Two-Stage Computing Budget Allocation Approach for Response Surface Method

PENG JI

NATIONAL UNIVERSITY OF SINGAPORE

2005

Two-Stage Computing Budget Allocation Approach for Response Surface Method

PENG JI

(B. Eng. USTC)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF ENGINEERING

DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2005

Acknowledgement

I would like to express my profound gratitude to my supervisors, Dr. Ng Szu Hui, Dr. Lee Chulung and Prof. Lee Loo Hay, for their invaluable guidance and support during my stay at NUS.

Special thanks to Dr. Ng Szu Hui and Prof. Lee Loo Hay, not only for their continuous help and advice throughout my research but also for their offering of financial support. Without them this thesis can never be finished.

I also would like to extend my gratitude to all my friends who have made my life in Singapore an experience I will never forget. Thank Yang Guiyu, Liu Bin, Xu Zhiyong, Lin Wei, Zeng Yifeng, Bao Jie, Han Yongbin and Liu Na (my seniors) and all other students in Logistics Lab where I spent most of the time in the past two years.

Finally, thank my family for their support, understanding and encouragement throughout the course of my study and research.

Contents

Summary

Currently simulation optimization techniques are widely used to identify the best levels of the input parameters that will yield the optimal expected performance of the stochastic system. Response Surface Methodology (RSM) is one of the main statistical approaches to search for the best input parameters. In the early stages of RSM, the steepest ascent is locally estimated and the iterative hill-climbing procedure is involved. To improve the method of steepest ascent, Kleijnen et al. [1] propose a technique which they call adapted steepest ascent (ASA). Although the search method for hill-climbing is efficient, simulation itself can be very time consuming and expensive. Moreover, when there are budget constraints little research is done on determining the best allocation of the replications at each design point.

In this thesis, we apply ASA technique to the simulation optimization problems, improve on it by considering the more realistic case where there are computing budget constraints, and look into the important question of experimental design. We assume the initial design structure for every iteration of hill-climbing is a two-level factorial design and propose a two-stage approach to determine the allocation of replications for this factorial design. In stage 1, a regular two-level factorial design is applied, and a small portion of the limited computing budget is used to estimate the true response function. In stage 2, the rest of the budget is allocated in the local region to maximize the lower bound of predicted response at the next design point, which is determined by the technique of Kleijnen et al. [1]. In order to demonstrate the advantages of our two-stage computing budget allocation approach, we compare it with the approach which allocates the simulation runs equally to each design point. The numerical results show that our two-stage allocation outperforms the equal allocation especially when the system noise is large, and if we have more replications to be allocated by the second stage, the efficiency of hill-climbing will be even higher.

(Computing Budget Allocation; Experimental Design; Response Surface Methodology; Simulation Optimization)

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Currently simulation is widely used to design a system that will yield optimal expected performance. In a typical simulation optimization study, it is assumed that the performance of the system of interest depends on the values of a few input parameters chosen for the system, and experimenters want to determine the optimal values of these input parameters using simulation. (Without loss of generality, we only consider maximization problems in this thesis.)

Sequential response surface methodology (RSM) is one of the main statistical procedures to help experimenters search for the optimal values of input parameters, see Box and Draper [2], Myers and Montgomery [3] and Khuri and Cornell [4]. In the early stages of RSM, two-level factorial or fractional factorial designs are extensively used to locally fit a first-order model, and an iterative steepest ascent (SA) search procedure is involved. The iterative steepest ascent search procedure is also known as hill-climbing, and it can be illustrated by a two-dimensional model. Figure 1.1 represents the contour plot of a two-dimensional response function, d_1 and d_2 are the two input parameters, and point **A** marked with '*' is the initial design point. In the first iteration, point **A** is the center of the region of experimentation, a 2^2 factorial design is used and four design points $\mathbf{d_1}$, $\mathbf{d_2}$, $\mathbf{d_3}$, $\mathbf{d_4}$ marked with '$\cdot$' are determined. The local square region that contains all the four design points is called the region of

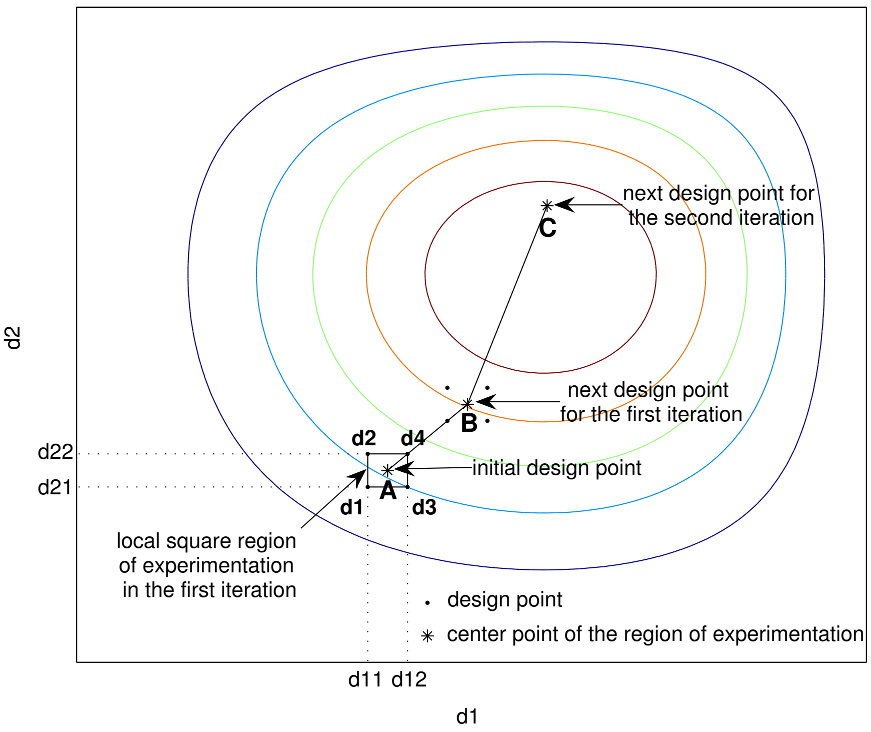


Figure 1.1: Two-dimensional model for the method of steepest ascent

experimentation. If low and high values of d_1 in the local region of experimentation are d_{11} and d_{12} respectively, and those of d_2 are d_{21} and d_{22}, then $\mathbf{d_1} = (d_{11}, d_{21})$, $\mathbf{d_2} = (d_{11}, d_{22})$, $\mathbf{d_3} = (d_{12}, d_{21})$ and $\mathbf{d_4} = (d_{12}, d_{22})$. With the observations at $\mathbf{d_1}$, $\mathbf{d_2}$, $\mathbf{d_3}$, and $\mathbf{d_4}$, a first-order regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 d_1 + \hat{\beta}_2 d_2$ is estimated, where β_0, β_1 and β_2 are the unknown coefficients and y is the response. Then the direction of steepest ascent $(\hat{\beta}_1, \hat{\beta}_2)$ is determined. In the direction of steepest ascent, an arbitrary step size is chosen and the next design point **B** is identified. In the second iteration, point **B** becomes the center of region of experimentation, and four new design points of the two-level factorial design are chosen. The same procedure to climb the response surface is then repeated and the next design point moves to point **C**. This hill-climbing procedure will continue until a termination criteria is met. Although steepest ascent is the most efficient path to improve response values based on the local observations, the direction of steepest ascent is scale-dependent and its step size is always chosen arbitrarily, which means if the scale of d_1 or d_2 is different, or a different step size is chosen, the values of next design points (**B** and **C** in this example) can be different (see Myers and Montgomery [3]).

To tackle the above two drawbacks of steepest ascent, Kleijnen et al. [1] suggest a novel technique which they call Adapted Steepest Ascent (ASA) technique. In their study, they consider the lower, one-sided $1 - \alpha$ confidence interval for the predictor $\hat{y}$ based on the first-order model, and this interval will range from infinity down to the lower bound $\hat{y}_{min}$. The authors prove that $\hat{y}_{min}$ is a concave function and they derive the design point $\mathbf{d}^+$ which maximizes $\hat{y}_{min}$. The authors refer to the maximal point $\mathbf{d}^+$ as the next design point because it includes both a search direction and a step size. Furthermore, the authors prove that their technique is scale-independent and they also illustrate that their suggested adapted steepest ascent (ASA) direction is better than the SA direction through Monte Carlo experiments. If we apply the ASA technique to conduct hill-climbing in Figure 1.1, then the next design point **B** after experimentation in the local region of **A** can be obtained mathematically by computing $\mathbf{d}^+$. Similarly, **C** can be determined by computing $\mathbf{d}^+$ after experimentation in local region of **B** is conducted. The details of this technique will be discussed in Chapter 3.

While the searching method to improve the response of a simulation model is very efficient, the simulation experiment itself can be very time consuming and expensive (Law and Kelton [5]). In order to obtain a good statistical estimate at each design point, a large number of simulation replications is usually required. The ultimate accuracy (typically expressed as a confidence interval) of a performance estimator cannot improve faster than $O(1/\sqrt{M})$, where M is the number of simulation replications (see Fabian [6]). If the accuracy requirement is high, the total number of simulation replications can easily become prohibitively large. Besides the large number of replications, one single replication can also be very time consuming for a large-scale simulation due to the large number of random occurrences and the long run length required to obtain stable estimates. Typical simulation studies of sea port operations or air traffic systems can take an average of 10 to 12 hours, and Kleijnen [7] also reported a simulation study of a manufacturing system where one design could take six hours of computer time. In addition, although the computer hardware is getting cheaper and faster, the cost of simulation software alone can also make simulation experiments very expensive. All of these make computing budget constraints a significant concern when conducting simulation experiments.

When there is insufficient budget to carry out all the necessary experiments, most of the literature seeks the designs to reduce the design points directly, such as the fractional factorial design (Kleijnen [8]), or to screen out unimportant factors and reduce the design points indirectly, such as the Plackett-Burman designs (Plackett and Burman [9] and Ahuja et al. [10]). Unfortunately, little research is done on determining the optimal allocation of replications at the fixed design points of factorial design when there are budget constraints.

Here we define the computing budget allocation problem as the experimental design problem specialized in allocating the replications among the design points of a two-level factorial design in every iteration of hill-climbing. Referring to Figure 1.1, if we consider the first iteration only (i.e. moving from region A to region B), the computing budget allocation problem means how to decide the number of replications at $\mathbf{d_1}$, $\mathbf{d_2}$, $\mathbf{d_3}$ and $\mathbf{d_4}$ given the total number of replications.

Thus it is important to study the computing budget allocation problem as simulation can be both time consuming and expensive, and the regular response surface designs rarely consider the problem of allocating replications at each design point when there are computing budget constraints. Even in Kleijnen et al. [1], although the authors show that their ASA technique improves the efficiency of the traditional SA method, the authors assume the design is fixed and do not consider the computing budget allocation problem either.

1.2 Problem Statement

In this thesis, we focus on the early stages of RSM and consider a more realistic case in which budget constraints are present. Instead of studying how to reduce design points to save computing budget, we study how to allocate the given number of replications for the two-level factorial designs and further improve the efficiency of the ASA technique. In short, while the ASA technique studies into how to identify the maximal point $\mathbf{d}^+$ of $\hat{y}_{min}$ without considering budget constraints, we will study how to design experiments that maximize $\hat{y}_{min}$ at the next design point $\mathbf{d}^+$ given a fixed budget. More specifically, we are considering the following problem:

(P) given design points $\mathbf{d_1}$, $\mathbf{d_2}$, $\cdots$, $\mathbf{d_m}$ of a factorial design, find the best allocation of $n_1, n_2, \cdots, n_m$, so that the lower bound of the predicted response $\hat{y}_{min}$ at the next design point $\mathbf{d}^+$ is maximized with the constraint that $n_1+n_2+\cdots+n_m = N$, where n_i is the number of replications at point $\mathbf{d}_i$, m is the number of design points, and N is the total number of replications for the factorial design in the region of experimentation.

Intuitively, if we know the next design point when we are still in the local region of experimentation, we can design experiments in the local region to maximize the lower bound of predicted response $\hat{y}_{min}$ at the next design point. In that case, when we move to the next design point, we can be more assured that the expected response $\hat{y}$ is better improved because $\hat{y}_{min}$ is the worst prediction of $\hat{y}$ at level α.

As the ASA technique, which is used to determine the next design point, can be applied only after the true response function is estimated, the main challenge in solving problem **(P)** is: how to estimate the true response function and determine the next design point, in order to find the best allocation of replications to maximize $\hat{y}_{min}$ at the next design point.

1.3 Research Contributions

The main contribution in this thesis is: we apply the ASA technique to simulation optimization problems, and consider the case where there are computing budget constraints. We develop a two-stage computing budget allocation approach for one single iteration of hill-climbing. In stage 1, a traditional two-level factorial design is used, a limited computing budget is equally distributed to all the design points in the region of experimentation and a linear response function is estimated; in stage 2, the rest of the budget is distributed among the design points of that factorial design to maximize the lower bound of predicted response at the next design point $\mathbf{d}^+$. A series of numerical experiments are carried out for linear models and nonlinear models to test the performance of our two-stage computing budget allocation approach. The numerical results show that our two-stage approach outperforms the equal allocation especially when the noise is large, and the efficiency of hill-climbing can be further improved if we leave more budget to be determined by the second stage of our two-stage approach.

1.4 Organization of this Thesis

The rest of this thesis is divided into 4 parts. In Chapter 2, a literature review of simulation optimization techniques is presented, followed by a survey of response surface designs, where we can find optimal design, robust design and some new developments, and then an important contribution in computing budget allocation, the Optimal Computing Budget Allocation (OCBA), is reviewed. Chapter 3 first introduces the main idea of adapted steepest ascent (ASA) technique proposed by Kleijnen et al. [1]. Then we develop our two-stage computing budget allocation approach, explain how it works for simulation optimization problems and do some pilot studies to validate its advantages. Chapter 4 contains the numerical experiments used to compare our two stage allocation approach to the equal allocation approach for nonlinear models. In Chapter 5, we summarize this thesis and propose the future work.

Chapter 2

Literature Review

As we focus on using RSM to optimize the simulation output under computing budget constraints, we review the following three topics: simulation optimization, response surface designs and computing budget allocation.

We first outline simulation optimization and its main techniques in section 2.1. Next we discuss the response surface designs for response surface methodology in section 2.2, and finally we review the research work to tackle the problem of computing budget allocation in section 2.3.

2.1 Simulation Optimization Techniques

Simulation has been recognized as a very powerful tool to evaluate and justify a stochastic system. In the last decade, however, 'optimization' routines have been prominently adopted by many simulation packages, and simulation optimization has thus become widespread. Fu [11] defined simulation optimization as 'optimization of performance measures based on outputs from stochastic simulations', and he divided the simulation techniques into the following main categories:

- Statistical procedures: sequential response surface methodology, ranking & selection procedures, and multiple comparison procedures;

- Metaheuristics: methods directly adopted from deterministic optimization search strategies, such as simulated annealing, tabu search, and genetic algorithms;

- Stochastic optimization: random search, stochastic approximation;
- Others, including ordinal optimization and sample path optimization.

The detailed introduction about these techniques can be found in Fu [12].

2.2 Response Surface Designs

Myers et al [13] defined response surface methodology (RSM) as a collection of tools in design or data analysis that enhance the exploration of a region of design variables in one or more responses. By this definition, it highlights two important aspects of RSM, one is response surface design and another is data analysis. Response surface design is the main concern in this thesis and it will be further discussed in the later part of this section. For data analysis, a distinction can be made between analysis whose goals are to explore the response surface and that whose goals are to estimate the optimal input levels. The method of steepest ascent (SA) is a viable technique for exploring the response surface and sequentially moving toward the optimum response. And it is also the main technique to do data analysis in this thesis. To study the optimal point, the canonical analysis is the most popular tool. One can write the true second-order response model as the canonical form $y = \beta_0 + \mathbf{x}^T\beta + \mathbf{x}^T\mathbf{B}\mathbf{x}$, where $\mathbf{x}$ denotes k control factors $\mathbf{x}^T = [x_1, x_2, \cdots, x_k]$, β is a $k \times 1$ vector containing the regression coefficients of the control factors, and $\mathbf{B}$ is a $k \times k$ matrix whose main diagonals are the regression coefficients associated with the pure quadratic effects of the control factors and whose off-diagonals are one-half of the mixed quadratic (interaction) effects of the control factors. With this canonical form, experimenters may compute the stationary point, the response at the stationary point and the confidence region for the location of the stationary point, as well as analyze the characterization of the stationary point (i.e., as a point of maximum or minimum response or a saddle point). Another important characteristic of RSM is that most of its applications are *sequential* in nature. At first, many factors or variables may be taken in account as potential important effects that affect the response, and a screening experiment is designed to investigate these factors with a view toward eliminating the unimportant ones. A response surface analysis should never be done until a screening experiment has been performed, and fractional factorial designs are powerful tools to identify the important factors. Once the important variables are identified, the next phase is to determine if the current setting of input variables results in a value of the response that is near the optimum or if the process is operating in some other region which is remote from the optimum. If the current setting is not consistent with the optimum performance, the experimenter must decide how to adjust the process variables that will move the response toward the optimum. This phase of response surface methodology makes considerable use of the first-order model and an optimization technique known as the method of steepest ascent. If the process is near the optimum, the final phase is carried out. Because the true response surface usually exhibits curvature near the optimum and the regression model must accurately approximate the true response function, a second-order or higher-order model will be used. Once an appropriate approximating model has been estimated, this model can be analyzed to determine the optimum conditions for the process. More details on RSM can be found in books like Box and Draper [2], Myers and Montgomery [3], and Khuri and Cornell [4].

While RSM is one of the main statistical procedures to maximize the process, response surface design is a critical issue within the context of RSM because it addresses the problem to fit the response surface and represent the surface mathematically. For good response surface designs, Box and Draper [14] suggested the following desirable properties. The design should:

1. Generate a satisfactory distribution of information throughout the region of interest, R.

2. Ensure that the fitted value be as close as possible to the true value.

3. Give good detectability of lack of fit.

4. Allow transformations to be estimated.

5. Allow experiments to be performed in blocks.
6. Allow designs of increasing order to be built up sequentially.
7. Provide an internal estimate of error.
8. Be insensitive to wild observations and to violation of the usual normal theory assumptions.
9. Require a minimum number of experimental runs.
10. Provide simple data patterns that allow ready visual appreciation.
11. Ensure simplicity of calculation.
12. Behave well when errors occur in the setting of the predictor variables, the x's.
13. Not require an impractically large number of levels of the predictor variables.
14. Provide a check on the "constant variance" assumption.

Although not all of the above properties are required in every RSM experience, most of them must be considered seriously. When we design the two-stage computing budget allocation approach, we also consider some of the above properties.

Since Box and Wilson [15], substantial progress has been made in the area of response surface designs for both first- and second-order models. The review first goes through the two main categories of response surface design - optimal design and robust design, and then briefly discusses two useful designs, sequential design and Bayesian design. Because this thesis concentrates on the earlier stage of RSM, we will pay more attention to those response surface designs for first-order model in the review.

2.2.1 Optimal Design

If we consider the linear model

$$y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \cdots + \beta_k d_k + \epsilon,$$

we can express it in matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where vector $\mathbf{Y}$ is an $n \times 1$ vector of observations; $\mathbf{X}$ is an $n \times q$ matrix, with row i containing $\mathbf{x}_i{}^T$, and $q = k + 1$; $\mathbf{x}_i$ is a $q \times 1$ vector of predictor variables for the i^{th} input combination $(1 \;\; d_{1i} \;\; d_{2i} \;\; \cdots \;\; d_{ki})^T$; $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown parameters $(\beta_0 \;\; \beta_1 \;\; \cdots \;\; \beta_k)^T$; $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of independently and identically distributed random variables, with mean zero and variance σ^2.

We assume that least squares estimates of the parameter $\boldsymbol{\beta}$ are to be obtained, so that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

and the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Then the predicted response at point $\mathbf{x}^* \in \chi$, where χ denotes design space, is

$$\hat{y}(\mathbf{x}^*) = \mathbf{x}^{*T}\hat{\boldsymbol{\beta}},$$

with variance

$$var(\hat{y}|\mathbf{x}^*) = \sigma^2\mathbf{x}^{*T}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*.$$

The design problem consists of selecting vectors $\mathbf{x}_i, i = 1, 2, \cdots, n$ from χ such that the design defined by these n vectors is, in some defined sense, optimal.

Smith [16] was one of the first to state a criterion and obtain optimal experimental designs for regression problems. She proposed the criterion

$$min_{x_i, i=1,2,\cdots,n} max_{\mathbf{x}\in\chi} var(\hat{y}(\mathbf{x})).$$

This criterion was later called G-optimality by Kiefer and Wolfowitz [17]. Wald [18] proposed the criterion of maximizing the determinant of $\mathbf{X}^T\mathbf{X}$ as a means of maximizing the local power of the F-ratio for testing a linear hypothesis on the parameters of certain fixed-effects analysis of variance model. Kiefer and Wolfowitz

[17] later called this criterion D-optimality and extended its use to regression models in general.

Later, Kiefer and Wolfowitz [19] proved the equivalence of D- and G-optimality. Based on their contributions, people could easily verify whether or not a specific design is D-optimal, and many efficient algorithms were proposed to construct D-optimal designs. A dedicated book, Fedorov [20], discussed extensively on constructing optimal designs.

Besides D- and G-optimality, there are a few other variance-optimal designs, such as A- and E-optimality, and sometimes they are called alphabetic optimality as a whole. More details about variance-optimal design in the response surface context can be found in Chapter 14 of Box and Draper [2].

Because optimal designs are only concerned with optimality of a very narrow kind and they assume the estimated model exactly represents the true model, Box, Hunter and Hunter [21] criticized that "in recent years the study of optimal design has become separated from real experimentation with the predictable consequence that its limitations have not been stressed, or, often, even realized" (p.472). Thus, a lot of works were proposed to study the robustness of response surface design, which is reviewed in the next section.

2.2.2 Robust Design

Steinberg and Hunter [22] divided the robust designs for RSM into the following categories: (a) protection against model misspecification, (b) designs for extrapolation under conditions of model misspecification, (c) robustness to errors in the design levels, and (d) robustness to outliers or missing observation. Aside from the above model-robust designs and error-robust designs, Steinberg and Hunter [22] also regarded the designs whose purpose was to discriminate among candidate models as a kind of robust designs and those designs were referred as model-sensitive designs.

Here we review the designs dealing with protection against model misspecification and the designs for extrapolation under conditions of model misspecification. As we mentioned before, this review focuses on the first-order model and model misspecification in this context means that the true model is a two- or higher-order model.

Box and Draper [23, 24] first introduced the notion of robustness of response surface design to model misspecification. The fundamental philosophy of their work is to consider the integrated weighted mean squared error (IMSE)

$$J = \frac{NK}{\sigma^2} \int_R w(x) E[\hat{y}(x) - g(x)]^2 \, \mathrm{d}x,$$

where $\hat{y}(x)$ is the fitted polynomial of order o_1 and $g(x)$ is a model of order o_2 ($o_2 > o_1$) which is regarded as the "true" response, R is the region of interest—that is, a region in which it is important for $\hat{y}$ to predict well, K is the reciprocal of the volume of R, N is the total number of observations, $w(x)$ is a weight function, and σ^2 is the error variance. One important work that extended the IMSE criterion to simulation design was done by Donohue, Houck and Myers [25]. They considered the strategy for the assignment of pseudorandom number streams proposed by Schruben and Margolin [26] and investigated how to select simulation designs so that bias due to possible model misspecification as well as error variance in first-order response surfaces could be reduced.

Another important topic is about designs for extrapolation that are insensitive to the possible bias from a higher-order model. This is particularly important in RSM since a response surface is often used for extrapolation purpose. Draper and Herzberg [27] studied a special type of extrapolation problem using "variance plus bias" methods. Later, the same authors [28] investigated if the region of extrapolation was a k-dimensional hyperspherical shell with inner radius one and outer radius Θ, how one could choose a design which would provide some protection against bias from a higher-order model and also would be suitable for extrapolation in all directions outside the k-dimensional hypersphere. In the last decade, researchers considered not only the model misspecification but also the heteroscedasticity in the errors. The details can be referred to the works, such as Wiens [29] and Fang [30].

2.2.3 Sequential Design and Bayesian Design

Sequential design is a very important and very effective approach. Within such a multiple stage design, additional experimental costs can be saved if no further experiments are needed and the experienced professional can modify the chosen design strategy at a certain stage. Thus this approach is very suitable for practical experiments and it often results in more efficient experiments. Sequential design can be dated back to Box and Wilson [15], who suggested that the central composite design be deployed sequentially, with the first stage being a 2-level factorial or fractional factorial design and the axial points forming a second stage. The axial points are used if curvature is found in the model by a lack-of-fit test. Some recent work also applies this sequential approach to screen factors and de-alias effects of potential interest. For example, Mee and Peralta [31] described semifolding, a technique using half of a standard fold-over design (see also Barnett et al [32]). Chipman and Hamada [33] advocated an effect-based approach and illustrated how the follow-up design selected depends on the family of models selected. Nelson et al. [34] compared augmentation strategies for both 2^{k-p} and Plackett-Burman designs.

When there are uncertainties in model selection and model parameters, Bayesian design might be necessary. DuMochel and Jones [35] assumed that there were two types of model terms, *certain* terms and *potential* terms, and set a prior distribution on the potential terms. Then they proposed a Bayesian D-optimal design that maximizes the determinant of the posterior information matrix. For the case of uncertainties in model parameters, Chanoler and Verdinelli [36] reviewed the Bayesian approach to design. Lin, Myers and Ye [37] utilized a two-stage approach to Bayesian design where the prior information was updated at the completion of the first stage.

2.3 Optimal Computing Budget Allocation

A technique known as the Optimal Computing Budget Allocation (OCBA) that tackles the computing budget allocation problem has been done within the context of Ranking and Selection (R&S).

The R&S procedures are developed to select the best system or a subset that contains the best system from a set of k competing alternatives (Goldsman and Nelson [38]). When the goal of the simulation study is to select the best system design from a finite set of competing alternatives, R&S procedures become applicable.

Generally, R&S defines selecting the best system or a subset that contains the best system as the correct selection. Chen [39] and Chen [40] proposed a technique to make a correct selection using a multistage approach and allocating the simulation runs in an optimal manner. Later they called it the Optimal Computing Budget Allocation, in which clearly inferior designs were identified and discarded in the earlier stage of sampling, and then those alternatives that might increase the probability of correct selection would be allocated with incremental computing budget. Chen et al. [41] extended this work by presenting a different method for estimating gradient information, and they also discussed how to choose the initial simulation replication number n_0 and one-time incremental computing budget Δ. Chen et al. [42] reported a further extension of this work that accounted for simulation experiments with different sampling costs. Through numerical experimentation, they observed this approach to be more efficient than the method discussed in Chen et al. [41]. Chen et al. [43] [44] offered an asymptotic allocation rule to enhance the efficiency of their allocation scheme.

Chapter 3

Two-stage Computing Budget Allocation Approach

Since the ASA technique is used as the fundamental technique to conduct the hill-climbing in this thesis, we first introduce it in section 3.1. Then we develop our two-stage computing budget allocation approach for the ASA technique in section 3.2. In order to justify the advantages of our two-stage computing budget allocation approach, a pilot study for a two-dimensional linear model is done in section 3.3.

We highlight the assumptions in this thesis as follows:

1. The cost to conduct one simulation run at any design point is similar, and we can thus measure the computing budget in terms of the number of replications;

2. The noise of each simulation replication all follows independent and identically distributed (i.i.d.) normal distribution with zero mean and constant variance;

3. For each two-level factorial design in the region of experimentation, the design points are fixed;

4. The size of the region of experimentation is the same for all the iterations;

5. In each iteration of hill-climbing, the total number of replications is the same.

3.1 Adapted Steepest Ascent Technique

Our two-stage computing budget allocation approach is mainly motivated by the Adapted Steepest Ascent (ASA) technique proposed by Kleijnen et al. [1]. We explain this technique first, and then propose our two-stage computing budget allocation approach in the next section.

The work of Kleijnen et al. [1] focuses on the early stages of Response Surface Methodology (RSM), in which RSM locally fits a first-order polynomial and the steepest ascent (SA) path is estimated by this polynomial. However, SA suffers from two well-known problems: (i) the search direction is scale-dependent; (ii) the step size along its path is selected intuitively (see Myers and Montgomery [3]). To tackle these two problems, Kleijnen et al. [1] derive the adapted steepest ascent (ASA) technique which is scale-independent, and mathematically obtain a step size in the ASA direction.

The local first-order polynomial approximation is given as:

$$y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \cdots + \beta_k d_k + \epsilon, \tag{3.1}$$

where y is the response or the observation of simulation run, β_i is the unknown coefficient, d_i is the regressor variable or the controllable factor, and ϵ is *white noise*, i.e. ϵ is normally, identically, and independently distributed with zero mean and constant variance σ^2.

Define the design point $\mathbf{d} = (d_1 \quad d_2 \quad \cdots \quad d_k)^T$, vector $\mathbf{x}^T = (1 \quad \mathbf{d}^T)$, vector $\boldsymbol{\beta} = (\beta_0 \quad \beta_1 \quad \cdots \quad \beta_k)^T$, and the model can then be written in the matrix form. Ordinary least squares (OLS) is a normal approach to estimate the coefficients $\beta_i's$, and the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{3.2}$$

with

d vector with the k regressor variables in the regression model

q number of regressor variables including the intercept β_0 ($q = 1 + k$ in Equation 3.1)

$\mathbf{x}$ vector with the q regressor variables including the 'dummy' variable d_0 with constant value 1

$\hat{\boldsymbol{\beta}}$ vector with the q estimated coefficients in the regression model

$\mathbf{X}$ the design matrix, an N$\times q$ matrix of independent regressor variables including the 'dummy' variable d_0; $\mathbf{X}$ is assumed to have linearly independent columns so $\mathbf{X}$ has full column rank

$\mathbf{Y}$ N$\times$1 vector, including all the observations of simulation runs

N $\sum_{i=1}^{m} n_i$: total number of replications in the simulation runs

n_i number of replications at input combination or point i. If $n_1 = n_2 = \cdots = n_m$, we refer to this special allocation as *equal allocation*

m number of different design points in the region of experimentation. If we consider a full factorial design, then $m = 2^k$.

The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} a & \mathbf{b}^T \\ \mathbf{b} & \mathbf{C} \end{pmatrix}, \tag{3.3}$$

where a is a scalar, $\mathbf{b}$ is a k-dimensional vector, and $\mathbf{C}$ is a $k \times k$ matrix.

The unknown parameter σ^2 can be estimated through the *mean squared residual* (MSR) or the *mean squared pure error* (MSPE) (see Myers and Montgomery [3]).

The predicted response $\hat{y}$ at point $\mathbf{d}$ is

$$\hat{y}(\mathbf{d}) = \begin{pmatrix} 1 & \mathbf{d}^T \end{pmatrix} \hat{\boldsymbol{\beta}}, \tag{3.4}$$

and the variance of this predictor is

$$var(\hat{y}|\mathbf{d}) = \begin{pmatrix} 1 & \mathbf{d}^T \end{pmatrix} cov(\hat{\boldsymbol{\beta}}) \begin{pmatrix} 1 \\ \mathbf{d} \end{pmatrix}, \tag{3.5}$$

Kleijnen et al. [1] proved that given a design point $\mathbf{d}$, the lower bound of one-sided $1-\alpha$ confidence interval for $\hat{y}$

$$\begin{aligned} \hat{y}_{min}(\mathbf{d}) &= \hat{y}(\mathbf{d}) - t^{\alpha}_{N-q}\sqrt{var(\hat{y}|\mathbf{d})} \\ &= \begin{pmatrix} 1 & \mathbf{d}^T \end{pmatrix} \hat{\boldsymbol{\beta}} - t^{\alpha}_{N-q}\sqrt{\begin{pmatrix} 1 & \mathbf{d} \end{pmatrix}^T (\mathbf{X}^T\mathbf{X})^{-1} \begin{pmatrix} 1 \\ \mathbf{d} \end{pmatrix} \cdot \hat{\sigma}^2}, \end{aligned} \tag{3.6}$$

is a concave function in $\mathbf{d}$, where t^{α}_{N-q} denotes the $1-\alpha$ quantile of the t distribution with $N-q$ degrees of freedom, and $\hat{\sigma}^2$ denotes the estimate of constant variance σ^2. The point $\mathbf{d}^+$ that maximizes the minimum output $\hat{y}_{min}(\mathbf{d})$ can be obtained easily by solving $\hat{y}'_{min}(\mathbf{d}) = 0$, and it is given by

$$\mathbf{d}^+ = -\mathbf{C}^{-1}\mathbf{b} + \lambda\mathbf{C}^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{-0}}, \tag{3.7}$$

where $-\mathbf{C}^{-1}\mathbf{b}$ is the *starting point* in the region of experimentation, $\mathbf{C}^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{-0}}$ is the *Adapted Steepest Ascent* (ASA) direction ($\hat{\boldsymbol{\beta}}_{\mathbf{-0}}$ equals $\hat{\boldsymbol{\beta}}$ excluding the intercept $\hat{\beta}_0$), and λ is the *step size* specified by

$$\lambda = \sqrt{\frac{a - \mathbf{b}^T\mathbf{C}^{-1}\mathbf{b}}{(t^{\alpha}_{N-q})^2\hat{\sigma}^2 - \hat{\boldsymbol{\beta}}_{\mathbf{-0}}^{T}\mathbf{C}^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{-0}}}}, \tag{3.8}$$

(see also Kleijnen et al. [1]).

The maximal point $\mathbf{d}^+$ gives both a search direction and a step size, and Kleijnen et al. [1] refer to it as the optimal input values of the next design point. Kleijnen et al. [1] prove that $\mathbf{d}^+$ is scale-independent attributed to the identical lower bound surfaces of predicted response in different scale systems. They also demonstrate the superiority of ASA compared to SA through Monte Carlo experiments. They first define a truly optimal search direction which is the vector starting at the initial design point and ending at the true optimum, and they apply SA technique and ASA technique to the same Input/Output (I/O) data. Then they find that the angle between the true search direction and ASA search direction is significantly smaller than the one between the true search direction and SA search direction.

The ASA technique offers a scale-independent next design point $\mathbf{d}^+$, and its search direction is superior to the traditional SA direction. However, when Kleijnen et al. [1]

propose the ASA technique, they assume the design matrix $\mathbf{X}$ is fixed, and their only concern is the point $\mathbf{d}^+$ that maximized the lower bound surface of predicted response. They do not consider the situation when there are computing budget constraints, and also the issue of experimental design.

3.2 The Algorithm of Two-stage Computing Budget Allocation

Here we apply the ASA technique to the simulation optimization problems, and consider a more realistic case where there is a limited computing budget. Then we develop the two-stage computing budget allocation approach.

Since we consider our approach in the simulation optimization scenario, it is reasonable to measure the computing budget in terms of the number of simulation replications. We define a *feasible region* for the input parameters of the simulation models, in which the simulation model is well defined. In most simulation studies, the inputs of a simulation model is valid only within a certain region. Outside this region, either the simulation model becomes invalid or there does not exist such a system in the real world. For example, the capacity of an inventory system must be finite, or the service time of a queueing model must be positive, etc.

Given the same estimates of $\boldsymbol{\beta}$ and σ^2, we note that in Equation 3.6 the value of $\hat{y}_{min}$ at a given design point $\mathbf{d}$ is determined by the design matrix $\mathbf{X}$, and the maximal point $\mathbf{d}^+$ of $\hat{y}_{min}$ in Equation 3.7 is also determined by the design matrix $\mathbf{X}$ through $\mathbf{b}$ and $\mathbf{C}$ in Equation 3.3. Since we consider the case where the design points in the region of experimentation are fixed, the different allocations $n_1, n_2, \cdots, n_m$ can construct different design matrixes, and this results in different values of $\mathbf{d}^+$ and $\hat{y}_{min}(\mathbf{d}^+)$. Thus after we have the estimation of $\boldsymbol{\beta}$ and σ^2 it is possible for us to compare all the allocations of $n_1, n_2, \cdots, n_m$ and pick the one that gives the maximal value of $\hat{y}_{min}(\mathbf{d}^+)$. Here we consider maximizing the lower bound of the predicted response as a criterion for comparison. The idea is that when the lower bound $\hat{y}_{min}$ is maximized at the next design point $\mathbf{d}^+$, the expected response $\hat{y}$ is likely to be better improved when the region of experimentation moves from the original region to the next one as $\hat{y}_{min}$ is the worst prediction of $\hat{y}$ at level α.

Let $\hat{y}_{min}(\mathbf{d}|n_1, n_2, \cdots, n_m)$ denote the lower bound of predicted response $\hat{y}_{min}$ at design point $\mathbf{d}$ with the allocation of $n_1, n_2, \cdots, n_m$. Then the computing budget allocation problem in every iteration of hill-climbing can be rewritten as

$$\begin{array}{rc} \text{(P)} & \max\limits_{n_1, n_2, \cdots, n_m} F(n_1, n_2, \cdots, n_m) \\ \text{with} & n_1 + n_2 + \cdots + n_m = N \\ \text{and} & F(n_1, n_2, \cdots, n_m) = \max\limits_{\mathbf{d}^+}(\hat{y}_{min}(\mathbf{d}^+|n_1, n_2, \cdots, n_m)) \end{array}$$

where $n_1, n_2, \cdots, n_m$ are the decision variables, the values of $\mathbf{d}^+$ and $\hat{y}_{min}(\mathbf{d}^+)$ are determined by the allocation of $n_1, n_2, \cdots, n_m$, and $F(n_1, n_2, \cdots, n_m)$ denotes the lower bound of predicted response $\hat{y}_{min}$ at its maximal point given the allocation of $n_1, n_2, \cdots, n_m$. The value of $F(n_1, n_2, \cdots, n_m)$ can be computed using Equations 3.6 and 3.7, where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are assumed to be known and $\mathbf{X}$ solely depends on $n_1, n_2, \cdots, n_m$ in the local region of experimentation.

The above discussion is based on the assumption that we have the estimation of $\boldsymbol{\beta}$ and σ^2, and therefore the major issue in addressing the problem (P) is before we determine the best allocation of $n_1, n_2, \cdots, n_m$, how we can estimate $\boldsymbol{\beta}$ and σ^2.

To tackle this issue, we propose a two-stage computing budget allocation approach.

Stage 1. A two-level factorial design is used, which equally distributes n_0 runs to all the design points in the region of experimentation. $\boldsymbol{\beta}$ and σ^2 are estimated using these n_0 observations;

Stage 2. We compare all the allocations for the rest of $N - n_0$ runs and pick the best one that gives the maximal value of $F(n_1, n_2, \cdots, n_m)$

After the allocation problem in stage 2 is settled, we will estimate the linear response function again using all the N observations, and the value of the next design point will be updated to be $\mathbf{d}^+$. For the next iteration, $\mathbf{d}^+$ is the center of the region of experimentation.

Intuitively, stage 1 helps to roughly estimate the true response function. In stage 2, given the estimation of $\boldsymbol{\beta}$ and σ^2 in stage 1, we can compare all the allocations and select the best one with the maximal value of $F(n_1, n_2, \cdots, n_m)$. Since $F(n_1, n_2, \cdots, n_m)$ is the worst prediction of expected response at the next design point at level α, if its value is improved, then when we move to the next design point, we can be more assured that the response at the next design point is improved. After we have all N observations, $\mathbf{d}^+$ is the point that maximizes the value of lower bound function $\hat{y}_{min}$, and therefore $\mathbf{d}^+$ becomes the center point of the region of experimentation in the next iteration.

We can also use Figure 1.1 to illustrate the basic idea of our two-stage approach. In the first iteration, point $\mathbf{A}$ is the center of the region of experimentation, and the four design points $\mathbf{d_1}$, $\mathbf{d_2}$, $\mathbf{d_3}$ and $\mathbf{d_4}$ of a 2^2 factorial design are determined. In stage 1, n_0 runs are equally distributed to these four design points. β_0, β_1, β_2 and σ^2 are estimated by these n_0 observations. In stage 2, we generate all the possible allocations of n_1, n_2, n_3, n_4 where $n_1 + n_2 + n_3 + n_4 = N - n_0$. For each allocation, we construct the design matrix and use the estimates of β_i and σ^2 in stage 1 to compute the values of $F(n_1, n_2, n_3, n_4)$. Then we select the best allocation which gives the maximal value of $F(n_1, n_2, n_3, n_4)$. Next we distribute the rest of $N - n_0$ runs according to this best allocation, run the simulations and compute the value of $\mathbf{d}^+$ based on all the N observations. Suppose the value of $\mathbf{d}^+$ is point $\mathbf{B}$, then we move to point $\mathbf{B}$ in the second iteration and make it as the center of the new region of experimentation. In the following iterations, the same procedure will be repeated until a terminating condition is met.

In the study of Kleijnen et al. [1], the authors assume the design matrix is given, the maximal point of $\mathbf{d}^+$ has offered an ASA search direction and a possible step size, and the ASA direction is shown to be superior to the traditional SA direction. In our approach, their design matrix is only one option to determine the allocation of $n_1, n_2, \cdots, n_m$, and thus we can expect our approach to further improve their ASA technique.

In this two-stage allocation, because we are trying to find the allocation with the maximal value of $F(n_1, n_2, \cdots, n_m)$, the variance of the predicted response $\hat{y}$ throughout the region of interest may suffer and the predicted response $\hat{y}$ in certain regions becomes inaccurate, which is not desired by good response surface designs. However, we focus on the early stages of RSM, whose main objective is to find the most effective path to improve the response. In our approach, we find the next design point to improve the response directly by Equation 3.7, and allocate the replications to improve the worst prediction of expected response at level α. Thus our approach is much more conservative. Moreover, the numerical results in Chapter 4 show that our two-stage approach performs better than the traditional approach.

In addition, according to the results of Kleijnen et al. [1], the next design point of the ASA technique might be at infinity in certain situations which makes $F(n_1, n_2, \cdots, n_m)$, the maximal value of the lower bound of predicted response, infinite also. In this case we are not able to identify the allocation that maximizes $F(n_1, n_2, \cdots, n_m)$. However, we focus on the early stages of RSM where the main objective is to find the most effective path to improve the response. We see from Equation 3.7 that since $\mathbf{b}$, $\mathbf{C}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{-0}}$ are determined from the design matrix $\mathbf{X}$, these values and $\mathbf{Y}$ are always finite. Thus the infinite next design point can only be due to the infinite step size λ. In this case, we propose an asymptotic approach to decide the allocation. We choose a design point $\mathbf{d}^*$ in the ASA direction, and study the behavior of $\hat{y}_{min}(\mathbf{d}^*)$ when the step size of $\mathbf{d}^*$ approaches infinity.

Let $\mathbf{d}^* = -\mathbf{C}^{-1}\mathbf{b} + \lambda^*\mathbf{C}^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{-0}}$, where $\mathbf{C}$, $\mathbf{b}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{-0}}$ are the same as in Equation 3.7, and the step size λ^* is fixed. Substitute $\mathbf{d}^*$ into $\hat{y}_{min}$ in Equation 3.6, we will get

$$\hat{y}_{min}(\mathbf{d}^*) = \hat{\beta}_0 - \hat{\boldsymbol{\beta}}_{\mathbf{-0}}^T\mathbf{C}^{-1}\mathbf{b} + \lambda^*\hat{\boldsymbol{\beta}}_{\mathbf{-0}}^T\mathbf{C}\hat{\boldsymbol{\beta}}_{\mathbf{-0}} - t^{\alpha}_{N-q}\sqrt{(a - \mathbf{b}^T\mathbf{C}^{-1}\mathbf{b} + \lambda^{*2}\hat{\boldsymbol{\beta}}_{\mathbf{-0}}^T\mathbf{C}^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{-0}})\hat{\sigma}^2}. \tag{3.9}$$

When the step size approaches infinity ($\lambda^* \to \infty$), the ratio of $\hat{y}_{min}(\mathbf{d}^*)$ and λ^* is

$$\lim_{\lambda^* \to \infty} \frac{\hat{y}_{min}(\mathbf{d}^*)}{\lambda^*} = \hat{\boldsymbol{\beta}}_{\mathbf{-0}}^T\mathbf{C}^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{-0}} - t^{\alpha}_{N-q}\sqrt{\hat{\boldsymbol{\beta}}_{\mathbf{-0}}^T\mathbf{C}^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{-0}} \cdot \hat{\sigma}^2}. \tag{3.10}$$

This ratio depends on the allocation as $\mathbf{C}$ is determined by $n_1, n_2, \cdots, n_m$ through the design matrix $\mathbf{X}$ in Equation 3.3, and thus we may choose a particular allocation to maximize this ratio.

In our two-stage approach, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are estimated in stage 1, and in stage 2 we can identify an allocation with the maximal value in Equation 3.10. We define this allocation as the dominating allocation as it makes the value of $\hat{y}_{min}$, the lower bound of predicted response, always larger or equal to the value of $\hat{y}_{min}$ of other allocations when the design point approaches infinity in the ASA direction. The idea of dominating allocation is consistent with our original idea to find the allocation that gives a better value of the lower bound of predicted response.

As a result, in the case that the next design point is determined to be at infinity, we formulate the computing budget allocation problem as

$$\begin{array}{lc} \text{(P1)} & \max\limits_{n_1,n_2,\cdots,n_m} \left(\hat{\boldsymbol{\beta}}_{\mathbf{-0}}{}^T \mathbf{C}^{-1} \hat{\boldsymbol{\beta}}_{\mathbf{-0}} - t^{\alpha}_{N-q} \sqrt{\hat{\boldsymbol{\beta}}_{\mathbf{-0}}{}^T \mathbf{C}^{-1} \hat{\boldsymbol{\beta}}_{\mathbf{-0}} \cdot \hat{\sigma}^2} \right) \\ \text{with} & n_1 + n_2 + \cdots + n_m = N \end{array}$$

where $\mathbf{C}$ is determined by $n_1, n_2, \cdots, n_m$ through design matrix $\mathbf{X}$, and $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are the estimates of $\boldsymbol{\beta}$ and σ^2.

If some allocations make the value of $F(n_1, n_2, \cdots, n_m)$ infinite in stage 2, we will allocate the rest of runs according to the dominating allocation. If the value of next design point $\mathbf{d}^+$ is determined to be at infinity or outside the feasible region after stage 2, we will set the intersection of the boundary of feasible region and the ASA direction as the next design point.

Our suggested procedure to conduct hill-climbing in the early stages of RSM with computing budget constraints is summarized as follows:

Step 1. A two-level factorial design is used, which equally distributes n_0 runs to all the design points. $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are estimated using these n_0 observations;

Step 2. Identify the best allocation for the rest of runs using brute force search;

Step 2.1 Generate all possible allocations of $\{n_1, n_2, \cdots, n_{m-1}, n_m\}$ for the remaining $N - n_0$ runs: $\boldsymbol{\Lambda} = \{\{0, 0, \cdots, 0, N - n_0\}, \{0, 0, \cdots, 1, N - n_0 - 1\}, \cdots, \{N - n_0, 0, \cdots, 0, 0\}\}$;

Step 2.2 Select an allocation from $\mathbf{\Lambda}$ and reconstruct the design matrix $\mathbf{X}$ using the total of N replications;

Step 2.3 Compute a, $\mathbf{b}$, $\mathbf{C}$ from $(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} a & \mathbf{b}^T \\ \mathbf{b} & \mathbf{C} \end{pmatrix}$;

Step 2.4 Use $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$ and current allocation to compute $F(n_1, n_2, \cdots, n_m)$.

Step 2.5 If the value of $F(n_1, n_2, \cdots, n_m)$ is infinite, add this allocation as a potential dominating allocation into the variable set 'n_temp'; otherwise, compare the value of $F(n_1, n_2, \cdots, n_m)$ with the previous maximal value of $F(n_1, n_2, \cdots, n_m)$. If the new $F(n_1, n_2, \cdots, n_m)$ is larger than the previous maximal value, replace the previous maximal value with the new $F(n_1, n_2, \cdots, n_m)$ and record the current allocation as the best allocation;

Step 2.6 If all the possible allocations in $\mathbf{\Lambda}$ have been searched, go to step 2.7; otherwise go to step 2.2 and select another allocation from $\mathbf{\Lambda}$ which is not previously selected;

Step 2.7 If 'n_temp' is not empty, use brute force search for all the allocations in 'n_temp' and choose the dominating allocation which maximizes Equation 3.10 from 'n_temp' as the best allocation;

Step 2.8 Return the best allocation;

Step 3. Conduct the rest of runs, estimate the true response function using all the observations, and compute the next design point $\mathbf{d}^+$. If an infinite $\mathbf{d}^+$ is identified or $\mathbf{d}^+$ is outside the feasible region, set the intersection of the boundary of feasible region and the ASA direction as the next design point;

Step 4. If the terminating condition is met, stop the procedure. Otherwise, move to the next design point, make it as the center of the two-level factorial design and go back to step 1.

Here, the terminating condition will be the maximal number of iterations is reached or the response at the final design point is within certain percentage of the true optimum, which will be discussed later in this thesis.

The flowchart for step 2, i.e. identifying the best allocation for the rest of runs using brute force search, is given in figure 3.1 at the end of this chapter.

A special case for our two-stage computing budget allocation approach is the one-dimensional model because there is only one dimension in the search direction. If the next design point of the ASA technique is at infinity, given a finite number of runs, we can find a unique dominating allocation, and this dominating allocation is the equal allocation. We prove this in Appendix C. Correspondingly, the flowchart to identify the best allocation for a one-dimensional model is given in figure 3.2, and the procedure to conduct hill-climbing with budget constraints is similar.

3.3 Two-dimensional Linear Model

Since our approach locally approximates the true response function using a linear model, we first apply our two-stage computing budget allocation approach to a linear model to validate its advantages. In this case, the effect of model misspecification is removed. In Chapter 4, we will test our approach for nonlinear models.

As the optimal point for a linear model is at infinity, we consider the infinite feasible input region. However, in order to compare the different allocation schemes, we consider only the cases where the noise is large enough for the ASA technique to yield a finite next design point. When the noise is too small, the next design point is infinite and no allocation comparisons can be made. In Equation 3.8, we can always choose a large σ^2 so that the step size λ is finite for the next design point $\mathbf{d}^+$.

The main concern here is the performance of different computing budget allocation schemes. We compare the traditional 2^k factorial design which allocates runs equally to each design point to our two-stage allocation. Both allocation schemes will use the ASA technique to determine the next design point. We also fix some general settings such as the true response function, the initial design point and the size of region of experimentation to make the testing conditions homogeneous.

Here we summarize the general settings for the numerical experiment:

(1). the true response function: $y = 0.5 + 0.25 \cdot d_1 + 0.25 \cdot d_2 + \epsilon$, where ϵ follows i.i.d. $N(0, \sigma^2)$;

(2). the initial point: $(0, 0)$, which is the initial level of input variables (d_1,d_2) for all the experiments;

(3). the length of the region of experimentation: $l = 2$. In the two-dimensional model, the four design points of 2^2 design are $(d_1^* - l/2, d_2^* - l/2)$, $(d_1^* - l/2, d_2^* + l/2)$, $(d_1^* + l/2, d_2^* - l/2)$ and $(d_1^* + l/2, d_2^* + l/2)$, where (d_1^*, d_2^*) is the center point of current region of experimentation;

(4). the standard deviation of noise: $\sigma = 50$. $\sigma = 50$ is a large noise compared to the coefficients of the true response function;

(5). $\alpha = 0.05$ which is used to determine the t-value when computing the step size λ in Equation 3.8;

and the adjustable factors:

(1). the number of iterations (N.O.I.): as the design moves from the original region to the future design point, we say the design moves one step. The number of iterations indicates how many steps the design moves;

(2). the total number of runs in each iteration (N), or the number of runs in stage 1 v.s. the number of runs in stage 2 $\{n_0, N - n_0\}$;

In simulation optimization study, the key output of interest is the final expected response after using up all the budget. Therefore we compare the final expected responses of two-stage allocation and the final expected responses of traditional equal allocation at the end of all the iterations.

The following hypothesis for each setting is tested at level 0.05:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ v.s. } H_1 : \mu_1 - \mu_2 > 0$$

where μ_1 is the average response at the final points for two-stage allocation and μ_2 is the average response at the final points for equal allocation. To compute μ_1 and μ_2, we replicate the experiment and obtain 100 final responses for both two-stage allocation and equal allocation. The final responses are computed from the true response function at the final points that each allocation scheme obtains. Similar as the study of Lawson, Keats and Montgomery [45], we assume that the outliers in our study are the samples which fall outside the 3% tail area of the sampling distribution, which means we treat the largest three and the smallest three samples as the outliers and delete them before we proceed to the statistical analysis. In this study, we use the two-sample t-test. Since we do not assume equal variance for this test, the degrees of freedom will be determined by not only the sample size but also the variances of the two samples (see Devore [46]). If the hypothesis H_0 is finally rejected, then we can conclude that the observations strongly suggest that the two-stage allocation improves the response much faster than the equal allocation.

We first investigate whether the two-stage allocation can improve significantly the expected response over the traditional equal allocation when the true response is used to compute $F(n_1, n_2, \cdots, n_m)$ (i.e. the true response function is known when we search for the best allocation, in step 1 and step 2). Since we use the true response function to compute $F(n_1, n_2, \cdots, n_m)$, there is no estimation error. The best allocation determined by the two-stage allocation is the theoretically optimal solution for problem **(P)**. Experiments are done based on this theoretically optimal allocation, the true response function is estimated and then the next design point is computed (i.e. in step 3). When we move to the next region of experimentation, the best allocation is always determined by the true response function. In this case, the two-stage allocation should be better than the equal allocation. If two-stage allocation does not perform significantly better we would expect it not to work well when the true response function is estimated. We give the numerical results for two-dimensional known models in section 3.3.1. Next we test its performance for fixed but unknown linear models since in reality most of the response functions are unknown and need to be estimated. The numerical results are given in section 3.3.2 for the two-dimensional unknown model.

3.3.1 Computing Budget Allocation with Known Model

Here we assume the true response function is known. In step 2.4 of the iterative procedure, we will compute the next design point based on the true $\boldsymbol{\beta}$ and σ^2 and then determine the best allocation. We run different settings to test the performance of the two-stage allocation, and the different levels of the adjustable factors are:

(1). the number of iterations (N.O.I.): 1 or 5. When N.O.I.=1, 100% of the experiments get finite $\mathbf{d}^+$. After 5 iterations (N.O.I.=5), about 30% of the experiments get infinite $\mathbf{d}^+$. Although we have set the noise to be large, the step size in Equation 3.8 is a random variable, and there is a nonzero probability that it gets an infinite value. We compute $\mathbf{d}^+$ in each iteration, and the more iterations we have, the larger the probability that the next design point is determined to be at infinity. Since we drop all the infinite observations in this study, hence if many experiments fail to get finite $\mathbf{d}^+$ at the end of all iterations, the comparison is biased because the sample sizes of μ_1 and μ_2 are different and we only keep those biased samples with smaller step sizes. To prevent our results from having such a large bias, we fix the maximal N.O.I. to be 5.

(2). the total number of runs in each iteration (N) : 20, or 40, or 80.

Table 3.1 shows the results for varying N.O.I., N, μ_1, μ_2, σ_1 - the standard deviation of μ_1 over 94 samples, σ_2 - the standard deviation of μ_2 over 94 samples, and gives the P-value of the hypothesis test. DOF is the degrees of freedom for the two-sample t-test with unequal variance. Here, DOF $= \frac{(\frac{\sigma_1^2}{94}+\frac{\sigma_2^2}{94})^2}{\frac{(\sigma_1^2/94)^2}{93}+\frac{(\sigma_2^2/94)^2}{93}}$. If the value of DOF is not an integer, it will be rounded down to the nearest integer (see page 366 of Devore [46]).

Table 3.1: The comparison of final responses for a known two-dimensional linear model

N.O.I.	σ	N	Two-stage Allocation		Equal Allocation		P-value	DOF
			μ_1	σ_1	μ_2	σ_2		
1	50	20	0.873	0.142	0.482	0.230	0.000	154
1	50	40	0.932	0.126	0.444	0.306	0.000	123
1	50	80	0.978	0.088	0.529	0.212	0.000	123
5	50	20	2.515	0.496	0.439	0.736	0.000	162
5	50	40	2.741	0.300	0.565	0.671	0.000	128
5	50	80	2.840	0.272	0.550	0.588	0.000	131

From table 3.1 we can observe that:

1. Two-stage allocation always gives higher final responses ($\mu_1 > \mu_2$), and the final design points of equal allocation remain close to the initial point (the initial point is $(0,0)$, the response at $(0,0)$ is 0.5 and μ_2 is always around 0.5).

2. For this particular model, $N = 40$ seems good. The difference between the final responses of $N = 40$ and the final responses of $N = 80$ is less than 10% while $N = 40$ only costs half of the runs for $N = 80$.

3. The variability of the final responses for the equal allocation is much larger than the variability of two-stage allocation ($\sigma_2 > \sigma_1$).

For observation 1, when the noise σ^2 is large ($\sigma^2 = 2500$ while the responses μ_1 and μ_2 are less than 3), the step size of the ASA technique in Equation 3.8 is very small, which makes the next design point $\mathbf{d}^+$ close to the starting point $-\mathbf{C}^{-1}\mathbf{b}$. However, the starting points $-\mathbf{C}^{-1}\mathbf{b}$ for equal allocation and two-stage allocation are quite different. For equal allocation, we show that the starting point is the center of the region of experimentation in Appendix B. For two-stage allocation, the starting point is close to the point with the largest response in the local region. First, when the next design point $\mathbf{d}^+$ is close to the starting point $-\mathbf{C}^{-1}\mathbf{b}$, the value of $\hat{y}(\mathbf{d}^+)$ is close to the value of $\hat{y}(-\mathbf{C}^{-1}\mathbf{b})$. Second, $\hat{y}_{min}(\mathbf{d}^{+})$ is likely to be large when $\hat{y}(\mathbf{d}^{+})$ is large. Therefore the value of $\hat{y}(-\mathbf{C}^{-1}\mathbf{b})$ is expected to be as large as possible so that $\hat{y}(\mathbf{d}^{+})$ would be large, and consequently $\hat{y}_{min}(\mathbf{d}^{+})$ would be large. However, the starting point $-\mathbf{C}^{-1}\mathbf{b}$ must be inside the region of experimentation because it is the point with minimal variance of predicted response (see Kleijnen et al. [1]), and therefore the 'ideal' starting point for two-stage allocation is the point with the largest response in the local region. Two-stage allocation will search for such an allocation that makes the starting point close to that 'ideal' point. In summary, for equal allocation, the small step size makes the next design point close to the center of region of experimentation, while for two-stage allocation, it makes the next design point close to the point with the largest response in the local region. After we move to the next design point and climb the response surface for several iterations, these two designs will show significant difference.

For observation 3, it implies that the performance of our two-stage computing budget allocation approach is much more stable. Since the true response function is linear, it also implies that the final design points of two-stage allocation are close to each other while the final design points of equal allocation are more widely spread out.

Since our two-stage allocation works well for the large noise case when the true response function is known, we apply it to the unknown case in section 3.3.2 to further show its advantages.

3.3.2 Two-stage Approach with Unknown Model

When the true response function is unknown, we will use the estimates of stage 1 as the true response function and determine the best allocation that makes the lower bound of predicted response maximized.

We run different settings to test the performance of the two-stage allocation, and the different levels of the adjustable factors are:

(1). the number of iterations (N.O.I.): 1 or 5.

(2). the total number of runs in stage 1 v.s. the total number of runs in stage 2, $\{n_0, N - n_0\}$: $\{12, 28\}$, $\{20, 20\}$, $\{28, 12\}$.

We choose 40 runs as the total number of runs in one iteration, N, based on the observations in section 3.3.1. The results of the experiment are given in table 3.2. The structure of this table is identical to table 3.1 except that $\{n_0, N - n_0\}$ replaces N.

Table 3.2: The comparison of final responses for an unknown two-dimensional linear model

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
1	50	12	28	0.615	0.184	0.521	0.236	0.001	175
1	50	20	20	0.583	0.150	0.521	0.236	0.016	157
1	50	28	12	0.536	0.172	0.521	0.236	0.314	169
5	50	12	28	0.973	0.410	0.502	0.735	0.000	145
5	50	20	20	0.790	0.346	0.502	0.735	0.000	132
5	50	28	12	0.610	0.418	0.502	0.735	0.109	147

From table 3.2, we can observe that

1. $\{12, 28\}$ seems to be the best combination among the three combinations of $\{n_0, N - n_0\}$. When noise is large in this case, the final responses of $\{12, 28\}$ are always the largest. $\{28, 12\}$ is the worst; there is no significant difference between $\{28, 12\}$ and the equal allocation.

2. The performance of two-stage allocation seems to be more robust than the equal allocation. When the setting is the same, the variability of final responses for two-stage allocation is always smaller than the equal allocation ($\sigma_1 < \sigma_2$).

For observation 1, it is not surprising that $\{28, 12\}$ performs similarly as equal allocation, because most of the runs (28 out of 40 runs) are equally distributed in stage 1. For observation 2, this may be due to the different step sizes of two-stage allocation and equal allocation.

We draw the following conclusions for this particular two-dimensional linear model:

When the true response function is known two-stage allocation is always the best. The final response is larger and the variability of final responses is smaller. This study helps us assure that the two-stage allocation will work. However, it is necessary to test the performance of the two-stage allocation for unknown response function since the true response function is always unknown.

When the true response function is unknown, our two-stage allocation outperforms the equal allocation when the noise is large. If we assign more runs in stage 1, the performance of two-stage allocation gets worse.

In this chapter, we introduce our two-stage computing budget allocation approach in detail. We also compare it to the traditional equal allocation for a linear model, and observe that our two-stage approach significantly outperforms the traditional equal allocation when the noise is large. In the following chapter, we will conduct more numerical studies to validate its advantages.

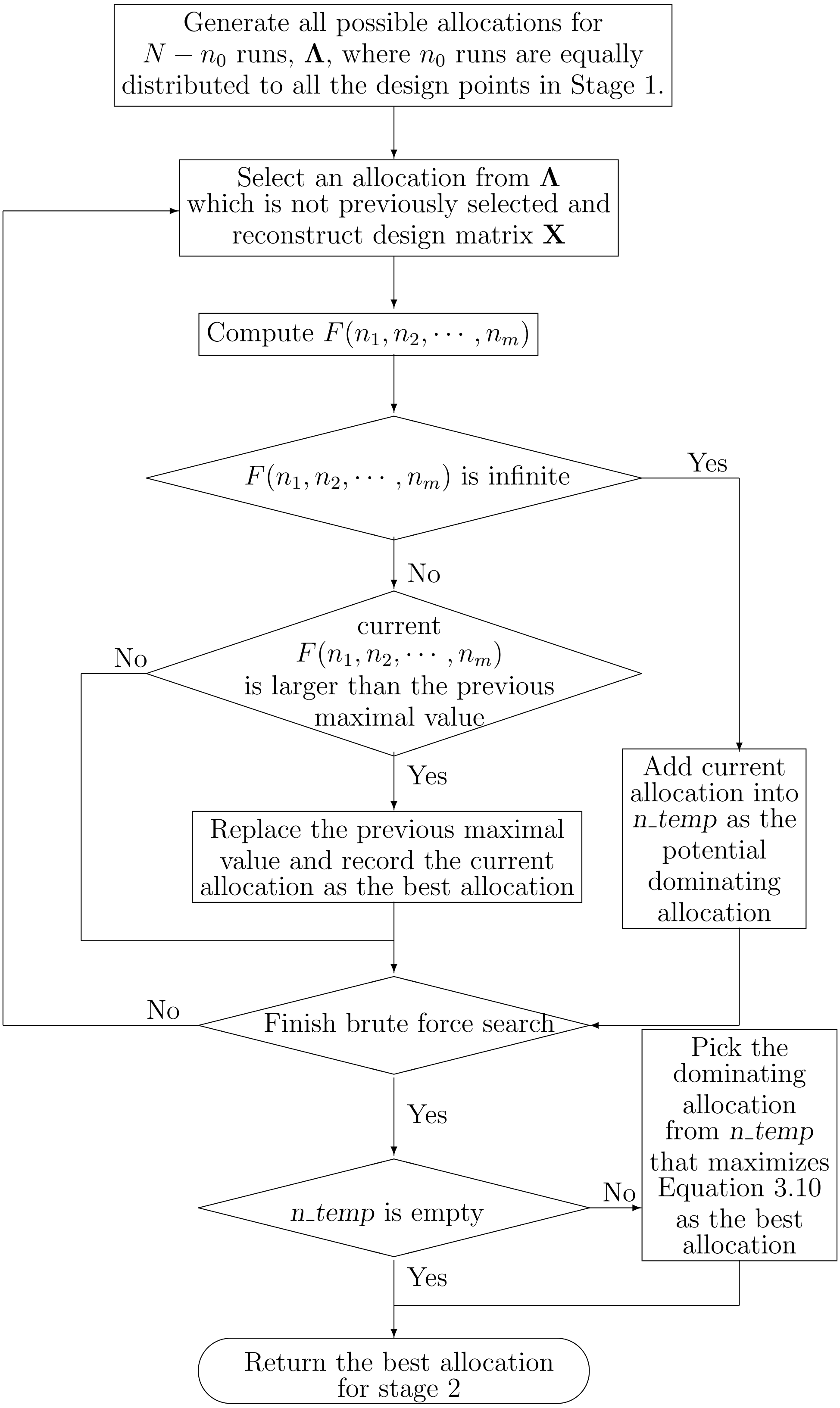

Generate all possible allocations for $N - n_0$ runs, $\mathbf{\Lambda}$, where n_0 runs are equally distributed to all the design points in Stage 1.
Select an allocation from $\mathbf{\Lambda}$ which is not previously selected and reconstruct design matrix $\mathbf{X}$
Compute $F(n_1, n_2, \cdots, n_m)$
$F(n_1, n_2, \cdots, n_m)$ is infinite
Yes
No
current $F(n_1, n_2, \cdots, n_m)$ is larger than the previous maximal value
No
Yes
Replace the previous maximal value and record the current allocation as the best allocation
Add current allocation into *n_temp* as the potential dominating allocation
Finish brute force search
No
Yes
n_temp is empty
No
Pick the dominating allocation from *n_temp* that maximizes Equation 3.10 as the best allocation
Yes
Return the best allocation for stage 2

Figure 3.1: The procedure to determine the best allocation for higher-dimensional model

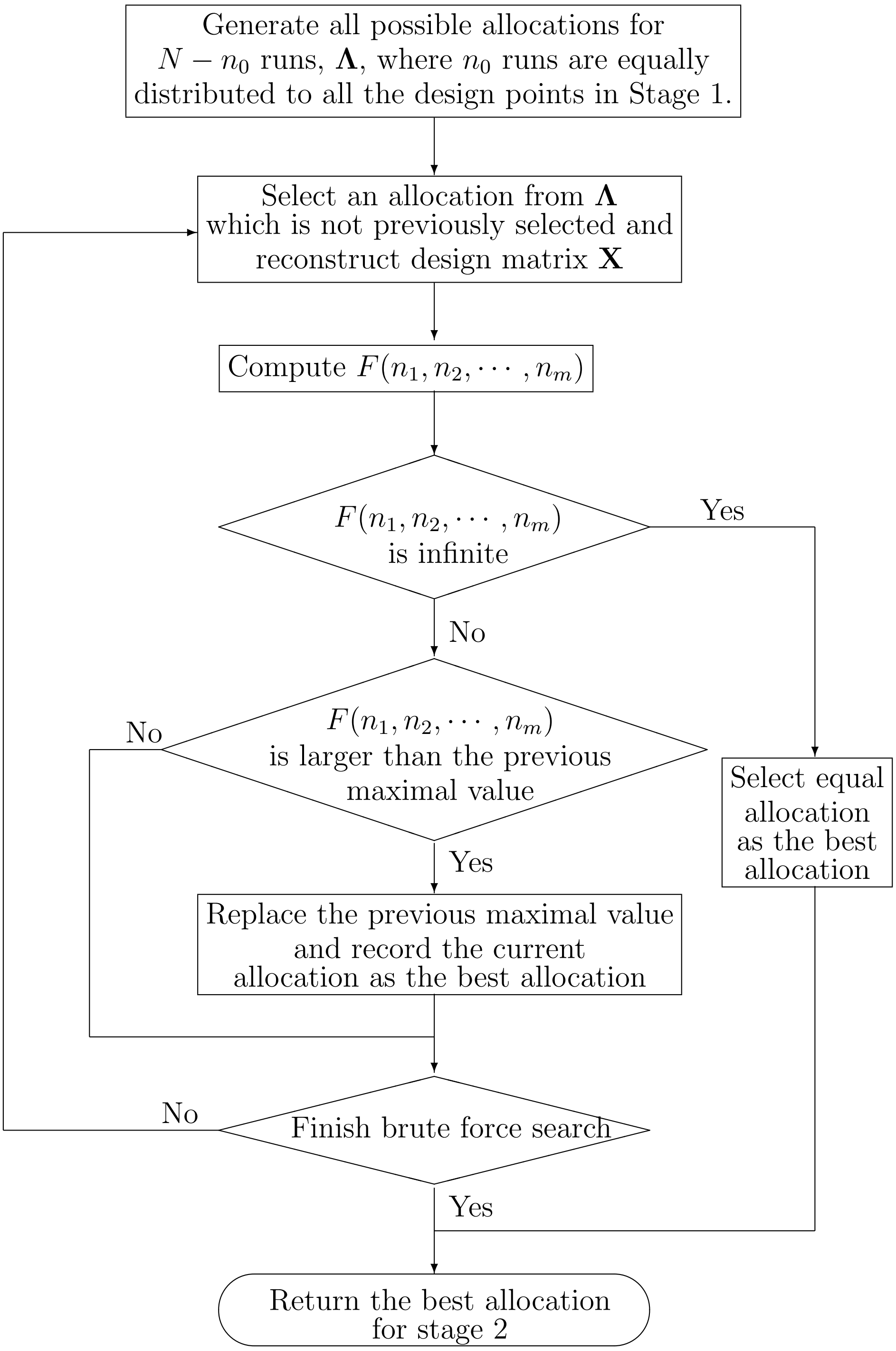

Generate all possible allocations for $N - n_0$ runs, $\mathbf{\Lambda}$, where n_0 runs are equally distributed to all the design points in Stage 1.
Select an allocation from $\mathbf{\Lambda}$ which is not previously selected and reconstruct design matrix $\mathbf{X}$
Compute $F(n_1, n_2, \cdots, n_m)$
$F(n_1, n_2, \cdots, n_m)$ is infinite
Yes
No
$F(n_1, n_2, \cdots, n_m)$ is larger than the previous maximal value
No
Yes
Select equal allocation as the best allocation
Replace the previous maximal value and record the current allocation as the best allocation
Finish brute force search
No
Yes
Return the best allocation for stage 2

Figure 3.2: The procedure to determine the best allocation for one-dimensional model

Chapter 4

Numerical Results for Nonlinear Model

It is important to investigate how our two-stage computing budget allocation approach performs with nonlinear models so that we can be more assured our approach works under different circumstances. This chapter lists all the experimental settings and results for the comparison of our two-stage allocation and traditional equal allocation. As there is no standard response function in the literature to test the different hill-climbing methods, we select several different shaped low order responses for the testing, such as $y = -\frac{(d-10)^2}{50} + 2$ and $y = e^{-\frac{(d-10)^2}{50}+2}$. In Kleijnen et al. [1], the authors also use an arbitrary second-order polynomial to compare their ASA technique with the SA technique.

As this thesis concentrates on the earlier stages of RSM, we use linear models to approximate the nonlinear response functions. Thus the main idea of two-stage computing budget allocation for nonlinear models is the same as the one in linear model case. The iterative procedure to conduct hill-climbing with computing budget constraints is also the same as stated in Chapter 3.

There is a feasible region for all the nonlinear models. Our approach does not intend to distinguish the local optimum, and therefore we consider the cases in which there is only one optimum in the feasible region. The general settings for numerical experiments are:

(1). the initial point: the origin;

(2). the length of the region of experimentation: $l = 2$;

(3). $\alpha = 0.05$ which is used to determine the t-value when computing the step size λ;

(4). the value of N: 20 for the one-dimensional models and 40 for the two-dimensional models. From table 3.1, we have found that $N = 40$ seems good for the two dimensional model when noise is large. On average there are 10 runs at each design point in one iteration, and this setting appears to work well for our numerical study. For the rest of experiments we will make this as the general setting. There are two design points for the one-dimensional models in one iteration, hence N equals 20 for the one-dimensional models.

For nonlinear cases, the key output of interest is still the final expected response after using up all the budget. In addition, we consider how many iterations each design needs to get to the true optimum. Since we consider only first-order model in this thesis, which will be inadequate when the design is close to the optimum, we define a neighborhood of the true optimum and compare how many iterations each design needs to get to this fixed neighborhood of the true optimum. In short, we consider:

Method 1. Which allocation obtains a better final response after using up the fixed budget.

Method 2. With an unlimited budget, which allocation requires less number of iterations to obtain a final response within $t\%$ of the true optimum.

For method 1, the adjustable factors are:

(1). the number of iterations (N.O.I.);

(2). the standard deviation of noise (σ);

(3). the total number of runs in stage 1 v.s. the total number of runs in stage 2.

For method 2, the adjustable factors are:

(1). the standard deviation of noise (σ);

(2). the total number of runs in stage 1 v.s. the total number of runs in stage 2.

We list all the experimental settings and the numerical results for one-dimensional models in section 4.1, and two-dimensional models in section 4.2. We can see that our two-stage computing budget allocation approach outperforms the traditional equal allocation approach when the system noise is large.

4.1 One-dimensional Nonlinear Model

We run different settings to test the performance of the two-stage allocation, and four different response surfaces are tested:

(1). **Model 1:** $y = -\dfrac{(d-10)^2}{50} + 2 + \epsilon = \dfrac{2}{5} \cdot d - \dfrac{1}{50} \cdot d^2 + \epsilon$

- the feasible region for d is $(-2, 22)$, the range of response is $(-0.88, 2)$, and the response at the starting point is 0.
- the gradient at the starting point is $\dfrac{2}{5}$, and it is continuously decreasing to 0 with the ratio $\frac{1}{25}$.

(2). **Model 2:** $y = -\dfrac{(d-10)^2}{20} + 5 + \epsilon = d - \dfrac{1}{20} \cdot d^2 + \epsilon$

- the feasible region for d is $(-2, 22)$, the range of response is $(-2.2, 5)$, and the response at the starting point is 0.
- the gradient at the starting point is 1, and it is continuously decreasing to 0 with the ratio $\frac{1}{10}$.

(3). **Model 3:** $y = e^{-\frac{(d-10)^2}{50}+2} + \epsilon = e^{\frac{2}{5}\cdot d - \frac{1}{50}\cdot d^2} + \epsilon$

- the feasible region for d is $(-2, 22)$, the range of response is $(0.415, 7.39)$, and the response at the starting point is 1.
- the gradient: 0.4 (starting point) → 0.896 (largest) → 0 (optimum).

(4). **Model 4:** $y = e^{-\frac{(d-10)^2}{20}+3} + \epsilon = e^{-2+d-\frac{1}{20}\cdot d^2} + \epsilon$

- the feasible region for d is $(-2, 22)$, the range of response is $(0.015, 20.09)$, and the response at the starting point is 0.14.
- the gradient: 0.14 (starting point) → 3.85 (largest) → 0 (optimum).

The response surfaces for these four models are given in figure 4.1.

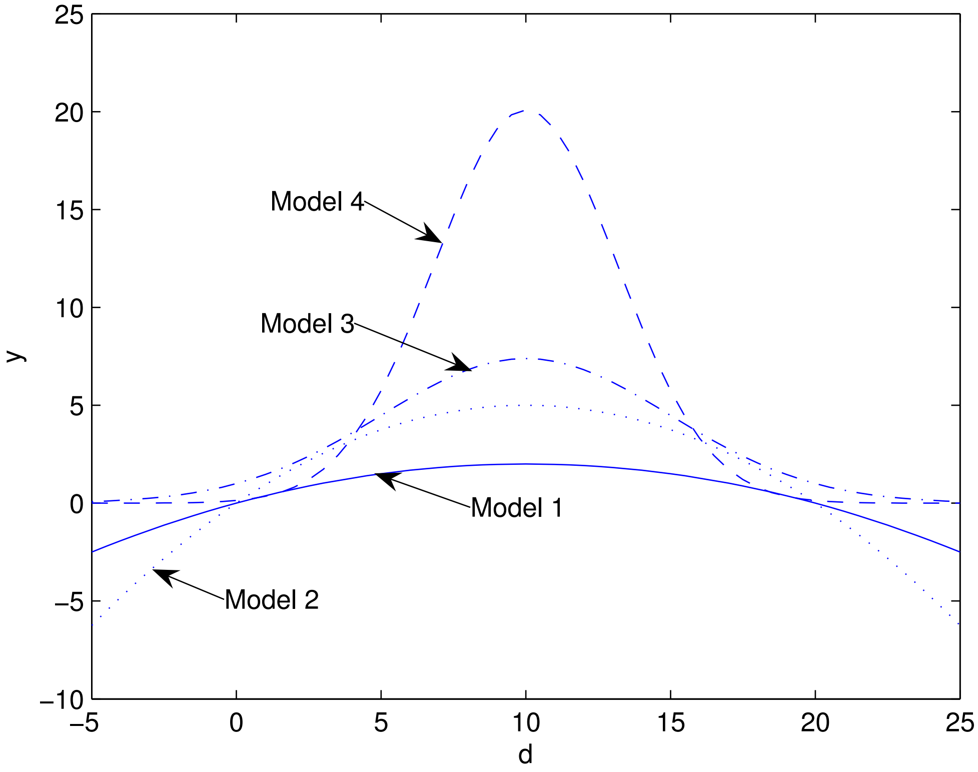


Figure 4.1: The response surfaces for one-dimensional models 1 ∼ 4

For Model 1 and Model 3, they both represent the flat response surfaces, whose gradient changes slowly near the optimum. But Model 1 is linear in $\boldsymbol{\beta}$ and Model 3 is nonlinear in $\boldsymbol{\beta}$.

For Model 2 and Model 4, they both represent the steep response surfaces, whose gradient changes quickly near the optimum. But Model 2 is linear in $\boldsymbol{\beta}$ and Model 4 is nonlinear in $\boldsymbol{\beta}$.

Although these four models are nonlinear, we use a linear model $y = \beta_0 + \beta_1 d$ to approximate the response surface locally. Through these four models, we are going to examine how our two-stage allocation improves the expected response compared to the traditional equal allocation.

For the comparison by method 1, the adjustable factors are:

(1). the number of iterations (N.O.I.): 10 or 50;

(2). the standard deviation of noise (σ): 10 or 50;

(3). the total number of runs in stage 1 v.s. the total number of runs in stage 2, $\{n_0, N - n_0\}$: $\{8, 12\}$, $\{12, 8\}$, $\{16, 4\}$.

The variable of interest is the expected response at the final point. And the following hypothesis for each setting is tested at level 0.05:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ v.s. } H_1 : \mu_1 - \mu_2 > 0$$

where μ_1 is the average response at the final points for two-stage allocation and μ_2 is the average response at the final points for equal allocation. To compute μ_1 and μ_2, we replicate the experiment and obtain 100 final responses for both two-stage allocation and equal allocation. We assume that the largest three samples and the smallest three samples are outliers, and they are deleted afterwards. If the hypothesis H_0 is rejected, then we can conclude that the observations strongly suggest that the two-stage allocation improves the response much faster than the equal allocation.

The results for the comparison using method 1 are given in tables 4.1 to 4.4. The structure of these four tables is identical to table 3.2.

(1). Model 1

Table 4.1: The comparison of final responses for one-dimensional nonlinear model 1

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
10	10	8	12	-0.068	0.767	-0.128	0.691	0.287	184
10	10	12	8	-0.110	0.737	-0.128	0.691	0.429	185

10	10	16	4	-0.152	0.755	-0.128	0.691	0.590	184
10	50	8	12	-0.290	0.645	-0.046	0.761	0.991	181
10	50	12	8	-0.253	0.693	-0.046	0.761	0.974	184
10	50	16	4	-0.250	0.686	-0.046	0.761	0.973	184
50	10	8	12	0.068	0.776	-0.340	0.578	0.000	171
50	10	12	8	-0.007	0.865	-0.340	0.578	0.001	162
50	10	16	4	-0.271	0.680	-0.340	0.578	0.225	181
50	50	8	12	-0.327	0.631	-0.536	0.487	0.006	174
50	50	12	8	-0.127	0.738	-0.536	0.487	0.000	161
50	50	16	4	-0.167	0.747	-0.536	0.487	0.000	159

(2). Model 2

Table 4.2: The comparison of final responses for one-dimensional nonlinear model 2

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
10	10	8	12	0.400	2.120	-0.580	1.660	0.000	175
10	10	12	8	0.050	2.000	-0.580	1.660	0.010	179
10	10	16	4	0.190	2.040	-0.580	1.660	0.003	178
10	50	8	12	-0.370	1.740	-0.640	1.820	0.143	185
10	50	12	8	-0.670	1.690	-0.640	1.820	0.535	184
10	50	16	4	-0.300	1.900	-0.640	1.820	0.102	185
50	10	8	12	0.200	2.090	0.120	2.160	0.398	185
50	10	12	8	0.460	2.310	0.120	2.160	0.155	185
50	10	16	4	0.560	2.370	0.120	2.160	0.097	184
50	50	8	12	-0.190	1.910	-0.770	1.550	0.011	178
50	50	12	8	-0.170	2.050	-0.770	1.550	0.012	173
50	50	16	4	-0.420	1.900	-0.770	1.550	0.080	178

(3). Model 3

Table 4.3: The comparison of final responses for one-dimensional nonlinear model 3

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
10	10	8	12	1.160	1.180	0.917	0.884	0.056	172
10	10	12	8	1.320	1.450	0.917	0.884	0.012	153
10	10	16	4	0.929	0.832	0.917	0.884	0.462	185
10	50	8	12	0.924	0.838	1.004	0.922	0.733	184
10	50	12	8	0.972	0.737	1.004	0.922	0.604	177
10	50	16	4	1.073	0.905	1.004	0.922	0.304	185
50	10	8	12	1.470	1.640	1.440	1.890	0.453	182
50	10	12	8	1.450	1.710	1.440	1.890	0.476	184
50	10	16	4	1.160	1.210	1.440	1.890	0.882	158
50	50	8	12	1.220	1.300	0.884	0.688	0.015	141
50	50	12	8	0.796	0.540	0.884	0.688	0.834	176
50	50	16	4	0.928	0.731	0.884	0.688	0.337	185

(4). Model 4

Table 4.4: The comparison of final responses for one-dimensional nonlinear model 4

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
10	10	8	12	0.318	0.984	0.237	0.686	0.255	166
10	10	12	8	0.286	0.659	0.237	0.686	0.307	185
10	10	16	4	0.231	0.586	0.237	0.686	0.524	181
10	50	8	12	0.660	2.090	0.390	1.140	0.136	143
10	50	12	8	0.335	0.833	0.390	1.140	0.645	170

10	50	16	4	0.206	0.608	0.390	1.140	0.914	141
50	10	8	12	0.241	0.666	0.126	0.283	0.063	125
50	10	12	8	0.161	0.382	0.126	0.283	0.238	171
50	10	16	4	0.231	0.482	0.126	0.283	0.036	150
50	50	8	12	1.180	2.920	0.520	1.410	0.025	134
50	50	12	8	0.750	2.370	0.520	1.410	0.215	151
50	50	16	4	0.313	0.919	0.520	1.410	0.883	159

From the above numerical results, although some P-values are significant, the conclusion is not very consistent. Unlike the linear models, there is only one optimum in the feasible region for the nonlinear models considered. However, in this thesis, we do not take the convergence to the optimum into account. When the design point reaches the optimum during the hill-climbing procedure, it still can jump away in the next iteration. Here we only compare a particular response when the design finishes moving its last ($10th$ or $50th$) step.

In the following experiments, we will compare the best response that each allocation ever gets during its 10-step or 50-step hill-climbing. Correspondingly, the variable of interest becomes the best response during the hill-climbing. It is reasonable as in reality the experimenter is only concerned with the best response during the hill-climbing. The following hypothesis is tested at level 0.05:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ v.s. } H_1 : \mu_1 - \mu_2 > 0$$

where μ_1 is the average response at the best points during hill-climbing for two-stage allocation and μ_2 is the average response at the best points for equal allocation. The rest of the settings are the same as the comparison of the final responses.

The results for comparing the best responses are given in tables 4.5 to 4.8.

(1). Model 1

Table 4.5: The comparison of best responses for one-dimensional nonlinear model 1

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
10	10	8	12	0.645	0.512	0.453	0.383	0.002	172
10	10	12	8	0.713	0.470	0.453	0.383	0.000	178
10	10	16	4	0.591	0.522	0.453	0.383	0.020	170
10	50	8	12	0.601	0.548	0.376	0.514	0.002	185
10	50	12	8	0.456	0.435	0.376	0.514	0.123	181
10	50	16	4	0.526	0.473	0.376	0.514	0.019	184
50	10	8	12	1.407	0.454	1.148	0.561	0.000	178
50	10	12	8	1.469	0.422	1.148	0.561	0.000	172
50	10	16	4	1.356	0.465	1.148	0.561	0.003	179
50	50	8	12	1.223	0.473	1.015	0.513	0.002	184
50	50	12	8	1.236	0.487	1.015	0.513	0.001	185
50	50	16	4	1.192	0.535	1.015	0.513	0.011	185

(2). Model 2

Table 4.6: The comparison of best responses for one-dimensional nonlinear model 2

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
10	10	8	12	2.080	1.270	1.480	1.370	0.001	184
10	10	12	8	2.040	1.400	1.480	1.370	0.003	185
10	10	16	4	1.930	1.440	1.480	1.370	0.014	185
10	50	8	12	1.520	1.220	0.990	1.140	0.001	185

10	50	12	8	1.330	1.280	0.990	1.140	0.027	183
10	50	16	4	1.290	1.270	0.990	1.140	0.045	183
50	10	8	12	3.992	0.901	3.710	1.170	0.032	174
50	10	12	8	4.064	0.768	3.710	1.170	0.008	160
50	10	16	4	3.965	0.970	3.710	1.170	0.052	179
50	50	8	12	3.340	1.240	2.750	1.420	0.001	182
50	50	12	8	3.160	1.160	2.750	1.420	0.014	178
50	50	16	4	2.960	1.080	2.750	1.420	0.126	173

(3). Model 3

Table 4.7: The comparison of best responses for one-dimensional nonlinear model 3

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
10	10	8	12	2.120	1.280	1.737	0.907	0.010	167
10	10	12	8	1.940	1.120	1.737	0.907	0.085	178
10	10	16	4	2.310	1.600	1.737	0.907	0.002	147
10	50	8	12	1.920	1.130	1.664	0.988	0.048	182
10	50	12	8	2.050	1.300	1.664	0.988	0.011	173
10	50	16	4	1.657	0.898	1.664	0.988	0.522	184
50	10	8	12	5.040	1.930	3.850	1.990	0.000	185
50	10	12	8	4.290	1.960	3.850	1.990	0.063	185
50	10	16	4	4.380	2.040	3.850	1.990	0.037	185
50	50	8	12	3.720	1.750	2.950	1.560	0.001	183
50	50	12	8	3.940	1.870	2.950	1.560	0.000	180
50	50	16	4	3.390	1.760	2.950	1.560	0.036	183

(4). Model 4

Table 4.8: The comparison of best responses for one-dimensional nonlinear model 4

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
10	10	8	12	1.470	2.300	1.030	2.020	0.081	182
10	10	12	8	1.030	1.710	1.030	2.020	0.498	181
10	10	16	4	1.000	1.670	1.030	2.020	0.553	179
10	50	8	12	1.360	2.800	1.090	2.250	0.235	177
10	50	12	8	1.560	2.720	1.090	2.250	0.101	179
10	50	16	4	1.200	2.490	1.090	2.250	0.379	184
50	10	8	12	6.720	6.530	4.860	5.370	0.017	179
50	10	12	8	5.860	6.650	4.860	5.370	0.130	178
50	10	16	4	6.120	6.360	4.860	5.370	0.072	180
50	50	8	12	6.200	6.100	4.660	5.440	0.035	183
50	50	12	8	6.600	6.570	4.660	5.440	0.014	179
50	50	16	4	6.470	6.920	4.660	5.440	0.024	176

From tables 4.5 to 4.8, we can observe that:

1. Two-stage allocation works very well for all the four response surfaces. For model 4, because the initial gradient is very flat and the step size in Equation 3.8 is very small when the gradient is flat, therefore the design points of both allocations move very slowly. After 10 iterations most of the design points are still in the flat region and there is no significant difference between two-stage allocation and equal allocation. However, after 50 iterations, two-stage allocation excels again.

2. $\{8, 12\}$ seems to be the best combination of $\{n_0, N-n_0\}$ for two-stage allocation. In most cases, $\{16, 4\}$ is the worst combination, and sometimes there is no

significant difference between $\{16, 4\}$ and equal allocation, which may be due to most of the runs being equally distributed in stage 1.

For the comparison by method 2, we choose $t\% = 90\%$, and the adjustable factors are:

(1). the standard deviation of noise (σ): 10 or 50;

(2). the total number of runs in stage 1 v.s. the total number of runs in stage 2, $\{n_0, N - n_0\}$: $\{8, 12\}$, $\{12, 8\}$, $\{16, 4\}$.

We fix an upper bound for the computing budget, which means for each setting of σ and $\{n_0, N - n_0\}$, a maximum of 300 iterations is carried out. If the allocation fails to improve the response within 90% of the true optimum after 300 iterations, then 300 is recorded as the lower bound for the number of iterations.

The variable of interest is the number of iterations required by each design to get to 90% of the true optimum. The following hypothesis is tested at level 0.05:

$$H_0 : \nu_1 - \nu_2 = 0 \text{ v.s. } H_1 : \nu_1 - \nu_2 < 0$$

where ν_1 is the average number of iterations to get to 90% of the true optimum for two-stage allocation and ν_2 is the average number of iterations to get to 90% of the true optimum for equal allocation. To compute ν_1 and ν_2, we replicate the experiment for 300 times for both two-stage allocation and equal allocation. We also delete the largest three samples and the smallest three samples as outliers. If the hypothesis H_0 is rejected, then we can conclude that the observations strongly suggest that the two-stage allocation improves the response much faster than the equal allocation.

Tables 4.9 to 4.12 show the results for varying σ, $\{n_0, N - n_0\}$, ν_1, ν_2, σ_1 - the standard deviation of ν_1 over 294 samples, σ_2 - the standard deviation of ν_2 over 294 samples, and give the P-value of the hypothesis test. Here, DOF $= \frac{(\frac{\sigma_1^2}{294} + \frac{\sigma_2^2}{294})^2}{\frac{(\sigma_1^2/294)^2}{293} + \frac{(\sigma_2^2/294)^2}{293}}$. If the value of DOF is not an integer, it will be rounded down to the nearest integer.

(1). Model 1

Table 4.9: The comparison of N.O.I. for one-dimensional nonlinear model 1

σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
			ν_1	σ_1	ν_2	σ_2		
10	8	12	139.5	97.9	170.0	107.0	0.000	581
10	12	8	138.1	97.3	170.0	107.0	0.000	580
10	16	4	157.0	103.0	170.0	107.0	0.065	585
50	8	12	168.0	104.0	206.0	102.0	0.000	585
50	12	8	183.0	107.0	206.0	102.0	0.004	584
50	16	4	185.0	107.0	206.0	102.0	0.008	584

(2). Model 2

Table 4.10: The comparison of N.O.I. for one-dimensional nonlinear model 2

σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
			ν_1	σ_1	ν_2	σ_2		
10	8	12	96.0	73.9	130.0	100.0	0.000	538
10	12	8	94.1	73.7	130.0	100.0	0.000	537
10	16	4	102.0	82.0	130.0	100.0	0.000	563
50	8	12	156.0	103.0	188.0	103.0	0.000	585
50	12	8	159.0	107.0	188.0	103.0	0.000	585
50	16	4	171.0	105.0	188.0	103.0	0.021	585

(3). Model 3

Table 4.11: The comparison of N.O.I. for one-dimensional nonlinear model 3

σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
			ν_1	σ_1	ν_2	σ_2		
10	8	12	131.2	94.3	180.0	106.0	0.000	578
10	12	8	143.0	100.0	180.0	106.0	0.000	584
10	16	4	151.0	103.0	180.0	106.0	0.000	585
50	8	12	194.0	105.0	226.5	97.7	0.000	583
50	12	8	197.0	104.0	226.5	97.7	0.000	583
50	16	4	204.0	107.0	226.5	97.7	0.004	581

(4). Model 4

Table 4.12: The comparison of N.O.I. for one-dimensional nonlinear model 4

σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
			ν_1	σ_1	ν_2	σ_2		
10	8	12	198.0	106.0	236.9	92.0	0.000	575
10	12	8	197.0	108.0	236.9	92.0	0.000	572
10	16	4	208.0	102.0	236.9	92.0	0.000	580
50	8	12	198.0	102.0	230.0	97.6	0.000	585
50	12	8	202.0	105.0	230.0	97.6	0.000	582
50	16	4	220.0	101.0	230.0	97.6	0.116	585

From tables 4.9 to 4.12, we can observe that:

1. Two-stage allocation needs less iterations to reach the fixed neighborhood of the true optimum.

2. When the noise becomes smaller, the ASA technique results in larger step sizes (see Equation 3.8). Hence when σ is smaller, both allocations require less iterations to reach the fixed region for Models 1, 2 and 3. However, the N.O.I. is larger for Model 4 when σ is smaller. It may be due to the particular shape of Model 4. Its fixed region is small because its gradient near the optimum is steep. The large step sizes may make the design points jump over the fixed region, and eventually fail to get into the fixed region, which can be seen also in table 4.13.

3. $\{16, 4\}$ is still the worst combination of $\{n_0, N - n_0\}$ for most of the cases.

Table 4.13 shows the results that in those 300 trials, how many trials successfully reach the fixed neighborhood of the true optimum within 300 iterations for varying σ and allocations.

Table 4.13: The comparison of the number of successful trials for one-dimensional nonlinear models

		Two-stage Allocation $\{n_0, N - n_0\}$			Equal Allocation
	σ	$\{8, 12\}$	$\{12, 8\}$	$\{16, 4\}$	
Model 1	10	253	259	234	220
	50	219	200	191	169
Model 2	10	291	290	284	254
	50	232	221	217	199
Model 3	10	258	248	234	205
	50	193	180	162	142
Model 4	10	180	174	164	125
	50	191	165	143	131

From table 4.13, we can observe that there are many trials for equal allocation that do not reach the fixed neighborhood after running out of 300 iterations. This may be due to the reason that it is easier for equal allocation to obtain an infinite step size, and since we have a finite feasible input region for those nonlinear models, if an infinite next design point is identified, then we will move to the boundary of the feasible region. Thus for equal allocation the design points will bounce on the boundaries of the feasible region and fail to get into the fixed neighborhood of the true optimum at the end of all iterations.

Since we record 300 as the lower bound for the number of iterations if the experiment fails to obtain 90% of the true optimum after 300 iterations, we can also conclude that the value of N.O.I. from tables 4.9 to 4.12 has been under estimated. However, from table 4.13 we can see that the underestimation for equal allocation is more serious than the case for two-stage allocation, and equal allocation may need much more runs to reach the fixed neighborhood of the true optimum than the value of ν_2 as shown from tables 4.9 to 4.12.

From tables 4.9 to 4.13, we can conclude that for those one-dimensional nonlinear models, our two-stage allocation can reach the optimal region much faster than the equal allocation, and thus the efficiency of our two-stage approach to conduct hill-climbing is higher than the equal allocation.

4.2 Two-dimensional Nonlinear Model

For two dimensional nonlinear model, the procedure to select the optimal allocation is identical as the procedure in two dimensional linear model case, and we also consider four different response surfaces besides the previous adjustable factors:

(1). **Model 5:** $y = -\dfrac{(d_1-10)^2+(d_2-10)^2}{40}+5+\epsilon = \dfrac{1}{2}\cdot d_1+\dfrac{1}{2}\cdot d_2-\dfrac{1}{40}d_1^2-\dfrac{1}{40}d_2^2+\epsilon$

- the feasible region is a circle with center $\{10, 10\}$ and radius $12\sqrt{2}$, the range of response is $(-2.2, 5)$, and the response at the starting point is 0.
- the gradient at the starting point is 0.71, and it is continuously decreasing to 0.

The response surface for this model is given in figure 4.2

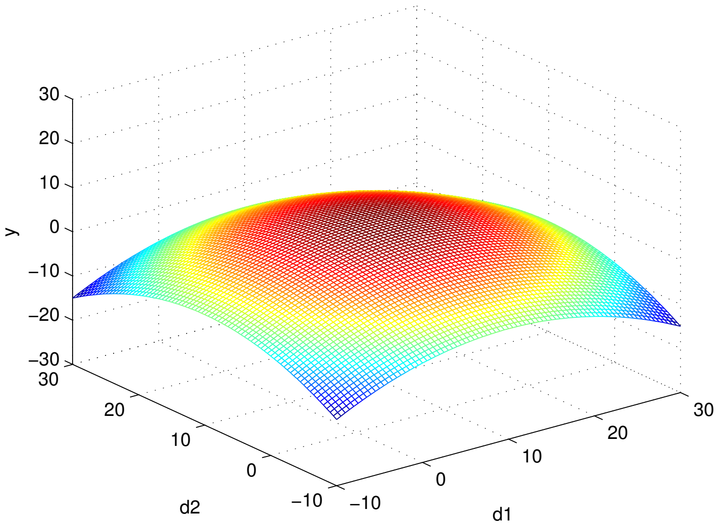


Figure 4.2: The response surface for two-dimensional model 5

(2). **Model 6:** $y = -\dfrac{(d_1-10)^2+(d_2-10)^2}{20} + 10 + \epsilon = d_1 + d_2 - \dfrac{1}{20}d_1^2 - \dfrac{1}{20}d_2^2 + \epsilon$

- the feasible region is a circle with center $\{10, 10\}$ and radius $12\sqrt{2}$, the range of response is $(-4.4, 10)$, and the response at the starting point is 0.
- the gradient at the starting point is 1.41, and it is continuously decreasing to 0.

The response surface for this model is given in figure 4.3

(3). **Model 7:** $y = e^{-\frac{(d_1-8)^2+(d_2-8)^2}{40}+2} + \epsilon = e^{-1.2+0.4\cdot d_1+0.4\cdot d_2-\frac{1}{40}d_1^2-\frac{1}{40}d_2^2} + \epsilon$

- the feasible region is a circle with center $\{8, 8\}$ and radius $10\sqrt{2}$, the range of response is $(0.05, 7.39)$, and the response at the starting point is 0.3.
- The gradient: 0.17 (starting point) → 1 (largest) → 0 (optimum).

The response surface for this model is given in figure 4.4

(4). **Model 8:** $y = e^{-\frac{(d_1-8)^2+(d_2-8)^2}{20}+3} + \epsilon = e^{-3.4+0.8\cdot d_1+0.8\cdot d_2-\frac{1}{20}d_1^2-\frac{1}{20}d_2^2} + \epsilon$

- the feasible region is a circle with center $\{8, 8\}$ and radius $10\sqrt{2}$, the range of response is $(0.001, 20.09)$, and the response at the starting point is 0.03.

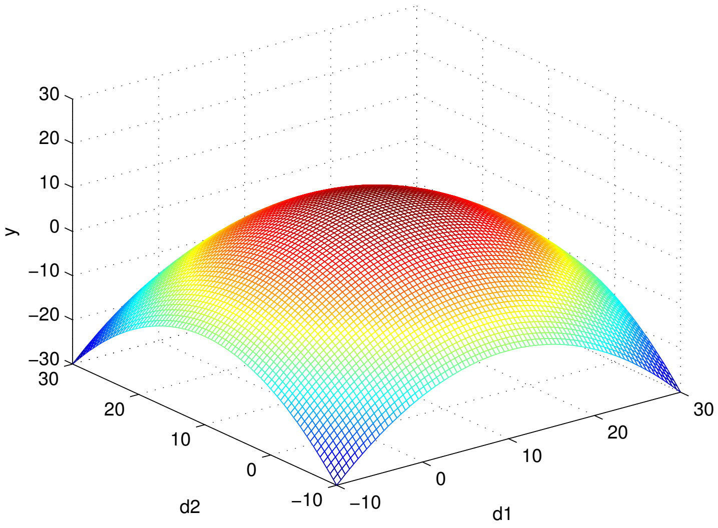


Figure 4.3: The response surface for two-dimensional model 6

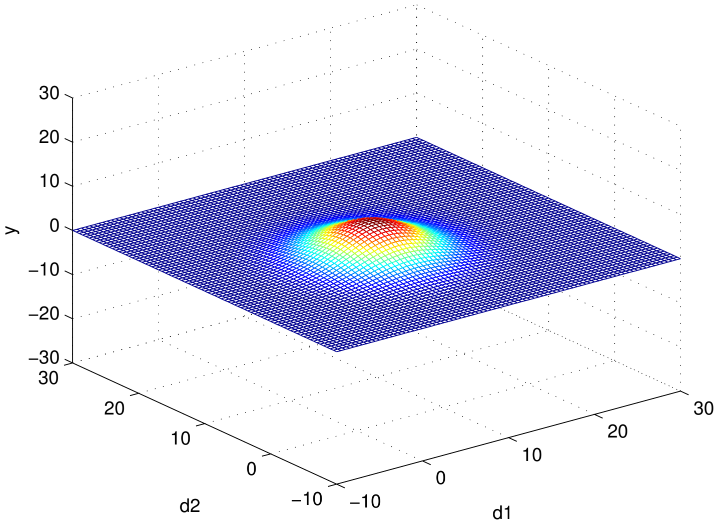


Figure 4.4: The response surface for two-dimensional model 7

- The gradient: 0.04 (starting point) → 3.85 (largest) → 0 (optimum).

The response surface for this model is given in figure 4.5

Similar to the one-dimensional case, Model 1 and Model 3 represent the flat response surfaces, whose gradient changes slowly near the optimum, while Model 2 and Model 4 represent the steep response surfaces, whose gradient changes quickly near the optimum. Also Model 1 and Model 2 are linear in $\boldsymbol{\beta}$, and Model 3 and

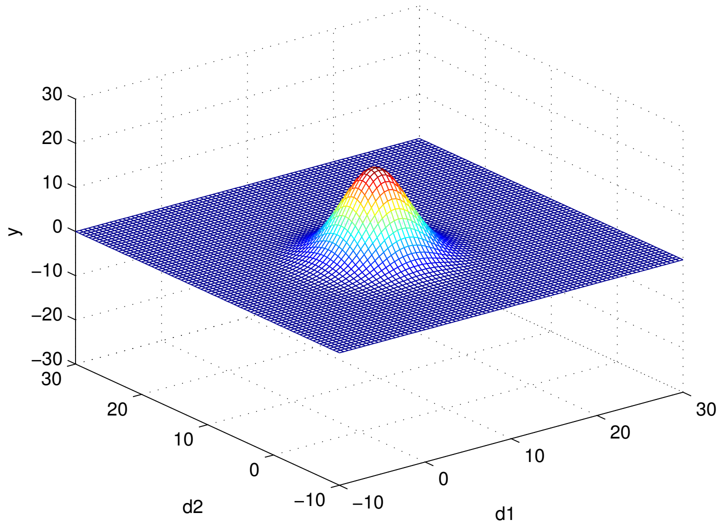


Figure 4.5: The response surface for two-dimensional model 8

Model 4 are nonlinear in $\boldsymbol{\beta}$.

The local approximation is $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2$. Through these four models, we are going to examine how our two-stage allocation improves the expected response compared to the traditional equal allocation.

For the comparison by method 1, the adjustable factors are:

(1). the number of iterations (N.O.I.): 10 or 50;

(2). the standard deviation of noise (σ): 10 or 50;

(3). the total number of runs in stage 1 v.s. the total number of runs in stage 2, $\{n_0, N - n_0\}$: $\{12, 28\}$, $\{20, 20\}$, $\{28, 12\}$.

The variable of interest is the best expected response during the hill-climbing procedure. The following hypothesis is tested at level 0.05:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ v.s. } H_1 : \mu_1 - \mu_2 > 0$$

where μ_1 is the average response at the best points during the hill-climbing procedure for two-stage allocation and μ_2 is the average response at the best points for equal allocation. To compute μ_1 and μ_2, we replicate the experiment and obtain 100 samples for both two-stage allocation and equal allocation. The largest three samples and the smallest three samples are deleted as outliers. If the hypothesis H_0 is rejected, then we can conclude that the observations strongly suggest that the two-stage allocation improves the response much faster than the equal allocation.

The results for comparison using method 1 are given in tables 4.14 to 4.17. The table structure is the same as the structure of tables 4.5 to 4.8

(1). Model 5

Table 4.14: The comparison of best responses for two-dimensional nonlinear model 5

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
10	10	12	28	2.032	0.921	0.560	1.290	0.000	178
10	10	20	20	1.631	0.657	0.560	1.290	0.000	146
10	10	28	12	1.435	0.957	0.560	1.290	0.000	182
10	50	12	28	1.768	0.839	0.140	1.030	0.000	190
10	50	20	20	1.281	0.739	0.140	1.030	0.000	179
10	50	28	12	1.066	0.869	0.140	1.030	0.000	192
50	10	12	28	4.052	0.876	1.500	1.300	0.000	173
50	10	20	20	4.019	0.721	1.500	1.300	0.000	154
50	10	28	12	3.520	1.060	1.500	1.300	0.000	190
50	50	12	28	3.866	0.925	1.130	1.410	0.000	170
50	50	20	20	3.566	0.975	1.130	1.410	0.000	176
50	50	28	12	2.110	1.310	1.130	1.410	0.000	196

(2). Model 6

Table 4.15: The comparison of best responses for two-dimensional nonlinear model 6

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
10	10	12	28	4.560	1.420	1.330	2.260	0.000	156
10	10	20	20	3.980	1.340	1.330	2.260	0.000	151
10	10	28	12	4.280	1.830	1.330	2.260	0.000	178
10	50	12	28	3.920	1.450	0.490	1.970	0.000	170
10	50	20	20	2.960	1.400	0.490	1.970	0.000	167
10	50	28	12	2.220	1.830	0.490	1.970	0.000	184
50	10	12	28	8.170	1.280	3.790	2.060	0.000	155
50	10	20	20	9.019	0.886	3.790	2.060	0.000	126
50	10	28	12	8.330	1.650	3.790	2.060	0.000	177
50	50	12	28	8.160	1.410	1.980	1.820	0.000	174
50	50	20	20	7.370	1.890	1.980	1.820	0.000	185
50	50	28	12	5.160	2.290	1.980	1.820	0.000	177

(3). Model 7

Table 4.16: The comparison of best responses for two-dimensional nonlinear model 7

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
10	10	12	28	1.345	0.844	0.387	0.242	0.000	108
10	10	20	20	1.035	0.575	0.387	0.242	0.000	125
10	10	28	12	0.842	0.515	0.387	0.242	0.000	132
10	50	12	28	1.331	0.634	0.492	0.618	0.000	185

10	50	20	20	0.882	0.506	0.492	0.618	0.000	178
10	50	28	12	0.746	0.439	0.492	0.618	0.001	167
50	10	12	28	5.130	1.830	0.896	0.992	0.000	143
50	10	20	20	4.320	2.310	0.896	0.992	0.000	126
50	10	28	12	3.590	2.820	0.896	0.992	0.000	115
50	50	12	28	4.340	2.020	0.697	0.760	0.000	118
50	50	20	20	3.540	1.930	0.697	0.760	0.000	121
50	50	28	12	2.080	1.800	0.697	0.760	0.000	125

(4). Model 8

Table 4.17: The comparison of best responses for two-dimensional nonlinear model 8

N.O.I.	σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
				μ_1	σ_1	μ_2	σ_2		
10	10	12	28	0.910	1.030	0.144	0.362	0.000	115
10	10	20	20	0.535	0.658	0.144	0.362	0.000	144
10	10	28	12	0.690	1.570	0.144	0.362	0.001	102
10	50	12	28	1.000	1.190	0.082	0.155	0.000	96
10	50	20	20	0.436	0.529	0.082	0.155	0.000	108
10	50	28	12	0.218	0.258	0.082	0.155	0.000	152
50	10	12	28	12.100	6.270	0.311	0.733	0.000	95
50	10	20	20	13.210	7.960	0.311	0.733	0.000	94
50	10	28	12	5.870	7.450	0.311	0.733	0.000	94
50	50	12	28	7.950	6.270	0.430	1.220	0.000	100
50	50	20	20	8.810	7.190	0.430	1.220	0.000	98
50	50	28	12	4.710	6.780	0.430	1.220	0.000	99

From tables 4.14 to 4.17, we can observe that:

1. Two-stage allocation outperforms equal allocation significantly.

2. $\{12, 28\}$ is still the best combination of $\{n_0, N - n_0\}$ for two-stage allocation. This observation is consistent with the one-dimensional case. If we leave more runs to be determined by the second stage of the two-stage allocation approach, the efficiency of improving the response is higher.

3. From the raw data (not shown here), we observe that for equal allocation many experiments end on the boundary of the feasible region at the end of all iterations. Since the response values on the boundary are all the same and very small, therefore the variability σ_2 is much smaller and μ_2 is small also. This observation may also be due to the reason that it is easier for equal allocation to obtain an infinite step size, and the feasible input region restricts the next design point of equal allocation.

For the comparison by method 2, we choose $t\% = 80\%$. Because the search space for two-dimensional models is very large, we choose a smaller $t\%$ so that the design point is easier to get into the fixed region. As all the settings are the same for two-stage allocation and equal allocation, the smaller $t\%$ will not affect the conclusions. The adjustable factors are:

(1). the standard deviation of noise (σ): 10 or 50;

(2). the total number of runs in stage 1 v.s. the total number of runs in stage 2, $\{n_0, N - n_0\}$: $\{12, 28\}$, $\{20, 20\}$, $\{28, 12\}$.

We also fix an upper bound for the computing budget, which means for the same σ and $\{n_0, N - n_0\}$, a maximum of 500 iterations is carried out. If the allocation fails to improve the response within 80% of the true optimum after 500 iterations, 500 is recorded as the lower bound for the number of iterations.

The variable of interest is the number of iterations required by each design to get to 80% of the true optimum. The following hypothesis is tested at level 0.05:

$$H_0 : \nu_1 = \nu_2 \text{ v.s. } H_1 : \nu_1 < \nu_2$$

where ν_1 is the average number of iterations to get to 80% of the true optimum for two-stage allocation, and ν_2 is the average number of iterations to get to 80% of the true optimum for equal allocation. To compute ν_1 and ν_2, we replicate the experiment for 300 times for both two-stage allocation and equal allocation. The largest three samples and the smallest three samples are deleted as outliers. If the hypothesis H_0 is rejected, then we can conclude that the observations strongly suggest that the two-stage allocation improves the response much faster than the equal allocation.

The results for comparison using method 2 are given in tables 4.18 to 4.21. The table structure is the same as the structure of tables 4.9 to 4.12

(1). Model 5

Table 4.18: The comparison of N.O.I. for two-dimensional nonlinear model 5

σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
			ν_1	σ_1	ν_2	σ_2		
10	12	28	73.9	93.2	103.7	94.8	0.000	585
10	20	20	94.0	114.0	103.7	94.8	0.126	567
10	28	12	98.9	99.8	103.7	94.8	0.274	584
50	12	28	103.0	132.0	133.0	129.0	0.003	585
50	20	20	150.0	138.0	133.0	129.0	0.930	583
50	28	12	230.0	183.0	133.0	129.0	1.000	526

(2). Model 6

Table 4.19: The comparison of N.O.I. for two-dimensional nonlinear model 6

σ	n_0	$N - n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
			ν_1	σ_1	ν_2	σ_2		
10	12	28	61.6	88.4	82.5	75.9	0.001	573
10	20	20	46.2	70.0	82.5	75.9	0.000	582
10	28	12	43.1	37.1	82.5	75.9	0.000	425

50	12	28	92.0	119.0	128.0	123.0	0.000	585
50	20	20	140.0	138.0	128.0	123.0	0.863	578
50	28	12	181.0	168.0	128.0	123.0	1.000	536

(3). Model 7

Table 4.20: The comparison of N.O.I. for two-dimensional nonlinear model 7

σ	n_0	$N-n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
			ν_1	σ_1	ν_2	σ_2		
10	12	28	145.0	174.0	275.0	172.0	0.000	585
10	20	20	126.0	156.0	275.0	172.0	0.000	580
10	28	12	184.0	185.0	275.0	172.0	0.000	582
50	12	28	160.0	177.0	310.0	181.0	0.000	585
50	20	20	254.0	200.0	310.0	181.0	0.000	580
50	28	12	319.0	193.0	310.0	181.0	0.726	583

(4). Model 8

Table 4.21: The comparison of N.O.I. for two-dimensional nonlinear model 8

σ	n_0	$N-n_0$	Two-stage Allocation		Equal Allocation		P-value	DOF
			ν_1	σ_1	ν_2	σ_2		
10	12	28	171.0	182.0	381.0	163.0	0.000	579
10	20	20	132.0	178.0	381.0	163.0	0.000	581
10	28	12	268.0	210.0	381.0	163.0	0.000	552
50	12	28	214.0	193.0	357.0	176.0	0.000	581
50	20	20	225.0	187.0	357.0	176.0	0.000	583
50	28	12	298.0	198.0	357.0	176.0	0.000	577

From tables 4.18 to 4.21, we can observe that:

1. For Models 5, 6 and 7, both allocations require less iterations to reach the fixed region of the true optimum when the noise is smaller. Similar to the one-dimensional case, this does not hold for Model 8 because its gradient near the optimum is steep also.

2. $\{12, 28\}$ performs consistently well. $\{28, 12\}$ is the worst combination of $\{n_0, N-n_0\}$ for two-stage allocation. In some cases of combination $\{28, 12\}$, we fail to reject H_0 and conclude that the data do not show significant difference between $\{28, 12\}$ and equal allocation. Since $\{28, 12\}$ allocates most of the runs (28 out of 40 runs) equally in stage 1, it will perform similarly as equal allocation.

Table 4.22 shows the results that in those 300 trials, how many trials successfully reach the fixed neighborhood of the true optimum within 500 iterations for varying σ and allocations.

Table 4.22: The comparison of the number of successful trials for two-dimensional nonlinear models

		Two-stage Allocation $\{n_0, N-n_0\}$			Equal Allocation
	σ	$\{12, 28\}$	$\{20, 20\}$	$\{28, 12\}$	
Model 5	10	287	287	296	298
	50	272	275	231	292
Model 6	10	289	295	300	299
	50	279	276	262	289
Model 7	10	243	263	243	233
	50	238	192	163	195
Model 8	10	235	245	188	133
	50	212	215	176	155

From table 4.22, we can observe that among the three combinations of $\{n_0, N-n_0\}$, $\{28, 12\}$ always has the worst performance when $\sigma = 50$ for all the models. The number of successful trials for $\{28, 12\}$ when $\sigma = 50$ is always the smallest, even smaller than the value of equal allocation sometimes. Therefore we must be very careful to choose the combination of $\{n_0, N - n_0\}$. Furthermore, the different shapes of the response surfaces seem to affect the performance. For model 5 and model 6, it is easier for equal allocation to reach the fixed neighborhood of the true optimum, while for model 7 and model 8, it becomes easier for two-stage allocation. The combination $\{12, 28\}$ always works well, as it is comparable to equal allocation for model 5 and model 6, and it is better than equal allocation for model 7 and model 8. Similar to table 4.13, from table 4.22, we can conclude that the values of N.O.I. from tables 4.18 to 4.21 are underestimated.

From tables 4.18 to 4.22, we can conclude that for those two-dimensional nonlinear models, our two-stage allocation can reach the optimal region faster than the equal allocation for most of the cases, and the combination $\{12, 28\}$ consistently works well.

In this chapter, we consider one- and two- dimensional nonlinear models. According to the numerical results, our two-stage allocation can always get a better response in a fixed number of iterations. Most of the time, it needs less iterations to reach the fixed neighborhood of the true optimum.

If we assign less runs in stage 1, and leave more runs to be decided by our two-stage allocation, the efficiency of hill-climbing is much higher. These observations are very obvious and consistent in these numerical results, such as $\{8, 12\}$ for one-dimensional nonlinear models and $\{12, 28\}$ for two-dimensional nonlinear models. Since our two-stage allocation can get a higher efficiency if it has more runs to be decided by its second stage, these observations also assure us that our two-stage approach is a very good design to distribute the simulation runs.

If we assign more runs in stage 1, the two-stage allocation is close to equal allocation since the runs in stage 1 are equally distributed. Thus it is not surprising that $\{16, 4\}$ for one-dimensional cases and $\{28, 12\}$ for two-dimensional cases perform similarly as equal allocation for most of the time.

In the next chapter, we will conclude this thesis and propose the research work that can be done in the future.

Chapter 5

Conclusion and Future Research

This chapter concludes this thesis in section 5.1 and proposes directions for future research in section 5.2.

5.1 Summary and Conclusion

When we apply the ASA technique to simulation optimization problems with computing budget constraints, our two-stage computing budget allocation approach becomes applicable. When we compare it to the approach which allocates the runs equally to each design point, the numerical results show that:

1. For most of the cases, our two-stage allocation outperforms the equal allocation. After using up the same number of iterations, our two-stage allocation obtains a larger response than equal allocation, and our two-stage allocation needs less number of iterations to reach the fixed neighborhood of the true optimum.

2. For two-stage allocation, if we assign less runs in stage 1 and more runs in stage 2, the performance of the two-stage allocation will be even better. These observations are very obvious and consistent, such as $\{n_0, N - n_0\} = \{8, 12\}$ for one-dimensional cases and $\{n_0, N - n_0\} = \{12, 28\}$ for two-dimensional cases. These observations assure us that our two-stage approach is a good experimental design to distribute the computing budget because its efficiency

to conduct hill-climbing is much higher if it can have more budget to be decided in its second stage.

3. For two-stage allocation, if we assign more runs in stage 1 and less runs in stage 2, we always draw the conclusion that there is no significant difference between two-stage allocation and equal allocation. This is not surprising because the runs in stage 1 of two-stage allocation are equally distributed. If most of the runs are assigned to stage 1, then computing budget allocation by our two-stage approach will be close to equal allocation.

The ASA technique considers the lower bound of predicted response, the worst case of prediction. Thus it is expected to work well in the worst case in which the system noise is very large. Our conclusion would be:

Our two-stage allocation approach is a good experimental design for the simulation optimization problems with computing budget constraints. When the system noise is observed or known to be large, our two-stage allocation approach will work significantly better than the equal allocation. If we allocate less runs in stage 1 and more runs in stage 2, the performance of our two-stage allocation will be even better.

5.2 Future Research

In our two-stage allocation approach, we apply brute force search to identify the best allocation, which will be very time-consuming when the number of design point increases or the linear model is a higher-dimensional one. In the optimization literature, there are a few alternatives for brute force search, such as genetic algorithms, tabu search or simulated annealing. These algorithms may be adopted by our two-stage allocation.

We do not discuss the ratio of n_0 and $N - n_0$ in this thesis. Based on the numerical results and the experience that we acquired in this study, we would suggest that when applying this two-stage computing budget allocation approach, less than half of the total budget should be used in the first stage. However, less n_0 may cause the estimates of $\boldsymbol{\beta}$ and σ^2 to be inaccurate. The tradeoff between accuracy of estimates and the efficiency of improving response values is an important and interesting area for further research.

The performance of the hill-climbing method may rely on the shape of the selected response surface (e.g. the Model 4 and the Model 8 in Chapter 4). In our experiments, only a few types of models that are representative of different shaped low order responses have been tested. Closer studies of the relationship with the shape of the response functions and further experiments on higher order more complicated models are avenues for further research.

In this thesis, we focus on the computing budget allocation problem, and we suggest a feasible allocation scheme when the next design point $\mathbf{d}^+$ is at infinity. However we do not solve the infinite $\mathbf{d}^+$ problem completely. When $\mathbf{d}^+$ is infinite, we drop the observations or use a feasible region to restrict the input variables. Regularization seems to be a good method to resolve this. Previously, we only maximize $\hat{y}_{min}$. When the noise is very small, or the gradient is very large, the maximal point $\mathbf{d}^+$ is infinite (see Appendix A for the one-dimensional case). We may add a penalty term to the objective function $\hat{y}_{min}$ to penalize the distance from the original center of the region of experimentation because the further away the design point is from the original center, the less reliable the predicted response. However, more work should be done to properly choose the penalty term and regularization parameter.

We always assume the true response function can be sufficiently approximated by a first-order homoscedastic model in this thesis. However, from the point of view of robust design, this design can be improved. There are two possible ways to improve the current approach. The first one is to consider the model misspecification in this design, and then determine how to allocate the simulation runs so that the lower bound of predicted response can be maximized. The second one is to consider the heteroscedastic case. Generalized linear models can be applied in this case. The main idea of two-stage allocation, which is to improve the lower bound of predicted response so that we can have more confidence to move to the next design point is the same, and our two-stage computing budget allocation can be applied similarly.

Our approach is a limited form of sequential design because it involves two stages. However, the information we gather during the hill-climbing is not utilized when we consider this two-stage approach. Bayesian designs may offer us an opportunity to consider the historical data along with the hill-climbing. These data can be regarded as the prior information, and we can update these prior information at the completion of stage 1 so that we can allocate the runs in stage 2 in a more optimal manner.

Bibliography

[1] Kleijnen, J.P.C., Hertog, D. and Angün, E. 2004. Response surface methodology's steepest ascent and step size revisited. European Journal of Operational Research, 159, 121-131.

[2] Box, G.E.P. and Draper, N.R. 1987. Empirical model-building and response surfaces. John Wiley & Sons, New York.

[3] Myers, R.H. and Montgomery, D.C. 2002. Response surface methodology: process and product optimization using designed experiments, second ed. John Wiley & Sons, New York.

[4] Khuri, A. I. and Cornell, J. A. 1996. Response Surface: Design and Analyses. 2nd edition. Marcel Dekker, New York.

[5] Law, A.M. and Kelton W.D. 2000. Simulation modeling and analysis, 3rd ed. McGraw-Hill, New York.

[6] Fabian, V. 1971. Stochastic approximation, optimization methods in statistics, Edited by J.S.Rustagi, Academic Press, New York.

[7] Kleijnen, J.P.C. 1995. Sensitivity analysis and optimization in simulation: design of experiments and case studies. Proceedings of the 1995 Winter Simulation Conference.

[8] Kleijnen, J.P.C. 1993. Simulation and optimization in production planning: a case study. Decision Support Systems 9: 269-280.

[9] Plackett, R.L., and Burman, J.P. 1946. The design of optimum multifactor experiments, Biometrika, 33: 305-325.

[10] Ahuja, S.K., Ferreira, G.M. and Moreira, A.R. 2004. Application of Plackett-Burman design and response surface methodology to achieve exponential growth for aggregated shipworm bacterium. Biotechnology and Bioengineering, vol. 85, issue 6, pages 666-675.

[11] Fu, M.C. 2001. Simulation optimization. In Encyclopedia of Operations Research and Management Science, 2nd edition, ed. S. Gass and C. Harris, 756-759. Boston: Kluwer Academic Publishers.

[12] Fu, M.C. 2001. Simulation optimization, Proceedings of the 2001 Winter Simulation Conference.

[13] Myers, R.H., Khuri. A., and Carter. W. 1989. Response Surface Methodology: 1966-1988. Technometrics, (31): 137-157.

[14] Box, G.E.P. and Draper, N.R. 1975. Robust designs. Biometrika, 62, 347-352.

[15] Box, G.E.P. and Wilson, K.B. 1951. On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society, series B, 13, 1-45.

[16] Smith, K. 1918. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. Biometrika, 12, 1-85.

[17] Kiefer, J. and Wolfowitz, J. 1959. Optimum designs in regression problems. The Annals of Mathematical Statistics, 30, 271-294.

[18] Wald, A. 1943. On the efficient design of statistical investigations. The Annals of Mathematical Statistics, 14, 134-140.

[19] Kiefer, J. and Wolfowitz, J. 1960. The equivalence of two extremum problems. Canadian Journal of Mathematics, 12, 363-366.

[20] Fedorov, V.V. 1972. Theory of Optimal Experiments. Translated and edited by W.J. Studden, and E.M. Klimko, Academic Press, New York.

[21] Box, G.E.P., Hunter, W.G., and Hunter, J.S. 1978. Statistics for Experimenters, New York: John Wiley.

[22] Steinberg, D.M. and Hunter, W.G. 1984. Experimental design: review and comment. Technometrics, vol.26, no.2, 71-97.

[23] Box, G.E.P. and Draper, N.R. 1959. A basis for the selection of a response surface design. Journal of the American Statistical Association, 54, 622-654.

[24] Box, G.E.P. and Draper, N.R. 1963. The choice of a second order rotatable design. Biometrika, 50, 335-352.

[25] Donohue, J.M., Houck, E.C. and Myers, R.H. 1993. Simulation designs and correlation induction for reducing second-order bias in first-order resposne surfaces. Operations Research, Vol. 41, No. 5, 880-902.

[26] Schruben, L.W. and Margolin, B.H. 1978. Pseudorandom number assignment in statistically designed simulation and distribution sampling experiments. Journal of the American Statistical Association, Vol. 73, 363, 504-520.

[27] Draper, N.R. and Herzberg, A.M. 1973. Some designs for extrapolation outside a sphere. Journal of the Royal Statistical Society, Series B, Vol. 35, No. 2, 268-276.

[28] Draper, N.R. and Herzberg, A.M. 1979. An investigation of first-order and second-order designs for extrapolation outside a hypersphere. The Canadian Journal of Statistics, Vol. 7, No. 1, 97-101.

[29] Wiens, D.P. 1998. Minimax robust designs and weights for approximately specific regression models with heteroscedastic errors. Journal of the American Statistical Association, 93, 1440-1450.

[30] Fang, Z.D. and Wiens, D.P. 1999. Robust extrapolation designs and weights for biased regression models with heteroscedastic errors. The Canadian Journal of Statistics, Vol. 27, No. 4, 751-770.

[31] Mee, R.W. and Peralta, M. 2000. Semifolding 2^{k-p} Designs. Technometrics, 42, pp. 122-134.

[32] Barnett, J., Czitrom, V., John, P.W.M. and Leon, R.V. 1997. Using fewer wafers to resolve confounding in screening experiments. Statistical Case Studies for Industrial Process Improvement, edited by V.Czitrom and P.D. Spagon, SIAM, Philadelphia, PA., pp.235-250.

[33] Chipman, H. and Hamada, M.S. 1996. Discussion: factor-based or effect-based modeling? Implications for design. Technometrics, 38, pp. 317-320.

[34] Nelson, B.J., Montgomery, D.C. Elias, R.J. and Maass, E. 2000. A comparison of several design augmentation strategies. Quality and Reliability Engineering International 16, pp. 435-449.

[35] Dumochel, W. and Jones, B. 1994. A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model. Technometrics, 43, pp. 682-688.

[36] Chanoler, K. and Verdinelli, I. 1995. Bayesian experimental design: a review. Statistical Science, 10, pp. 273-304.

[37] Lin, H.F., Myers, R.H. and Ye, K.Y. 2000. Bayesian two-stage optimal design for mixture models. Journal of Statistical Computation and Simulation, 66, pp. 209-231.

[38] Goldsman, D. and Nelson, B.L. 1994. Ranking, selection and multiple comparisons in computer simulation. In Proceedings of the 1994 Winter Simulation Conference J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, Eds. IEEE, Piscataway, N.J., 192-199.

[39] Chen, C.H. 1995. An effective approach to smartly allocate computing budget for discrete event simulation. Proceedings of the 34th IEEE Conference on Decision and Control.

[40] Chen, C.H. 1996. A lower bound for the correct subset-selection probability and its application to discrete event system simulations. IEEE Transactions on Automatic Control 41(8).

[41] Chen, H.C., Chen, C.H., Dai, L., and Yücesan, E. 1997. New development of optimal computing budget allocation for discrete event simulation. Proceedings of the 1997 Simulation Conference.

[42] Chen, C.H., Yuan Y., Chen, H.C., Yücesan, E. and Dai, L. 1998. Computing budget allocation for simulation experiments with different system structures, Proceedings of the 1998 Winter Simulation Conference.

[43] Chen, C.H., Lin, J., Yücesan, E. and Chick, S.E. 2000. Simulation budget allocation for further enhancing the efficiency of ordinal optimization, Journal of Discrete Event Dynamic Systems: Theory and Applications, Vol.10, pp.251-270.

[44] Chen, H.C., Chen, C.H. and Yücesan, E. 2000. Computing efforts allocation for ordinal optimization and discrete event simulation, IEEE Transactions on Automatic Control 45(5).

[45] Lawson, C., Keats, J.B. and Montgomery, D.C. 1997. Comparison of robust and least-squares regression in computer-generated probability plots, IEEE Transactions on Reliability 46(1).

[46] Devore, J.L. 2000. Probability and statistics for engineering and the sciences, 5th ed. pp.366-367. Duxbury.

Appendix A One-dimensional case

We study the general solution when the ASA technique is applied to the one-dimensional linear model, and investigate why and how the step size will be infinity.

The linear model is:

$$y = \beta_0 + \beta_1 d + \epsilon \qquad \epsilon \sim i.i.d.N(0, \sigma^2),$$

where y is the response, d is the regressor variable, and β_0 and β_1 are the unknown coefficients.

Assume the two levels of the regressor variable d are d_1 and d_2, and n_1 and n_2 runs are allocated to d_1 and d_2 respectively. Hence we have the observations y_{11}, y_{12}, $\cdots$, y_{1n_1}, and y_{21}, y_{22}, $\cdots$, y_{2n_2}. The regression model and OLS solution become:

$$\begin{pmatrix} \mathbf{Y_1} \\ \mathbf{Y_2} \end{pmatrix} = \mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \varepsilon = \begin{pmatrix} \boldsymbol{I_1} & \mathbf{D_1} \\ \boldsymbol{I_2} & \mathbf{D_2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \varepsilon,$$

with

$$\mathbf{Y_1} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \end{pmatrix}_{n_1 \times 1} \qquad \mathbf{Y_2} = \begin{pmatrix} y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \end{pmatrix}_{n_2 \times 1}$$

$$\boldsymbol{I_1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n_1 \times 1} \qquad \mathbf{D_1} = \begin{pmatrix} d_1 \\ d_1 \\ \vdots \\ d_1 \end{pmatrix}_{n_1 \times 1} \qquad \boldsymbol{I_2} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n_2 \times 1} \qquad \mathbf{D_2} = \begin{pmatrix} d_2 \\ d_2 \\ \vdots \\ d_2 \end{pmatrix}_{n_2 \times 1}$$

The OLS estimates are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \frac{1}{n_1 n_2 (d_1 - d_2)^2} \begin{pmatrix} \sum_{i=1}^{n_1} y_{1i} n_2 d_2 (d_1 - d_2) - \sum_{j=1}^{n_2} y_{2j} n_1 d_1 (d_1 - d_2) \\ \sum_{i=1}^{n_1} y_{1i} n_2 (d_1 - d_2) - \sum_{j=1}^{n_2} y_{2j} n_1 (d_1 - d_2) \end{pmatrix}.$$

Therefore,

$$\hat{\beta}_1 = \frac{1}{n_1 n_2 (d_1 - d_2)^2} (n_2 (d_1 - d_2) \sum_{i=1}^{n_1} y_{1i} - n_1 (d_1 - d_2) \sum_{j=1}^{n_2} y_{2j}) = \frac{1}{d_1 - d_2} \left(\frac{\sum_{i=1}^{n_1} y_{1i}}{n_1} - \frac{\sum_{j=1}^{n_2} y_{2j}}{n_2} \right)$$

and

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n_1n_2(d_1-d_2)^2}\begin{pmatrix} n_1d_1^2+n_2d_2^2 & -(n_1d_1+n_2d_2) \\ -(n_1d_1+n_2d_2) & n_1+n_2 \end{pmatrix} = \begin{pmatrix} a & b \\ b & C \end{pmatrix}.$$

The lower bound of the one-sided $1-\alpha$ confidence interval for the predictor $\hat{y}$ at d is

$$\hat{y}_{min} = \hat{\beta}_0 + \hat{\beta}_1 d - t^{\alpha}_{N-2}\hat{\sigma}\sqrt{(1\ d)(\mathbf{X}^T\mathbf{X})^{-1}\begin{pmatrix} 1 \\ d \end{pmatrix}}$$

where t^{α}_{N-2} denotes the $1-\alpha$ quantile of the t distribution with $N-2$ degrees of freedom and $N = n_1 + n_2$.

Because $\hat{y}_{min}$ is concave in d, we can find d^+ which is the point that maximizes $\hat{y}_{min}$.

$$\frac{\partial \hat{y}_{min}}{\partial d}|_{d^+} = \hat{\beta}_1 - \frac{t^{\alpha}_{N-2}\hat{\sigma}}{\sqrt{a + 2bd^+ + Cd^{+2}}}\cdot(b + Cd^+) = 0.$$

Solving for d^+, we get

$$d^+ = -\frac{b}{C} + \frac{\hat{\beta}_1}{C}\sqrt{\frac{aC - b^2}{t^{\alpha}_{N-2}{}^2\hat{\sigma}^2 C - \hat{\beta}_1^2}}.$$

Substitute a, b and C into d^+, then

$$d^+ = \frac{n_1d_1 + n_2d_2}{n_1+n_2} + \hat{\beta}_1\frac{n_1n_2(d_1-d_2)^2}{n_1+n_2}\sqrt{\frac{1}{(n_1+n_2)t^{\alpha}_{N-2}{}^2\hat{\sigma}^2 - \hat{\beta}_1^2 n_1n_2(d_1-d_2)^2}}.$$

From the above formula, we can know

starting point	$\frac{n_1d_1+n_2d_2}{n_1+n_2}$
ASA direction	$\hat{\beta}_1\frac{n_1n_2(d_1-d_2)^2}{n_1+n_2}$
step size	$\sqrt{\frac{1}{(n_1+n_2)t^{\alpha}_{N-2}{}^2\hat{\sigma}^2 - \hat{\beta}_1^2 n_1n_2(d_1-d_2)^2}}$

From the formula of step size, it is easy to see that if noise σ is too small or the gradient β_1 is too big, the step size is infinite. Intuitively, if the noise is too small, the linear regression model becomes a deterministic linear model, and the maximum point is at infinity, so that the step size can be very large. When the gradient is very steep, it is reasonable to believe that the maximal point is far away from the current region, and hence the step size can also be very large.

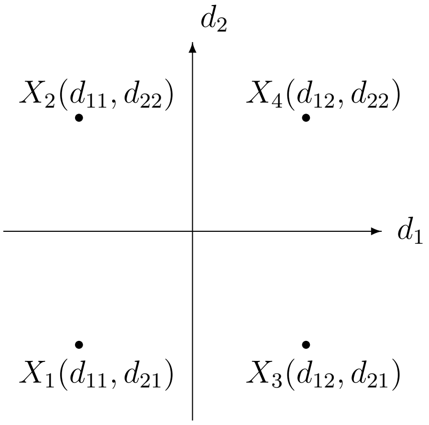
Appendix B Examples for unequal allocation in two-dimensional case

We study into how we would allocate the simulation runs when the ASA direction happens to be four special directions, and we conclude that equal allocation might not be the best choice when d^+ is at infinity for the two-dimensional case.

The regression model is assumed to be

$$y = \beta_0 + \beta_1 \cdot d_1 + \beta_2 \cdot d_2 + \epsilon \quad \epsilon \sim i.i.d.N(0, \sigma^2).$$

In the 2^2 factorial design, the two levels of d_1 are d_{11} and d_{12}, and the two levels of d_2 are d_{21} and d_{22}, which are all fixed in the region of experimentation. Thus the four design points are X_1, X_2, X_3 and X_4, and there are n_1, n_2, n_3 and n_4 runs done at each design point respectively.



Then it is easy to know that the design matrix is

$$\mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{D_1} \\ \mathbf{1} & \mathbf{D_2} \\ \mathbf{1} & \mathbf{D_3} \\ \mathbf{1} & \mathbf{D_4} \end{pmatrix},$$

where

$$\mathbf{D_1} = \begin{pmatrix} d_{11} & d_{21} \\ d_{11} & d_{21} \\ \vdots & \vdots \\ d_{11} & d_{21} \end{pmatrix}_{n_1\times 1} \qquad \mathbf{D_2} = \begin{pmatrix} d_{11} & d_{22} \\ d_{11} & d_{22} \\ \vdots & \vdots \\ d_{11} & d_{22} \end{pmatrix}_{n_2\times 1}$$

$$\mathbf{D_3} = \begin{pmatrix} d_{12} & d_{21} \\ d_{12} & d_{21} \\ \vdots & \vdots \\ d_{12} & d_{21} \end{pmatrix}_{n_3\times 1} \qquad \mathbf{D_4} = \begin{pmatrix} d_{12} & d_{22} \\ d_{12} & d_{22} \\ \vdots & \vdots \\ d_{12} & d_{22} \end{pmatrix}_{n_4\times 1}$$

The observations are given as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y_1} \\ \mathbf{Y_2} \\ \mathbf{Y_3} \\ \mathbf{Y_4} \end{pmatrix},$$

where

$$\mathbf{Y_1} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \end{pmatrix}_{n_1\times 1} \quad \mathbf{Y_2} = \begin{pmatrix} y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \end{pmatrix}_{n_2\times 1} \quad \mathbf{Y_3} = \begin{pmatrix} y_{31} \\ y_{32} \\ \vdots \\ y_{3n_3} \end{pmatrix}_{n_3\times 1} \quad \mathbf{Y_4} = \begin{pmatrix} y_{41} \\ y_{42} \\ \vdots \\ y_{4n_4} \end{pmatrix}_{n_4\times 1}$$

The OLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

and given (d_1, d_2), the predictor $\hat{y}$ is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 d_1 + \hat{\beta}_2 d_2.$$

The lower bound of the one-sided $1-\alpha$ confidence interval for the predictor $\hat{y}$ at $(d_1\ d_2)$ is

$$\hat{y}_{min} = \hat{\beta}_0 + \hat{\beta}_1 d_1 + \hat{\beta}_2 d_2 - t^{\alpha}_{N-3}\hat{\sigma}\sqrt{\begin{pmatrix} 1 & d_1 & d_2 \end{pmatrix}(\mathbf{X}^T\mathbf{X})^{-1}\begin{pmatrix} 1 \\ d_1 \\ d_2 \end{pmatrix}}$$

where t_{N-3}^{α} denotes the $1-\alpha$ quantile of the t distribution with $n_1+n_2+n_3+n_4-3$ degrees of freedom and $(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{det}\begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix} = \begin{pmatrix} a & \mathbf{b}^T \\ \mathbf{b} & \mathbf{C} \end{pmatrix}$ with

$$\begin{aligned}
det &= (d_{11}-d_{12})^2(d_{21}-d_{22})^2(n_2n_3n_4+n_1n_3n_4+n_1n_2n_3+n_1n_2n_4) \\
b &= (d_{22}(d_{11}n_1+d_{12}n_3)(n_2+n_4)-d_{21}(d_{11}n_2+d_{12}n_4)(n_1+n_3))\,(d_{21}-d_{22}) \\
c &= (d_{12}(d_{21}n_1+d_{22}n_2)(n_3+n_4)-d_{11}(d_{21}n_3+d_{22}n_4)(n_1+n_2))(d_{11}-d_{12}) \\
d &= (n_1+n_3)(n_2+n_4)(d_{21}-d_{22})^2 \\
e &= (n_2n_3-n_1n_4)(d_{11}-d_{12})(d_{21}-d_{22}) \\
f &= (n_1+n_2)(n_3+n_4)(d_{11}-d_{12})^2
\end{aligned}$$

Simplifying $\hat{y}_{min}$, we get

$$\hat{y}_{min} = \hat{\beta}_0+\hat{\beta}_1d_1+\hat{\beta}_2d_2-t_{N-3}^{\alpha}\hat{\sigma}\sqrt{(a+2bd_1+2cd_2+2ed_1d_2+dd_1^2+fd_2^2)/det}.$$

The next design point $\mathbf{d}^+$ is

$$\mathbf{d}^+ = -\mathbf{C}^{-1}\mathbf{b}+\lambda\mathbf{C}^{-1}\hat{\boldsymbol{\beta}}_{-0},$$

and the starting point $-\mathbf{C}^{-1}\mathbf{b}$ can be simplified as

$$-\mathbf{C}^{-1}\mathbf{b} = \begin{pmatrix} \dfrac{d_{11}(n_1+n_2)+d_{12}(n_3+n_4)}{n_1+n_2+n_3+n_4} \\ \dfrac{d_{21}(n_1+n_3)+d_{22}(n_2+n_4)}{n_1+n_2+n_3+n_4} \end{pmatrix}.$$

If it is the equal allocation $n_1=n_2=n_3=n_4$, then the starting point is the center of region of experimentation

$$-\mathbf{C}^{-1}\mathbf{b} = \begin{pmatrix} \dfrac{d_{11}+d_{12}}{2} \\ \dfrac{d_{21}+d_{22}}{2} \end{pmatrix}.$$

Now we consider when d^+ is at infinity and given the same estimators β_0, β_1, β_2 and σ, how to select n_1, n_2, n_3 and n_4 so that the $\hat{y}_{min}$ of this allocation will dominate that of any other allocations along the following four special directions:

1. d_1 is fixed and $d_2 \to \infty$

2. d_2 is fixed and $d_1 \to \infty$

3. $d_1 = d_2 = r \to \infty$

4. $d_1 = -d_2 = r \to \infty$

Case 1.

$$\hat{y}_{min}|_{d_1=h,\ d_2\to\infty} = \hat{\beta}_2 d_2 - t^{\alpha}_{N-3}\hat{\sigma}\sqrt{\frac{f}{det}}|d_2|$$

where $\dfrac{f}{det} = \dfrac{(n_1+n_2)(n_3+n_4)}{(n_2n_3n_4+n_1n_3n_4+n_1n_2n_3+n_1n_2n_4)} \cdot \dfrac{1}{(d_{21}-d_{22})^2}$ and h denotes a fixed but finite value.

Given $\hat{\beta}_2$ and $\hat{\sigma}$, the dominating allocation has the minimal value of $\dfrac{f}{det}$. Because $f > 0$, $det > 0$ and $n_1 + n_2 + n_3 + n_4 = N$, to solve this mathematical programming problem, we get when $n_1 = n_2$ and $n_3 = n_4$, $min(\dfrac{f}{det}) = \dfrac{2}{N} \cdot \dfrac{1}{(d_{21}-d_{22})^2}$ and $\hat{y}_{min}|_{d_1=h,\ d_2\to\infty}$ is maximized.

Case 2.

$$\hat{y}_{min}|_{d_1\to\infty, d_2=h} = \hat{\beta}_1 d_1 - t^{\alpha}_{N-3}\hat{\sigma}\sqrt{\frac{d}{det}}|d_1|$$

where $\dfrac{d}{det} = \dfrac{(n_1+n_3)(n_2+n_4)}{(n_2n_3n_4+n_1n_3n_4+n_1n_2n_3+n_1n_2n_4)} \cdot \dfrac{1}{(d_{11}-d_{12})^2}$ and h denotes a fixed but finite value.

Given $\hat{\beta}_1$ and $\hat{\sigma}$, the dominating allocation has the minimal value of $\dfrac{d}{det}$.

Because $d > 0$, $det > 0$ and $n_1 + n_2 + n_3 + n_4 = N$, to solve this mathematical programming problem, we get when $n_1 = n_3$ and $n_2 = n_4$, $min(\dfrac{d}{det}) = \dfrac{2}{N} \cdot \dfrac{1}{(d_{11}-d_{12})^2}$ and $\hat{y}_{min}|_{d_1\to\infty, d_2=h}$ is maximized.

Case 3.

$$\hat{y}_{min}|_{r\to\infty} = (\hat{\beta}_1 + \hat{\beta}_2)r - t^{\alpha}_{N-3}\hat{\sigma}\sqrt{(\frac{2e+d+f}{det})}|r|$$

where

$$\begin{aligned} &\frac{2e+d+f}{det} \\ = \; &\frac{2(n_2n_3-n_1n_4)(d_{11}-d_{12})(d_{21}-d_{22})}{(n_2n_3n_4+n_1n_3n_4+n_1n_2n_3+n_1n_2n_4)(d_{11}-d_{12})^2(d_{21}-d_{22})^2} + \\ &\frac{(n_1+n_3)(n_2+n_4)(d_{21}-d_{22})^2+(n_1+n_2)(n_3+n_4)(d_{11}-d_{12})^2}{(n_2n_3n_4+n_1n_3n_4+n_1n_2n_3+n_1n_2n_4)(d_{11}-d_{12})^2(d_{21}-d_{22})^2}. \end{aligned}$$

When $d_{11} - d_{12} = d_{21} - d_{22}$, then

$$\frac{2e+d+f}{det} = \frac{4n_2n_3 + n_1n_2 + n_3n_4 + n_1n_3 + n_2n_4}{(n_2n_3n_4 + n_1n_3n_4 + n_1n_2n_3 + n_1n_2n_4)} \cdot \frac{1}{(d_{11} - d_{12})^2}.$$

Because X_2 and X_3 are symmetric along this direction, we assume $n_2 = n_3 = N_1$, then

$$\frac{2e+d+f}{det} = \frac{4N_1 + 2(N - N_1)}{(N - N_1)N_1 + 2n_1n_4} \cdot \frac{1}{(d_{11} - d_{12})^2}.$$

To find the dominating allocation is equivalent to minimize $\dfrac{2e+d+f}{det}$.
When N_1 is fixed, it is easy to know when $n_1 = n_4 = \dfrac{N - 2N_1}{2}$, $\dfrac{2e+d+f}{det}$ is minimized.
As a result, when $n_2 = n_3 = N_1 = 0$ and $n_1 = n_4 = N/2$, $min(\dfrac{2e+d+f}{det}) = \dfrac{4}{N} \cdot \dfrac{1}{(d_{11} - d_{12})^2}$ and $\hat{y}_{min}|_{r \to \infty}$ is maximized.

Case 4. Similar to case 3, when $n_1 = n_4 = 0$ and $n_2 = n_3 = N/2$, $min(\dfrac{-2e+d+f}{det}) = \dfrac{4}{N} \cdot \dfrac{1}{(d_{21} - d_{22})^2}$ and $\hat{y}_{min}|_{r \to \infty}$ is maximized.

Conclusion: from the above four special cases, we can conclude that given a particular direction, equal allocation might not be the optimal allocation which makes the lower bound of predictor dominate the other lower bounds of any other allocations. Moreover, given a direction, we can determine the dominating allocation by using mathematical programming.

Appendix C The dominating allocation for one-dimensional model

Theorem 1 *For one-dimensional linear model, given the same estimates of β_i and σ^2, when the selected design point* d *approaches infinity, the lower bound of predicted response of equal allocation always dominates the lower bound of predicted response of any other allocations.*

Proof. Since this is a one-dimensional linear model, there are only two design points. Assume the two levels of local design point to be d_1 and d_2, n_1 runs and n_2 runs are allocated to d_1 and d_2 respectively and there are a total of N runs. For equal allocation, $n_1 = n_2 = N/2$; for unequal allocation, $n'_1 \neq n'_2$ and $n'_1 + n'_2 = N$.
Denote $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$ to be the estimates of β_0, β_1 and σ^2 respectively. Then the response at the given point d is

$$\hat{y}(d) = \hat{\beta}_0 + \hat{\beta}_1 d.$$

The lower bound of predicted response at d can be expressed as

$$\hat{y}_{min}(d) = \hat{y}(d) - t^{\alpha}_{N-q}\sqrt{var(y|d)} = \hat{\beta}_0 + \hat{\beta}_1 d - t^{\alpha}_{N-q}\sqrt{\begin{pmatrix} 1 & d \end{pmatrix} (\mathbf{X}^T\mathbf{X})^{-1} \begin{pmatrix} 1 \\ d \end{pmatrix} \cdot \hat{\sigma}^2}$$

where $(\mathbf{X}^T\mathbf{X})^{-1} = \dfrac{1}{n_1 n_2 (d_1 - d_2)^2} \begin{pmatrix} n_1 d_1^2 + n_2 d_2^2 & -(n_1 d_1 + n_2 d_2) \\ -(n_1 d_1 + n_2 d_2) & n_1 + n_2 \end{pmatrix}$.
Therefore $\hat{y}_{min}(d)$ can be simplified as

$$\hat{y}_{min}(d) = \hat{\beta}_0 + \hat{\beta}_1 d - t^{\alpha}_{N-q}\sqrt{\frac{(n_1+n_2)d^2 + (-2n_1 d_1 - 2n_2 d_2)d + (n_1 d_1^2 + n_2 d_2^2)}{n_1 n_2 (d_1 - d_2)^2} \cdot \hat{\sigma}^2}.$$

When $d \to \infty$,

$$\hat{y}_{min}(d) = \hat{\beta}_0 + \hat{\beta}_1 d - t^{\alpha}_{N-q}\sqrt{\frac{n_1+n_2}{n_1 n_2 (d_1 - d_2)^2} \cdot \hat{\sigma}^2}|d| = \hat{\beta}_0 + \hat{\beta}_1 d - t^{\alpha}_{N-q}\sqrt{\frac{N}{n_1 n_2 (d_1 - d_2)^2} \cdot \hat{\sigma}^2}|d|.$$

Compare the $\hat{y}_{min}(d)$ of equal allocation with the $\hat{y'}_{min}(d)$ of unequal allocation given the selected design point d and the same estimates of β_i and σ^2:

$$\begin{aligned}
\hat{y}_{min}(d) - \hat{y'}_{min}(d) &= \hat{\beta}_0 + \hat{\beta}_1 d - t^{\alpha}_{N-q}\sqrt{\frac{N}{n_1 n_2 (d_1 - d_2)^2} \cdot \hat{\sigma}^2 |d|} \\
&\quad - \left(\hat{\beta}_0 + \hat{\beta}_1 d - t^{\alpha}_{N-q}\sqrt{\frac{N}{n'_1 n'_2 (d_1 - d_2)^2} \cdot \hat{\sigma}^2 |d|} \right) \\
&= t^{\alpha}_{N-q}\sqrt{\frac{N}{(d_1 - d_2)^2}\hat{\sigma}^2 |d|} \left(\sqrt{\frac{1}{n'_1 n'_2}} - \sqrt{\frac{1}{n_1 n_2}} \right)
\end{aligned}$$

$\because n_1 n_2 > n'_1 n'_2 \quad \therefore \sqrt{\frac{1}{n'_1 n'_2}} - \sqrt{\frac{1}{n_1 n_2}} > 0$

As a result, $\hat{y}_{min}(d)$ is always greater than $\hat{y'}_{min}(d)$ when $d \to \infty$. □