=\; & \mathbb{E}\left[\left(\beta_1 + \sum_{j=3}^{q} \beta_{1j}\right) x_1^{(3)}\right] + \mathbb{E}\left[\left(\beta_2 + \sum_{j=3}^{q} \beta_{2j}\right) x_2^{(3)}\right] \\ & + \mathbb{E}\left[\beta_{12} x_1^{(3)} x_2^{(3)}\right] + \mathbb{E}\left[\sum_{i=3}^{q} \beta_i + \sum_{i=3}^{q-1} \sum_{j=i+1}^{q} \beta_{ij}\right]\end{aligned}$$

in which the last term reduces to zero by the prior assumption Eqs.(2.6). The first two terms could be obtain by applying Lemma 1 and find both amounts to $\mathbb{E}\left[\beta_1 x_1^{(3)}\right] +$

$(q-2)\mathbb{E}\left[\beta_{1j}x_1^{(3)}\right]$. It remains to find the term $\mathbb{E}\left[\beta_{12}x_1^{(3)}x_2^{(3)}\right]$. By letting $\sigma_X^2 = \sigma_{\text{INT}}^2$, $\sigma_Y^2 = \frac{\sigma_\epsilon^2}{16}$, and $\sigma_Z^2 = \sigma_{\text{ME}}^2 + (q-2)\sigma_{\text{INt}}^2 + \frac{\sigma_\epsilon^2}{4}$, and applying Lemma 3, we obtain

$$\mathbb{E}\left[\beta_{12}x_1^{(3)}x_2^{(3)}\right] = \sqrt{\frac{2}{\pi}}\frac{\sigma_{\text{INT}}^2}{\sqrt{\sigma_{\text{ME}}^2 + (q-1)\sigma_{\text{INT}}^2 + \frac{\sigma_\epsilon^2}{2}}}.$$

□

Proof: (Theorem 4) The probability to be determined could be expanded as

$$\begin{aligned} P\left(\beta_{12}x_1^{(3)}x_2^{(3)} \geq 0\right) &= P\left(\beta_{12} \geq 0,\, x_1^{(3)} = +1,\, x_2^{(3)} = +1\right) + P\left(\beta_{12} \geq 0,\, x_1^{(3)} = -1,\, x_2^{(3)} = -1\right) \\ &+ P\left(\beta_{12} \leq 0,\, x_1^{(3)} = +1,\, x_2^{(3)} = -1\right) + P\left(\beta_{12} \leq 0,\, x_1^{(3)} = -1,\, x_2^{(3)} = +1\right) \end{aligned}$$

Consider the first term of the expression above first. Substituting in the values of the final settings of the variables $x_1^{(3)}$ and $x_2^{(3)}$, we have

$$\begin{aligned} &P\left(\beta_{12} \geq 0,\, x_1^{(3)} = +1,\, x_2^{(3)} = +1\right) \\ &= P\left(\beta_{12} \geq 0,\, \beta_{12} + \beta_1 + \sum_{j=3}^{q}\beta_{1j} + \frac{\epsilon_0 - \epsilon_1}{2} \geq 0,\, \beta_{12} + \beta_2 + \sum_{j=3}^{q}\beta_{2j} + \frac{\epsilon_0 - \epsilon_2}{2} \geq 0\right) \end{aligned}$$

By (Tong, 1990), we know that, if $\mathbf{X} = (X_1,\, X_2,\, X_3)^T \sim \mathcal{N}(\mathbf{0},\, \Sigma)$, where Σ is a 3×3 matrix, then we have

$$P(\min\{X_1, X_2, X_3\} \geq 0) = \frac{1}{8} + \frac{1}{4\pi}\left(\sin^{-1}\left(\frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}}\right) + \sin^{-1}\left(\frac{\Sigma_{13}}{\sqrt{\Sigma_{11}\Sigma_{33}}}\right) + \sin^{-1}\left(\frac{\Sigma_{23}}{\sqrt{\Sigma_{22}\Sigma_{33}}}\right)\right)$$

We apply this result to the first term with $\Sigma_{1j} = \sigma_{\text{INT}}^2$, $\forall\, j = 1, 2, 3$, $\Sigma_{23} = \sigma_{\text{INT}}^2 + \frac{\sigma_\epsilon^2}{4}$ and $\Sigma_{22} = \Sigma_{33} = \sigma_{\text{ME}}^2 + (q-1)\sigma_{\text{INT}}^2 + \frac{\sigma_\epsilon^2}{2}$, and thus obtain

$$\begin{aligned} &\frac{1}{8} + \frac{1}{2\pi}\arctan\left(\frac{\sigma_{\text{INT}}}{\sqrt{\sigma_{\text{ME}}^2 + (q-2)\sigma_{\text{INT}}^2 + \frac{\sigma_\epsilon^2}{2}}}\right) \\ &\qquad + \frac{1}{4\pi}\arctan\left(\frac{\sigma_{\text{INT}}^2 + \frac{\sigma_\epsilon^2}{4}}{\left(\sigma_{\text{ME}}^2 + (q-2)\sigma_{\text{INT}}^2 + \frac{\sigma_\epsilon^2}{2}\right)\sqrt{\sigma_{\text{ME}}^2 + q\sigma_{\text{INT}}^2 + \frac{\sigma_\epsilon^2}{2}}}\right) \end{aligned}$$

Applying the same procedure to the three other three terms gives expressions in a similar form with only the signs of the terms changing. Summing up the four expressions, we get the expression in Theorem 4. □

Proof: (Theorem 5) For some $j \in \{2, 3, \ldots, q\}$, we consider interaction β_{ij} and β_{kj} for some $i, k < j$. Then the probability that interaction β_{ij} is exploited could divided into 2^{j-1} cases which is identical to the cases that we consider in the probability of exploiting β_{kj}. Substituting β_{ij} by β_{kj} completes the proof. □

Proof: (Theorem 6) By Eqs.(2.7), we have

$$x_3^{(q)} = sgn\left(\beta_3 + \sum_{i=1}^{2}\beta_{i3}x_i^{(q)} + \sum_{j=4}^{q}\beta_{3j} + \frac{\frac{1+x_1^{(q)}}{2}\frac{1+x_2^{(q)}}{2}\epsilon_0 + \frac{1-x_1^{(q)}}{2}\frac{1+x_2^{(q)}}{2}\epsilon_1 + \frac{1-x_2^{(q)}}{2}\epsilon_2 - \epsilon_3}{2}\right)$$

The probability $P(\beta_{13}x_1^{(q)}x_3^{(q)})$ could be extended to four terms as we did in the proof of Theorem 4. The sum of the first two terms would be larger than

$$2\times P\left(\beta_{13} \geq 0,\ \beta_1 - \beta_{12} - \beta_{13} + \sum_{j=2}^{q}\beta_{1j} + \frac{\epsilon_0 - \epsilon_1}{2} \geq 0,\ \beta_2 + \beta_{12} - \beta_{23} + \sum_{j=4}^{q}\beta_{3j} + \frac{\epsilon_3 - \epsilon_2}{2} \geq 0\right)$$

because substituting the term $\beta_{13} \geq 0$ by $\beta_{13} \leq 0$ in the second term makes the joint event less likely. Hence we have $P(\beta_{13}x_1^{(q)x_3^{(q)}}) \geq P(\beta_{12}x_1^{(q)x_2^{(q)}})$ because the substitution of β_{13} for β_{12} makes the joint event less likely. A similar process holds for every case in which the interaction β_{13} is exploited. By Theorem 6, we have $P(\beta_{i3}x_i^{(q)x_3^{(q)}}) \geq P(\beta_{12}x_1^{(q)x_2^{(q)}})$ for all $i \leq 3$. Repeating reasoning applied to β_{13} for every β_{1j} for all $3 \leq j \leq q$, we have $P(\beta_{1j}x_1^{(q)x_j^{(q)}}) \geq P(\beta_{12}x_1^{(q)x_2^{(q)}})$. Again, by repeating the application of Theorem 6, we complete the proof. □

Proof: (Theorem 7) From Eq.(2.5), we have

$$\begin{aligned}
\mathbb{E}\left[y(\mathbf{x}_{\text{OFAT}})\right] &= \mathbb{E}\left[\beta_0\right] + \sum_{i=1}^{q}\mathbb{E}\left[\beta_i x_i^{(q)}\right] + \sum_{i=1}^{q-1}\sum_{j=i+1}^{q}\mathbb{E}\left[\beta_{ij}x_i^{(q)}x_j^{(q)}\right] + \mathbb{E}\left[\epsilon\right] \\
&= \sum_{i=1}^{q}\left(\mathbb{E}\left[|\beta_i|\right]\mathbf{P}\left(\beta_i x_1^{(q)} \geq 0\right) + \mathbb{E}\left[-|\beta_i|\right]\mathbf{P}\left(\beta_i x_1^{(q)} \leq 0\right)\right) + \sum_{i=1}^{q-1}\sum_{j=i+1}^{q}\mathbb{E}\left[\beta_{ij}x_i^{(q)}x_j^{(q)}\right] \\
&= \sum_{i=1}^{q}\left(\left(2\mathbf{P}\left(\beta_i x_i^{(q)} \geq 0\right) - 1\right)\cdot\mathbb{E}\left[|\beta_i|\right]\right) + \sum_{i=1}^{q-1}\sum_{j=i+1}^{q}\left(\left(2\mathbf{P}\left(\beta_{ij}x_i^{(q)}x_j^{(q)} \geq 0\right) - 1\right)\cdot\mathbb{E}\left[|\beta_{ij}|\right]\right) \\
&\geq \sum_{i=1}^{q}\frac{2}{\pi}\arctan\left(\frac{\sigma_{\text{ME}}}{\sqrt{(q-1)\sigma_{\text{INT}}^2 + \frac{\sigma_\epsilon^2}{2}}}\right)\times\sqrt{\frac{2}{\pi}}\sigma_{\text{ME}} \\
&\qquad + \sum_{i=1}^{q-1}\sum_{j=1}^{q}\frac{2}{\pi}\arctan\left(\frac{\sigma_{\text{INT}}}{\sqrt{\sigma_{\text{ME}}^2 + (q-2)\sigma_{\text{INT}}^2 + \frac{\sigma_{\text{ME}}^2}{2}}}\right)\times\sqrt{\frac{2}{\pi}}\sigma_{\text{INT}} \\
&= q\times\left(\frac{2}{\pi}\right)^{\frac{3}{2}}\times\arctan\left(\frac{\sigma_{\text{ME}}}{\sqrt{(q-1)\sigma_{\text{INT}}^2 + \frac{\sigma_\epsilon^2}{2}}}\right)\times\sigma_{\text{ME}} \\
&\qquad + \frac{q(q-1)}{2}\times\left(\frac{2}{\pi}\right)^{\frac{3}{2}}\times\arctan\left(\frac{\sigma_{\text{INT}}}{\sqrt{\sigma_{\text{ME}}^2 + (q-2)\sigma_{\text{INT}}^2 + \frac{\sigma_{\text{ME}}^2}{2}}}\right)\times\sigma_{\text{INT}}
\end{aligned}$$

The inequality come from the conclusion of Theorem 6 and the fact that $\mathbb{E}\left[|\beta_i|\right] = \sqrt{\frac{2}{\pi}}\sigma_{\text{ME}}$ and $\mathbb{E}\left[|\beta_{ij}|\right] = \sqrt{\frac{2}{\pi}}\sigma_{\text{INT}}$ for all possible i and j. □

Proof: (Theorem 8) Given $q+1$ observations of responses as defined in Eqs.(2.5) are made using a saturated resolution III fractional factorial experiment with a design matrix $\mathbf{X}$. It follows that the estimate for each main effect factor β_j would be

$$\hat{\beta}_j = \beta_j + \sum_{i=1}^{q+1}\epsilon_i\mathbf{X}(i,\,j) + \frac{1}{q+1}\sum_{i=1}^{q+1}\sum_{j=1,\,j\neq i}^{q}\beta_{ij}\mathbf{X}(i,\,j)$$

The choice of factor levels is set according to the sign of the main effect estimates. Therefore the expected value is

$$\mathbb{E}\left[\beta_i \cdot sgn\left(\beta_j + \sum_{i=1}^{q+1} \epsilon_i \mathbf{X}(i,j) + \frac{1}{q+1}\sum_{i=1}^{q+1}\sum_{j=1, j\neq i}^{q} \beta_{ij}\mathbf{X}(i,j)\right)\right]$$

By Lemma 1 and the properties of the design matrix X, the result of Theorem 8 follows. □

Proof: (Theorem 9) Since the experiment is saturated, the total number runs of the experiment is $q+1$. Then, the probability that the main effect β_i will be exploited is

$$\begin{aligned}
&\mathbf{P}\left(\beta_1 \geq 0, (q+1)\beta_i + (q+1)\sum_{i=1}^{q}\sum_{j=i+1}^{q+1}\beta_{ij} + \sum_{j=1}^{q+1}\epsilon_j \geq 0\right) + \mathbf{P}\left(\beta_1 \leq 0, (q+1)\beta_i + (q+1)\sum_{i=1}^{q}\sum_{j=i+1}^{q+1}\beta_{ij} + \sum_{j=1}^{q+1}\epsilon_j \leq 0\right) \\
&= 2\mathbf{P}\left(\beta_1 \geq 0,\ (q+1)\beta_i + (q+1)\sum_{i=1}^{q}\sum_{j=i+1}^{q+1}\beta_{ij} + \sum_{j=1}^{q+1}\epsilon_j \geq 0\right) \\
&= 2\mathbf{P}\left(\beta_1 \geq 0,\ \beta_i + \sum_{i=1}^{q}\sum_{j=i+1}^{q+1}\beta_{ij} + \sum_{j=1}^{q+1}\frac{\epsilon_j}{q+1} \geq 0\right) \\
&= \mathbf{P}\left(\beta_1 \geq 0 \,\middle|\, \beta_i + \sum_{i=1}^{q}\sum_{j=i+1}^{q+1}\beta_{ij} + \sum_{j=1}^{q+1}\frac{\epsilon_j}{n+1} \geq 0\right) \\
&= \frac{1}{2} + \frac{1}{\pi}\arctan\frac{\sigma_{\mathrm{ME}}}{\sqrt{\frac{(q-1)(q-2)}{2(q+1)}\sigma^2_{\mathrm{INT}} + \frac{\sigma^2_\epsilon}{q+1}}}
\end{aligned}$$

The first equality comes from the symmetry assumption Eq.(2.6), and the third equality follows the Bayes' rule and the last equality follows Lemma 2 in our paper. □

Chapter 3

The Optimization Scheme Through Sequential Experiments in Compact Design Space

In Chapter 2, we study the sequential experiments in the finite design space, propose the adaptive OFAT scheme, and present the (pre-)posterior analysis of its outcome. However, although easily implemented, the scheme is restricted at least in two ways in addition to the restriction of finite design space. The first restriction is that the selection of input level $\mathbf{x}$ at each iteration depends only on the observations, but not on the posterior analysis at all, implying that the information from the whole experiment history might not be well utilized. The second restriction is that the normal linear model for the posterior analysis is confined to be a first-order polynomial model, implying that the model considers a relatively small design space. In this section, we will remove these two restrictions, and consider the problem of optimization through sequential experiments in compact design space. We will first present the model considered in this chapter and then propose a general algorithm. Two important properties of the algorithm will be investigated. One is the consistency issue, which justifies the use of the algorithm in the limiting behavior. The other issue is the expected improvement we could expect from one experiment to the next. This issue would help experimenter to determine when the experiments should stop. To begin

with, we will present the model considered in this chapter.

3.1 The Fundamental Model

We consider the following linear regression model

$$y(\mathbf{x}_k,\, \beta) = \mathbf{f}(\mathbf{x}_k)^T\beta + \epsilon_k, \quad \forall k = 1, 2, \ldots$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ are the unobservable *i.i.d.* experimental errors, $\beta = (\beta_1, \ldots, \beta_p)^T$ are unknown parameters, and $y_i \triangleq y(\mathbf{x}_i,\, \beta)$ is the observed response corresponding to the input levels $\mathbf{f}(\mathbf{x}_i) = (f_1(\mathbf{x}_i), \ldots, f_p(\mathbf{x}_i))$ in which the input level $\mathbf{x}_i^T = \left(x_1^{(i)}, \ldots, x_q^{(i)}\right)$ is selected in a nonempty compact set $\mathcal{X} \subset \mathbb{R}^q$, and the function $f_i(\cdot)$ is assumed to be continuous for all i. Following the Bayesian embedding approach, we assume that the unknown parameter β has a prior

$$\beta \,|\, \sigma_\epsilon^2 \sim \mathcal{N}\left(\mathbf{m}_0,\, \sigma_\epsilon^2\mathbf{V}_0\right),$$

and we denote $\bar{\beta}$ as the true value of β. Note that, unlike what we assume in Chapter 2, the prior of β given σ_ϵ^2 has a mean $\mathbf{m}_0$ which is not necessarily zero. Moreover, we also assume the matrix $\mathbf{V}_0$ to be positive definite to avoid triviality.

3.2 The Proposed Algorithm

Ideally, we would like to locate the optimal input level with a tradeoff between its expected response and the corresponding variance. In summary, an algorithm to achieve our goal will go through the following steps:

1. run the kth experiment

2. update the posterior analysis of parameter β, the mean $\mathbf{m}_k$ and the variance $\mathbf{V}_k$, with the information collected from the kth experiment

3. locate the optimal input level $\mathbf{c}_{k+1}$ according to current estimation of mean $\mathbf{m}_k$ and variance $\mathbf{V}_k$

4. determine appropriate input level $\mathbf{x}_{k+1}$ for the next run of experiment

5. ensure the consistency of the proposed algorithm

6. investigate the expected improvement of each iteration to determine whether to conduct next run of experiment or not

To present the adaptive experiment strategy explicitly, we propose the following algorithm

$$\mathbf{c}_{k+1} = \arg\max_{\mathbf{c}\in\mathcal{X}} \quad \mathbf{f}(\mathbf{c})^T\mathbf{m}_k - \frac{\alpha_k}{2}\Big(\mathbf{f}(\mathbf{c})^T\mathbf{V}_k\mathbf{f}(\mathbf{c})\Big) \tag{3.1}$$

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \mathbf{f}(\mathbf{c}_{k+1})^T\mathbf{V}_{k+1}\mathbf{f}(\mathbf{c}_{k+1})^T = \arg\max_{\mathbf{x}\in\mathcal{X}} \frac{\Big(\mathbf{f}(\mathbf{c}_{k+1})^T\mathbf{V}_k\mathbf{f}(\mathbf{x})\Big)^2}{1+\mathbf{f}(\mathbf{x})^T\mathbf{V}_k\mathbf{f}(\mathbf{x})} \tag{3.2}$$

$$\mathbf{V}_k = \Big(\mathbf{V}_{k-1}^{-1} + \mathbf{f}(\mathbf{x}_k)\mathbf{f}(\mathbf{x}_k)^T\Big)^{-1} \tag{3.3}$$

$$\begin{aligned}\mathbf{m}_k &= \mathbf{V}_k\Big(\mathbf{V}_{k-1}^{-1}\mathbf{m}_{k-1} + y_k\mathbf{f}(\mathbf{x}_k)\Big) \\ &= \left(I_p - \frac{\mathbf{V}_{k-1}\mathbf{f}(\mathbf{x}_k)\mathbf{f}(\mathbf{x}_k)^T}{1+\mathbf{f}(\mathbf{x}_k)^T\mathbf{V}_{k-1}\mathbf{f}(\mathbf{x}_k)}\right)\Big(\mathbf{m}_{k-1} + y_k\left(\mathbf{V}_{k-1}\mathbf{f}(\mathbf{x}_k)\right)\Big)\end{aligned} \tag{3.4}$$

We would like to make several remarks on this algorithm to facilitate the discussion of this proposed algorithm.

1. Eq.(3.1) considers the tradeoff between the high mean and low variance, and the sequence of strictly positive coefficients $\{\alpha_k\}$ denotes the tradeoff between the mean, $\mathbf{f}(\mathbf{c})^T\mathbf{m}_k$, and variance, $\mathbf{f}(\mathbf{c})^T\mathbf{V}_k\mathbf{f}(\mathbf{c})$, at step k

2. Eq.(3.2) minimizes the volatility corresponding to the point of interest $\mathbf{c_k}$; furthermore, the numerator $\mathbf{f}(\mathbf{c})^T\mathbf{V}_k\mathbf{f}(\mathbf{x})$ denotes the covariance between point $\mathbf{c}$ and $\mathbf{x}$ after the kth experiment

3. $\mathbf{V}_k$ denotes the posterior covariance matrix of parameter β given known σ_ϵ^2

4. Eq.(3.4) represents the least square estimator, which obtains the minimum variance among all the linear estimator; in this case, since the unobserved errors

are all *i.i.d.* normally distributed random variables, the LSE and the Bayesian estimator coincide

3.2.1 A Brief Discussion On The Problem of Optimization Through Sequential Experiments

Lai (2001) has discussed this problem in Section 5, in which he treated this problem as a self-tuning problem in the control field as we discussed in Chapter 1. Later, in the comment of this paper, Wei extended the problem discussed in Lai (2001) to the multivariate case and said that *"... the mean response is maximized while controlling the variance at a specific level. It is not clear what is the effect of the roles of the means and variance on the associated procedure."* We are motivated by this paper and would like to investigate an approach which could locate optimum input level with acceptable variance in finite number of experiments. Pronzato (2000) proposed an iteration algorithm

$$\mathbf{x}_{k+1} = \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{ \mathbf{f}(\mathbf{x})^T \mathbf{m}_k + \alpha_k \mathbf{f}(\mathbf{x})^T \mathbf{V}_k^{-1} \mathbf{f}(\mathbf{x}) \right\} \tag{3.5}$$

In the paper, although the convergence rate remains unknown, Pronzato (2000) proved that the sequence $\{\mathbf{x}_k\}$ generated by above algorithm (3.5) would *asymptotically* achieve the optimal input level $\mathbf{x}^*$ which maximizes the objective function

$$\mathbf{f}(\mathbf{x})^T \mathbf{m}_\infty = \mathbf{f}(\mathbf{x})^T \overline{\beta}.$$

Recall that $\overline{\beta}$ indicates the real value of unknown β. Later, Pronzato and Thierry (2003) discussed the finite case by applying the concept of dynamic programming, although the issue of dimensionality has not been discussed. The algorithm that Pronzato (2000) proposed is indeed an adaptive D-optimum experimental design. The idea of the approach is that the algorithm needs to be repeated *infinitely* many times to have the event $\left\{ \mathbf{V}_k^{-1} \to \mathbf{0} \right\}$ occurs almost surely. However, since the estimation is consistent, the achievement of locating the optimum input levels should be straight-

forward.

We will proceed the analysis of the proposed algorithm in two categories: one with known σ_ϵ^2 and the other with unknown σ_ϵ^2. In the former case, we will focus on the consistency issue and the step-wise improvement. In the other case, we will derive an explicit presentation for the stepwise improvements.

3.3 Bayesian Analysis with Known σ_ϵ^2

3.3.1 Consistency of the Bayesian Estimator $\mathbf{m}_k$

Lai and Wei (1982) have shown that the necessary and sufficient condition for the least square estimator to be consistent is that

$$\lambda_{\max}(\mathbf{V}_k) \to 0, \quad \log(\lambda_{\min}(\mathbf{V}_k)) = o(\lambda_{\max}(\mathbf{V}_k)) \qquad \text{a.s.}$$

where $\lambda_{\max}(\mathbf{V})$ denotes the maximum eigenvalue of matrix $\mathbf{V}$, and $\lambda_{\min}(\mathbf{V})$ denotes the minimum eigenvalue of matrix $\mathbf{V}$. Moreover, Hu (1996) also argues that the necessary and sufficient condition for the Bayesian estimator to be strongly consistent almost surely is that

$$\lambda_{\max}(\mathbf{V}_k) \to 0 \tag{3.6}$$

We are going to demonstrate the consistency property of the Bayesian estimator presented in Eq.(3.4) by showing that the sequences $\{\mathbf{c}_k\}$ and $\{\mathbf{x}_k\}$ generated by proposed algorithm satisfy the condition (3.6) in the following theorem.

Theorem 10 (Consistency) *If $f(\mathbf{x}_k) \neq \mathbf{0}$ for all possible k, then*

$$\lim_{k\to\infty} \mathbf{V}_k \mathbf{f}(\mathbf{x}_{k+1}) = 0 \qquad \text{and} \qquad \lambda_{max}(\mathbf{V}_k) \to 0 \qquad a.s.$$

Proof: Pick some $\mathbf{c} \in \mathcal{X}$ and define the value function

$$J_{k+1}(\mathbf{c}) \triangleq \mathbf{f}(\mathbf{c})^T \mathbf{V}_{k+1} \mathbf{f}(\mathbf{c}) = \mathbf{f}(\mathbf{c})^T \mathbf{V}_k \mathbf{f}(\mathbf{c}) - \frac{\left(\mathbf{f}(\mathbf{c})^T \mathbf{V}_k \mathbf{f}(\mathbf{x}_{k+1})\right)^2}{1 + \mathbf{f}(\mathbf{x}_{k+1})^T \mathbf{V}_k \mathbf{f}(\mathbf{x}_{k+1})}, \quad \forall k = 0, 1, 2, \ldots$$

Then, we have

$$J_k(\mathbf{c}) - J_{k-1}(\mathbf{c}) = -\frac{\left(\mathbf{f}(\mathbf{c})^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k)\right)^2}{1 + \mathbf{f}(\mathbf{x}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k)} \leq 0$$

Hence, after each iteration, the estimation volatility of each point in the design space $\mathcal{X}$ will decrease. By iteration, we have

$$J_k(\mathbf{c}) = J_0(\mathbf{c}) - \sum_{s=1}^{k} \frac{\left(\mathbf{f}(\mathbf{c})^T \mathbf{V}_{s-1} \mathbf{f}(\mathbf{x}_s)\right)^2}{1 + \mathbf{f}(\mathbf{x}_s)^T \mathbf{V}_{s-1} \mathbf{f}(\mathbf{x}_s)}$$

Notice that this is valid for all $\mathbf{c} \in \mathcal{X}$ and k. Thus, we have

$$\infty > \sum_{s=1}^{\infty} \frac{\left(\mathbf{f}(\mathbf{c})^T \mathbf{V}_{s-1} \mathbf{f}(\mathbf{x}_s)\right)^2}{1 + \mathbf{f}(\mathbf{x}_s)^T \mathbf{V}_{s-1} \mathbf{f}(\mathbf{x}_s)} \geq \sum_{s=1}^{\infty} \frac{\left(\mathbf{f}(\mathbf{c})^T \mathbf{V}_{s-1} \mathbf{f}(\mathbf{x}_s)\right)^2}{1 + \lambda_{\max}(\mathbf{V}_{i-1}) B^2} \geq \frac{\sum_{s=1}^{\infty} \left(\mathbf{f}(\mathbf{c})^T \mathbf{V}_{s-1} \mathbf{f}(\mathbf{x}_s)\right)^2}{1 + \lambda_{\max}(\mathbf{V}_0) B^2}$$

in which, by the compactness/discreteness of $\mathcal{X}$ and the continuity of all f_i's, there exists a number $B > 0$ such that $\mathbf{f}(\mathbf{x})^T \mathbf{f}(\mathbf{x}) \leq B$ for all $\mathbf{x} \in \mathcal{X}$. This implies that

$$\lim_{k \to \infty} \mathbf{f}(\mathbf{c})^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) = 0, \quad \forall \mathbf{c} \in \mathcal{X} \tag{3.7}$$

Therefore, if $\mathbf{f}(\mathbf{x}_k) \neq \mathbf{0}$ for all possible k, we could have

$$\lim_{k \to \infty} \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) = \mathbf{0}.$$

Next, we show that $\lambda_{\max}(\mathbf{V}_k) \to 0$. Since Eq.(3.7) is valid for all $\mathbf{c} \in \mathcal{X}$, we replace $\mathbf{c}_k$ by $\mathbf{x}_k$, and have

$$\lim_{k \to \infty} \mathbf{f}(\mathbf{x}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) = 0$$

Since we do not restrict the generation of the sequence $\{\mathbf{x}_k\}$ but only assume that $\mathbf{f}(\mathbf{x}_k) \neq \mathbf{0}$, $\forall k$, there must exist a sufficient small coefficient $\delta > 0$ such that

$$\mathbf{f}(\mathbf{x})^T \mathbf{f}(\mathbf{x}) > \delta, \quad \forall \mathbf{x} \in \tilde{\mathcal{X}} = \mathcal{X} \setminus \{\mathbf{x} \in \mathcal{X} \,|\, \mathbf{f}(\mathbf{x}) = \mathbf{0}\}$$

By Eq.(3.7), we also have $\mathbf{f}(\mathbf{x})^T \mathbf{V}_k \mathbf{f}(\mathbf{x}) \to 0$ for all $\mathbf{x} \in \tilde{\mathcal{X}}$. Note that, it implies that

$$\max_{\mathbf{x} \in \tilde{\mathcal{X}}} \mathbf{f}(\mathbf{x})^T \mathbf{V}_k \mathbf{f}(\mathbf{x}) > \delta \lambda_{\max}(\mathbf{V}_k)$$

Since the left hand side approaches to zero asymptotically, we have

$$\lambda_{\max}(\mathbf{V}_k) \to 0$$

. □

It remains to show that $\mathbf{f}(\mathbf{x}_k) \neq \mathbf{0}$ for all possible k. Consider the sequence $\mathbf{x}_k$ generated by Eq.(3.2) and suppose that, for some k, we have $\mathbf{f}(\mathbf{x}_k) = \mathbf{0}$. In the event of $\mathbf{f}(\mathbf{c}_k) = \mathbf{0}$ or $\mathbf{V}_{k-1} = \mathbf{0}$, we could choose $\mathbf{f}(\mathbf{x}_k) \neq \mathbf{0}$ without changing the corresponding variance. If the events $\mathbf{f}(\mathbf{c}_k) \neq \mathbf{0}$ or $\mathbf{V}_{k-1} \neq \mathbf{0}$ occur, then we have a contradiction. Consequently, we conclude that the algorithm we proposed in Eqs.(3.1) to (3.4) is consistent almost surely.

3.3.2 Discussion for Selection of Sequence $\{\alpha_k\}$

We only require that the sequence $\{\alpha_k\}$ be positive in our algorithm. However, in the literature of stochastic control, the selection of sequence $\{\alpha_k\}$ plays an important role in ensuring the consistency of the estimator. Usually, the selection of $\{\alpha_k\}$ is with a strict constraint. For example, Pronzato and Thierry (2003) requires

(i) $\frac{\alpha_k}{k} \log \alpha_k$ decreases monotonically, and

(ii) $\frac{\alpha_k}{(\log k)^{1+\delta}}$ increases monotonically to ∞ for some $\delta > 0$.

to ensure the consistency of the estimator in the proposed algorithm. The reason why the sequence $\{\alpha_k\}$ is crucial in ensuring the consistency of the estimator is because the selection of the sampling points and the selection of optimal inputs are bundled together in the form of Eq.(3.5). Therefore, in our discussion concerning the consist property of estimator $\mathbf{m_k}$, we do not see the sequence $\{\alpha_k\}$ crucial in any perspective. However, the selection of $\{\alpha_k\}$ remains certain practical meanings.

In the kth stage, we consider the following optimization problem

$$\begin{aligned} &\text{minimize}_{\mathbf{c}\in\mathcal{X}} \quad -\mathbf{f}(\mathbf{c})^T\mathbf{m}_k \\ &\text{subject to} \quad \frac{1}{2}\Big(\mathbf{f}(\mathbf{c})^T\mathbf{V}_k\mathbf{f}(\mathbf{c})\Big) \leq \nu_k \end{aligned} \tag{3.8}$$

where the exogenous parameter ν_k satisfies the practical condition $0 < \nu_k \leq \sup_{\mathbf{c}\in\mathcal{X}} \mathbf{f}(\mathbf{c})^T\mathbf{V}_k\mathbf{f}(\mathbf{c})$. Then the Lagrangian function for the optimization becomes

$$\mathcal{L}(\nu_k) = -\mathbf{f}(\mathbf{c})^T\mathbf{m}_k + \alpha\left(\frac{1}{2}\mathbf{f}(\mathbf{c})^T\mathbf{V}_k\mathbf{f}(\mathbf{c}) - \nu_k\right)$$

where α denotes the Lagrangian multiplier. Following the Lagrangian method, we need to minimize the Lagrangian function $\mathcal{L}(\nu_k)$, which is equivalent to

$$\max_{\mathbf{c}\in\mathcal{X}} \quad \mathbf{f}(\mathbf{c})^T\mathbf{m}_k - \frac{\alpha}{2}\Big(\mathbf{f}(\mathbf{c})^T\mathbf{V}_k\mathbf{f}(\mathbf{c})\Big)$$

which is exactly the same with Eq.(3.1) except that α_k is replaced by the Lagrangian multiplier α. Hence, we realize that Eq.(3.1) is indeed a mean-variance optimization problem with a trade-off (Lagrangian) parameter α_k.

By the Lagrangian theorem, we know that the Lagrangian multiplier α is a function of the exogenous parameter ν_k. From the above analysis, we ignore the constraint $\mathbf{c} \in \mathcal{X}$ temporarily and know that optimal solution $\mathbf{c}^*$ would satisfy

$$\mathbf{f}(\mathbf{c}^*) = \frac{1}{\alpha}\mathbf{V_k}^{-1}\mathbf{m}_k$$

Therefore, considering the constraint in the optimization problem (3.8), we have

$$\alpha(\nu_k) = \frac{1}{\sqrt{2\nu_k}}\,\mathbf{m}_k^T\mathbf{V}_k^{-1}\mathbf{m}_k$$

Recall that ν_k is the exogenous parameter of the optimization problem (3.8), and that ν_k represents the experimenter's allowance of the estimation variance at the points of interests. Hence, the selection of ν_k would be cope with the practical need. By

setting $\alpha_k = \alpha(\nu_k)$, we know how to choose α_k in Eq.(3.1).

3.3.3 Expected Improvement in Each Iteration

To begin with, we define the following conditional expected value operator

$$\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathbf{c}_k, \mathbf{x}_k, \mathbf{m_k}, \mathbf{V}_k].$$

Then by Eq.(3.1), we redefine the value function

$$J_{k+1}(\mathbf{c}_{k+1}) \triangleq \mathbf{f}(\mathbf{c}_{k+1})^T \mathbf{m}_k - \frac{\alpha_k}{2} \left\{ \mathbf{f}(\mathbf{c}_{k+1})^T \mathbf{V}_k \mathbf{f}(\mathbf{c}_{k+1}) \right\}$$

which indicates the value we will obtain after the kth experiment. The following lemma will help us to evaluate the improvement from one experiment to the next.

Lemma 4 *For all possible $k \geq 1$, we have*

$$J_{k+1}(\mathbf{c}_{k+1}) - J_k(\mathbf{c}_k) \geq \frac{\mathbf{f}(\mathbf{c}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k)}{1 + \mathbf{f}(\mathbf{x}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k)} \Big(y_k - \mathbf{f}(\mathbf{x}_k)^T (\mathbf{m}_k) \Big) + \frac{\alpha_k}{2} \frac{\Big(\mathbf{f}(\mathbf{c}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) \Big)^2}{1 + \mathbf{f}(\mathbf{x}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k)}$$

Proof: Recall that

$$\begin{aligned}
J_{k+1}(\mathbf{c}_{k+1}) &= \mathbf{f}(\mathbf{c}_{k+1})^T \mathbf{m}_k - \frac{\alpha_k}{2} \left\{ \mathbf{f}(\mathbf{c}_{k+1})^T \mathbf{V}_k \mathbf{f}(\mathbf{c}_{k+1}) \right\} \\
&= \mathbf{f}(\mathbf{c}_{k+1})^T \left\{ \mathbf{m}_{k-1} + y_k \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) - \frac{\mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) \mathbf{f}(\mathbf{x}_k)^T}{1 + \mathbf{f}(\mathbf{x}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k)} \Big(\mathbf{m}_k + y_k \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) \Big) \right\} \\
&\quad - \frac{\alpha_k}{2} \left\{ \mathbf{f}(\mathbf{c}_{k+1})^T \left(\mathbf{V}_{k-1} - \frac{\mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) \mathbf{f}(\mathbf{x}_k)^T \mathbf{V}_{k-1}}{1 + \mathbf{f}(\mathbf{x}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k)} \right) \mathbf{f}(\mathbf{c}_{k+1}) \right\} \\
&\geq \mathbf{f}(\mathbf{c}_k)^T \mathbf{m}_{k-1} - \frac{\alpha_{k-1}}{2} \mathbf{f}^T(\mathbf{c}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{c}_k) + y_k \mathbf{f}(\mathbf{c}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) \\
&\quad - \frac{\mathbf{f}(\mathbf{c}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) \mathbf{f}(\mathbf{x}_k)^T}{1 + \mathbf{f}(\mathbf{x}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k)} \Big(\mathbf{m}_k + y_k \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) \Big) + \frac{\alpha_k}{2} \frac{\Big(\mathbf{f}(\mathbf{c}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) \Big)^2}{1 + \mathbf{f}(\mathbf{x}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k)}
\end{aligned}$$

Hence, we obtain

$$J_{k+1}(\mathbf{c}_{k+1}) - J_k(\mathbf{c}_k) \geq \frac{\mathbf{f}(\mathbf{c}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k)}{1 + \mathbf{f}(\mathbf{x}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k)} \Big(y_k - \mathbf{f}(\mathbf{x}_k)^T (\mathbf{m}_k) \Big) + \frac{\alpha_k}{2} \frac{\Big(\mathbf{f}(\mathbf{c}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k) \Big)^2}{1 + \mathbf{f}(\mathbf{x}_k)^T \mathbf{V}_{k-1} \mathbf{f}(\mathbf{x}_k)}$$

□

Theorem 11 (Stepwise Improvement) *If the initial prior of β is unbiased, i.e., $\mathbb{E}_0[\mathbf{m}_0] = \bar{\beta}$, we obtain*

$$\mathbb{E}_k\left[J_{k+1}(\mathbf{c}_{k+1}) - J_k(\mathbf{c}_k)\right] \geq \frac{\alpha_k}{2}\frac{\left(\mathbf{f}(\mathbf{c}_k)^T\mathbf{V}_{k-1}\mathbf{f}(\mathbf{x}_k)\right)^2}{1+\mathbf{f}(\mathbf{x}_k)^T\mathbf{V}_{k-1}\mathbf{f}(\mathbf{x}_k)} > 0$$

Proof: By the property of least square estimator, if $\mathbf{m}_0$ is unbiased, then we have

$$\mathbb{E}_k\left[\mathbf{m}_k\right] = \bar{\beta}$$

Then, by Lemma 4, we have

$$\begin{aligned}
&\mathbb{E}_k\left[J_{k+1}(\mathbf{c}_{k+1}) - J_k(\mathbf{c}_k)\right] \\
&\geq \frac{\mathbf{f}(\mathbf{c}_k)^T\mathbf{V}_{k-1}\mathbf{f}(\mathbf{x}_k)}{1+\mathbf{f}(\mathbf{x}_k)^T\mathbf{V}_{k-1}\mathbf{f}(\mathbf{x}_k)}\left(\mathbf{f}(\mathbf{x}_k)^T\mathbb{E}_k[\bar{\beta} - (\mathbf{m}_k) + \epsilon_k]\right) + \frac{\alpha_k}{2}\frac{\left(\mathbf{f}(\mathbf{c}_k)^T\mathbf{V}_{k-1}\mathbf{f}(\mathbf{x}_k)\right)^2}{1+\mathbf{f}(\mathbf{x}_k)^T\mathbf{V}_{k-1}\mathbf{f}(\mathbf{x}_k)} \\
&= \frac{\alpha_k}{2}\frac{\left(\mathbf{f}(\mathbf{c}_k)^T\mathbf{V}_{k-1}\mathbf{f}(\mathbf{x}_k)\right)^2}{1+\mathbf{f}(\mathbf{x}_k)^T\mathbf{V}_{k-1}\mathbf{f}(\mathbf{x}_k)}
\end{aligned}$$

which is guaranteed to be positive because $\mathbf{f}(\mathbf{x}_k)^T\mathbf{V}_{k-1}\mathbf{f}(\mathbf{x}_k) \geq 0$. □

3.3.4 An Illustrative Example

Ford and Silvey (1980) considered a design problem for locating the optimal input level of a linear model $y_k = \theta_1 x_1^{(k)} + \theta_2 x_2^{(k)} + \epsilon_k$ with design space $\mathcal{X} = [-1, +1]^2$. We first implement our algorithm in Eq.(3.1)-(3.4). We assume the prior distribution of θ is

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0.5 \\ 0.2 \end{pmatrix}, \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}\right)$$

and the real $\overline{\theta} = (0.21, -0.3)^T$. For simplicity, we also define the estimator error as

$$e_k = \sqrt{(\overline{\theta} - \hat{\theta}_k)^T(\overline{\theta} - \hat{\theta}_k)}.$$

We first run our algorithm for the case $\sigma_\epsilon^2 = 1$. Notice that volatility of the noises is relatively large. We present the simulation results in Fig. 3-1 and Fig. 3-2. In Fig. 3-1, one could observe that the estimator error approaches to zero fast after 200 iterations and the selected input level also approaches to the optimal level. Figure 3-2 shows the output value of the selected input level. One could observe that the output value, i.e., $\mathbf{f}(\mathbf{c}_k)^T\overline{\theta}$, closely attain the optimal performance fast within 100 iterations. Next consider the case when $\sigma_\epsilon^2 = 0.5$. Notice that the volatility here is more moderate. One could observe the associated result presented in Figure 3-3 and find that the estimation error diminishes to zero very fast and also the input level hits the optimal level very quickly.

One would also notice that the the sampling sequence $\{\mathbf{x}_k\}$ is either at $+1$ or -1. This is similar with the optimal experimental design constructed in (Ford and Silvey, 1980).

We next consider the algorithm (3.5) proposed by Pronzato (2000). The results are shown in Fig. 3-4 to Fig. 3-6, in which we assume that $\sigma_\epsilon^2 = 1$. The simulation shows that the performance of the algorithm is not stable. Around half of the repeated trials have a bad performance, as the one shown in Figure 3-4. It seems that the estimator would loose strong consistency property. On the other hand, sometimes the algorithm performs very well. As indicated in Figure 3-5 and 3-6, the estimation error converges to zero very fast and the output performance would also converge to the optimal level fast, while the final result is not as good as our proposed algorithm.

3.4 Bayesian Analysis with Unknown σ_ϵ^2

In this section, we assume that σ_ϵ^2 is uncertain and has prior

$$f(\sigma_\epsilon^2) = \frac{\left(\frac{a}{2}\right)^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \left(\sigma_\epsilon^2\right)^{-\frac{d+2}{2}} \exp\left(-\frac{a}{2\sigma_\epsilon^2}\right), \qquad \text{where } a,\, d > 0.$$

and all other conditions are identical to the case when σ_ϵ^2 is known. Then, from Appendix A, we have

$$f(\beta,\, \sigma_\epsilon^2 |\, y_1, \ldots, y_k) = \frac{\left(\frac{a_k}{2}\right)^{\frac{d+1}{2}}}{(2\pi)^{\frac{p}{2}} |\mathbf{V}_k|^{\frac{1}{2}} \Gamma(\frac{d+1}{2})} \left(\sigma_\epsilon^2\right)^{-\frac{d+p+3}{2}} \exp\left(-\frac{(\beta - \mathbf{m}_k)^T (\mathbf{V}_k)^{-1} (\beta - \mathbf{m}_k) + a_k}{2\sigma_\epsilon^2} \right)$$

where the posterior analysis results are

$$\begin{aligned}
\mathbf{m}_{k+1} &= \left(\mathbf{V}_k^{-1} + \mathbf{f}(\mathbf{x}_k)\mathbf{f}(\mathbf{x}_k)^T\right)^{-1} \left(\mathbf{V}_k^{-1}\mathbf{m}_k + \mathbf{f}(\mathbf{x}_k)^T y_k\right) \\
\mathbf{V}_{k+1} &= \left(\mathbf{V}_k^{-1} + \mathbf{f}(\mathbf{x}_k)\mathbf{f}(\mathbf{x}_k)^T\right)^{-1} \\
a_{k+1} &= a_k + y_k^2 + \mathbf{m}_k^T \mathbf{V}_k^{-1} \mathbf{m}_k - (\mathbf{m}_{k+1})^T (\mathbf{V}_{k+1})^{-1} (\mathbf{m}_{k+1})
\end{aligned}$$

After the kth experiment , we want to find the optimal input level $\mathbf{c}_{k+1}$ by solving

$$\begin{aligned}
\mathbf{c}_{k+1} &= \arg\max_{\mathbf{c}\in\mathcal{X}} \mathbf{f}(\mathbf{c})^T \mathbf{m}_k - \frac{\alpha_k}{2} \left(\frac{a_k \Big(1 + \mathbf{f}(\mathbf{c})^T \mathbf{V}_k \mathbf{f}(\mathbf{c})\Big)}{d-1} \right) \\
&= \arg\max_{\mathbf{c}\in\mathcal{X}} \mathbf{f}(\mathbf{c})^T \mathbf{m}_k - \frac{\alpha_k a_k}{2(d-1)} \Big(\mathbf{f}(\mathbf{c})^T \mathbf{V}_k \mathbf{f}(\mathbf{c})\Big) \qquad (3.9)
\end{aligned}$$

Comparing Eq.(3.1) and Eq.(3.9), one could easily find that they are in the same structure but the tradeoff coefficient α_k is replaced by $\frac{\alpha_k a_k}{d-1}$. Then, the series of $\{\alpha_k\}$ in the case when α_ϵ^2 is unknown must decrease faster than the case in which α_ϵ^2 is known. However, it is convenient to maintain the assumption that the sequence $\{\frac{\alpha_k a_k}{d-1}\}$ is nonincreasing, for example, by setting $d_k = \frac{1}{\log(k) a_k}$.

Next, we determine the next sampling input $\mathbf{x}_{k+1}$ by considering

$$\begin{aligned}
\mathbf{x}_{k+1} &= \arg\min_{\mathbf{x}\in\mathcal{X}} \frac{a_{k+1}}{d-1}\left(1+\mathbf{f}(\mathbf{c}_{k+1})^T\left(\mathbf{V}_k^{-1}+\mathbf{f}(\mathbf{x}_k)\mathbf{f}(\mathbf{x}_k)^T\right)^{-1}\mathbf{f}(\mathbf{c}_{k+1})\right) \\
&= \arg\min_{\mathbf{x}\in\mathcal{X}} \left(a_k+\mathbf{m}_k\mathbf{V}_k^{-1}\mathbf{m}_k+y_k^2-\mathbf{m}_k^T(\mathbf{V}_k^{-1}+\mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x})^T)\mathbf{m}_k\right) \\
&\qquad \times\left(1+\mathbf{f}(\mathbf{c}_{k+1})^T\left(\mathbf{V}_k^{-1}+\mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x})^T\right)^{-1}\mathbf{f}(\mathbf{c}_{k+1})\right) \\
&= \arg\min_{\mathbf{x}\in\mathcal{X}} \left(a_k+y_k^2-(\mathbf{m}_k^T\mathbf{f}(\mathbf{x}))^2\right)\left(1+\mathbf{f}(\mathbf{c}_k)^T\mathbf{V}_k\mathbf{f}(\mathbf{c}_{k+1})-\frac{\left(\mathbf{f}(\mathbf{c}_{k+1})^T\mathbf{V}_k\mathbf{f}(\mathbf{x})\right)^2}{1+\mathbf{f}(\mathbf{x})^T\mathbf{V}_k\mathbf{f}(\mathbf{x})}\right)
\end{aligned} \tag{3.10}$$

Observing Eq.(3.10), we know that the sequence of $\mathbf{x}_k$ will not incur the event $\{\mathbf{f}(\mathbf{x}_k)=\mathbf{0}\}$.

Now, we apply the value function in the previous section and discuss the consistency issue and expected improvement for each iteration. Observing the value function in this unknown σ_ϵ^2 case

$$J_{k+1}(\mathbf{c}_{k+1}) = \mathbf{f}(\mathbf{c}_{k+1})^T\mathbf{m}_k - \frac{\alpha_k\, a_k}{2(d-1)}\left\{\mathbf{f}(\mathbf{c}_{k+1})^T\mathbf{V}_k\mathbf{f}(\mathbf{c}_{k+1})\right\}$$

we know that the rest of the derivation are identical to that in the previous section, except that $\alpha_k := \frac{\alpha_k a_k}{2(d-1)}$ in this section. Hence, we could attain the same conclusion as we made in the case when σ_ϵ^2 is known.

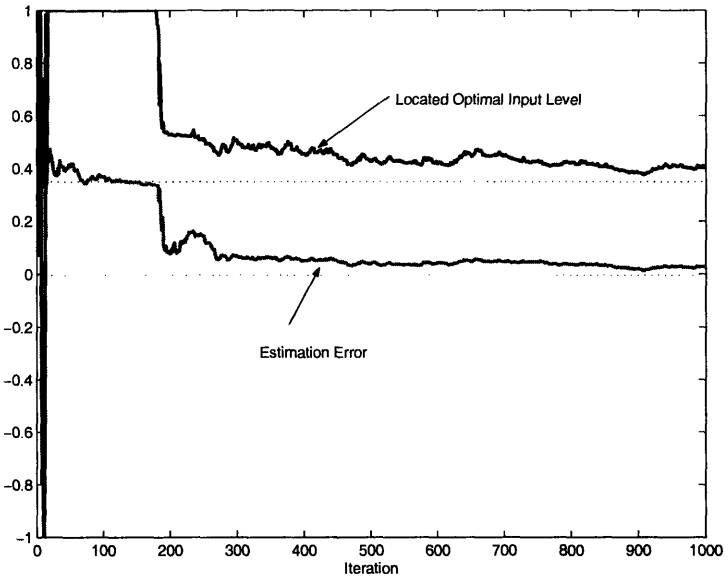

Figure 3-1: The Performance of Proposed Algorithm with $\sigma_\epsilon^2 = 1$

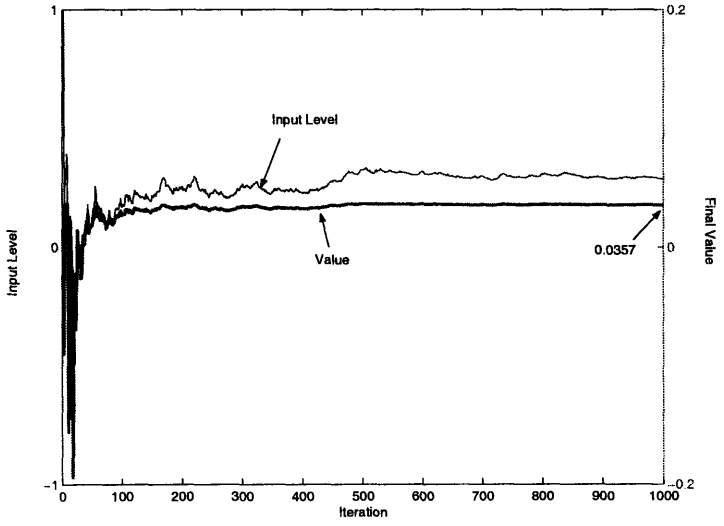

Figure 3-2: The Performance of Proposed Algorithm with $\sigma_\epsilon^2 = 1$

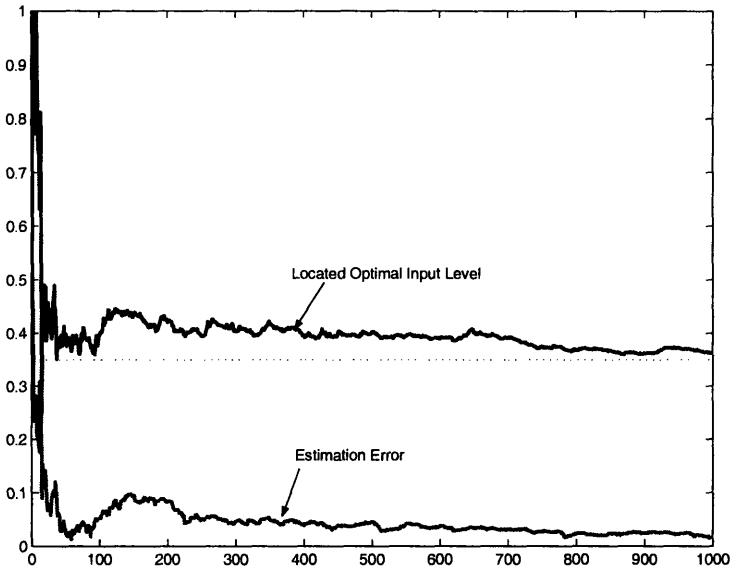

Figure 3-3: The Performance of Proposed Algorithm with $\sigma_\epsilon^2 = 0.5$

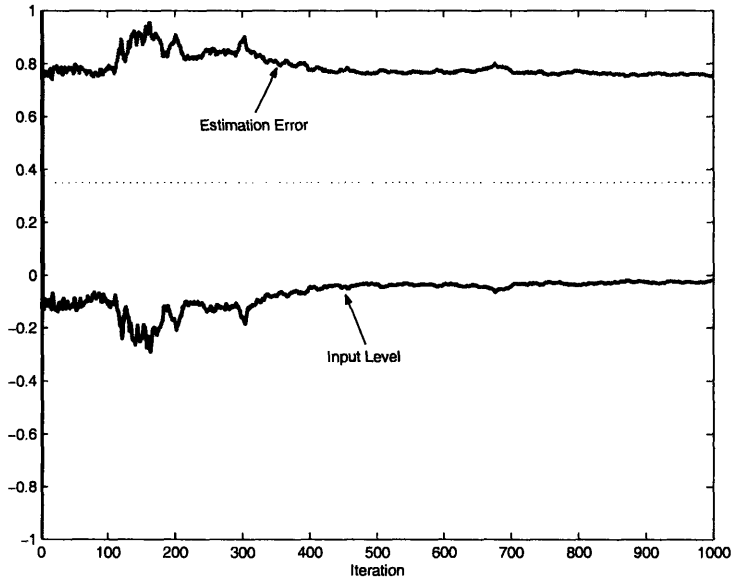

Figure 3-4: The Performance of Algorithm Proposed byPronzato (2000) with $\sigma_\epsilon^2 = 1$: Bad performance

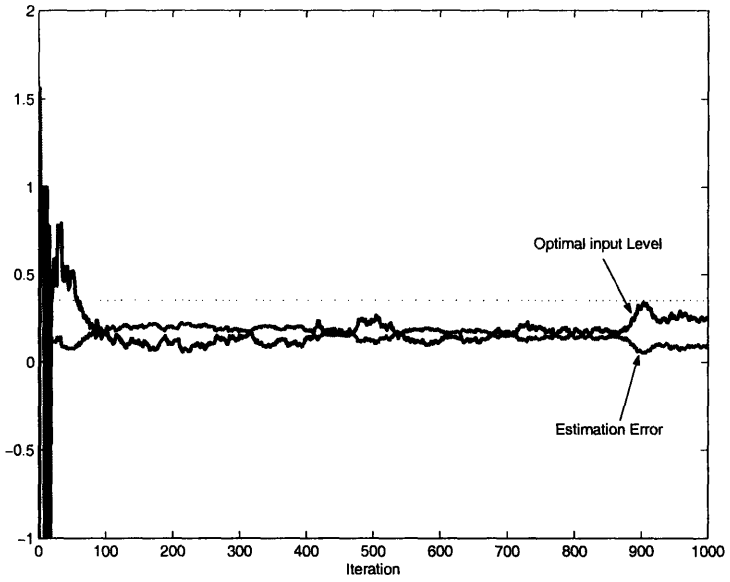

Figure 3-5: The Performance of Algorithm Proposed byPronzato (2000) with $\sigma_\epsilon^2 = 1$: Good performance

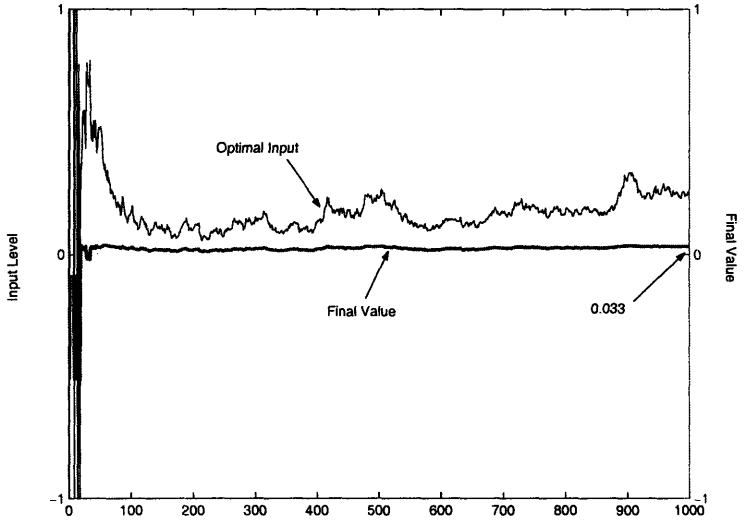

Figure 3-6: The Performance of Algorithm Proposed byPronzato (2000) with $\sigma_\epsilon^2 = 1$

Chapter 4

A Case Study for Adaptive OFAT Scheme and Sequential Response Optimization Scheme

The purpose of this chapter is to convey two case studies for the following two approaches: the adaptive OFAT scheme described by Eqs.(2.7) in Chapter 2, and the Sequential Response Optimization scheme described by Eqs.(3.1) to (3.4) in Chapter 3. I will begin with introducing the background of each case studied in this chapter, and then present and discuss the performances of the two proposed schemes.

4.1 Electric Powered Aircraft Flight Duration

4.1.1 Background for the case study

This case study concerns improvement in the flight duration of electric powered aircrafts. To build an electric powered aircraft, one needs to determine the seven factors listed in Table 4.1. One could read from the table that each of the seven factors has two possible selections, which amounts to a total of 128 possible combinations. We encode each factor as letters and the corresponding selections as -1 and +1. For example, the selection of 450 in^2 for wing area is coded as $D = -1$, whereas the

selection of 600 in^2 is encoded as $D = +1$.

		Level	
Coded Factor	Description	-	+
A	Propeller Diameter	7 in	8 in
B	Propeller Pitch	4 in	5 in
C	Gear Ratio	1:1	1:1.85
D	Wing Area	450 in^2	600 in^2
E	Size of Battery	7	8
F	Motor Type	SP400	SP480
G	Number of Motor	1	2

Table 4.1: The factors and levels in the electric powered aircraft experiment

In this case study, the performance measure of interest is the maximum flight duration defined here as the battery life at the lowest throttle setting capable of level flight or at slight a positive rate of climb. A computer aided analysis package, Electricalc Version 1.0E, has been developed for simulating the performance of the aircrafts and the prediction is quite accurate when used properly. A full factorial 2^7 experiment was conducted using Electricalc based on all the combinations of the factors and levels presented in Table 4.1 by manually feeding the software with associated aircraft weights and wing loading with the following equations

$$\begin{aligned}\text{Aircraft Weight} &= 5\,\text{oz} + 0.01\,\text{oz} \times \text{Wing Area} + 0.5\,\text{oz} \times \text{Size of Battery} \\ &\quad + \text{Number of Motors} \times \left(\begin{pmatrix} 2.3\text{oz if SP400} \\ 3.5\text{oz if SP480} \end{pmatrix} + \begin{pmatrix} 0.0\text{oz if } 1:1.00 \\ 0.5\text{oz if } 1:1.85 \end{pmatrix} \right) \\ \text{Wing Loading} &= \frac{\text{Aircraft Weight}}{\text{Wing Area}}\end{aligned}$$

The complete results of the full factorial experiments are shown in Table 4.3, which is attached to the end of this chapter for completeness. We regard the full factorial responses as the true data for this case study.

Based on the data from the full factorial 2^7 experiment, the main effects and interactions were computed. The twelve largest effects are listed in Table 4.2. Table 4.2 suggests that this system is dominated by main effects (especially factor C, the gear

ratio), but also has a number of interactions of substantial practical significance. For example, the D×G interaction represents the influence of wing area on the benefit of adding an additional motor to the aircraft. This interaction accounts for about two minutes of flight time, which is practically significant. It appears that this is a design problem in which interactions influence the outcomes in an important way. This observation also implies that the flight duration is a complicated combination of the seven factors and will be difficult to derive an analytical expression for the relationship between the flight duration and the seven factors. Therefore, a careful design of experiment is needed to select the combination of the factors to achieve larger flight duration.

Term	C	G	E	F	DG	AC	CFG	EG	BC	DEG	CDEF	B	BG	AF
Coefficient	9.71	5.10	3.58	-3.24	1.91	1.43	-1.13	0.9	0.83	0.83	0.79	-0.79	0.38	0.35

Table 4.2: The largest 12 effects in magnitude

4.1.2 Adaptive OFAT Scheme

Imagine a scenario in which each design evaluation involves building and evaluating an electric airplane. Moreover, since the experiment duration is very limited, the experimenter could only select the two choices for each factor, as shown in Table 4.1. Under these conditions, a design method that requires only eight design variants seems feasible and any method that requires much more prototyping might be ruled out depending on the budget and schedule. Under such circumstance, we consider to conduct the proposed adaptive OFAT scheme, which satisfies the resource and schedule constraints imposed by the environment.

The data from the full factorial experiment were used to simulate adaptive OFAT. For each trial, a starting point design and an order in which to toggle the factors were selected at random. Then the adaptive OFAT process was simulated by looking up the response in the tabulated data. Experimental error was simulated by adding a normally distributed pseudo-random number to the tabulated value to create a simulated observation. After the adaptive OFAT process selected the seven factor

levels, the response at that set of levels without simulated error was stored as the outcome of the trial. This was repeated 10,000 times for each of 8 different amounts of simulated experimental error.

The results of the simulations are presented in Figure 4-1. The maximum flight time within this design space of 128 possible discrete factor settings is 50.5 minutes and the mean is 36.3 minutes, which are both indicated by heavy lines. The circles represent the average flight time achieved by adaptive OFAT over different starting point designs and orderings of the factors. The bars indicate the range exhibited from one empirical standard deviation. When experimental error was low, adaptive OFAT provided an average flight time of 48.02 minutes. This observation represents 83% of the potential improvement (from starting design to final design) within this discrete space of factor settings. This is consistent with the prediction made in Chapter 2 for low experimental error and for systems with $\frac{\sigma_{\text{INT}}}{\sigma_{\text{ME}}} \approx \frac{1}{3.5}$.

One way to understand why the improvement made by adaptive OFAT is high is to record the probability that effects are exploited when the strength of experimental error σ_ϵ^2 is low. The main effects were exploited with probability 82%. The two-factor interactions were exploited with probability 62%. The largest two-factor interaction $D \times G$ was exploited with probability 72%. These values are all consistent with the prediction we derived in Chapter 2 with low experimental error and $\frac{\sigma_{\text{INT}}}{\sigma_{\text{ME}}} \approx \frac{1}{3.5}$.

When experimental error becomes high, the performance of adaptive OFAT declined. With experimental error having a standard deviation of 3.0 minutes, which is a very large error for such an engineering experiment, adaptive OFAT provided an average flight time of 45.4 minutes or 64.5% of the potential improvement. This is consistent with our prediction for high experimental error and moderate interactions.

Comparison with Saturated Fractional Factorial Experimental Design

Under the same strict resource and schedule limitations, another popular strategy is the saturated fractional factorial experiment design strategy. To provide a basis of comparison, the data from the full factorial experiment were used to simulate factorial experimentation. Since adaptive OFAT required 8 experiments, a 2_{III}^{7-4} experimental

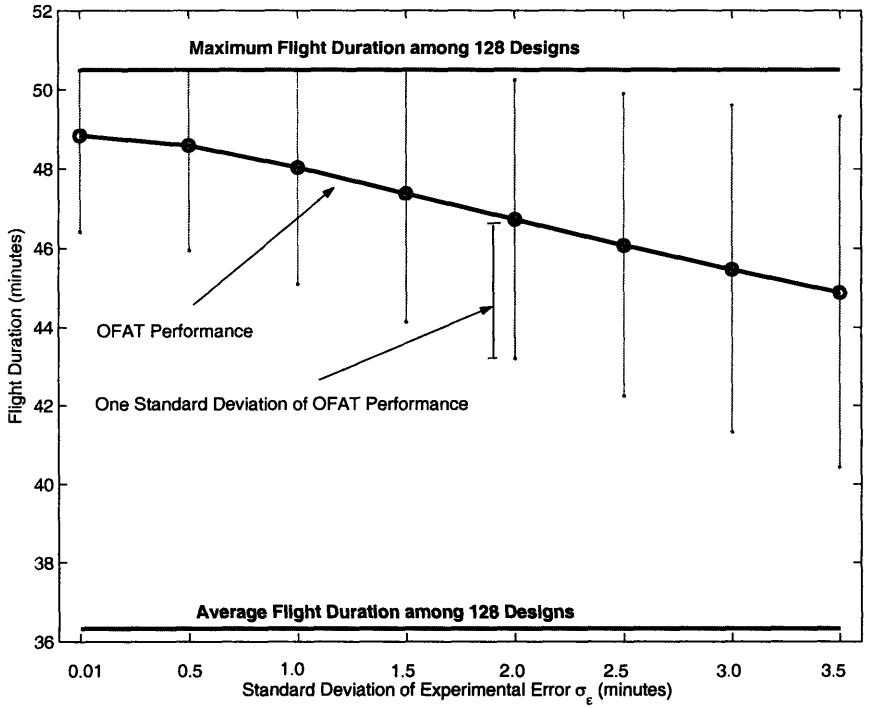

Figure 4-1: Flight Duration from OFAT under Various Experimental Error

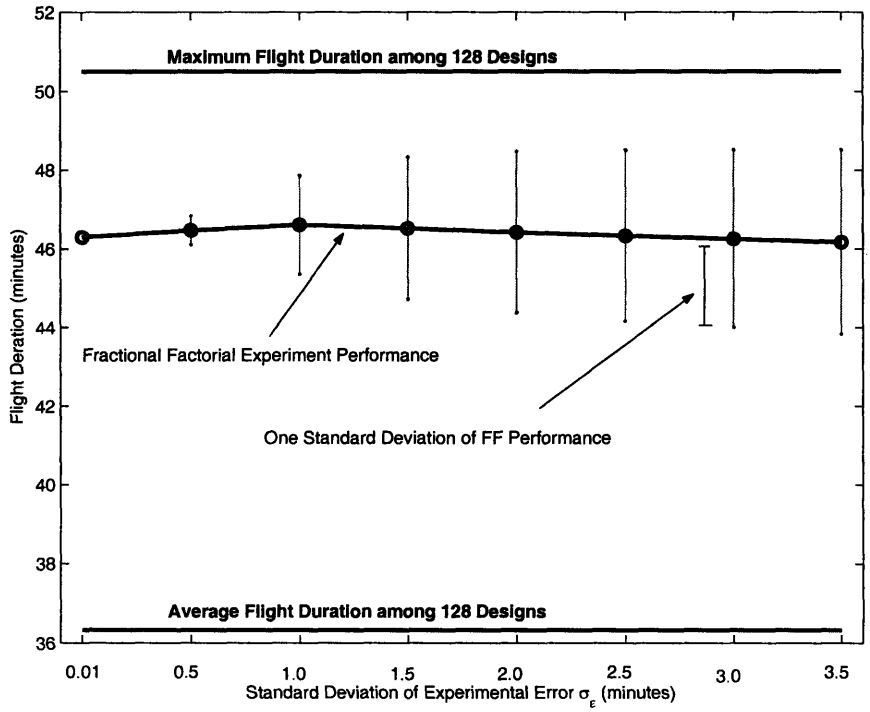

Figure 4-2: Flight Duration from Fractional Factorial Experiments under Various Experimental Error

design was used to maintain an equivalence of resource requirements. For each trial, one of 35 possible fractions was selected at random. Then the experiment was simulated by looking up the responses in the tabulated data and adding random variables to simulate experimental error as before. After the data collection was complete, the main effect estimates were used to select factor levels. The response at that set of levels without simulated error was stored as the outcome of the trial. This was repeated 10,000 times for each of 8 different amounts of simulated experimental error.

The results of the simulations are presented in Figure 4-2. The circles connected by the solid lines represent the average flight time achieved by the fractional factorial experimentation. The bars indicate the range exhibited from one standard deviation. When experimental error was low, the process provided an average flight time of 46.3 minutes. That represents 74% of the potential improvement. This is substantially less than the 83% improvement provided by adaptive OFAT under the same conditions. This is consistent with our prediction in Chapter 2 for low experimental error and for systems with $\frac{\sigma_{\text{INT}}}{\sigma_{\text{ME}}} \approx \frac{1}{3.5}$. It is expected that the volatility of outcomes for the factorial design would be much smaller than the adaptive OFAT; however, the range of the performance includes mostly the worse outcomes on the low end than that of adaptive OFAT and the high end is still much smaller than the counterpart of adaptive OFAT.

When experimental error was introduced, the performance of fractional factorial experimentation was relatively consistent. With error having a standard deviation of 3 minutes, the process still provided an average flight time of 46.2 minutes or 69% of the potential improvement. This was superior to the performance of adaptive OFAT under very noisy experimental environments.

It is worth considering how much experimental error is present when factorial design provides more improvement than adaptive OFAT. In this case study the curves for average flight time in Figure 4-1 and Figure 4-2 cross at an experimental error of around 2.5 minutes. The theorems derived in Chapter 2 suggest that the break even point of the degree of error is about the same size as that of a typical main effect coefficient. In this case study, the median main effect coefficient is about 3.5 minutes and the crossing point is about 2.5 minutes. The crossing points vary widely from

case to case, and so this seems close enough to be consistent with our result.

4.1.3 Sequential Response Optimization Scheme

Now imagine a scenario in which the experimental resources are abundant and more selections are available for the factors in Table 4.1. One key feature of this scenario is that the experiments will be conducted sequentially, instead of in a parallel manner. This is very common when an experimenter investigates a large scale engineering systems, e.g. satellites, or when he considers the ethical issues, e.g. clinical experiments.

Since the experiments are conducted sequentially, at the end of each experiment, the experimenter will make Bayesian estimations, locate the optimal levels of the factors, and determine whether to run next experiment. I will simulate the sequential response optimization scheme described by Eq.(3.1) to Eq.(3.4) in this section. More specifically, at the end of each runs, I will estimate the main effect and the interaction effects by the least square estimator. Given the updated estimates, I determine the optimum levels of the factors from the design space to achieve a longer flight duration. The design space, for simplicity, is assumed to be the product of the closed interval between the two levels for each factors.

To be consistent with the case study investigated in the adaptive OFAT scheme, I select the regressor as

$$\mathbf{f}(x_1, \ldots, x_7) = (1,\, x_1, \ldots,\, x_7,\, x_1x_2,\, x_1x_3, \ldots,\, x_6x_7)$$

where $\mathbf{f}(\mathbf{x})$ above includes the main factors and interaction factors identical to those in the OFAT scheme. Note that the choice of regressors is free in the sequential response optimization scheme. The priors for the coefficients are *i.i.d.* normal random variables with mean zero and standard deviation for the constant effect is 30, the main effects are 15 and the interaction coefficients are 5, which is close to our prediction $\frac{\sigma_{\text{INT}}}{\sigma_{\text{ME}}} \approx \frac{1}{3.5}$ in the adaptive OFAT case. We simulate this case with eight different experiment error standard deviations. For the purpose of presentation, we present the following three cases: $\sigma_\epsilon = 0.01$, $\sigma_\epsilon = 1.5$, and $\sigma_\epsilon = 3.50$.

Consistency of the Bayesian Estimator

We first consider the consistency property of this sequential experimentation scheme. As we discussed in Chapter 3, if the event that the covariance matrix $\mathbf{V}_k$ of a sequential estimator approaches to zero occurs, i.e. $\{\mathbf{V}_k \rightarrow \mathbf{0}\}$, then the estimator is consistent. Therefore, we track the maximum eigenvalue of the covariance matrix at each stage.

The maximum eigenvalue of matrix $\mathbf{V}_k$ at each stage k for all $k = 1, 2, \ldots, 500$, is presented in Fig 4-3. It is clear that the maximum eigenvalues of the three different experimentation error cases, i.e., $\sigma_\epsilon = 0.01$, $\sigma_\epsilon = 1.5$, and $\sigma_\epsilon = 3.50$, all converge to zero with almost the identical speed, which implying that the estimators are consistent. That the maximum eigenvalues converge in the same pattern for three different cases meets our expectation because the generation of the covariance matrix $\mathbf{V}$ is influenced directly by the selection of estimation points chosen from the design space, but much less influenced by the observation errors. More importantly, the maximum eigenvalues begin to decay around the 25th iteration, and it decays dramatically fast between the 25th and 100th iterations, implying that the estimator $\mathbf{m}$ of the unknown coefficient β for two consecutive iterations changes dramatically during these iterations. Fig. 4-4 presents the norm of the difference between two consecutive estimators, and the decreases of the norms for the three cases presented are consistent with our observation in the maximum eigenvalues. That is, the difference would be dramatic before the 100th iteration, and the difference would be relatively mild afterwards. Hence, the maximum eigenvalue of the covariance matrix $\mathbf{V}$ could be an indicator to determine whether to terminate the sequential response optimization scheme.

Performance of the Sequentially Response Optimization Scheme

The goal of this scheme is to select the optimal factor levels to achieve longer flight duration. Hence, we present the flight duration determined at each iteration in this subsection. Given the Bayesian estimation after each iteration, the experimenter determines optimal levels of the seven factors. The associated *observed* optimum

response is presented in Fig. 4-5. For the purpose of illustration, the case of $\sigma_\epsilon = 1.5$ is taken off from this figure because it overlaps with the case of $\sigma_\epsilon = 3.50$. The result shown in Fig. 4-5 meets our expectation: the larger the experimentation error is, the more chaos the observed optimal response would be. However, when the case of $\sigma_\epsilon = 0.01$ is regarded as the mean of the other larger experimentation error case, we could see that the observed response follows the trend which approaches to the optimum response. Hence, we know that, although the sampling points and the estimated optimal factors levels are different in each case, the estimation of the optimal response behaves almost the same in all cases and the experimental error is the only source of perturbation in observation. Moreover, after the 100th iteration, the volatility of the observed responses becomes obviously smaller than those in previous iterations, which is consistent with our observation in the maximum eigenvalue of the covariance matrix $\mathbf{V}$.

The true response, the observed response minus the experimentation error, of each iteration is presented in Figure 4-6 and 4-7.

Figure 4-6 presents the first 150 iteration and Fig 4-7 is for the first 500 iterations. In Fig. 4-6, we observe that, the smaller the experimentation error is, the faster the process of response achieves to the optimal level and less volatile the process is. One spike is shown in the case of $\sigma_\epsilon = 1.5$, we suspect it is from the computation error because the optimization tool box terminated improperly at that iteration.

Figure 4-7 presents longer term behavior. We again observe that the true response becomes stable after the 100th iteration, when the maximum eigenvalue of the covariance matrix becomes small. Moreover, the less volatile the experimentation error is, the faster the true response would become stable. It is clear that the final optimal response is around 50.5 the maximum flight duration in this case study. Finally, from Figure 4-7, we suspect that the response process behaves as a submartingle process.

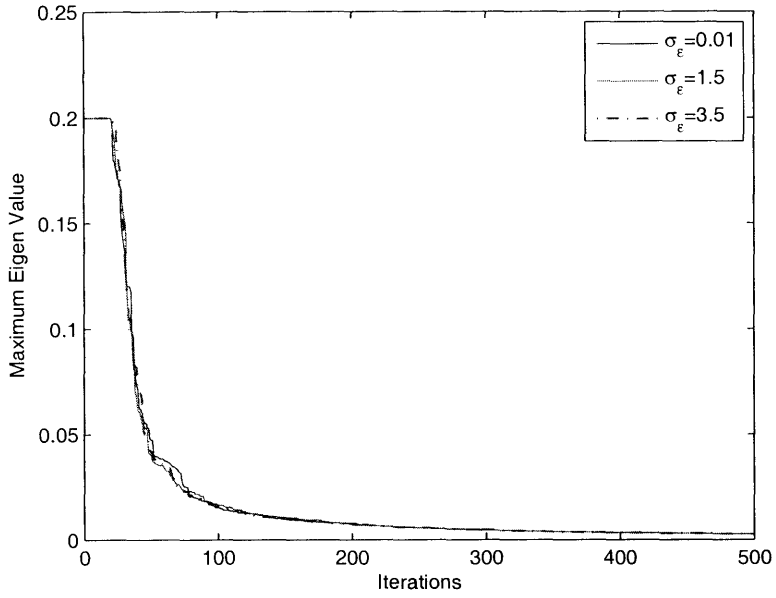

Figure 4-3: The Maximum Eigenvalue of Covariance Matrix $\mathbf{V}_k$ at Each Stage

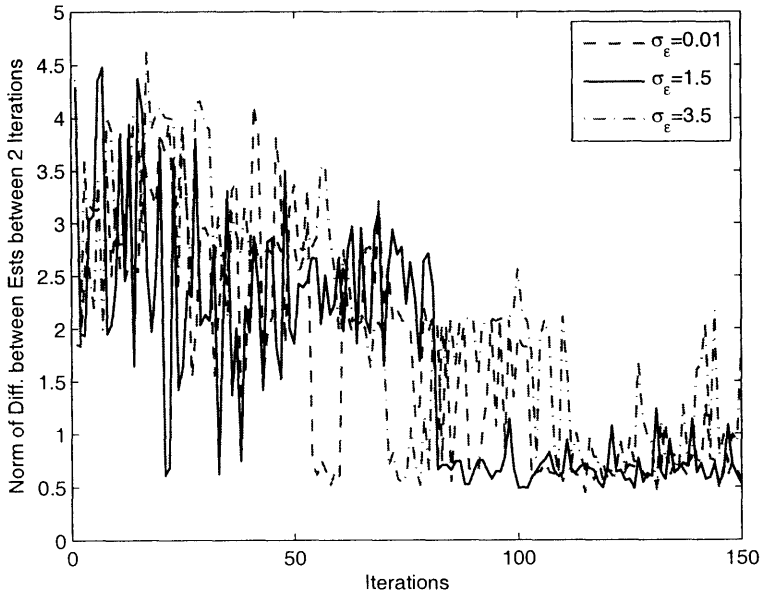

Figure 4-4: Norm of Differences between Estimators of Two Consecutive Iterations

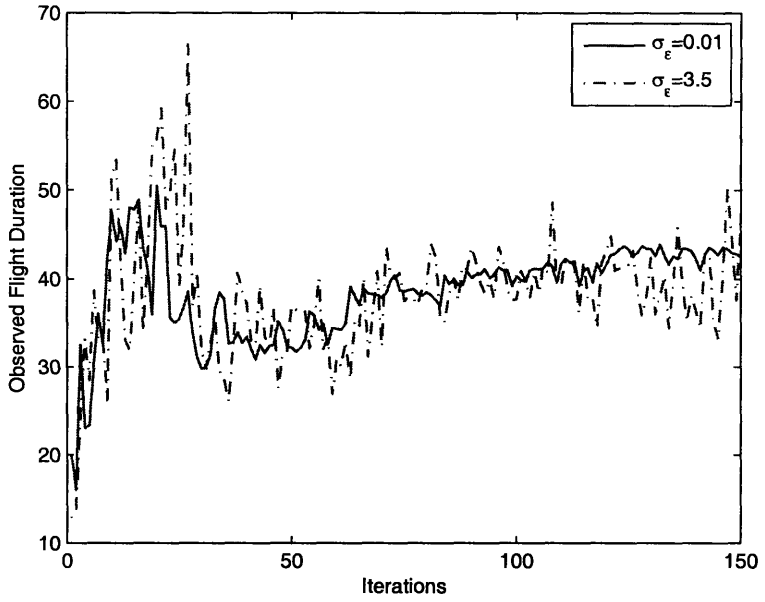

Figure 4-5: Estimated Optimal Response at each Iteration

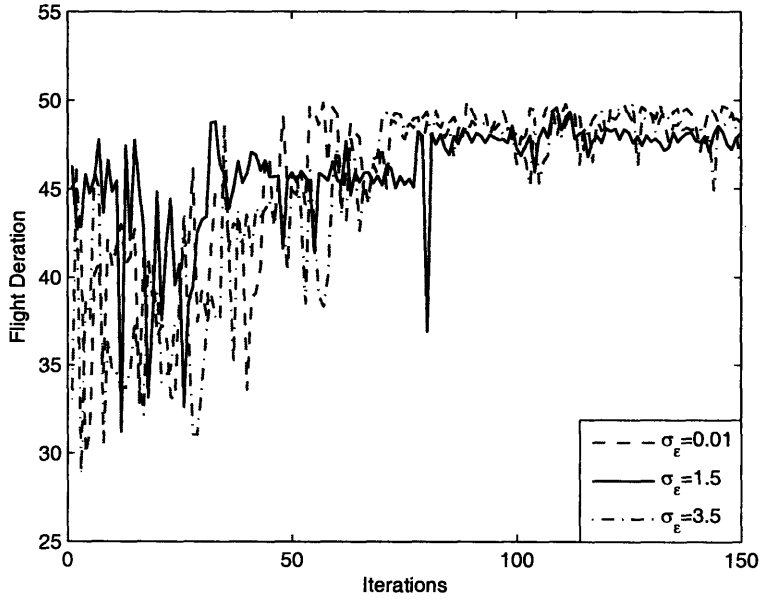

Figure 4-6: True Response of the Estimated Optimum Levels of Factors

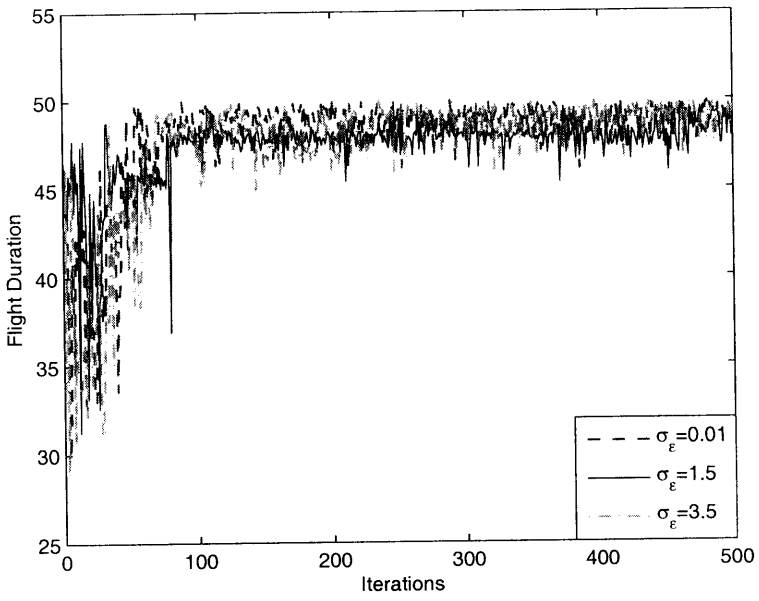

Figure 4-7: True Response of the Estimated Optimum Levels of Factors

A	B	C	D	E	F	G	Response	A	B	C	D	E	F	G	Response
-1	-1	-1	-1	-1	-1	-1	33.7	+1	-1	-1	-1	-1	-1	-1	31.4
-1	+1	-1	-1	-1	-1	-1	30.4	+1	+1	-1	-1	-1	-1	-1	30.0
-1	-1	+1	-1	-1	-1	-1	39.4	+1	-1	+1	-1	-1	-1	-1	40.1
-1	+1	+1	-1	-1	-1	-1	36.6	+1	+1	+1	-1	-1	-1	-1	40.3
-1	-1	-1	+1	-1	-1	-1	32.6	+1	-1	-1	+1	-1	-1	-1	28.9
-1	+1	-1	+1	-1	-1	-1	28.5	+1	+1	-1	+1	-1	-1	-1	29.0
-1	-1	+1	+1	-1	-1	-1	37.4	+1	-1	+1	+1	-1	-1	-1	38.5
-1	+1	+1	+1	-1	-1	-1	34.8	+1	+1	+1	+1	-1	-1	-1	37.3
-1	-1	-1	-1	+1	-1	-1	36.1	+1	-1	-1	-1	+1	-1	-1	32.8
-1	+1	-1	-1	+1	-1	-1	33.1	+1	+1	-1	-1	+1	-1	-1	31.4
-1	-1	+1	-1	+1	-1	-1	42.9	+1	-1	+1	-1	+1	-1	-1	43.8
-1	+1	+1	-1	+1	-1	-1	42.9	+1	+1	+1	-1	+1	-1	-1	42.7
-1	-1	-1	+1	+1	-1	-1	33.1	+1	-1	-1	+1	+1	-1	-1	32.8
-1	+1	-1	+1	+1	-1	-1	30.5	+1	+1	-1	+1	+1	-1	-1	30.4
-1	-1	+1	+1	+1	-1	-1	40.6	+1	-1	+1	+1	+1	-1	-1	42.1
-1	+1	+1	+1	+1	-1	-1	38.1	+1	+1	+1	+1	+1	-1	-1	41.0
-1	-1	-1	-1	-1	+1	-1	27.6	+1	-1	-1	-1	-1	+1	-1	26.4
-1	+1	-1	-1	-1	+1	-1	25.4	+1	+1	-1	-1	-1	+1	-1	23.7
-1	-1	+1	-1	-1	+1	-1	37.1	+1	-1	+1	-1	-1	+1	-1	37.4
-1	+1	+1	-1	-1	+1	-1	35.1	+1	+1	+1	-1	-1	+1	-1	37.5
-1	-1	-1	+1	-1	+1	-1	27.6	+1	-1	-1	+1	-1	+1	-1	25.2
-1	+1	-1	+1	-1	+1	-1	23.6	+1	+1	-1	+1	-1	+1	-1	23.7
-1	-1	+1	+1	-1	+1	-1	34.4	+1	-1	+1	+1	-1	+1	-1	34.8
-1	+1	+1	+1	-1	+1	-1	34.6	+1	+1	+1	+1	-1	+1	-1	34.9
-1	-1	-1	-1	+1	+1	-1	29.8	+1	-1	-1	-1	+1	+1	-1	27.4
-1	+1	-1	-1	+1	+1	-1	28.7	+1	+1	-1	-1	+1	+1	-1	26.9
-1	-1	+1	-1	+1	+1	-1	40.5	+1	-1	+1	-1	+1	+1	-1	39.4
-1	+1	+1	-1	+1	+1	-1	39.5	+1	+1	+1	-1	+1	+1	-1	39.8
-1	-1	-1	+1	+1	+1	-1	29.8	+1	-1	-1	+1	+1	+1	-1	26.5
-1	+1	-1	+1	+1	+1	-1	26.2	+1	+1	-1	+1	+1	+1	-1	24.7
-1	-1	+1	+1	+1	+1	-1	37.5	+1	-1	+1	+1	+1	+1	-1	38.0
-1	+1	+1	+1	+1	+1	-1	36.6	+1	+1	+1	+1	+1	+1	-1	38.3
-1	-1	-1	-1	-1	-1	+1	32.8	+1	-1	-1	-1	-1	-1	+1	31.6
-1	+1	-1	-1	-1	-1	+1	33.4	+1	+1	-1	-1	-1	-1	+1	30.8
-1	-1	+1	-1	-1	-1	+1	41.4	+1	-1	+1	-1	-1	-1	+1	42.9
-1	+1	+1	-1	-1	-1	+1	41.7	+1	+1	+1	-1	-1	-1	+1	44.5
-1	-1	-1	+1	-1	-1	+1	31.3	+1	-1	-1	+1	-1	-1	+1	30.1
-1	+1	-1	+1	-1	-1	+1	31.8	+1	+1	-1	+1	-1	-1	+1	29.4
-1	-1	+1	+1	-1	-1	+1	44.1	+1	-1	+1	+1	-1	-1	+1	43.7
-1	+1	+1	+1	-1	-1	+1	42.4	+1	+1	+1	+1	-1	-1	+1	47.3
-1	-1	-1	-1	+1	-1	+1	34.0	+1	-1	-1	-1	+1	-1	+1	33.9
-1	+1	-1	-1	+1	-1	+1	33.1	+1	+1	-1	-1	+1	-1	+1	33.1

A	B	C	D	E	F	G	Response	A	B	C	D	E	F	G	Response
-1	-1	+1	-1	+1	-1	+1	46.8	+1	-1	+1	-1	+1	-1	+1	46.3
-1	+1	+1	-1	+1	-1	+1	47.4	+1	+1	+1	-1	+1	-1	+1	47.5
-1	-1	-1	+1	+1	-1	+1	45.7	+1	-1	-1	+1	+1	-1	+1	42.9
-1	+1	-1	+1	+1	-1	+1	41.8	+1	+1	-1	+1	+1	-1	+1	39.6
-1	-1	+1	+1	+1	-1	+1	46.8	+1	-1	+1	+1	+1	-1	+1	49.3
-1	+1	+1	+1	+1	-1	+1	47.4	+1	+1	+1	+1	+1	-1	+1	50.5
-1	-1	-1	-1	-1	+1	+1	33.9	+1	-1	-1	-1	-1	+1	+1	30.1
-1	+1	-1	-1	-1	+1	+1	33.1	+1	+1	-1	-1	-1	+1	+1	29.6
-1	-1	+1	-1	-1	+1	+1	39.7	+1	-1	+1	-1	-1	+1	+1	38.3
-1	+1	+1	-1	-1	+1	+1	38.8	+1	+1	+1	-1	-1	+1	+1	39.4
-1	-1	-1	+1	-1	+1	+1	33.9	+1	-1	-1	+1	-1	+1	+1	30.1
-1	+1	-1	+1	-1	+1	+1	33.1	+1	+1	-1	+1	-1	+1	+1	31.7
-1	-1	+1	+1	-1	+1	+1	40.4	+1	-1	+1	+1	-1	+1	+1	40.5
-1	+1	+1	+1	-1	+1	+1	41.1	+1	+1	+1	+1	-1	+1	+1	39.4
-1	-1	-1	-1	+1	+1	+1	36.6	+1	-1	-1	-1	+1	+1	+1	32.7
-1	+1	-1	-1	+1	+1	+1	35.8	+1	+1	-1	-1	+1	+1	+1	30.7
-1	-1	+1	-1	+1	+1	+1	42.7	+1	-1	+1	-1	+1	+1	+1	41.2
-1	+1	+1	-1	+1	+1	+1	43.8	+1	+1	+1	-1	+1	+1	+1	44.1
-1	-1	-1	+1	+1	+1	+1	36.6	+1	-1	-1	+1	+1	+1	+1	35.1
-1	+1	-1	+1	+1	+1	+1	35.8	+1	+1	-1	+1	+1	+1	+1	32.2
-1	-1	+1	+1	+1	+1	+1	45.4	+1	-1	+1	+1	+1	+1	+1	43.6
-1	+1	+1	+1	+1	+1	+1	46.5	+1	+1	+1	+1	+1	+1	+1	46.7

Table 4.3: The Response of the 2^7 Full Factorial Experiments

4.2 Uniformity in the Manufacture of Integrated Circuits

4.2.1 Background for the case study

The second case study considers the experiments which Buckner et al. (1997) reported on characterizing the uniformity of a tungsten deposition tool used in the manufacture of integrated circuits. As indicated by Buckner et al. (1997), poor uniformity in the deposition of tungsten thin films has been the significant factor for the loss of yield in the manufacture of integrated circuits using tungsten Chemical Vapor Deposition (CVD). They carefully selected seven factors presented in Table 4.4. According to (Buckner et al., 1997), the factors A to E are selected by a commonly accepted kinetic rate equation for tungsten deposition. Factor F, the ratio of backslide H_2 and argon flow is unique to the reactor. Last, the factor G, backslide total flow, is used to prevent deposition on the backslide of the wafer.

		Level		
Coded Factor	Description	-	+	Units
A	Temperature	405	425	$°C$
B	Pressure	35	45	torr
C	$(P_{H_2})^{\frac{1}{2}}$	2.61	4.51	$\text{torr}^{\frac{1}{2}}$
D	WF_6 Flow	360	440	sccm
E	Argon Flow	10	14	slm
F	Backslide H_2/Argon Flow	1.5	3.5	(unitless)
G	Backside Total Flow	5	9	slm

Table 4.4: The factors and levels in the uniformity experiment

In this case study, we assume that the measure of interest is the maximum uniformity of the manufacture of integrated circuits. Buckner et al. (1997) reported a 2_{IV}^{7-4} experiment plus three center points spaced at the beginning, middle, and the end of the sequence of 19 runs. Later, Mee and Peralta (2000) conducted a semifolding experiment analysis based on the experimental data provided in (Buckner et al., 1997). For completeness, we present the the analysis results for combined analysis of the initial 2_{IV}^{7-3} experiment plus the semifolding design in Table 4.5. It is obvious that

the interaction coefficients AC, BC, CD, and CF are statistically significant and the factor C plays a key role among these interaction coefficients. For the purpose of our study, we will use the analysis results in Table 4.5 to simulate the adaptive OFAT scheme and optimization algorithm for the compact design space.

Term	Estimated coefficient	Standard error	t	p value
A	.192	.051	3.74	.007
B	.121	.044	2.72	.030
C	.226	.051	4.40	.003
D	.061	.051	1.18	.276
E	-.048	.051	-.94	.379
F	-.078	.051	-1.69	.134
G	.279	.049	5.63	<.001
AC	.406	.051	7.92	<.001
$BE = DG$	.108	.051	2.11	.073
BC	-.548	.044	-12.34	<.001
$AE = DF$	-.017	.044	-.37	.720
CD	-.215	.051	-.419	.004
$BF = AG$	-.019	.051	-.38	.717
CE	.042	.051	.83	.435
$AB = FG$	-.133	.051	-2.60	.036
CF	-.661	.051	-12.90	<.001
$BD = EG$	.009	.051	.18	.860
CG	.099	.051	1.94	.094
$AD = EF$	.098	.051	1.90	.099
$DE(= AF = BG)$	-.066	.044	-1.48	.183

Table 4.5: Analysis of Buckner et al. (1997) 2_{IV}^{7-3} Experiment source: (Mee and Peralta, 2000)

4.2.2 Adaptive OFAT Scheme

Imagine a scenario in which each design evaluation involves building and evaluating a wafer. Moreover, since the experiment duration is very limited, the experimenter could only select the two choices for each factor, as shown in Table 4.4. Under these conditions, a design method that requires only eight design variants seems feasible and any method that requires much more experimental resources might be ruled

out depending on the budget and schedule. Under such circumstance, we consider to conduct the proposed adaptive OFAT scheme, which satisfies the resource and schedule constraints imposed by the environment.

We apply the information presented in Table 4.5 to simulate adaptive OFAT. For each trial, a starting point design and an order in which to toggle the factors were selected at random. Then the adaptive OFAT process was simulated by the value generated from Table 4.5. Experimental error was simulated by adding a normally distributed pseudo-random number to the simulated value to create an experimental observation. After the adaptive OFAT process selected the seven factor levels, the response at that set of levels without simulated error was stored as the outcome of the trial. This was repeated 10,000 times for each of 8 different amounts of simulated experimental error.

The results of the simulations are presented in Figure 4-8. The maximum uniformity capability within this design space of 128 possible discrete factor settings is 2.951(%) with the corresponding input $\mathbf{x} = (+1, -1, +1, +1, -1, -1m + 1)$, and the average uniformity is approximately 0%, which are both indicated by heavy horizontal lines. The circles represent the average uniformity capability achieved by adaptive OFAT over different starting point designs and orderings of the factors. The bars indicate the range exhibited from one empirical standard deviation. When experimental error was low, adaptive OFAT provided an average 1.8% uniformity. This observation represents approximately 148% of the potential improvement (from starting design to final design) within this discrete space of factor settings. This is consistent with the prediction made in Chapter 2 for low experimental error and for systems with $\frac{\sigma_{\text{INT}}}{\sigma_{\text{ME}}} = \frac{1}{10}$.

When experimental error becomes high, the performance of adaptive OFAT declined. With experimental error having a standard deviation of $\sqrt{2}$ (100%), which is a very large error for such an engineering experiment, adaptive OFAT provided an average uniformity capability of 1.02% or 98% of the potential improvement. This is consistent with our prediction for high experimental error and strong interactions.

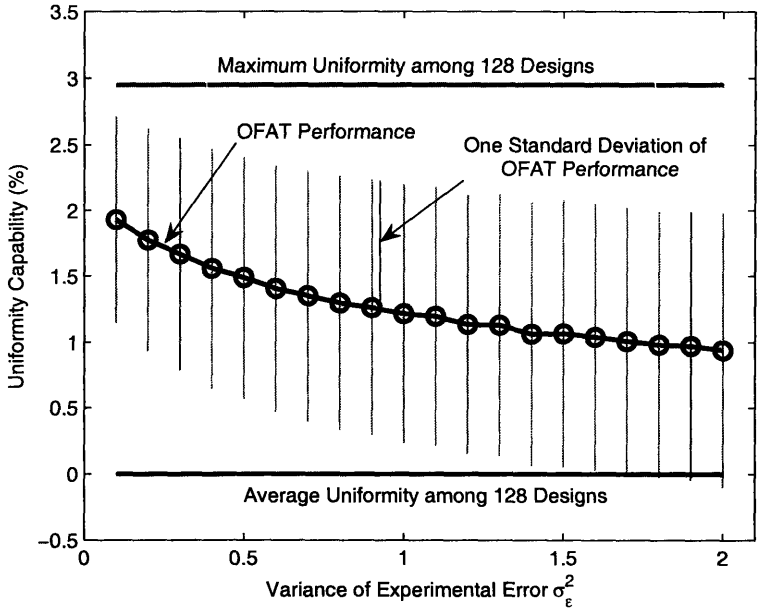

Figure 4-8: Uniformity Capability from OFAT under Various Experimental Error

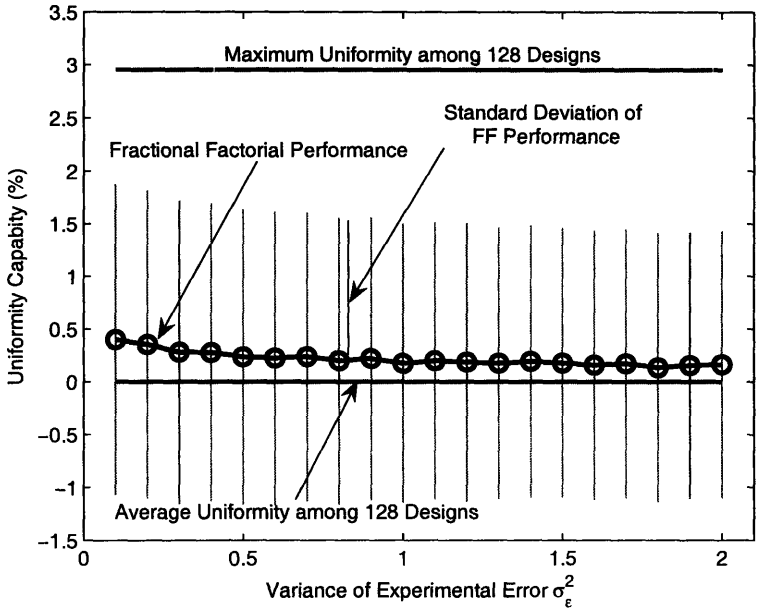

Figure 4-9: Uniformity Capability from Fractional Factorial Experiments under Various Experimental Error

Comparison with Saturated Fractional Factorial Experimental Design

Under the same strict resource and schedule limitations, another popular strategy is the saturated fractional factorial experiment design. To provide a basis for comparison, the data presented in Table 4.5 are used to simulate factorial experimentation. Since adaptive OFAT required 8 experiments, a 2^{7-4}_{III} experimental design was used to maintain an equivalence of resource requirements. For each trial, one of 35 possible fractions was selected at random. Then the experiment was simulated by the approximated function with coefficients in the tabulated data and adding random variables to simulate experimental error as before. After the data collection was complete, the main effect estimates were used to select factor levels. The response at that set of levels without simulated error was stored as the outcome of the trial. This was repeated 10,000 times for each of 8 different amounts of simulated experimental error.

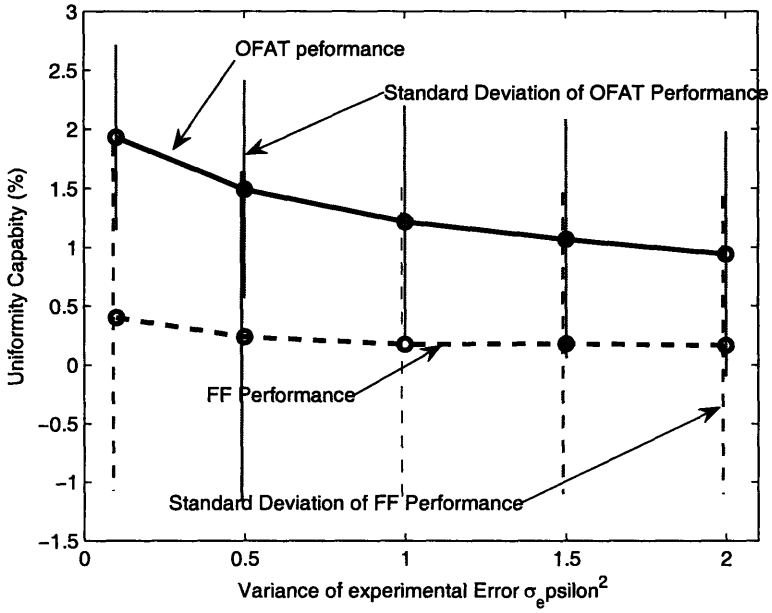

Figure 4-10: Performance Comparison between OFAT and Fractional Factorial Schemes

The results of the simulations are presented in Figure 4-10. The circles connected by the solid lines represent the average uniformity capability achieved by the sat-

urated fractional factorial experimentation. The bars indicate the range exhibited from one standard deviation. When experimental error was low, the process provided an average uniformity of merely 0.3%. That represents only 7.5% of the potential improvement. This is substantially less than the 148% improvement provided by adaptive OFAT under the same conditions. This is consistent with our discussion in Chapter 2, since the interactions are much more significant than the main factor as indicated in Table 4.5.

Figure 4-10 presents the performance comparison of the OFAT and the fractional factorial experiment design scheme. It is evident that the range of the performance includes mostly the worse outcomes on the low end than that of adaptive OFAT and the high end is still way smaller than the lower part of adaptive OFAT. This observation matches the discussion in Chapter 2. The performance of OFAT will seriously outperform when the main effect is weak and the interaction factors are strong, which is the case we study in this subsection.

When experimental error was introduced, the performance of fractional factorial experimentation was relatively consistent. With error having a standard deviation of $\sqrt{2}$, the fractional factorial remains an average 0.3% uniformity or 7.5% of the potential improvement. This is still lower to the performance of adaptive OFAT under very noisy experimental environments, where the performance down half to approximately 102%. Finally, notice that, since the interaction factors, instead of the main factors, dominate the response of the experiments, the expected performance of the OFAT scheme remains much better than that of the saturated fractional factorial scheme. This observation matches our discussion in Chapter 2.

4.2.3 Sequential Response Optimization Scheme

Now imagine a scenario in which the experimental resources are abundant and more selections are available for the factors in Table 4.4. One key feature of this scenario is that the experiments will be conducted sequentially, instead of in a parallel manner. This is very common when an experimenter investigates a large scale engineering systems, e.g. satellites, or when he considers the ethical issues, e.g. clinical experiments.

Since the experiments are conducted sequentially, at the end of each experiment, the experimenter will make Bayesian estimations, locate the optimal levels of the factors, and determine whether to run next experiment. I will simulate the sequential response optimization scheme described by Eq.(3.1) to Eq.(3.4) in this section. More specifically, at the end of each runs, I will estimate the main effect and the interaction effects by the least square estimator. Given the updated estimates, I determine the optimum levels of the factors from the design space to achieve a higher uniformity capability. The design space, for simplicity, is assumed to be the product of the closed interval between the two levels for each factors presented in Table 4.4.

Following the data presented in Table 4.5, I set the regressor as

$$\mathbf{f}(x_1, x_2, \ldots, x_7) = \left(x_1, x_2, \ldots, x_7, x_1x_2, x_1x_3, \ldots, x_6x_7\right)$$

The priors for the coefficients are *i.i.d.* normal random variables with mean zero and standard deviation for the main effects are $\sqrt{5}$ and the interaction coefficients are $\sqrt{10}$ since we know that the interaction would be statistically significant in this uniformity experiment. For the purpose of presentation, we present the following three cases: $\sigma_\epsilon^2 = 0.1$, $\sigma_\epsilon^2 = 0.5$, and $\sigma_\epsilon^2 = 1.0$.

Consistency of the Bayesian Estimator

We first consider the consistency property of this sequential experimentation scheme. As we discussed in Chapter 3, if the event that the covariance matrix $\mathbf{V}_k$ of a sequential estimator approaches to zero occurs, i.e. $\{\mathbf{V}_k \to \mathbf{0}\}$, then the estimator is consistent. Therefore, we track the maximum eigenvalue of the covariance matrix at each stage.

The maximum eigenvalue of matrix $\mathbf{V}_k$ at each stage k for all $k = 1, 2, \ldots, 600$, is presented in Fig 4-11. It is clear that the maximum eigenvalues of the three different experimentation error cases, i.e., $\sigma_\epsilon = 0.1$, $\sigma_\epsilon = 0.5$, and $\sigma_\epsilon = 1.0$, all converge to zero with almost the identical speed, which implying that the convergence process are insensitive to the experimental error strength σ_ϵ. That the maximum eigenvalues

converge in the same pattern for three different cases meets our expectation because the generation of the covariance matrix $\mathbf{V}$ is influenced directly by the selection of estimation points chosen from the design space, but much less influenced by the experimental errors. More importantly, the maximum eigenvalues begin to decay around the $50th$ iteration, and it decays dramatically fast between the $50th$ and $150th$ iterations, implying that the estimator $\mathbf{m}$ of the unknown coefficient β for two consecutive iterations changes dramatically during these iterations. Fig. 4-12 presents the norm of the difference between two consecutive estimators, and the decreases of the norms for the three cases presented are consistent with our observation in the maximum eigenvalues. That is, the difference would be dramatic before the $50th$ iteration, and the difference would be relatively mild afterwards. Hence, the maximum eigenvalue of the covariance matrix $\mathbf{V}$ could be an indicator to determine whether to terminate the sequential response optimization scheme.

Figure 4-13 presents the norm of the difference between parameter estimator and the true parameters for each iteration, i.e., $||\mathbf{m}_k - \bar{\beta}||_2, \forall k = 1, 2, \ldots, 600$. The figure shows that the deviation of the estimators and the true parameters decreases dramatically after the $50th$ iteration, which is consistent with our previous observation. However, as indicated in Figure 4-13, the larger the experimental errors are the smaller the convergence rate of the estimators would be. Table 4.6 presents all the estimators for the coefficients after the 600 iteration for the three different experimental error variances. It is clear that all the statistically significant terms, such as AC, BC and CF, are captured by the algorithm; that is, the estimated parameters are also significant as well. However, the accuracy would be deteriorated as the experimental error increases, an observation which meets our intuition.

Performance of the Sequentially Response Optimization Scheme

Figure 4-14 presents the actual uniformity capability of the estimated optimal input for each iteration. As we expected, the response would become more stable after the $50th$ iteration since the change in estimated parameters would become smaller as we discussed earlier. It is obvious that the response would hit the optimality frequently,

while a noticeable portion of outcome stays far from the optimal level. We conjecture that it is because the optimization algorithm terminates without reaching the real optimality. Since the parameter estimations changes in a rather small magnitude after the $150th$ iteration, there exists no reason for a dramatic change in the selection of optimal inputs with a compact design space.

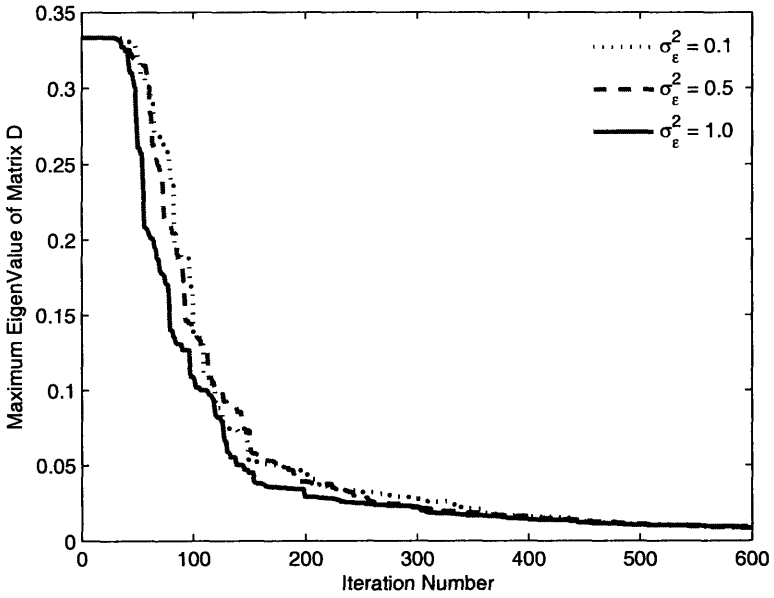

Figure 4-11: The Maximum Eigenvalue of Matrix $\mathbf{V}$ for Every Iteration

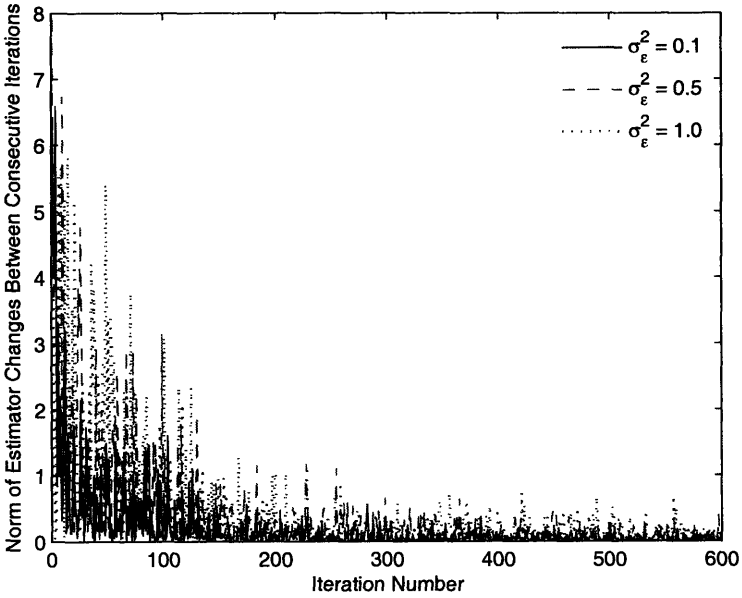

Figure 4-12: Norm of Differences between Estimators of Two Consecutive Iterations

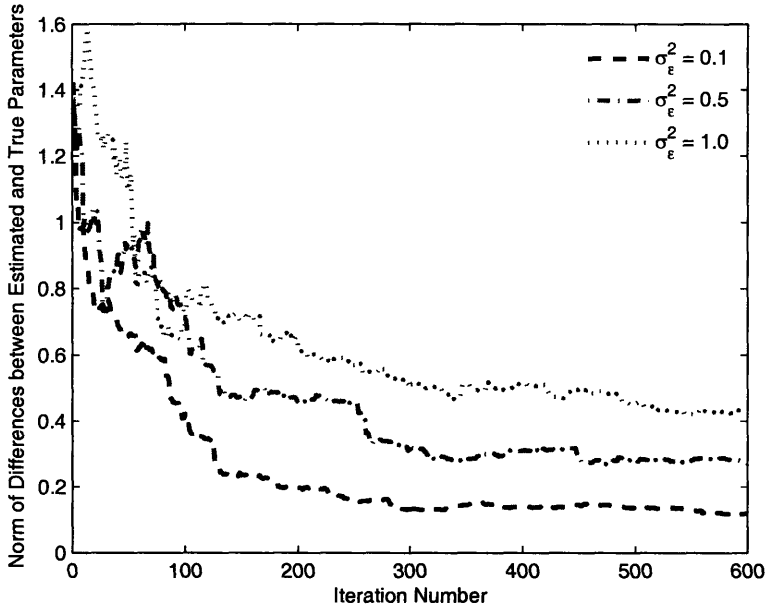

Figure 4-13: Norm of Differences between Estimated and True Parameters for Each Iteration

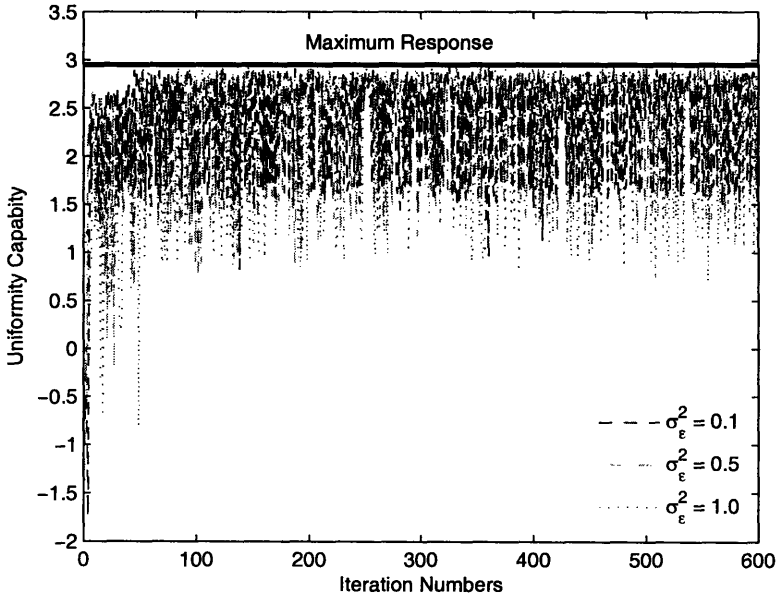

Figure 4-14: True Response of the Estimated Optimum Levels of Factors

Term	True Parameter	Estimator $\sigma_\epsilon^2 = 0.1$	Estimator $\sigma_\epsilon^2 = 0.5$	Estimator $\sigma_\epsilon^2 = 1.0$
A (x_1)	0.192	0.195	0.143	0.182
B (x_2)	0.121	0.108	0.110	0.115
C (x_3)	0.226	0.203	0.195	0.263
D (x_4)	0.061	0.023	0.035	0.120
E (x_5)	-0.048	-0.064	-0.024	0.030
F (x_6)	-0.087	-0.072	-0.147	-0.271
G (x_7)	0.279	0.306	0.274	0.145
$AB\,(x_1x_2)$	-0.133	-0.159	-0.094	-0.141
$AC\,(x_1x_3)$	0.406	0.410	0.430	0.359
$AD\,(x_1x_4)$	0.098	0.098	0.098	0.131
$AE\,(x_1x_5)$	-0.017	-0.035	0.027	0.044
$AF\,(x_1x_6)$	-0.006	0.021	0.030	0.116
$AG\,(x_1x_7)$	-0.019	0.022	0.016	0.099
$BC\,(x_2x_3)$	-0.548	-0.539	-0.601	-0.668
$BD\,(x_2x_4)$	0.009	0.013	-0.119	-0.010
$BE\,(x_2x_5)$	0.108	0.075	-0.005	0.080
$BF\,(x_2x_6)$	-0.019	-0.005	0.067	0.092
$BG\,(x_2x_7)$	-0.006	0.018	0.006	0.059
$CD\,(x_3x_4)$	-0.215	-0.202	-0.171	-0.161
$CE\,(x_3x_5)$	0.042	0.051	0.000	0.012
$CF\,(x_3x_6)$	-0.661	-0.629	-0.560	-0.491
$CG\,(x_3x_7)$	0.099	0.073	0.087	0.023
$DE\,(x_4x_5)$	-0.066	-0.043	-0.072	-0.015
$DF\,(x_4x_6)$	-0.017	-0.064	0.023	-0.144
$DG\,(x_4x_7)$	0.108	0.119	0.133	0.052
$EF\,(x_5x_6)$	0.098	0.124	0.158	0.102
$EG\,(x_5x_7)$	0.009	0.007	0.013	-0.009
$FG\,(x_6x_7)$	-0.133	-0.121	-0.140	-0.103

Table 4.6: True and Estimated Parameters under Different Experiment Error Variance σ_ϵ^2

Chapter 5

Concluding Remarks

Achieving better improvement through proper sequential experiments is an exciting area of research. The Bayesian analysis approach presented in this thesis has much to explore in the Design of Experiment field, in which the updated information could be used to determine the next stage of experiment, balance the tradeoff between exploration and exploitation, and validate the accuracy of estimations.

In this thesis, we start with considering a problem of achieving higher system performance with scarce experimental resources through analyzing a normal linear regression model with a Bayesian embedding prior information. To make the best use of information collected from stage-by-stage Bayesian analysis, we propose an adaptive one-factor-at-a-time (OFAT) experimental scheme for the case of discrete design space. We provide a mathematical argument to prove the superiority of this OFAT scheme in terms of expected improvement and its exploitation capability on factors. It is shown that, although this adaptive scheme could not achieve the "minimum" covariance matrix for estimation, the outcome of OFAT scheme is sustained by both the main as well as the interaction factors (see Theorem 7), while the conventional fractional factorial experimental design is only sustained by the main factor (see theorem 8) under the same experimental resource restriction. That is, the OFAT scheme could achieve a better improvement not only when the main factors but also the interaction factors are strong, while the fractional factorial design would perform poorly once the interaction factors become stronger. Figure 2.5 illustrates this com-

parison. The exploitation capability of OFAT scheme provides a further insight about its superiority in achieving better performance improvement. The OFAT scheme is shown to have a higher probability to exploit the interaction factor than does the saturated fractional experimental design. Moreover, Theorem 5 shows that the latter interaction factor would be more likely to be exploited than those exploited earlier. This property implies the stage-by-stage improvement capability of OFAT scheme, see (Frey and Wang, 2006) for detailed presentation.

We extended the Bayesian analysis to the case of generalized normal linear regression model within the compact design space. We apply the concepts of **c**-optimum experimental design and Bayesian experimental design to propose an algorithm Eq.(3.1) to Eq.(3.4) for the purpose of achieving optimum through a sequence of experiments. This experimental design has been shown to be consistent in its limiting behavior. Moreover, we also derive the expected stage improvement achieved by this algorithm for the analysis of its intermediate behavior. Through the analysis, we learn that there is a tradeoff between the variance of exploration (estimation) and the predicted mean response (exploitation). However, it is not necessary to minimize the covariance matrix for the whole design space, i.e. minimizing $\lambda_{\max}(\mathbf{V})$, to achieve our goal. Instead, we show that it is sufficient to minimize the uncertainty at the points which are of interests. Moreover, all the sequential optimum schemes developed so far are considered in their limiting behavior. The expected value of stepwise improvement could help experimenters determine whether to terminate the experiments or not.

This work is intended as an early step in a broader research program regarding adaptive experimentation schemes. Moreover, we are interested in potential applications of adaptive experimental scheme in engineering, economics and medicine areas. We therefore conduct a case study in the engineering design field. In addition, the adaptive mechanisms explored here are only the simplest kind requiring no physical knowledge. A richer theory would include consideration of the experimenter's mental models, the ways such models influence the planning of experiments, such as the order in which factors are considered, and the ways that the experimental data alter one's mental models. It is also possible that adaptive experimental scheme will prove useful

in computer experiments wherein some model errors may be present since physically reasonable predictions are more easily made when only one factor is changed or when the experimental resource is scarce. These topics are all interesting possibilities for future research.

Appendix A

A Brief Summary of Bayesian Analysis for Normal Linear Regression Models

A.1 Known Error Variance σ_ϵ^2

We consider the following normal linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where $\mathbf{y}$ is an $(b \times 1)$ vector, $\mathbf{X}$ is an $(b \times p)$ matrix, β is an $(p \times 1)$ column vector, and ϵ is an $(b \times 1)$ vector of random errors. The elements of ϵ is assumed to be *i.i.d.* normal random variables with zero mean and known variance σ_ϵ^2, i.e., $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 I_b)$, where I_b denotes the $(b \times b)$ identity matrix.[1] Moreover, the ith row of matrix $\mathbf{X}$ is denoted by $\mathbf{x}_i^T = (x_1, x_2, \ldots, x_q)$, where T stands for the transpose of a matrix. Thus the ith element of $\mathbf{y}$ is $y_i = \mathbf{x}_i^T\beta + \epsilon_i$. Note that b stands for the number of experimental runs, q denotes the number of factors, and p represents the number of elements in the coefficient vector β.

[1] If the variance of the error is unknown, then the natural conjugate of σ_ϵ^2 is the Inverse Gamma distribution.

The prior of the coefficients β is assumed to be

$$\beta \sim \mathcal{N}(\mathbf{m}, \mathbf{V})$$

where we assume that $\mathbf{V}$ is positive definite. We also define $\hat{\beta} \triangleq \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$, which will be used soon.

The posterior distribution of β would be

$$\begin{aligned} f(\beta\,|\,\mathbf{y}) &\propto f(\mathbf{y}|\beta)f(\beta) \\ &\propto \exp\left(-\frac{(\beta-\hat{\beta})^T\mathbf{X}^T\mathbf{X}(\beta-\hat{\beta})}{2\sigma_\epsilon^2}\right) \times \exp\left((\beta-\mathbf{m})^T\mathbf{V}^{-1}(\beta-\mathbf{m})\right) \\ &\propto \exp\left((\beta-\mathbf{m}^*)^T(\mathbf{V}^*)^{-1}(\beta-\mathbf{m}^*)\right) \end{aligned}$$

where

$$\begin{aligned} \mathbf{m}^* &= \left(\mathbf{V}^{-1}+\sigma_\epsilon^2\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{V}^{-1}\mathbf{m}+\sigma_\epsilon^2\mathbf{X}^T\mathbf{y}\right) \\ \mathbf{V}^* &= \left(\mathbf{V}^{-1}+\sigma_\epsilon^2\mathbf{X}^T\mathbf{X}\right)^{-1} \end{aligned}$$

Therefore we re-scale the prior of the coefficient β and set

$$\mathbf{V} = \sigma_\epsilon^2\Sigma$$

Since the posterior of the coefficient β is still normal, we have the associated posterior distribution

$$f(\beta|\,\mathbf{y}) \sim \mathcal{N}(\mathbf{m}^*, \mathbf{V}^*) \tag{A.1}$$

$$\mathbf{m}^* = \left(\Sigma^{-1}+\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\Sigma^{-1}\mathbf{m}+\mathbf{X}^T\mathbf{y}\right) \tag{A.2}$$

$$\mathbf{V}^* = \sigma_\epsilon^2\left(\Sigma^{-1}+\mathbf{X}^T\mathbf{X}\right)^{-1} \tag{A.3}$$

We make several remarks here.

1. If $(\mathbf{X}^T\mathbf{X})$ is singular, then there will not exist a unique solution of $\hat{\beta}$, which is

the well known *identification* issue. However, since we assume that $\Sigma > 0$, i.e., the matrix Σ is positive definite, we could still obtain a unique $\mathbf{m}^*$ in (A.2), and therefore have the posterior distribution (A.1) proper.

2. If the prior is very diffuse (weak), i.e., $\Sigma^{-1} \to \mathbf{0}$, then it is necessary that $(\mathbf{X}^T\mathbf{X})$ is non-singular; otherwise, the posterior distribution would become improper. As indicated in Chapter 2, the standard selection of matrix $\mathbf{X}$ in the *two-level* fractional factorial design would make the matrix $(\mathbf{X}^T\mathbf{X})$ nonsingular. Therefore, a diffuse prior would not cause any problem in obtaining the posterior distribution in the fractional factorial experiment design case.

3. Notice that

$$\begin{aligned}
\mathbf{m}^* &= \left(\Sigma^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1}\left(\Sigma^{-1}\mathbf{m} + \mathbf{X}^T\mathbf{y}\right) \\
&= \left(\Sigma^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1}\left(\Sigma^{-1}\mathbf{m} + \mathbf{X}^T\mathbf{X}\hat{\beta}\right) \\
&= \left(I_p - \Phi\right)\mathbf{m} + \Phi\hat{\beta}
\end{aligned}$$

where $\Phi = \left(\Sigma^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}$.

   (a) if the prior is very weak, i.e. Σ^{-1} is very "small", then the posterior mean $\mathbf{m}^*$ would put more weights on the data obtained from the experiments

   (b) if the prior $\mathbf{m}$ is "unbiased", then the posterior estimator $\mathbf{m}^*$ is unbiased

4. Suppose we are interested in the estimation of the responses with input matrix $\mathbf{X}_0$, and thus we denote

$$\mathbf{y}_0 = \mathbf{X}_0\beta + \epsilon_0$$

Then the distribution of the responses $\mathbf{y}_0$ is

$$\mathbf{y}_0 \sim \mathcal{N}\left(\mathbf{X}_0\mathbf{m}^*,\, \sigma_\epsilon^2\left(I + \mathbf{X}_0(\mathbf{V}^*)\mathbf{X}_0^T\right)\right)$$

A.2 Unknown Error Variance σ_ϵ^2

In this section, we consider the case in which the parameter σ_ϵ^2 is unknown. For simplicity, we assume that the prior distribution of σ_ϵ^2 is

$$f(\sigma_\epsilon^2) = \frac{\left(\frac{a}{2}\right)^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \left(\sigma_\epsilon^2\right)^{-\frac{d+2}{2}} \exp\left(-\frac{a}{2\sigma_\epsilon^2}\right) \tag{A.4}$$

which is the inverse Gamma distribution with parameter a, $d > 0$. As indicated earlier, given σ_ϵ^2, the prior distribution of β is $\beta|\sigma_\epsilon^2 \sim \mathcal{N}(\mathbf{m}, \mathbf{V})$. Thus, by Eq.(A.4), the joint prior of $(\beta, \sigma_\epsilon^2)$ is

$$f(\beta, \sigma_\epsilon^2) = \frac{\left(\frac{a}{2}\right)^{\frac{d}{2}}}{(2\pi)^{\frac{p}{2}}|\mathbf{V}|^{\frac{1}{2}}\Gamma(\frac{d}{2})} \left(\sigma_\epsilon^2\right)^{-\frac{d+p+2}{2}} \exp\left(-\frac{(\beta-\mathbf{m})^T\mathbf{V}^{-1}(\beta-\mathbf{m})+a}{2\sigma_\epsilon^2}\right) \tag{A.5}$$

where $|\mathbf{V}|$ denotes the determinate of matrix $\mathbf{V}$. Hence, the marginal prior distribution of β is

$$f(\beta) = \int f(\beta, \sigma_\epsilon^2)\, d(\sigma_\epsilon^2) = \frac{(a)^{\frac{d}{2}}\Gamma(\frac{d+p}{2})}{|\mathbf{V}|^{\frac{1}{2}}(\pi)^{\frac{p}{2}}\Gamma\left(\frac{d}{2}\right)} \Big(a + (\beta-\mathbf{m})^T\mathbf{V}^{-1}(\beta-\mathbf{m})\Big)^{-\frac{d+p}{2}}$$

Finally, we have

$$f(\mathbf{y}|\,\sigma_\epsilon^2, \beta) = \left(2\pi\sigma_\epsilon^2\right)^{-\frac{n}{2}} \exp\left(-\frac{(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)}{2\sigma_\epsilon^2}\right)$$

Therefore, we have the posterior distribution:

$$\begin{aligned} f(\beta, \sigma_\epsilon^2|\,\mathbf{y}) &\propto f(\mathbf{y}|\,\beta, \sigma_\epsilon^2) f(\beta, \sigma_\epsilon^2) \\ &\propto \left(\sigma_\epsilon^2\right)^{-\frac{n}{2}} \exp\left(-\frac{(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)}{2\sigma_\epsilon^2}\right) \\ &\qquad \times \left(\sigma_\epsilon^2\right)^{-\frac{d+p+2}{2}} \exp\left(-\frac{(\beta-\mathbf{m})^T\mathbf{V}^{-1}(\beta-\mathbf{m})+a}{2\sigma_\epsilon^2}\right) \\ &\propto \left(\sigma_\epsilon^2\right)^{-\frac{n+d+p+2}{2}} \exp\left(-\frac{(\beta-\mathbf{m}^\star)^T(\mathbf{V}^\star)^{-1}(\beta-\mathbf{m}^\star)+a^\star}{2\sigma_\epsilon^2}\right) \end{aligned}$$

that is equivalent to

$$f(\beta,\, \sigma_\epsilon^2|\, \mathbf{y}) = \frac{\left(\frac{a^\star}{2}\right)^{\frac{d+n}{2}}}{(2\pi)^{\frac{p}{2}}|\mathbf{V}^\star|^{\frac{1}{2}}\Gamma(\frac{d+n}{2})}\left(\sigma_\epsilon^2\right)^{-\frac{n+d+p+2}{2}}\exp\left(-\frac{(\beta-\mathbf{m}^\star)^T(\mathbf{V}^\star)^{-1}(\beta-\mathbf{m}^\star)+a^\star}{2\sigma_\epsilon^2}\right)$$

where

$$\begin{aligned}
\mathbf{m}^\star &= \left(\mathbf{V}^{-1}+\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{V}^{-1}\mathbf{m}+\mathbf{X}^T\mathbf{y}\right)\\
\mathbf{V}^\star &= \left(\mathbf{V}^{-1}+\mathbf{X}^T\mathbf{X}\right)^{-1}\\
a^\star &= a+\mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}+\mathbf{y}^T\mathbf{y}-(\mathbf{m}^\star)^T(\mathbf{V}^\star)^{-1}(\mathbf{m}^\star)
\end{aligned}$$

Finally, we consider the estimation distribution of the response $\mathbf{y}_0 = \mathbf{X}_0\beta + \epsilon_0$ at an $(r \times p)$ matrix $\mathbf{X_0}$. Notice that

$$\begin{aligned}
\mathbf{y}_0|\, \sigma_\epsilon^2 &\sim \mathcal{N}\left(\mathbf{X}_0\mathbf{m}^*, \sigma_\epsilon^2\left(I_r+\mathbf{X}_0\mathbf{V}^\star\mathbf{X}_0^T\right)\right)\\
f(\sigma_\epsilon^2|\, \mathbf{y}) &= \frac{\left(\frac{a^\star}{2}\right)^{\frac{d^\star}{2}}}{\Gamma\left(\frac{d^\star}{2}\right)}\left(\sigma_\epsilon^2\right)^{-\frac{d^\star+2}{2}}\exp\left(-\frac{a^\star}{2\sigma_\epsilon^2}\right)
\end{aligned}$$

Therefore, the posterior distribution of the response $\mathbf{y}_0$ is

$$f(\mathbf{y}_0|\, \mathbf{y}) = \frac{(a^*)^{\frac{n+d}{2}}\Gamma\left(\frac{n+d+r}{2}\right)}{|I_r+\mathbf{X}_0\mathbf{V}^\star\mathbf{X}_0^T|^{\frac{1}{2}}(\pi)^{\frac{r}{2}}\Gamma\left(\frac{n+d}{2}\right)}\left(a^\star+(\mathbf{y}_0-\mathbf{X}_0\mathbf{m}^\star)^T(I_r+\mathbf{X}_0\mathbf{V}^\star\mathbf{X}_0^T)^{-1}(\mathbf{y}_0-\mathbf{X}_0\mathbf{m}^\star)\right)^{-\frac{d+n+r}{2}} \tag{A.6}$$

which is a student t distribution.

Bibliography

Peter Auer. An improved on-line algorithm for learning linear evaluation functions. In *Proc. 13th Annu. Conference on Comput. Learning Theory*, pages 118–125. Morgan Kaufmann, San Francisco, 2000.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331. IEEE Computer Society Press, Los Alamitos, CA, 1995.

J. M. Bernardo. Expected Information as Expected Utility. *The Annals of Statistics*, 7(3):686–690, 1979.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, Cambridge, MA, 2001.

A. Blum, V. Kumar, A. Rudra, and F. Wu. Online learning in online auctions, 2003.

J. Blum. Approximation Method which Converge with Probability One. *Ann. Math. Statist.*, 25:382–386, 1954.

G. E. P. Box. Statistics as a Catalyst to Learning by Scientific Method: Part II - A Discussion. *Journal of Quality Technology*, 31(1):16–29, January 1999.

G. E. P. Box and N. R. Draper. Robust Designs. *Biometrika*, 62(2):347–352, 1975.

G. E. P. Box and P. T. Y. Liu. Statistics as a Catalyst to Learning by Scientific Method: Part I - An Example. *Journal of Quality Technology*, 31(1):1–15, January 1999.

G. E. P. Box and K. B. Wilson. On the Experimental Attainment of Optimum Conditions. *Journal of the Royal Statistical Society: Series B*, 13(1):1–38, 1951.

G. E. P. Box, W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building.* Wiley-Interscience, New York, June 1978.

George E. P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis.* Addison-Wiley, New York, 1973.

M. Brezzi and T. L. Lai. Incomplete Learning from Endogenous Data in Dynamic Allocation. *Econometrica*, 68(6):1511–1516, 2000.

James Buckner, Richard Huang, Kenneth A. Monnig, Eliot K. Broadbent, Barry L. Chin, Jon Henri, Mark Sorell, and Kevin Venor. *Modeling A Uniformity Bulls-Eye Inversion*, chapter 16, pages 235–250. Statistical Case Studies for Industrial Process Improvement. SIAM, Philadelphia, 1997. Verinica Czitrom and PatricK D. Spagon (Eds.).

K. Chaloner. Optimal Bayesian Experimental Design for Linear Models. *Annals of Statistics*, 12(1):283–300, 1984.

K. Chaloner and I. Verdinelli. Bayesian Experimental Design: A Review. *Statistical Science*, (3):273–304, August 1995.

H. M. Chipman, M. Hamada, and C. F. J. Wu. Bayesian Variable-Selection Approach for Analyzing Designed Experiments with Complex Aliasing. *Technometrics*, 39(4): 372–381, November 1997.

David A. Cohn. Neural Network Exploration Using Optimal Experiment Design. *Technical Report, MIT, AI Lab*, 1994.

V. Czitrom. One-Factor-at-a-Time Versus Designed Experiments. *The American Statistician*, 53(2):126–131, May 1999.

C. Daniel. One-at-a-Time Plans. *Journal of the American Statistical Association*, 68 (342):353–360, 1973.

M. H. DeGroot. Uncertainty, Information and Sequential Experiments. *The Annals of Mathematical Statistics*, 33:404–419, 1962.

H. Dette. A Note on Bayesian c- and d-Optimal Designs in Nonlinear Regression Models. *Annals of Statistics*, 24(3):1225–1234, June 1996.

J. Engel and A. F. Huele. A Generalized Linear Modeling Approach to Robust Design. *Technometrics*, 38(4):365–373, November 1996.

V. V. Fedorov. *Theory of Optimal Experiments.* Academic Press, New York, NY, 1972.

I. Ford and S. D. Silvey. A Sequentially Constructed Design for Estimating a Non-linear Parametric Function. *Biometrika*, 67(2):381–8, 1980.

Ian Ford, D.M. Titterington, and C.P. Kitsos. Recent Advances in Nonlinear Experimental Design. *Technometrics*, 31(1):46–90, Feburary 1989.

D. Foster and R. Vohra. Regret in the on-line decision problem, 1997.

D. D. Frey, F. Engelhardt, and E. M. Greitzer. A Role for One Factor at a Time Experimentation in Parameter Design. *Research in Engineering Design*, 14(2): 65–74, May 2002.

Daniel D. Frey. How One-Factor-at-a-Time Experimentation Can Lead to Greater Improvements than Orthogonal Arrays. Chicago, IL, September 2003. DETC 2003/DTM48646, ASME Design Engineering Technical Conferences.

Daniel D. Frey and Hungjen Wang. Adaptive One-Factor-at-a-Time Experimentation and Expected Value of Improvement. *Technometrics*, 48(3):418–431, August 2006.

M. Friedman and L. J. Savage. Planning Experiments Seeking Maxima. *Techniques of Statistical Analysis*, pages 365–372, 1947.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman & Hall, second edition, 2003.

J. C. Gittins. Bandit Processes and Dynamic Allocation Indices. *Journal of Royal Statitstical Scocirty Series B*, 41(2):148–177, 1979.

T. L. Graves and T. L. Lai. Asymptotically Efficient Adaptive Choice of Control Laws in Controlled Markov Chains. *SIAM J. Control and Optimization*, 35(3):715–743, May 1997.

E. J. Gumbel. *Statistics of Extremes.* Dover Publications.

R.F. Gunst and G.C. McDonald. The Importance of Outcome Dynamics, Simple Geometry and Pragmatic Statistical Arguments in Exposing Deficiencies of Experimental Design Strategies. *American Statistican*, 50(1):44–50, February 1996.

G. A. Hazelrigg. A Framework for Decision-Based Engineering Design. *Journal of Mechanical Design*, 120(4):653–658, December 1998.

W. J. Hill and W. G. Hunter. A Review of Response Surface Methodology: A Literature Survey. *Technometrics*, 8(4):571–590, November 1966.

Yu-Chi Ho and Kai Ching Chu. Team Decision Theory and Information Structures in Optimal Control Problems - Part I. *IEEE Transaction on Automatic Control*, 23:305–311, 1972.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, New York, 1990.

Inchi Hu. Strong Consistency of Bayes Estimates in Stochastic Regression Models. *Journal of Multivariate Analysis*, 57(2):215–227, May 1996.

J. Kiefer and J. Wolfowitz. Stochastic Estimation of the Maximum of a Regression Function. *Annals of Mathematical Statistics*, 23:462–466, 1952.

R. Koita. Strategies for Sequential Design of Experiments. Master's thesis, Massachusetts Institute of Technology, 1994.

G. A. Korn and T. M. Korn. *Mathematical Handbook for Scientists and Engineers.* Dover Publications, Mineola, NY, June 2000.

P. R. Kumar. A Survey of Some Results in Stochastic Adaptive Control. *SIAM J. Control and Optimization*, 23(3):329–380, 1985.

H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications.* Springer, New York Berlin Heidelberg, 2003.

T. L. Lai and H. Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advanced Applied Mathematics*, 6(1):4–22, 1985.

T. L. Lai and C. Z. Wei. Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems. *Ann. of Statist.*, 10(1):154–166, 1982.

Tze Leung Lai. Sequential Analysis: Some Classical Problems and New Challenges. *Statisca Sinica*, 11(2):303–408, April 2001.

R. V. Lenth. Quick and Easy Analysis of Unreplicated Factorials. *Technometrics*, 31 (4):469–473, November 1989.

Xiang Li, Nandan Sudarsanam, and Daniel D. Frey. Regularities in Data from Factorial Experiments: Research Articles. *Complex.*, (5):32–45, 2006. ISSN 1076-2787. doi: http://dx.doi.org/10.1002/cplx.v11:5.

D. K. J. Lin and W. Tu. Dual Response Surface Optimization. *Journal of Quality Technology*, 26(1):248–260, January 1995.

D. V. Lindley and A.F.M. Smith. Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society, Series B*, 34(1):1–41, 1972.

N. Logothetis and H. P. Wynn. *Quality Through Design.* Clarendon Press, Oxford, January 1995.

J. M. Lucas. How to Achieve a Robust Process using Response Surface Methodology. *Journal of Quality Technology*, 26(4):248–260, October 1994.

David J. C. MacKay. Information Based Objective Functions for Active Data Selection. *Neural computation*, 4(4):589–603, July 1992.

David J. C. MacKay. Searching for "Optimal" Inputs with an Empirical Regression Model. *Lecture Notes*, November 1997.

David J. C. MacKay. Introduction to Gaussian Processes. *Lecture Notes*, 1998.

Jacob Marschak and Roy Radner. *Economic Theory of Teams*. Yale University Press, New Heaven and London, 1972.

W. R. McDaniel and B. E. Ankenman. Comparing Experimental Design Strategies for Quality Improvement with Minimal Changes to Factor Levels. *Quality and Reliability Engineering International*, 16:355–362, November 2000.

Robert W. Mee and Marta Peralta. Semifolding 2^{k-p} Designs. *Technometrics*, 42(2): 122–134, May 2000.

R. H. Myers. Response Surface Medology: Current Status and Future Directions. *Journal of Quality Technology*, 31(1):30–44, January 1999.

R. H. Myers, A. I.Khuri, and W. Carter. Response Surface Methodology: 1966-1988. *Technometrics*, 31(2):137–157, May 1989.

R. H. Myers, A. I. Khuri, and G. Vining. Response Surface Alternatives to the Taguchi Robust Parameter Design Approach. *The American Statistician*, 46(2): 131–139, May 1992.

V. N. Nair. Taguchi's Parameter Design: A Panel Discussion. *Technometrics*, 34(2): 127–161, May 1992.

D.C. Parkes and B.A. Huberman. Multi-Agent Cooperative Search for Portfolio Selection. *Games and Economic Behavior*, 35(1-2):124–165, April 2001.

A. Pázman. A Convergence Theorem in the Theory of D-Optimum Experimental Designs. *The Annals of Statistics*, 2(1):216–218, 1974.

Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001, Amsterdam, The Netherlands, July 2001, Proceedings*, pages 208–223. Springer, Berlin, 2001. URL `citeseer.ist.psu.edu/piccolboni00discrete.html`.

L. Pronzato and E. Thierry. Sequential Experimental Design for Estimating the Optimum Point in a Response Surface. *Satistical Models and Applications*, 11(3): 277–292, 2003.

Luc Pronzato. Adaptive Optimization and D-Optimum Experimental Design. *Annals of Statistics*, 28(6):1743–1761, December 2000.

X. Qu and C. F. J Wu. One-Factor-at-a-Time Designs of Resolution V. *Journal of Statistical Planning and Inference*, 131(2):407–416, May 2005.

Roy Radner. Team Decision Problems. *Annals of Mathematical Statistics*, 33(3): 857–881, 1962.

H. Robbins. Some Aspects of the Sequential Design of Experiments. *Bull. Amer. Math. Soc.*, 58:527–535, 1952.

H. Robbins and S. Monro. A Stochastic Approximation Method. *Ann. Math. Stat.*, 22:400–407, 1951.

M. Rothschild. A Two-Armed Bandit Theory of Market Pricing. *Journal of Economic Theory*, 9(2):185–202, 1974.

D. Ruppert. *Stochastic Approximations*, pages 503–529. In Handbook of Sequential Analysis. Deffer, New York, 1991.

X.T. Shen, H.C. Huang, and J. Ye. Adaptive Model Selection and Assessment for Exponential Family Distributions. *TECHNOMETRICS*, 46(3):306–317, August 2004.

S. D. Silvey. *Optimal Design.* Chapman & Hall, London, 1980.

J. C. Spall. Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximations. *IEEE Transactions on Automatic Control*, 37(3): 332–341, March 1992.

J. C. Spall and J. A. Cristion. Nonlinear Adaptive Control using Neural Networks: Estimation with a Smoothed Form of Simultaneous Perturbation Gradient Approximation. *American Control Conference*, 3(1):2560– 2564, June 1994.

G. Taguchi. *Systems of Experimental Design.* Kraus, New York, 1987.

Y. L. Tong. *The Multivariate Normal Distribution.* Springer Series in Statistics, 1990.

Nair V. N, . Taam W, and K. Q. Ye. Analysis of Functional Responses from Robust Design Studies. *Journal of Quality Technology*, 34(4):355–370, October 2002.

I. Verdinelli. *Advances in Bayesian Experimental Design*, pages 467–481. Bayesian Statistics 4. Bernardo, J., Berger, J., Dawid, A., Smith, A. (Eds.). Clarendon Press, Oxford, 1992.

I. Verdinelli and J.B. Kadane. Information and Outcome. *Journal of the American Statistical Association*, 87(418):510–515, June 1992.

V. Vovk. Competitive on-line statistics, 1999.

C. F. J. Wu and M. Hamada. *Experiments: Planning, Analysis, and Parameter Design Optimization.* Wiley-Interscience, NY, April 2000.