TWO STAGE ADAPTIVE OPTIMAL DESIGN WITH APPLICATIONS TO DOSE-FINDING CLINICAL TRIALS

A Dissertation examination presented to the Faculty of the Graduate School at the University of Missouri

In Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

by

Adam Lane

Dr. Nancy Flournoy, Dissertation Supervisor

MAY 2013
The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

TWO STAGE ADAPTIVE OPTIMAL DESIGN WITH APPLICATIONS TO DOSE-FINDING CLINICAL TRIALS

presented by Adam Lane,
a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

____________________________________
Dr. Nancy Flournoy

____________________________________
Dr. Sounak Chakraborty

____________________________________
Dr. Wade Davis

____________________________________
Dr. Zhuoqiong (Chong) He

____________________________________
Dr. Jianguo (Tony) Sun
To my wife, LeAnna, and daughter, Mae Pearl
ACKNOWLEDGMENTS

This work could not have been accomplished without the assistance and support of my advisor, Dr. Nancy Flournoy. I greatly appreciate all that she has done to make this possible. I owe a great deal to her instruction and dedication to myself and my research topic.

I would also like to acknowledge Dr. Jianguo (Tony) Sun for his supervision on the NIH training grant for three semesters. This opportunity allowed me to focus on research without the time commitment of teaching.

I would also like to thank Dr. Rosemary Bailey and the Isaac Newton Institute for Mathematical Sciences for hosting me during the 2011 fall semester. A considerable amount of research for this dissertation was done during this time.

Dr. Valerii Fedorov’s many helpful research comments and suggestions were also greatly appreciated.

I am grateful to Dr. Sounak Chakraborty, Dr. Wade Davis, Dr. Zhuoqiong (Chong) He, and Dr. Jianguo (Tony) Sun for serving on my committee.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ................................................................. ii
LIST OF TABLES ........................................................................ vi
LIST OF FIGURES ....................................................................... viii
ABSTRACT ................................................................................... xiii

CHAPTER

1 Introduction ................................................................. 1
   1.1 Motivating Problem ......................................................... 1
   1.2 Optimal Design ............................................................. 6
      1.2.1 Background ............................................................. 6
      1.2.2 General Details ....................................................... 8
      1.2.3 Optimal Design Criteria ......................................... 10
      1.2.4 General Equivalence Theorem ................................. 14
      1.2.5 Design Efficiency ................................................... 14
      1.2.6 Adaptive Designs in Dose-Finding Clinical Trials ........ 15

2 Preamble ................................................................................. 27
   2.1 A Brief Review of Maximum Likelihood Theory for Nonlinear Regression with Independent Normal Errors .................. 28
      2.1.1 Normality of $\hat{\theta}_n$ .............................................. 29
   2.2 A Fixed One Point Design for a Regression Model with an Exponential Mean Function .............................................. 31
3 Information in a Two Stage Adaptive Optimal Design .......... 38

3.1 Information Bound in a Two-Stage Experiment ............... 38

3.2 The Model and the Adaptive Procedure ................... 40

3.2.1 The Adaptive Stage 2 Treatment .......................... 41

3.2.2 Fisher’s Information ....................................... 43

3.3 Optimal Selection of the Stage 1 Sample Size ............... 44

3.3.1 Approximation of the locally optimal stage one sample size .... 46

4 Inference in a Two Stage Adaptive Optimal Design .......... 51

4.1 Information Alternatives ..................................... 51

4.2 Example: An Exponential Mean Function ..................... 55

4.3 The Final MLE $\hat{\theta}_n$ of $\theta$ .......................... 57

4.3.1 Some Comparisons of Estimators of $\theta$ .................. 61

4.3.2 Comparison of $n\text{Var}[\hat{\theta}_n]$ to its approximations $M^{-1}$ and $M^{*-1}$. 69

4.3.3 Comparison of $n\text{Var}[\hat{\theta}_n]$ to its approximations $M^{-1}$ and $M_{ind}^{-1}$. 69

4.3.4 Comparison of estimates of information $\hat{M}^{-1}$, $\hat{M}_{ind}^{-1}$ and $\hat{M}_{obs}^{-1}$. 72

5 Two-Stage Adaptive Optimal Design with Fixed First Stage Sample Size ........................................ 81

5.1 Asymptotic Properties ....................................... 81

5.1.1 Large Stage 1 and Stage 2 Sample Sizes .................. 82

5.1.2 Distribution of the MLE if Only Second Stage Data are Considered ........................................ 82

5.1.3 Fixed First Stage Sample Size; Large Second Stage Sample Size 83

5.2 Example: One Parameter Exponential Mean Function ........ 84
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Probability of a Boundary Value, Variance, Bias, and Lower Bound for the MLE of a Fixed One Point Design. Values $\theta = 1.0$, $x = 2.0$, $\sigma = 1$ and $\bar{\theta} = 4.0$ were used.</td>
</tr>
<tr>
<td>4.1</td>
<td>Summary of notation for information measures and their estimates.</td>
</tr>
<tr>
<td>5.1</td>
<td>Integrated absolute difference of the cumulative distributions ($\times 100$) of $T_1 \sim N(0, [M(\xi^*, \theta)]^{-1})$, $T_2 \sim N(0, [M(\xi_A, \theta)]^{-1})$ and $T_3 \sim UQ$ versus the approximate cumulative distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ obtained via Monte Carlo simulations for various $n_1$ and various moderate sizes of $n$. Values $\theta = 1$, $x_1 = 2$, $\sigma = 0.5$, $a = 0.25$ and $b = 4$ were used.</td>
</tr>
<tr>
<td>5.2</td>
<td>Integrated absolute difference of the cumulative distributions ($\times 100$) of $T_1 \sim N(0, [M(\xi^*, \theta)]^{-1})$, $T_2 \sim N(0, [M(\xi_A, \theta)]^{-1})$ and $T_3 \sim UQ$ versus the approximate cumulative distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ obtained via Monte Carlo simulations for various $n_1$ and various large sizes of $n$. Values $\theta = 1$, $x_1 = 2$, $\sigma = 0.5$, $a = 0.25$ and $b = 4$ were used.</td>
</tr>
</tbody>
</table>
5.3 Quantiles from the distribution of $\hat{\theta}_n$ obtained via a Monte Carlo simulation are represented as “True”. Median quantiles for $T_2 \sim \mathcal{N}(0, [M(\xi_A, \theta)]^{-1})$ and $T_3 \sim UQ$ are presented for comparison. $x_1 = 2, \sigma = 0.5, a = 0.25$ and $b = 4$ were used.
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Example of underlying objectives functions in phase I, phase II and phase I/II, $p_1(x, \theta)$, $p_1(x, \theta)$ and $p_{01}(x, \theta)$, respectively. Figure courtesy of Fedorov and Wu. The solid lines connecting each curve to the $x$-axis represent potential targets for each phase the MTD$<em>{33}$, the ED$</em>{80}$ and the $\arg\max_x p_{10}(x, \theta)$.</td>
<td>18</td>
</tr>
<tr>
<td>2.1 The approximation $\sqrt{n M \left(x, \hat{\theta}_n\right)}^{-1}$ (Dotted line) and $\sqrt{\text{Var} \left[\hat{\theta}_n\right]}$ (solid line) plotted as a functions of $\hat{\theta}_n$. Values $n = 100$, $\theta = 1.0$, $x = 2.0$, $\sigma = 1$ and $\bar{\theta} = 4.0$ were used.</td>
<td>35</td>
</tr>
<tr>
<td>2.2 The approximation $\sqrt{n M \left(x, \hat{\theta}_n\right)}^{-1}$ (Dotted line) and $\sqrt{\text{Var} \left[\hat{\theta}_n\right]}$ (solid line) plotted as a functions of $\hat{\theta}_n$. Values $n = 1600$, $\theta = 1.0$, $x = 2.0$, $\sigma = 1$ and $\bar{\theta} = 4.0$ were used.</td>
<td>36</td>
</tr>
<tr>
<td>3.1 Optimal allocation to stage 1 by $x_1 \in [a, b] = [.25, 10]$. The dashed, solid and dot-dashed lines represent $[n_1^*(\theta)</td>
<td>\theta]$ for $\theta = 0.2, 0.5$ and 1, respectively, using $n = 100$ and $\sigma = 1$. The three dotted vertical dotted lines represent the locally corresponding locally optimal design points for the given $\theta$ values, $x^<em>(0.2) = 5$, $x^</em>(0.5) = 2$, $x^*(1) = 1$.</td>
</tr>
</tbody>
</table>
3.2 Ratio of the average expected information from an experiment with \( n_1 = \{1, \ldots, n\} \) to average expected information from an experiment with \( n_1 = n_1^* \) (solid line). Ratio of average expected information with \( n_1 = \sqrt{n} = 10 \) to average expected information from an experiment with \( n_1 = n_1^* \) (dashed line). Dotted line at 1.0 is provided and is used represent the average expected information from an experiment with \( n_1 = n_1^* \) since it is the value used in the denominator.

4.1 Histogram of \( \bar{y}_1 \). Mean Responses from 10,000 simulations of Model (3.3) for \( n = 100, \theta = 1.0, \bar{\theta} = 0.1, \tilde{\theta} = 4.0, x_1 = 2 \) and \( w_1 = 0.2 \).

4.2 Histogram of \( \bar{y}_2 \) from Stage 2. Mean Responses from 10,000 simulations of Model (3.3) for \( n = 100, \theta = 1.0, \bar{\theta} = 0.1, \tilde{\theta} = 4.0, x_1 = 2 \) and \( w_1 = 0.2 \).

4.3 Histogram of \( \hat{\theta}_{ss} \) from a single stage fixed design. Parameter Estimates from 10,000 simulations of Model (3.3) for \( n = 100, \theta = 1.0, \bar{\theta} = 0.1, \tilde{\theta} = 4.0, x_1 = 2 \) and \( w_1 = 0.2 \).

4.4 Histogram of \( \hat{\theta}_{n2} \) from Stage 2. Parameter Estimates from 10,000 simulations of Model (3.3) for \( n = 100, \theta = 1.0, \bar{\theta} = 0.1, \tilde{\theta} = 4.0, x_1 = 2 \) and \( w_1 = 0.2 \).

4.5 Histogram of \( \hat{\theta}_n \) from a two-stage adaptive procedure. Parameter Estimates from 10,000 simulations of Model (3.3) for \( n = 100, \theta = 1.0, \bar{\theta} = 0.1, \tilde{\theta} = 4.0, x_1 = 2 \) and \( w_1 = 0.2 \).
4.6 Smooth density histograms of $\hat{\theta}_{ss}$ (solid line), $\hat{\theta}_{n2}$ (dotted line) and $\hat{\theta}_n$ (dashed line). Parameter Estimates from 10,000 simulations of model (3.3) for $n = 100$, $\theta = 1.0$, $\bar{\theta} = 0.1$, $\bar{\theta} = 4.0$, $x_1 = 2$ and $w_1 = 0.2$. The median, mean and variance of simulated values are shown at the top of each histogram. ................................. 66

4.7 Smooth density histograms of $\hat{\theta}_{ss}$ (solid line), $\hat{\theta}_{n2}$ (dotted line) and $\hat{\theta}_n$ (dashed line). Parameter Estimates from 10,000 simulations of model (3.3) for $n = 1600$, $\theta = 1.0$, $\bar{\theta} = 0.1$, $\bar{\theta} = 4.0$, $x_1 = 2$ and $w_1 = 0.2$. The median, mean and variance of simulated values are shown at the top of each histogram. ................................. 67

4.8 Bias of $\hat{\theta}_{ss}$ (solid line), $\hat{\theta}_{n2}$ (dotted line) and $\hat{\theta}_n$ (dashed line) from model (3.3) plotted by the proportion of subjects in stage 1. Values $n = 100$, $\theta = 1.0$, $\bar{\theta} = 0.1$, $\bar{\theta} = 4.0$, $x_1 = 2$ were used. ................................. 68

4.9 Probability that parameter estimates are at the boundaries $\hat{\theta}_{ss}$ (solid line), $\hat{\theta}_{n2}$ (dotted line) and $\hat{\theta}_n$ (dashed line) from model (3.3) plotted by the proportion of subjects in stage 1. Values $n = 100$, $\theta = 1.0$, $\bar{\theta} = 0.1$, $\bar{\theta} = 4.0$, $x_1 = 2$ were used. ................................. 70

4.10 Variance approximations by proportion allocated to stage 1. The solid, dashed, and dotted lines represent $M^{*-1}$, $M^{-1}$ and $n\widehat{\text{Var}}[\hat{\theta}_n]$ by $w_1$, respectively. Values $x_1 = 2$, $n = 100$ and $\sigma = 1$ were used. ................................. 71

4.11 $P\left(\left|M^{-1}_{ind} - n\widehat{\text{Var}}[\hat{\theta}_n]\right| < \left|M^{-1} - n\widehat{\text{Var}}[\hat{\theta}_n]\right|\right)$ given $\theta = 1$, $x_1 = 2$, $a = .24$, $b = 4$, and $n = 100$ by proportion allocated to stage 1. ................................. 73
4.12 Variance approximations divided by $n \hat{\text{Var}} \left[ \hat{\theta}_n \right]$ as a function of the proportion allocated to stage 1. The solid lines represent 1.0 is for reference and represents $n \hat{\text{Var}} \left[ \hat{\theta}_n \right]$. The dotted, dot-dashed, and dashed lines are the $25^{th}$, $50^{th}$ and $75^{th}$ quantiles of the three information measure estimates.

4.13 Variance approximations evaluated at the $25^{th}$, $50^{th}$ and $75^{th}$ quantiles $\hat{\theta}_n$ divided by $n \hat{\text{Var}} \left[ \hat{\theta}_n \right]$ as a function of the proportion allocated to stage 1. The solid lines represent 1.0 is for reference and represents $n \hat{\text{Var}} \left[ \hat{\theta}_n \right]$. The dotted, dot-dashed, and dashed lines are the $25^{th}$, $50^{th}$ and $75^{th}$ quantiles of the three information measure estimates. Values $\theta = 1$, $x_1 = 2$, $\sigma = 1$, $X \in (25, 4)$ and $n = 100$ were used in the simulation.

4.14 Variance approximations evaluated at the $25^{th}$, $50^{th}$ and $75^{th}$ quantiles $\hat{\theta}_n$ divided by $n \hat{\text{Var}} \left[ \hat{\theta}_n \right]$ as a function of the proportion allocated to stage 1. The solid lines represent 1.0 is for reference and represents $n \hat{\text{Var}} \left[ \hat{\theta}_n \right]$. The dotted, dot-dashed, and dashed lines are the $25^{th}$, $50^{th}$ and $75^{th}$ quantiles of the three information measure estimates. Values $\theta = 1$, $x_1 = 2$, $\sigma = 0.25$, $X \in (25, 4)$ and $n = 100$ were used in the simulation.

4.15 Proportion of times one information measure estimate is closer to $n \hat{\text{Var}} \left[ \hat{\theta}_n \right]$ than another. The dotted line, $\hat{M}^{-1}$ is closer than $\hat{M}_{\text{ind}}^{-1}$, the dashed line, $\hat{M}^{-1}$ is closer than $\hat{M}_{\text{obs}}^{-1}$ and the solid line, $\hat{M}_{\text{obs}}^{-1}$ is closer than $\hat{M}_{\text{ind}}^{-1}$. 

xi
5.1 In each plot the CDF of \( \sqrt{n}(\hat{\theta}_n - \theta) \) was obtained via Monte Carlo simulations. The \( P(T_1 \leq t) \), where \( T_1 \sim \mathcal{N}(0, [M(\xi^*, \theta)]^{-1}) \). The \( P(T_2 \leq t) \), where \( T_2 \sim \mathcal{N}(0, [M(\xi_A, \theta)]^{-1}) \). The \( P(T_3 \leq t) \), where \( T_3 \sim UQ \). Values \( \theta = 1, x_1 = 2, n_1 = 5, \sigma = 0.5, a = 0.25 \) and \( b = 4 \) were used.

C.1 Map of \( z = -x_1 / \log \bar{y}_1 \) for \( \theta = 1, a = 0.25 \) and \( b = 4 \).

C.2 CDF of U for \( \theta = 1, x_1 = 2, n_1 = 5, \sigma^2 = 0.5, a = 0.25 \) and \( b = 4 \).
ABSTRACT

In adaptive optimal designs, each stage uses an estimate of the optimal design derived using cumulative data from all prior stages. This dependency on prior stages affects the properties of maximum likelihood estimates. To illuminate these effects, we assume for simplicity a nonlinear regression model with normal errors and that there are only two stages with a fixed first stage. Fisher’s information is motivated for adaptive designs by deriving the Cramér-Rao lower bound for such experiments. Then the usefulness of Fisher’s information is shown from both a design and analysis perspective. From a design perspective Fisher’s information is used in a procedure that is developed to select the proportion of observations assigned to the first stage. From an analysis perspective the information measure most commonly used in the optimal design literature is compared with Fisher’s information. Several estimates of information are compared and a procedure for selecting the proportion of subjects allocated to stage 1 is recommended.

Asymptotics for regular models with fixed number of stages are typically motivated by assuming the sample size of each stage goes to infinity as the overall sample size goes to infinity. However, it is not uncommon for a small pilot study of fixed size to be followed by a much larger experiment. We show that the distribution of the maximum likelihood estimates converges to a scale mixture family of normal random variables in such cases.

Throughout illustrative, numeric and simulated examples are provided using an exponential mean function. For this distribution Fisher’s information is derived explicitly and then used in a numeric example to demonstrate the usefulness of the
proposed procedure that allocates the number of subjects to stage 1 compared to alternatives. Monte Carlo simulations are used to compare different information measures as approximates to the variance of the maximum likelihood estimate. Limiting distributions assuming $n_1, n_2 \to \infty$ and $n_1$ fixed, $n_2 \to \infty$ are derived and compared against the simulated distribution of the maximum likelihood estimate.
Chapter 1

Introduction

1.1 Motivating Problem

Adaptive designs are well suited for use in phase I, phase I and phase I/II dose-finding clinical trials because they can be directed toward an experimental goal. This goal typically lies on a spectrum between ethically treating patients enrolled in the study and maximizing the amount of information collected. This trade off is often referred to as the treatment versus experimentation dilemma; see, for example Bartroff and Lai (2010), Baldi Antognini and Giovagnoli (2010) and Azriel, Mandel, and Rinott (2011). Examples of designs that attempt to maximize the ethical treatment of enrolled patients can be found in Li, Durham, and Flournoy (1995), Whitehead and Williamson (1998) and Thall and Cook (2004). These types of designs remain popular despite examples in Lai and Robbins (1982), Pronzato (2000), Chang and Ying (2009), Oron, Azriel, and Hoff (2011) and Azriel (2012) where such designs lead
to inconsistent estimates of the model parameters.

The purpose of this dissertation is to examine designs at the other end of the spectrum which use classical methods from the theory of optimal design to produce precise experiments. The emergence of current optimal design traces to Elfving (1952) where a geometric approach for determining a c-optimal design for linear regression models was introduced. Kiefer and Wolfowitz (1960) developed the celebrated equivalence theorem which provides an efficient method for verifying if a design is D-optimal, again for a linear model. These two results were generalized by Chernoff (1953) and White (1973) to include nonlinear models, respectively. See Bartoff (2012), O’Brien and Funk (2003) and references therein for extensions to the geometric and equivalence approaches.

One reason for the prevalence of the linear assumption in optimal design is that the problem can be explicitly described, as follows. Define an approximate design, proposed by Kiefer (1959), as \( \xi = \{ \lambda_i, x_i \}_{i=1}^K \), where \( \xi \) is a probability measure on \( \mathcal{X} \) consisting of support points \( x_i \in \mathcal{X} \) and corresponding design weights \( \lambda_i \), where \( \lambda_i \) is rational and defined on the interval \([0, 1]\) and \( \sum \lambda_i = 1 \). Then the optimal design problem is to locate the design that maximizes the precision for a specific experimental interest. Typically, this precision is achieved by minimizing some concave function, \( \phi \), of Fisher’s information matrix. For example when estimation of all the parameters is the primary interest then the D-optimality criteria, where \( \phi \) is equal to the determinant of the inverse of Fisher’s information, is the most popular method.

In general a design that is optimal with respect to the criteria \( \phi \) is referred to as the \( \phi \)-optimal design and is denoted \( \xi^*_\phi \). For examples of different optimality criteria and their corresponding concave functions see Section 1.2 and Pukelsheim (2006).
There is a wealth of literature on optimal designs for linear models [cf. Fedorov (1972), Silvey (1980) and Atkinson, Done, and Tobias (2007)]. However, when the underlying model is nonlinear, Fisher’s information matrix, and as a result optimal designs, will depend on the model parameters. This dependence represents a major challenge for the implementation of optimal designs in nonlinear models. For nonlinear models the term *locally optimal design* is often used to indicate that such designs are optimal only in the neighborhood of the true parameters. There have been many suggestions on how to deal with the locally optimal design problem. Fisher (1947, Chapter 68) and Chernoff (1953) suggest optimal designs be approximated by guessing the parameter values; however this method may be inefficient when the guess is far from the true parameter values. Ford, Torsney, and Wu (1992) argue that the procedure of guessing provides an appropriate benchmark by which to gauge the performance of alternate methods. Kitsos, Titterington, and Torsney (1988) suggest the use of a non-optimal design that have the property of being insensitive to the true parameter values. Dette and Sahm (1998) developed a minimax optimal design for use in nonlinear models. For a review of methods for nonlinear models see Ford, Titterington, and Kitsos (1989) and O’Brien and Funk (2003).

Others have suggested adaptive methods. Atkinson, Done, and Tobias (2007, Chapter 17) suggest a sequential procedure where the model is linearized, using an expansion; then the optimal design of the approximate linear model is used in the first stage and updated for consequent stages. Haines, Perevozkaya, and Rosenberger (2003) develop a sequential Bayesian optimal design procedure. Many researchers, including Box and Hunter (1965), Fedorov (1972), White (1975) and Silvey (1980) advocate what we refer to as adaptive optimal designs. An *adaptive optimal design*
is a procedure where the first stage is initialized, using expert opinion or prior data. Then each successive stage is allocated according to the estimated optimal design obtained using all data from the previous stages. Recently Dragalin and Fedorov (2005), Dragalin, Fedorov, and Wu (2007) and Dragalin, Hsuan, and Padmanabhan (2008) among others have proposed and analyzed such designs. For a comparison and further discussion of designs from both ends of the treatment/experimental spectrum see Fedorov, Flournoy, Wu, and Zhang (2011).

In Section 1.2.1 and 1.2.2 background information, motivation and the general procedures of optimal design are reviewed. Common optimality criteria and the usefulness of each is explained in Section 1.2.3.

The primary experimental motivation for this dissertation is designs in dose-finding clinical trials. The importance and procedure of clinical trials in general and specifically for dose-finding studies is reviewed in Section 1.2.6. Also in this section the usefulness of adaptive design in dose-finding studies is explained. The adaptive optimal design along with popular competing adaptive methods, up and down and best intention designs, are introduced. The positives and negatives of each adaptive method, from the perspective of dose-finding studies is examined.

In Section 2.1 maximum likelihood theory for a general one parameter nonlinear regression model is reviewed. The proofs of convergence in distribution for different situations occurring in adaptive optimal design adapt the procedure used in this simple case. In Section 2.2 the maximum likelihood procedure and performance for a non-linear regression model with an exponential mean function is examined. This simple model is used throughout as an illustrative example of the proposed methods.

In classical, non-adaptive, optimal design the information bound, the Cramèr-Rao
lower bound, provides a small sample justification for the use of Fisher’s information to design optimal experiments. In Section 3.1 we derive the Cramèr-Rao lower bound for a general two-stage adaptive experiment. In Section 3.2 we define the model and the two-stage adaptive optimal procedure examined in this exposition.

One issue present in current adaptive optimal design literature is that in place of constructing a likelihood from the joint density of responses and support points, responses have been treated as independent conditional on the design. Silvey (1980) and others point out that the information employed is not by definition Fisher’s information. To make this issue explicit in Section 3.2 we develop Fisher’s information for a two-stage adaptive optimal design.

We attempt to clarify the benefit of the unconditional Fisher information from two different perspectives. First, from a design perspective we define the locally optimal stage one sample size for a two-stage adaptive experiment in Section 3.3. Then we propose a method to approximate locally optimal stage one sample size when the parameter is unknown.

Second, in Section 4.1 from an analysis perspective we examine the effect of using Fisher’s information against commonly used alternatives. The alternatives we examine are an approximation based on the information measure derived under conditional independence and the observed information. In Section 4.2 a simulation is done to compare the performance of the different information measures and their estimates.

In Section 3.3 show that the optimal stage one sample size is of the order \( \sqrt{n} \), where \( n \) is the overall sample size, in a two stage regression model. Luc Pranzato obtains this relationship for a more general model (personal communication, 2012). However, in certain experiments, for example early phase clinical trials or bioassay studies, it
is common to use designs with very small stage one sample sizes. Current literature
has characterized the adaptive optimal design procedure under the assumption that
both stage one and stage two sample sizes are large.

In Chapter 5 we characterize the asymptotic distribution of the maximum likeli-
hood estimate (MLE) when the stage one sample size is fixed. The distribution for
a nonlinear regression model with normal errors and a one parameter exponential
mean function is derived explicitly. Then for a numeric example the inference when
the limiting distribution is derived assuming the first stage sample size is finite is
compared with other candidate approximate distributions.

Throughout this exposition we assume for simplicity that there are only two stages
and that the first stage treatment (support point) is fixed. Responses are assumed
to be normal with a nonlinear mean function. It is convenient for our narrative to
use such a simple set-up with the understanding that lessons learned apply to more
complex scenarios.

1.2 Optimal Design

1.2.1 Background

Smith (1918) examined the allocation of observations for a polynomial model that
minimized the standard deviation of the parameter estimates. This represents the
earliest work in what is today considered optimal design. Elfving (1952) showed, for
a $p$ dimensional linear regression model, that the optimal design depends on at most
$\frac{1}{2}p(p + 1)$ sample points. Chernoff (1953) generalized Elfving’s result to show that
the optimal design for any model with \( p \) parameters, including possible nonlinear functions, depends on at most \( p + (p - 1) + \cdots + (p - k + 1) \) sample points, where \( k \) represents the number of non-nuisance parameters.

The results in Elfving (1952) and Chernoff (1953) represent a major theoretical step forward, however, practically it is extremely difficult to design an algorithm which can determine an optimal design based only on an upper bound that is not overly computationally expensive. Kiefer (1959) introduced the concept of approximate (or continuous) designs and showed that such designs can be considered probability measures. Allowing designs to be probability measures no longer requires sample point allocation proportions that are exact multiples of the sample size, \( n \), but instead it is only necessary to allocate a rational proportion to each sample point.

Kiefer and Wolfowitz (1960) introduced the most important theorem in optimal design theory, the Equivalence Theorem. This theorem gives a criteria, for a linear model, which if satisfied implies that a design is D-optimal and G-Optimal. Where D-optimality is achieved by minimizing the determinant and G-optimality is achieved by minimizing the maximum diagonal element of Fisher’s Information. This theorem, which uses the concept of approximate designs, made the attainment of optimal designs realistic for most linear models since algorithms could be designed using this criteria to find approximately optimal designs quickly. Once this approximate design is found one can use a rounding procedure for any finite sample size.

The Equivalence theorem has been generalized in many significant ways. Fedorov (1972) derived a criteria, once again for a linear model, for A-optimality. Where A-optimality is achieved by minimizing the trace of Fisher’s information. Kiefer (1974), also for linear models, found a general criteria by which a design could be shown to be
γ-optimal, where γ is some concave function under various conditions that include A- and D-optimality along with other common optimality criteria. The first generalized criteria that included potentially nonlinear functions was developed by White (1973).

The majority of the optimal design literature has been developed under the assumption of a linear framework. The linear design problem is fairly straightforward since an optimal design is independent of the model parameter values. For nonlinear problems an optimal design will depend on the parameter. This dependence severely limits the usefulness of the concepts developed under the linear condition, since in order to have an efficient design an approximate value of the parameter must be known. The use of adaptive optimal design has been endorsed as a solution to this dependency by Box and Hunter (1965), Fedorov (1972) and White (1975).

1.2.2 General Details

Consider a regression model with normal errors, linear or nonlinear, with E(y) = η(x, θ). Then an approximate design, ξ, on Ξ, the set of all possible designs, is defined as

\[
ξ = \left\{ \lambda_1, \ldots, \lambda_m \right\},
\]

where \( x_i, i = 1, \ldots, m \) are the support points and \( \lambda_i, i = 1, \ldots, m \) are the corresponding proportions of the total sample size observed at the \( i^{th} \) support point. The fact that ξ can be viewed as a probability measure on the support space of \( x = (x_1, \ldots, x_m), \mathcal{X} \), implies that \( \int_{\mathcal{X}} \xi(dx) = 1 \), \( \lambda_i \in [0, 1] \) and \( \sum \lambda_i = 1 \). To observe the distinction between an approximate design and an exact design, one that
can be administered for a specific integer $n$, we denote an exact design as

$$\xi_n = \left\{ \frac{x_1}{n}, \ldots, \frac{x_m}{n}, \frac{n_1}{n}, \ldots, \frac{n_m}{n} \right\},$$

where $n_i$ is the integer number of support points at $x_i$ and $\sum n_i = n$. In practice for moderate to large $n$ exact designs can be found by relating $\xi$ to $\xi_n$ by an integer approximation.

Optimal design is, typically, concerned with minimizing some concave function, $\phi$, of Fisher’s information matrix, given by

$$M(\xi, \theta) = \sum \lambda_i \frac{\partial \eta(x_i, \theta)}{\partial \theta} \frac{\partial \eta(x_i, \theta)}{\partial \theta^T}.$$ 

Under standard regularity conditions it is well established that

$$\lim_{n \to \infty} n \text{Var} \left[ \hat{\theta}_n \right] \longrightarrow [M(\xi, \theta)]^{-1}, \quad (1.1)$$

where $\hat{\theta}_n$ is the maximum likelihood estimate. Note (1.1) requires $M(\xi, \theta)$ to be positive definite. For a description on how to find optimal designs when the positive definite assumption is not valid see Silvey (1978). Throughout this exposition we consider only cases where $M(\xi, \theta)$ is nonsingular.

An optimal design, for a given criteria function $\phi$, is defined by the design that maximizes, $\phi(M(\xi, \theta))$, with respect to $\xi \in \Xi$, i.e.,

$$\xi^*_\phi = \arg \min_{\xi \in \Xi} \phi(M(\xi, \theta))$$
We refer to $\xi^*_\phi$ as a $\phi$-optimal design.

### 1.2.3 Optimal Design Criteria

Below are more formal definitions and usages of several important design criteria [cf. Atkinson, Donev, and Tobias (2007) and Pukelsheim (2006)].

**A-optimality**

An A-optimal design minimizes the trace of the inverse of Fisher’s information, *i.e.*, 

$$
\xi^*_A = \arg\min_{\xi \in \Xi} \text{tr} \left( \left[ M(\xi, \theta) \right]^{-1} \right).
$$

Using this criteria minimizes the variances of the parameter estimates. Thus A-optimality is particularly useful when orthogonal linear combinations of the parameters are desired.

**D-optimality**

A D-optimal design minimizes the determinant of the inverse of Fisher’s information. In most cases the D-optimal criteria is expressed as the log of the determinant of the inverse of Fisher’s information, *i.e.*, 

$$
\xi^*_D = \arg\min_{\xi \in \Xi} \log \left| \left[ M(\xi, \theta) \right]^{-1} \right|.
$$

The use of the log determinant is used to ensure concavity.

D-optimality is the most widely studied optimal design criteria. This is due to
several advantageous design properties. For example the D-optimality criteria is
scale invariant with respect to the predictor variables, $x$. D-optimal designs are
not singular, i.e., if there model that contains more than a single parameter then
the D-optimal design will contain more than a single support point. In general,
for regular models, it is known that the D-optimal design will depend on between
$p$ and $\frac{1}{2}p(p + 1) \neq 1$ support points. Other optimal design criteria may not have
such appealing properties. Perhaps the most important factor in D-optimal design’s
popularity is the fact that this criteria minimizes the (asymptotic) confidence ellipsoid
containing the maximum likelihood estimates. It is the most efficient when estimation
of all the parameters is the primary goal.

\section*{c-optimality}

A c-optimal design minimizes, for some numeric vector $c$, $c^T [M(\xi, \theta)]^{-1} c$, i.e.,

$$
\xi^*_c = \arg \min_{\xi \in \Xi} c^T [M(\xi, \theta)]^{-1} c.
$$

This criteria minimizes the variance of the linear combination $c^T \theta$. When employing
c-optimal designs it is necessary to use caution because they can lead to singular
designs even in cases where $\dim(\theta) > 1$. A singular design will provide no information
about other aspects of the model.
**E-optimality**

An E-optimal design minimizes the maximum eigenvalue of $M(\xi, \theta)$, i.e.,

$$\xi^*_E = \arg \min_{\xi \in \Xi} \max_{i} \frac{1}{\tau_i},$$

where $\tau_i$ is the $i^{th}$ eigenvalue of $M(\xi, \theta)$. The E-optimality criterium minimizes the variance of the least well-estimated linear contrast of $a^T \theta$ with $a^T a = 1$.

**G-optimal**

Atkinson *et al.* (2007) discuss the standardized predicted variance function

$$d(x, \xi, \theta) = \frac{\partial \eta(x, \theta)}{\partial \theta^T} M^{-1}(\xi, \theta) \frac{\partial \eta(x, \theta)}{\partial \theta}.$$  

A G-optimal design minimizes the maximum of $d(x, \xi, \theta)$ with respect to $x$, i.e.,

$$\xi^*_G = \arg \min_{\xi \in \Xi} \max_{x \in X} d(x, \xi, \theta).$$

Recall that the D and G-optimal designs are equivalent.

**Compound Design Criteria**

The preceding design criteria are all directed at accomplishing a single research interest. However, it may be desired to find a design that will perform reasonably well for a variety of research interests. Compound design criteria allow for a linear combination of design criteria to be optimized. Suppose a design that addresses $q$ research interests, each with optimality criteria $\phi_i$ and weight $c_i$, $i = 1, \ldots, q$, then
the compound optimal design minimizes of the linear combination of the optimality criteria and design weight, \( i.e., \)

\[
\xi_{CD}^* = \arg \min_{\xi \in \Xi} \sum_{i}^{q} c_i \phi_i \left( [M(\xi, \theta)]^{-1} \right).
\]

Note \( \phi_i, i = 1, \ldots, q \) must each be concave functions of Fisher’s information.

**Penalized Optimal Designs**

Certain research goals cannot be optimized by simply obtaining a minimum of a concave function of Fisher’s information. For example in dose-finding studies the ethical treatment of the patients enrolled in the study may be of great concern. Using a traditional optimal design derived by the methods discussed may violate ethical standards of the experiment. For this reason penalized optimal designs have been proposed. A penalized optimal design minimizes for some concave function, \( \phi \), where \( \phi \) can be any standard optimal design criteria, the ratio of Fisher’s information to \( \Psi(x, \theta) \), \( i.e., \)

\[
\xi_{P}^* = \arg \min_{\xi \in \Xi} \phi \left( \frac{M(\xi, \theta)}{\Psi(x, \theta)} \right).
\]

\( \Psi(\cdot) \) should be selected to address the additional concern in the experiment. For example the ethical treatment of patients in clinical trials.
1.2.4 General Equivalence Theorem

Below is stated the general equivalence theorem, as it appears in White (1973), which includes both linear and nonlinear models.

**Theorem 1.** The following conditions on a design measure, \( \xi \), are equivalent

(i.) \( \xi \) is D-optimal

(ii.) \( \xi \) is G-optimal

(iii.) \( \sup d(\xi, \theta) = p \)

where \( p = \text{dim}(\theta) \).

1.2.5 Design Efficiency

The relative efficiency of any arbitrary design, \( \xi \), relative to a \( \phi \)-optimal design, \( \xi^*_\phi \), can be assessed via a comparison of the information matrices of the two designs. The \( \phi \)-optimal relative efficiency for any design \( \xi \) is

\[
\phi_{\text{eff}} = \frac{\phi(M(\xi, \theta))}{\phi(M(\xi_D, \theta))}.
\]

Assessing the relative efficiency is very informative when determining how much information is lost when an integer approximation of the optimal design, \( \xi_n \), is used for a finite sample size \( n \). It can also be very helpful in determining the loss in precision when penalized optimal design or an ad hoc design is used.
1.2.6 Adaptive Designs in Dose-Finding Clinical Trials

Clinical Trials

A clinical trial is a medical research study conducted on human volunteers (subjects). The subjects receive a treatment as determined by the researchers. In general this treatment could be any one of a variety of medical related interventions; a new drug, a procedural change, a medical device, behavior changes, etc. In this exposition treatment is thought to be a drug; however, there is no restriction in the concepts put forth that restrict their application to a specific intervention method.

When a clinical trial for a novel drug treatment is conducted the clinical trial is completed in phases. Phase I, referred to as the screening for safety (toxicity), assesses the seriousness and frequency of the drug’s adverse effects. Phase II assesses the drug’s effectiveness, or efficacy, at treating patients with a certain disease or condition. Phase III examines a range of objectives including further information on toxicity and efficacy, comparison to the standard of care or to a placebo group and interactions with other drugs for a range of dose levels. The final phase, phase IV, studies food and drug administration (FDA) approved drugs for long term assessments of toxicity, efficacy and optimal use.

Dose-Finding Studies

Dose-finding for toxicity and efficacy has been traditionally explored in phase I and phase II clinical trials, respectively. Recently there has been interest in assessing toxicity and efficacy jointly in a phase I/II clinical trial.

In phase I a common goal is to locate the maximum tolerated dose (MTD). In this
context we define the MTD as the dose that will result in \( r \)% of patients experiencing a toxic response. Once this dose is found the clinical trial proceeds to phase II. A common procedure in phase II is to search, on the dose domain \([0, \text{MTD}_r]\) for the effective dose (ED). The ED is defined as the dose that will result in \( s \)% of the patients experiencing an efficacious response and may not exist on \([0, \text{MTD}_r]\). Classically, the goal of dose-finding studies, phase I and II, is to use a relatively small number of subjects to identify a dose or a set of potential doses for use in the much larger phase III clinical trial. Phase III is not a dose-finding study.

An approach to dose-finding that has recently gained popularity is to use a phase I/II study instead of the traditional phase I and phase II setup. A stated goal of phase I/II studies is examine efficacy and toxicity jointly by searching for the dose that maximizes the probability of the patients experiencing an efficacious response without experiencing a toxic response.

To understand the different objectives from each dose-finding phase consider an example where responses are assumed to be from a multinomial experiment, i.e., \( E[y] = \eta(x, \theta) \), where \( \eta(x, \theta) = \{ p_{00}(x, \theta), p_{01}(x, \theta), p_{10}(x, \theta), p_{11}(x, \theta) \} \). The probabilities of the outcomes are defined as; \( p_{00}(x, \theta) \) probability of no efficacy and no toxicity, \( p_{01}(x, \theta) \) probability of toxicity without efficacy, \( p_{10}(x, \theta) \) probability of efficacy without toxicity and \( p_{11}(x, \theta) \) probability of toxicity and efficacy. Then the objective functions of phase I, phase II and phase I/II can each be described in terms of \( p_{ij}(x, \theta) \), \( i, j = 0, 1 \). For phase I the MTD is the dose that corresponds to the \( r^{th} \) percentile of the marginal probability of toxicity, \( p_1(x, \theta) = p_{10}(x, \theta) + p_{11}(x, \theta) \). Similarly for phase II the ED is the dose that corresponds to the \( s^{th} \) percentile of the marginal probability of efficacy, \( p_1(x, \theta) = p_{01}(x, \theta) + p_{11}(x, \theta) \). Finally for phase
In II the target is the dose that maximizes the probability of efficacy without toxicity, i.e., \( \text{argmax}_x p_{10}(x, \theta) \). In each case \( x \) is the dose and \( \theta \) is the parameter common to all probabilities.

Figure 1.1 plots each of the objective functions \( p_1(x, \theta) \), \( p_{10}(x, \theta) \) and \( p_{10}(x, \theta) \) from a hypothetical experiment. For each curve the vertical line connected to the \( x \)-axis represents a potential dose target for the corresponding objective function. In this example the MTD_{33}, the ED_{80} and the \( \text{argmax}_x p_{10}(x, \theta) \) and are used as illustrations.

This figure helps to illustrate how significantly the target doses can differ for the different objectives of each phase. For this reason it is important to design an experiment that is appropriate for accomplishing a specific research interest or interests.

**Adaptive Designs**

Adaptive designs are characterized by using data from the current study to determine treatment (dose) allocations, stopping rules, sample size allocations, etc. The use of adaptive or sequential designs is intuitive when subjects (observations) enter a study at different times, potentially in batches or one at a time. The means in which subjects enter may limit the practicality of fully or nearly fully adaptive designs. For example in the case of many phase I experiments it requires months for subjects to elicit a toxic response and it is impractical to wait for each subject to finish the treatment before proceeding to the next subject. The use of two or few stage designs, like the designs examined in this dissertation, are an appealing middle ground in such scenarios.

There are two sources of information available for use in an adaptive design;
Figure 1.1: Example of underlying objectives functions in phase I, phase II and phase I/II, $p_1(x, \theta)$, $p_0(x, \theta)$ and $p_{01}(x, \theta)$, respectively. Figure courtesy of Fedorov and Wu. The solid lines connecting each curve to the $x$-axis represent potential targets for each phase the MTD$_{33}$, the ED$_{80}$ and the argmax$_x p_{10}(x, \theta)$. 
treatment allocation frequencies (the number of subjects assigned to different dose levels) and treatment outcomes (subject responses). Most adaptive designs only use one of the two sources.

Designs that use only treatment allocations include some adaptive randomization procedures. A landmark example of such designs is the biased coin design proposed by Efron (1971). This design addresses the problem known to occur when the number and times of subjects entering the study is unknown. When a simple randomization procedure is employed there is a tendency for treatment allocations to become unbalanced. This unbalanced design will result in loss of power in testing the equality of treatment means. However, using a deterministic allocation rule in place of a random procedure is typically unacceptable. The biased coin design in Efron (1971) assigns a probability greater than $1/2$ to the treatment that has been allocated less often. This probability should be selected in order to maintain balance and achieve a random treatment allocation. Efron (1971) suggested the use of a $2/3$ probability. Markaryan and Rosenberger (2010) find the exact distribution of this biased coin design. Wei (1978) evaluates the procedure outlined in Efron (1971) along with other extensions and alternative adaptive randomization procedures. Adaptive randomization procedure have been suggested for use in phase III clinical trials.

Designs that incorporate subject responses have become popular in the dose-finding phases of clinical trials because of their ability to be tailored to accomplish complex experimental goals. For an example consider a motivating example from certain clinical trials in cancer research. Patients have failed, or do not meet the requirements, of the standard methods of care. The disease, if left untreated, will result in a severe negative outcome, often the death of the patient. The proposed
treatment (drug) is expected to have little or no effect at low dose levels, however, at high dose levels there is a severe risk of toxicity. For examples of such studies see Pisters et al. (2004), Mathew et al. (2004) and Neuenschwander, Brandson, and Gsponer (2008). In such experiments designs that not only collect information about the dose-response relationship, but also treat patients ethically are highly desirable.

The conflict that arises when a design is used to attempt to balance between treating patients enrolled in the study ethically and collecting information in a manner that will lead to good estimates of the model parameters has been referred to as the treatment versus experimentation dilemma by Bartroff and Lai (2010) and addressed in Baldi Antognini and Giovagnoli (2010) and Azriel, Mandel, and Rinott (2011). It has become customary for dose-finding designs to take into consideration the ethical treatment of enrolled patients.

Many of the response adaptive designs proposed for use in dose finding studies fall into one of three general frameworks up and down, best intention and adaptive optimal.

**Up and Down Designs**

Up and down design work on a discrete design space, \( \{d_1, \ldots, d_m\} \), and use a set of rules based on the responses to determine whether the next treatment will be increased by one, decreased by one or remain the same. Most up and down designs assume a monotonic dose-response relationship. Dixon and Mood (1948) developed the classical example of up and down designs. The Dixon and Mood (1948) procedure is as follows: suppose the \( i^{th} \) subject was treated at dose \( d_j \); if the \( i^{th} \) subject experiences a toxicity (failure) then subject \( i+1 \) is treated at \( d_{j-1} \); otherwise subject \( i+1 \) is treated at \( d_{j+1} \).
The Dixon and Mood (1948) procedure is most useful for dose-finding experiments where the median, e.g., MTD\textsubscript{50} or the ED\textsubscript{50}, is the target.

Derman (1957) extended the procedure put forth in Dixon and Mood (1948) by basing the dose escalation or de-escalation decision on not only the response but also on the flip of a biased coin. In the Derman (1957) procedure if the \textit{i}th subject, treated at \textit{d}_\textit{j}, does not experience a toxic response (success) and a biased coin lands on heads then the dose is increased to \textit{d}_{\textit{j}+1}; otherwise the dose is decreased to \textit{d}_{\textit{j}-1}. The bias of the coin, \( P\{\text{heads}\} \in [0.5, 1.0] \), can be selected in order to ensure that the treatment allocations will be centered around an arbitrary desired quantile.

In Derman (1957) subject \textit{i} could experience a non-toxic response and yet the dose level given to subject \textit{i} + 1 could decrease. This is an unappealing design property in dose-finding clinical trials. Durham and Flournoy (1994) proposed an extension to Derman (1957) that does not have this undesirable design property and still can center treatment allocations around any arbitrary quantile. The Durham and Flournoy (1994) biased coin design rule is to increase the dose by one if a non-toxic response occurs and the biased coin land on heads; keep the same dose if a non-toxic response occurs and the biased coin lands on tails; otherwise increase the dose by one. Durham and Flournoy (1994) also provide a method to select the bias of the coin such that the treatment allocations will center around the desired quantile.

Since Durham and Flournoy (1994) there have been many extensions and novel up and down design procedures proposed for use in dose-finding studies. Gezmu and Flournoy (2006), Ivanova (2006) and Antognini, Bortot, and Giovagnoli (2008) developed group up and down designs which allow for more than one subject to be treated at each stage. Wetherill (1963), Wetherill and Levitt (1965), Ivanova,

Up and down designs are nonparametric and thus have great practical appeal. They are also defined by a set of simple rules and thus are easy to implement and understand. Up and down designs naturally take into consideration the ethical treatment of patients in the study by defining the rules for escalation to ensure that doses far from the target occur infrequently. Under common regularity conditions up and down procedures are consistent.

Up and down designs are not without practical limitations. These designs work only on a discrete dose space, which is often not a severe limitation in dose-finding studies since it is common to only consider a discrete number of possible dose concentrations. However, if the dose space contains a very large number of points, then the usefulness of many up and down procedures are severely inhibited. Several suggestions for consistent point estimates, including isotonic regression, have been made. However, the proposed estimates are without a standard error unless an underlying parametric assumption is made about the dose response relationship. For a review of up and down procedures see Flournoy and Oron (2013).

**Best Intention Designs**

We define any design where the primary goal is to treat patients enrolled in the study at a target dose as a best intention design. These methods are often proposed under the assumption that it is ethical to treat patients at the target dose. For example in
a toxicity study when the target is the MTD$_r$, then it will be considered ethical to all patients at the estimate, based on all available data at the MTD$_r$.

For examples of designs that use the data from the previous stages to estimate the MTD$_r$ and treat the next cohort of patients at that level see, O’Quigley, Pepe, and Fisher (1990), O’Quigley and Shen (1996), Bapp, Rogatko, and Zacks (1998), Gasparini and Eisele (2000), Leung and Wang (2001) and Whitehead, Thygesen, and Whitehead (2010). Most of the best intention designs proposed are developed within the Bayesian framework. Dose-finding studies are often considered a decision based problem, eg., deciding which dose is the MTD$_r$. However, although the classical Bayesian framework is grounded in decision theory this theory is commonly ignored in many modern Bayesian applications. Best intention designs have a strong clinical appeal since the ethical treatment of individuals in the study are given priority. However, each of the procedures suggested are inconsistent with respect to the MTD$_r$ as shown by Azriel, Mandel, and Rinott (2011). Azriel, Mandel, and Rinott (2011) provide a proof that any experiment where each patient is treated at the estimate of the MTD$_r$ cannot lead to a strongly consistent estimate of the MTD$_r$.

In addition to inconsistent estimates, best intention designs are typically characterized by several drawbacks. They provide very little information about the dose-response relationship at levels other than the target dose. If a researcher wanted to have information about a dose range or more than one dose for use in a later phase a best intention procedure may not be an appropriate choice. When a parametric model assumption is made the best intention procedures provide poor parameter estimates. For a detailed comparison of popular best intention and up and down procedures see Oron and Hoff (2013).
Adaptive Optimal Designs

An adaptive optimal design is a procedure where the first stage, or stages, is initialized, using expert opinion or prior data. Then each successive stage is allocated according to an estimate of the locally optimal design, where the estimate is obtained using all the data from the previous stages. Box and Hunter (1965), Fedorov (1972), White (1975), Silvey (1980) and others proposed adaptive optimal designs for use in a variety of fields.

Traditional adaptive optimal design procedures are unappealing from a clinical perspective in dose-finding studies. The problem arises since optimal design have the primary goal of maximizing the information collected, i.e., a purely experimentation driven goal with no regard to the treatment of patients in the current study. In order to maximize information, optimal designs tend to include extreme treatment allocations. For example consider the simple linear regression model with normal errors, i.e., \(E[y] = \beta_0 + \beta_1 x\), where the dose, \(x\), has support on a the bounded interval \([-1, 1]\). Then the D-optimal design will assign half of the subjects to dose \(-1\) and half the subjects to dose \(1\). This would be an unacceptable design for use in a dose-finding study with human subjects.

Recently Dragalin and Fedorov (2005), Dragalin, Hsuan, and Padmanabhan (2008), Dragalin, Fedorov, and Wu (2007) among others have extended traditional adaptive optimal design mechanisms for use in dose-finding studies by adding a penalty function to the optimal design criteria that accounts for the ethical treatment of patients currently enrolled in the study along with the costs associated with a dose level. For example, Dragalin and Fedorov (2005) propose using, for a phase I/II study where
responses are from the multinomial experiment described earlier, the optimal design

\[ \xi^* = \arg \max_{\xi \in \Xi} \phi \left[ \frac{M(\xi, \theta)}{\Psi(\xi, \theta)} \right], \quad (1.2) \]

where \( \phi \) a traditional optimal design criteria, see section 1.2, and

\[ \Psi(\xi, \theta) = \int_x \psi(x, \theta) \xi d(x) \quad (1.3) \]

with

\[ \psi(x, \theta) = \{p_{10}(x, \theta)\}^{C_E} \{1 - p_{11}(x, \theta)\}^{-C_T}. \quad (1.4) \]

\( C_E \) and \( C_T \) are tuning parameters to be set by the practitioners in order to control the desired ethical and cost considerations for the study.

The design that results from the penalized procedure will no longer be optimal from an experimental perspective, \( i.e., \) estimating the model parameters or estimating the MTD\(_r\), ED\(_s\) or \( \arg \max \) \( p_{10}(x, \theta) \), but instead will be an optimal balance between treatment and experimentation as controlled by the researcher.

Due to the flexibility of the optimal design framework, this procedure can be adapted for use in phase I, phase II and phase I/II dose-finding clinical trials. It can also account for experimentation costs that up and down designs and best intention designs do not take into consideration.

Adaptive optimal designs do have some limitations. A parametric assumption must be made in order to determine the optimal design. At each stage the model parameters must be estimated in order to determine the estimate of the locally optimal
design. Typically maximum likelihood theory is used. Thus it is necessary to not only initialize the first stage but potentially initialize until the maximum likelihood estimates exist. For a detailed comparison of adaptive optimal designs and best intention procedures see Fedorov, Flournoy, Wu, and Zhang (2011).
Chapter 2

Preamble

Throughout this dissertation a nonlinear regression model with normal errors is assumed. This is done so that problems and recommendations for the complex adaptive optimal procedure can be highlighted without the unnecessary distractions of a complex model. This chapter reviews maximum likelihood theory for a general nonlinear model without adaptation. Then for the specific example, an exponential mean function, the procedure along with challenges of maximum likelihood theory in nonlinear regression when sample sizes are small are examined.
2.1 A Brief Review of Maximum Likelihood Theory for Nonlinear Regression with Independent Normal Errors

Suppose \( n_i \) subjects are treated at a fixed treatment level \( x_i, i = 1, \ldots, K \), with responses

\[
y_{ij} = \eta(x_i, \theta) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, K; \quad j = 1, \ldots, n_i, \quad (2.1)
\]

where \( \eta(x_i, \theta) \) is a nonlinear function of \( x \) and \( \theta \) with \( \theta \in \Theta \). The total sample size is \( n = \sum_{i=1}^{K} n_i \). Defining \( w_i = n_i/n \), the total design is denoted by \( \xi = \{w_i, x_{ij}\}_{i=1}^{K} \).

In this section, \( \xi \) is taken to be fixed and responses \( y = (y_1^T, \ldots, y_K^T)^T, y_i = (y_{i1}, \ldots, y_{in_i})^T, i = 1, \ldots, K \) are assumed to be independent, so the likelihood is

\[
\mathcal{L}_n(\theta) = \mathcal{L}(\theta | y) = \prod_{i=1}^{K} \prod_{j=1}^{n_i} f(y_{ij} | x_i, \theta) \propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \eta(x_i, \theta))^2 \right\}
\]

\[
\propto \exp \left\{ \sum_{i=1}^{K} \frac{1}{\sigma^2 w_i} \left( \eta(x_i, \theta)\bar{y}_i - \frac{1}{2} \eta^2(x_i, \theta) \right) \right\},
\]

where \( \bar{y}_i = \sum_{j=1}^{n_i} y_{ij}, i = 1, \ldots, K \). The maximum likelihood estimate (MLE)

\[
\hat{\theta}_n = \arg \max_{\theta \in \Theta} \mathcal{L}_n(\theta).
\]

Assuming \( f(y|x, \theta) \) is continuous and twice differentiable with \( \theta \in \Theta \), the score function for a single subject treated at \( x_i \),

\[
s_{ij} = s(y_{ij} | x_i, \theta) = \frac{\partial}{\partial \theta} \log f(y_{ij} | x_i, \theta) = \frac{1}{\sigma^2} [y_{ij} - \eta(x_i, \theta)] \frac{\partial}{\partial \theta} \eta(x_i, \theta),
\]
The maximum of \( \log L_n(\theta) \) occurs where the derivative of (2.3) with respect to \( \Delta = \hat{\theta}_n - \theta \) equals zero, i.e., \( S + \Delta \frac{\partial}{\partial \theta} S = 0 \). Taking this derivative and rearranging terms,
for \( \hat{\theta}_n \) in the neighborhood of \( \theta \),

\[
\sqrt{n} \left( \hat{\theta}_n - \theta \right) = -\frac{1}{n} \left[ \frac{\partial}{\partial \theta} S + \frac{1}{2} \left( \hat{\theta}_n - \theta \right) \frac{\partial^2}{\partial \theta^2} S_{\theta=\hat{\theta}} \right],
\]

(2.4)

where

\[
\frac{1}{\sqrt{n}} S = \sum_{j=1}^{K} \frac{1}{\sigma^2} w_i \left( \bar{y}_i - \eta(x_i, \theta) \right) \frac{\partial \eta(x_i, \theta)}{\partial \theta}, \sim \mathcal{N} \left( 0, M(\xi, \theta) \right),
\]

(2.5)

and by the law of large numbers,

\[
-\frac{1}{n} \frac{\partial}{\partial \theta} S \xrightarrow{a.s.} \frac{1}{n} \frac{\partial}{\partial \theta} S = \mathbb{E} \left[ -\frac{1}{n} \frac{\partial}{\partial \theta} S \right] = \mathbb{E} \left[ \sum_{j=1}^{K} w_i \left( -\frac{\partial}{\partial \theta} s_{ij} \right) \right]
\]

\[
= -\sum_{j=1}^{K} w_i \left( \mathbb{E} \left[ \bar{y}_i - \eta(x_i, \theta) \right] \frac{\partial^2 \eta(x_i, \theta)}{\partial \theta^2} + \frac{1}{\sigma^2} \left( \frac{\partial \eta(x_i, \theta)}{\partial \theta} \right)^2 \right)
\]

(2.6)

\[
= \sum_{j=1}^{K} \frac{1}{\sigma^2} w_i \left( \frac{\partial \eta(x_i, \theta)}{\partial \theta} \right)^2 = M(\xi, \theta).
\]

For any consistent \( \hat{\theta}_n \) and provided standard regularity conditions we have by (2.4), (2.5), (2.6) and Slutsky’s theorem,

\[
\sqrt{n} \left( \hat{\theta}_n - \theta \right) \xrightarrow{n \to \infty} \mathcal{N} \left( 0, M(\xi, \theta)^{-1} \right).
\]

The preceding proof was adapted from Lehmann (1999).
2.2 A Fixed One Point Design for a Regression Model with an Exponential Mean Function

In model (2.1), let \( \eta(x, \theta) = e^{-\theta x}, \theta \in (0, \infty), k = 1 \) and \( x \in (0, \infty) \) be a fixed constant; \( \bar{y} \) is a complete and sufficient statistic for \( \theta \) given \( x \) and \( n \). Therefore inference can be based on the likelihood

\[
\mathcal{L}(\theta|x, \bar{y}) = f(\bar{y}|x, \theta) = \left(\frac{n}{2\pi \sigma}\right)^{1/2} \exp\left\{-\frac{n}{2\sigma^2} (\bar{y} - e^{-\theta x})^2\right\}.
\]

The corresponding per-subject expected information is

\[
M(x, \theta) = \frac{1}{\sigma^2} x^2 e^{-2\theta x}.
\]

Since the mean function \( e^{-\theta x} \) is bounded in \((0, 1)\), the likelihood must maximized be separately for \( \bar{y} < 0, \bar{y} \in (0, 1) \) and \( \bar{y} > 1 \):

1. If \( \bar{y} \in (0, 1) \), then the MLE is the unique solution to

   \[
   S = \frac{\partial}{\partial \theta} \log f(\bar{y}|x, \theta) = -\frac{n}{2\sigma^2} (\bar{y} - e^{-\theta x}) x e^{-\theta x} = 0. \tag{2.7}
   \]

2. If \( \bar{y} > 1 \), then the left side of (4.5) is a decreasing function of \( \theta \) and \( x \). Thus the MLE of \( \theta \) is \( \arg\max_\theta \exp\{-\theta x\} = 0 \).

3. If \( \bar{y} < 0 \), then the left side of (4.5) is a increasing function of \( \theta \) and \( x \). The MLE of \( \theta \) is \( \arg\min_\theta \exp\{-\theta x\} = \infty \). The divergence of the MLE to infinity necessitates the restriction of the search to be less than some predetermined constant, \( \tilde{\theta} \).
In summary,
\[
\hat{\theta}_n = \begin{cases} 
-\log \bar{y} / x, & \text{if } \bar{y} \in (e^{-\bar{y}x}, 1), \\
0 & \text{if } \bar{y} \geq 1, \\
\bar{\theta} & \text{if } \bar{y} \leq e^{-\bar{y}x}.
\end{cases}
\]
(2.8)

The estimator \( \hat{\theta}_n \) is a smooth function of \( \bar{y} \) if \( \bar{y} \in (e^{-\bar{y}x}, 1) \) and if \( \theta \in (0, \bar{\theta}) \) then \( \bar{y} \in (e^{-\bar{y}x}, 1) \) a.s. as \( n \to \infty \). Thus it follows that \( \hat{\theta}_n \to \theta \) as \( n \to \infty \) if \( \theta \in (0, \bar{\theta}) \).

The selection of \( \bar{\theta} \) is arbitrary thus the MLE is consistent for \( \theta \in (0, \infty) \). Further,
\[
\sqrt{n} \left( \hat{\theta}_n - \theta \right) \to \mathcal{N} \left( 0, [M(\xi, \theta)]^{-1} \right)
\]
(2.9) as \( n \to \infty \).

It is common practice to use the MLE, \( \hat{\theta}_n \), to estimate \( \theta \) and \( (\text{Var}[S])^{-1} = [nM(x, \theta)]^{-1} \), to approximate \( \text{Var} \left[ \hat{\theta}_n \right] \). We introduce three concerns. First, for small samples sizes \( \hat{\theta}_n \) has large bias. Second, the inverse of the variance of the score function, \([nM(x, \theta)]^{-1}\), is quite different from \( \text{Var} \left[ \hat{\theta}_n \right] \) for small to moderate sample sizes. And third, the approximation \([nM(x, \theta)]^{-1}\) is a function of the parameter, \( \theta \), and thus must be estimated.

For finite samples the probability of the MLE equaling a boundary value, 0 and \( \bar{\theta} \) will be nonzero. Since \( \bar{y}|x \) is distributed \( \Phi \left( \sqrt{n} \left( \bar{y} - e^{-\bar{y}x} \right) / \sigma \right) \), where \( \Phi(\cdot) \) denotes the cumulative standard normal distribution function, the probabilities that the MLE
will equal the boundary points are given by

$$\pi_0 = P\{\hat{\theta}_n = 0\} = P\{\bar{y}_1 \geq 1\} = 1 - \Phi\left(\sqrt{n} \frac{1 - e^{-\theta x}}{\sigma}\right),$$

$$\pi_{\overline{\theta}} = P\{\hat{\theta}_n = \overline{\theta}\} = P\{\bar{y}_1 \leq e^{-\overline{\theta}x}\} = \Phi\left(\sqrt{n} \frac{e^{-\overline{\theta}x} - e^{-\theta x}}{\sigma}\right)$$

and the probability the MLE will be within $(0, \overline{\theta})$, $\pi_{(0,\overline{\theta})} = 1 - (\pi_0 + \pi_{\overline{\theta}})$. Table 2.1 reports $\pi_0$ and $\pi_{\overline{\theta}}$ for $n = \{100, 400, 900, 1600\}$ given values $x = 2$, $\theta = 1$, $\sigma = 1$ and $\overline{\theta} = 4$. When either $\pi_0$ or $\pi_{\overline{\theta}}$ is significantly different from 0, as is the case for $n = 100$ and $n = 400$, the bias and variance of $\hat{\theta}_n$ are greatly increased. To illustrate this effect, Table 2.1 gives the total bias and variance along with conditional contributions to the total given $\hat{\theta}_n = 0$, $\hat{\theta}_n \in (0, \overline{\theta})$ and $\hat{\theta}_n = \overline{\theta}$. For example, when $n = 100$ and $n = 400$ the ratio of the variance of $\hat{\theta}_n$ that can be attributed to the boundary $\overline{\theta}$, $\text{Var}[\hat{\theta}_n|\hat{\theta}_n = \overline{\theta}]$, to the total variance, $\text{Var}[\hat{\theta}_n]$, is $0.627/0.8554 = 0.75$ and $0.03/0.0851 = 0.35$, respectively. Similarly, the bias attributed to the boundary $\overline{\theta}$, $\text{Bias}[\hat{\theta}_n|\hat{\theta}_n = \overline{\theta}]$ to the total bias, $\text{Bias}[\hat{\theta}_n]$, is $0.265/0.0317 = 0.84$ and $0.01/0.051 = 0.10$, for $n = 100$ and $n = 400$, respectively. This represents a significant problem for maximum likelihood estimates for small sample sizes; however, $\pi_0$ and $\pi_{\overline{\theta}}$ go to 0 at a rate much faster than $n$, the rate at which $n\text{Var}[\hat{\theta}_n] \to [M(x, \theta)]^{-1}$ thus this problem is present for only small $n$.

More concerning is the fact that $n\text{Var}[\hat{\theta}_n]$ and $[M(x, \theta)]^{-1}$ differ significantly for small and moderate sample sizes. Table 2.1 collects $\text{Var}[\hat{\theta}_n]$ and $[nM(x, \theta)]^{-1}$ for the same values mentioned previously. From this table we can see that for $n = 900$ and $n = 1600$, $[M(x, \theta)]^{-1}$ is 23% and 10% less than $n\text{Var}[\hat{\theta}_n]$, respectively.
In the preceding discussion calculations have been done assuming the true parameter $\theta$ is known. The most concerning issue when conducting inference for the current model occurs when the approximation $[M(x, \theta)]^{-1}$ is evaluated at $\hat{\theta}_n$, i.e., using $[M(x, \hat{\theta}_n)]^{-1}$ to approximate $n \text{Var}[\hat{\theta}_n]$. Figure 2.1 shows a plot of the approximation $\sqrt{[nM(x, \hat{\theta}_n)]^{-1}}$ as a function of $\hat{\theta}_n$ and $n \text{Var}[\hat{\theta}_n]$ with all values the same as previous. In an ideal situation, the approximation, $\sqrt{[nM(x, \hat{\theta}_n)]^{-1}}$, would lie on a nearly straight line close to $\text{Var}[\hat{\theta}_n]$. Instead we see how sensitive the approximation is to the value of $\hat{\theta}_n$. This indicates a severe risk of underdispersion when $\hat{\theta}_n$ is small and a severe risk of overdispersion when $\hat{\theta}_n$ is large. For example when $\hat{\theta}_n = 3.0$, the 90th percentile, obtained from a Monte Carlo simulation $\sqrt{[nM(x, \hat{\theta}_n)]^{-1}}$ is over 20 times greater than $\text{Var}[\hat{\theta}_n]$. When $\hat{\theta}_n = 0.55$, the 2.5th percentile, $\sqrt{[nM(x, \hat{\theta}_n)]^{-1}}$ is over 7 times smaller than $\text{Var}[\hat{\theta}_n]$. In Figure 2.1 the sample size is increased to 1600; however, we still see that the risk of overdispersion or underdispersion is severe. For this example, when $\hat{\theta}_n = 1.3$, the 97.5th percentile, $\sqrt{[nM(x, \hat{\theta}_n)]^{-1}}$ is over 1.67 times greater the $\text{Var}[\hat{\theta}_n]$. And when $\hat{\theta}_n = 3.0$, the 2.5th percentile, $\sqrt{[nM(x, \hat{\theta}_n)]^{-1}}$ is over 1.5 times less than $\text{Var}[\hat{\theta}_n]$. 

The problems discussed are concerning since they indicate that even for moderate sample sizes traditional Wald confidence intervals may not be dependable. This discussion also highlights the importance of designing experiments, like those presented in this dissertation, that can improve the performance of the maximum likelihood estimates.
Figure 2.1: The approximation $\sqrt{\frac{nM(x, \hat{\theta}_n)}{\hat{\theta}_n}}^{-1}$ (Dotted line) and $\sqrt{\text{Var} [\hat{\theta}_n]}$ (solid line) plotted as a functions of $\hat{\theta}_n$. Values $n = 100$, $\theta = 1.0$, $x = 2.0$, $\sigma = 1$ and $\hat{\theta} = 4.0$ were used.
Figure 2.2: The approximation $\sqrt{\frac{nM(x, \hat{\theta}_n)}{\bar{\theta}^2}}$ (Dotted line) and $\sqrt{\text{Var} \left[ \hat{\theta}_n \right]}$ (solid line) plotted as a functions of $\hat{\theta}_n$. Values $n = 1600$, $\theta = 1.0$, $x = 2.0$, $\sigma = 1$ and $\bar{\theta} = 4.0$ were used.
<table>
<thead>
<tr>
<th>$n$</th>
<th>100</th>
<th>400</th>
<th>900</th>
<th>1600</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_0 = P{\hat{\theta}_n = \bar{\theta}}$</td>
<td>0.089</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$\pi_{\bar{\theta}} = P{\hat{\theta}_n = \bar{\theta}}$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$\pi_0 \text{Var}{\hat{\theta}_n</td>
<td>\hat{\theta}_n = \bar{\theta}}$</td>
<td>0.637</td>
<td>0.030</td>
<td>0.000</td>
</tr>
<tr>
<td>$\pi_{(0,\bar{\theta})} \text{Var}{\hat{\theta}_n</td>
<td>0 &lt; \hat{\theta}_n &lt; \bar{\theta}}$</td>
<td>0.218</td>
<td>0.055</td>
<td>0.019</td>
</tr>
<tr>
<td>$\pi_{\bar{\theta}} \text{Var}{\hat{\theta}_n</td>
<td>\hat{\theta}_n = 0}$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$\text{Var}{\hat{\theta}_n}$</td>
<td>0.8554</td>
<td>0.0851</td>
<td>0.0187</td>
<td>0.0094</td>
</tr>
<tr>
<td>$\pi_1 \text{Bias}{\hat{\theta}_n</td>
<td>\hat{\theta}_n = \bar{\theta}}$</td>
<td>0.265</td>
<td>0.010</td>
<td>0.000</td>
</tr>
<tr>
<td>$\pi_2 \text{Bias}{\hat{\theta}_n</td>
<td>0 &lt; \hat{\theta}_n &lt; \bar{\theta}}$</td>
<td>0.051</td>
<td>0.041</td>
<td>0.017</td>
</tr>
<tr>
<td>$\pi_3 \text{Bias}{\hat{\theta}_n</td>
<td>\hat{\theta}_n = 0}$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$\text{Bias}{\hat{\theta}_n}$</td>
<td>0.317</td>
<td>0.051</td>
<td>0.017</td>
<td>0.009</td>
</tr>
<tr>
<td>$\text{Cov}{S, \hat{\theta}_n}$</td>
<td>2.063</td>
<td>0.750</td>
<td>1.150</td>
<td>1.078</td>
</tr>
<tr>
<td>Lower Bound</td>
<td>0.2817</td>
<td>0.0256</td>
<td>0.0174</td>
<td>0.0092</td>
</tr>
<tr>
<td>$[nM(x, \theta)]^{-1}$</td>
<td>0.1365</td>
<td>0.0341</td>
<td>0.0152</td>
<td>0.0085</td>
</tr>
</tbody>
</table>

Table 2.1: Probability of a Boundary Value, Variance, Bias, and Lower Bound for the MLE of a Fixed One Point Design. Values $\theta = 1.0$, $x = 2.0$, $\sigma = 1$ and $\bar{\theta} = 4.0$ were used.
Chapter 3

Information in a Two Stage Adaptive Optimal Design

3.1 Information Bound in a Two-Stage Experiment

Attainment of the Cramèr-Rao lower bound provides a small sample justification for the use of Fisher’s information in the optimum design of experiments in linear models. We now consider this argument from the viewpoint of a two-stage adaptive design for a nonlinear model.

In the first stage, a vector of independent responses, \( y_1 \) from a distribution containing a single parameter \( \theta \) is observed from \( n_1 \) subjects at a fixed treatment level \( x_1 \). To determine the second stage treatment, a deterministic onto function of the first stage data is used, \( i.e., x_2 = x_2(x_1, y_1) \). Then a vector of responses, \( y_2 \), is observed from \( n_2 \) subjects at the adapted point, \( x_2 \). Note the vectors \( y_1 \) and \( y_2 \) are composed of independent observations, but are not independent of one another. Rather it is
assumed that $y_1$ and $y_2$ are from a joint density

$$f_{y_1,y_2|x_1}(y_1,y_2|x_1,\theta) = f_{y_2|y_1,x_1}(y_2|y_1,x_1,\theta)f_{y_1|x_1}(y_1|x_1,\theta), \quad (3.1)$$

which is bounded and twice differentiable with respect to $\theta \in \Theta$, where $\theta$ is an interior point of $\Theta$.

Let $\tilde{\theta}_n$ be an estimator of $\theta$ based on the $n = n_1 + n_2$ total subjects from stage one and stage two, with finite expectation $E[\tilde{\theta}_n] = \theta + b(x_1, \theta)$. The following derivation of the information inequality for adaptive experiments is based on the derivations in Cox and Hinkley (1974, p. 254) and Hogg, McKean, and Craig (2005, p. 322). Let $S = \partial \log f_{y_1,y_2|x_1}(y_1,y_2|x_1,\theta) / \partial \theta$ denote the score function. Then

$$\text{Cov} [\tilde{\theta}_n, S] = E[\tilde{\theta}_n S] = E \left[ \tilde{\theta}_n \frac{\partial}{\partial \theta} f_{y_1,y_2|x_1} \right] = \frac{\partial}{\partial \theta} E [\tilde{\theta}_n] = 1 + \frac{\partial}{\partial \theta} b(x_1, \theta).$$

By the Cauchy-Schwartz inequality, $\left[ \text{Cov} [\tilde{\theta}_n, S] \right]^2 \leq \text{Var} [\tilde{\theta}_n] \text{Var} [S]$ and therefore

$$\text{Var} [\tilde{\theta}_n] \geq \frac{\left[ 1 + \frac{\partial}{\partial \theta} b(x_1, \theta) \right]^2}{\text{Var} [S]}, \quad (3.2)$$

provided $\text{Var} [S] > 0$.

Most estimators will not attain the lower bound when the sample size $n$ is finite, as equality in (3.2) requires $\tilde{\theta}_n$ to be perfectly linearly correlated with $S$. However, $(\text{Var} [S])^{-1}$ is asymptotically equivalent to $\text{Var} [\hat{\theta}_n]$, where $\hat{\theta}_n$ is the maximum likelihood estimate (MLE) based on the data collected from both stages, provided $\hat{\theta}_n \to \theta$ as $n \to \infty$. The expected information, $\text{Var} [S]$, also known as Fisher’s information, is an approximation of $\left[ \text{Var} [\hat{\theta}_n] \right]^{-1}$. From (3.2) it can be seen that maximizing $\text{Var} [S]$
will result in an optimal lower bound for $\text{Var}[\hat{\theta}_n]$.

### 3.2 The Model and the Adaptive Procedure

This exposition is motivated by phase I dose-finding studies. For such studies the experimental goal is to collect data in order to maximize the information of the dose-response curve. We assume that the dose-response curve follows a known structure, i.e., $n = n_1 + n_2$ subjects are treated with responses

\[
\text{MODEL: } y_{ij} = \eta(x_i, \theta) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad x_i \in \mathcal{X} = [a, b],
\]

where $\eta(x, \theta)$ is a nonlinear function of $x$ and $\theta$. As is typically necessary, due to experimental or practical considerations, the design space is assumed to be a known bounded interval.

It is expected that complications encountered with model (3.3) evaluated at a single point are likely to exist, or be exaggerated, in the presence of more complicated models or more complicated mean functions. Extensions to such situations are straightforward, but would distract from the current exposition.

One complication of interest for finite samples occurs when the mean function, $\eta(x, \theta)$, is bounded on some interval and responses are on $(-\infty, \infty)$. Functions with this property include variations of the logistic, for example, the widely used $E_{\text{MAX}}$ model designs discussed in Fedorov and Leonov (2001), Dragalin, Hsuan, and Padmanabhan (2008), Fedorov and Leonov (2005) and Leonov and Miller (2009).
3.2.1 The Adaptive Stage 2 Treatment

Even though a primary motivation for studying a two-stage design here is clarity of exposition, there is also significant practical appeal for such designs. For instance despite the intuitive appeal of experiments with a large number of stages, Hardwick and Stout (2002), Dragalin, Fedorov, and Wu (2007), Fedorov, Wu, and Zhang (2012) and others have each shown two-stage or few-stage designs to be robust against alternatives.

The parameter, $\theta$, in model (3.3) is an unknown constant. A practical approach is to use stage 1 data to estimate the locally optimal design for use in stage 2. This procedure is similar to the procedures proposed in Dragalin and Fedorov (2005) and Dragalin, Fedorov, and Wu (2007) except that we consider only a two stage procedure, whereas they allow for a large number of stages. When the number of stages is large but the number of observations within each stage is not, proving consistency of parameter estimates based on all of the data is a difficult problem; see Wu (1985), Ford, Titterington, and Wu (1985) and Hu (1998) for examples. In contrast, when there are a finite number of stages, each with large within-stage sample sizes, consistency and normality of $\hat{\theta}_n$ follows assuming standard regularity conditions.

To develop the adaptive procedure, first consider the likelihood derived as if $x_1$ and $x_2$ were both fixed

$$
\mathcal{L}(\theta|y_1,y_2,x_1) \propto \exp \left\{ -\frac{n_1}{2\sigma^2} (\bar{y}_1 - \eta(x_1,\theta))^2 - \frac{n_2}{2\sigma^2} (\bar{y}_2 - \eta(x_2,\theta))^2 \right\},
$$

(3.4)
where $\bar{y}_i = \sum_{j=1}^{n_j} y_{ij}$. Then the score function for subject $j$ in stage $i$ is

$$s_{ij} = \frac{1}{\sigma^2} [y_{ij} - \eta(x_i, \theta)] \frac{\partial \eta(x_i, \theta)}{\partial \theta}$$

with variance

$$\text{Var}[s_{ij}] = \frac{1}{\sigma^2} \left( \frac{\partial \eta(x_i, \theta)}{\partial \theta} \right)^2. \quad (3.5)$$

Let $S_i = \sum_{j=1}^{n_i} s_{ij}$. Then the average expected information from stage $i$, $\text{Var}[S_i]/n_i$, is also given by (3.5) and the locally optimal design point for stage $i$ is the treatment that maximizes the average expected information, i.e.,

$$x^* (\theta) = \arg \max_{x \in \mathcal{X}} \frac{1}{n_i} \text{Var}[S_i] = \arg \max_{x \in \mathcal{X}} \left( \frac{\partial \eta(x, \theta)}{\partial \theta} \right)^2. \quad (3.6)$$

The adaptive optimal procedure assigns the first stage treatment using expert opinion or prior information to approximate the locally optimal design. Then for the second stage treatment an estimate of the locally optimal design based on the stage 1 data is used. A conventional method for estimating $x^* (\theta)$ is to use $x^* (\theta)|_{\hat{\theta}_{n_1}}$, where $\hat{\theta}_{n_1}$ is the MLE based only on the first stage data. Note $\hat{\theta}_{n_1}$ is a function of the stage 1 sufficient statistic $\bar{y}_1$ and the first stage treatment $x_1$ and thus can be expressed $\hat{\theta}_{n_1} = \hat{\theta}_{n_1} (x_1, \bar{y}_1)$ Further it is emphasized that because the second stage treatment depends on $\bar{y}_1$, and thus is dependent on $\varepsilon_{1j}, j = 1, \ldots, n_1$; however, it is independent of $\varepsilon_{2j}, j = 1, \ldots, n_2$.

The presence of boundary points must be addressed when using this procedure. Although $x^* (\theta) \in [a, b]$ there is a positive probability that $x^* (\theta)|_{\hat{\theta}_{n_1}}$ will map to a
point outside of the design restriction \([a, b]\). Therefore, it is necessary to restrict the second stage allocation, \(i.e.,\)

\[
x_2(\hat{\theta}_{n_1}) = \begin{cases} 
  x^*(\theta)|\hat{\theta}_{n_1}, & \text{if } x^*(\theta)|\hat{\theta}_{n_1} \in (a, b) \\
  a, & \text{if } x^*(\theta)|\hat{\theta}_{n_1} \leq a \\
  b, & \text{if } x^*(\theta)|\hat{\theta}_{n_1} \geq b.
\end{cases}
\] (3.7)

The allocation procedure given the first stage data is deterministic and does not depend on the unknown parameter. Thus, the likelihood that results from this procedure is equivalent to (3.4).

### 3.2.2 Fisher’s Information

Let \(I_z\) denote the indicator function (1 if \(z=1\), 0 otherwise), \(w_i = n_i/n, i = 1, 2, \pi_a = P\{x_2(\hat{\theta}_{n_1}) = a\}\) and \(\pi_b = P\{x_2(\hat{\theta}_{n_1}) = b\}\). Then the average Fisher, or average expected, information is

\[
M(\xi_A, \theta) = \frac{1}{n} \text{Var}[S] = \frac{1}{n} \text{Var}[S_1] + \frac{1}{n} \text{Var}[S_2] \\
= \frac{1}{\sigma^2} \left[ w_1 \left( \frac{\partial \eta(x_1, \theta)}{\partial \theta} \right)^2 + w_2 \pi_a \left( \frac{\partial \eta(a, \theta)}{\partial \theta} \right)^2 + w_2 \pi_b \left( \frac{\partial \eta(b, \theta)}{\partial \theta} \right)^2 \\
+ w_2 E_{x_2(\hat{\theta}_{n_1})} \left[ \left( \frac{\partial \eta(x_2(\hat{\theta}_{n_1}), \theta)}{\partial \theta} \right)^2 \cdot I_{a < x_2(\hat{\theta}_{n_1}) < b} \right] \right],
\] (3.8)

where \(\xi_A = \{n_i, x_i\}_1^2\) denotes a two-stage design with \(x_2(\hat{\theta}_{n_1})\) selected adaptively as in (3.7). The boundary probabilities \(\pi_a, \pi_b \to 0\) as \(n_1 \to \infty\); however, for finite
samples these probabilities can be significant. Note the equality in (3.8) uses the fact that \( \text{Cov}[S_1, S_2] = E[S_1 S_2] = E_{y_2 | y_1} [S_2 E_{y_1} [S_1]] = 0. \)

### 3.3 Optimal Selection of the Stage 1 Sample Size

The existing adaptive optimal design literature has primarily developed and examined procedures for updating the estimates of the locally optimal design. Little attention has been devoted to the question of how to allocate the total sample size to each stage. For a two-stage experiment with fixed sample size \( n \), this is accomplished by the selection of \( n_1 \), because \( n = n_1 + n_2 \). The target allocation advocated is the \textit{locally optimal stage one sample size} which we define to be the number that maximizes the average expected information, \( i.e., \)

\[
n_1^* = \underset{n_1 \in (1, n)}{\text{arg max}} \quad M(\xi_A, \theta).
\]  

(3.9)

The term \textit{locally optimal} is used to indicate its dependence on the unknown parameter \( \theta \). Note \( n_1^* \) is defined as the continuous approximation of \( \text{arg max}_{n_1 \in \{1, \ldots, n\}} M(\xi_A, \theta) \). This approximation is motivated by the concept of approximate designs first proposed by Kiefer (1959).

A justification for the usefulness of (3.9) is that \( n \text{Var}(\hat{\theta}_n) \) and \( [M(\xi_A, \theta)]^{-1} \) are asymptotically equivalent. Therefore \( n_1^* \) converges to the first stage sample size that minimizes the variance of \( \hat{\theta}_n \). In general it is difficult to determine \( \lim_{n \to \infty} M(\xi_A, \theta) \); however, when \( d\eta(x, \theta)/d\theta \) is bounded and continuous with respect to \( x \), \( \lim_{n \to \infty} M(\xi_A, \theta) = M(\xi^*, \theta) \), where \( \xi^* \) has the same fixed first stage \( \{n_1, x_1\} \) as \( \xi_A \), but the second stage design is \( \{n_2, x^*(\theta)\} \), \( i.e., \) the second stage uses the unknown optimal treatment.
Consider the following proposition (the proof is in the appendix):

**Theorem 2.** For model (3.3), if \( x^*(\theta) \) an interior point of \( \mathcal{X} \) and the second derivative of \( \left( \frac{\partial \eta(x,\theta)}{\partial \theta} \right)^2 \) is bounded and continuous with respect to \( x \in [a, b] \) then \( n_1^*(\theta) = O(\sqrt{n}) \).

Proposition 2 states that the asymptotically optimal allocation of \( n_1 \) is proportional to \( \sqrt{n} \) (this relationship was noted by Luc Pronzato for a more general model, personal communication, 2012). However, for finite samples using \( n_1 = \sqrt{n} \) will not necessarily be approximately optimal. This allocation rule will be compared to an alternative method proposed later in this section.

For finite sample sizes consider the following two points. First, it can be shown there exists a neighborhood for \( x_1 \) around \( x^*(\theta) \) such that \( n_1^*(\theta) = n \). This may seem counter intuitive at first because it indicates that if the experimenter’s guess of the locally optimal design point is within some small interval around the true locally optimal design point, then a single stage experiment will have greater expected information than any adaptive two stage experiment. To elaborate on this point, insert the representation \( M(\xi_A, \theta) = \frac{n_1}{n} \frac{1}{n_1} \text{Var}[S_1] + \left( 1 - \frac{n_1}{n} \right) \frac{1}{n_2} \text{Var}[S_2] \) into (3.9) to see that if the average expected information from the first stage, \( \frac{1}{n_1} \text{Var}[S_1] \), is greater than the average expected information from the second stage, \( \frac{1}{n_2} \text{Var}[S_2] \), then \( n_1^*(\theta) = n \). In order for \( \text{Var}[S_2] > \text{Var}[S_1] \) it is necessary that \( E[x_2 \left( \hat{\theta}_{n_1} \right)] \) be nearer to \( x^*(\theta) \) than \( x_1 \). In other words, \( n_1 \) must be large enough to ensure that \( x_2 \left( \hat{\theta}_{n_1} \right) \) has converged to within the interval \((x_1, x^*(\theta))\).

Second, provided \( \mathcal{X} \) is a sufficiently large interval, there exists a set of points \( x' < x^*(\theta) \) and \( x'' > x^*(\theta) \) such that for all \( x_1 < x' \) and \( x_1 > x'' \), \( n_1^*(\theta) < n \). This second point justifies the use of adaptive designs since it shows that outside some
small interval of \( x_1 \) around \( x^*(\theta) \) where \( n_1^*(\theta) = n \), there exists a two-stage design with greater average expected information than a single stage design. See Appendix A.2 for further details.

The discussion in the preceding paragraph will apply to any likelihood based adaptive design, meaning that one should not consider a second stage unless the expected value of the second stage treatment is nearer to the target than the first stage treatment. This logic follows regardless of whether the target be defined in terms of the ethical treatment of subjects in the study or, as it is in this paper, the minimization of a concave function of Fisher’s information. To meet this condition a moderate to large number of samples in the first stage may be required.

For an illustration consider model (3.3) with \( \eta(x, \theta) = \exp\{-\theta x\} \), \( n = 100 \), \( \sigma = 1 \) and \( \mathcal{X}^- = [0.25, 10] \). Figure 3.1 shows a plots of \( n_1^*(\theta) \) given \( \theta \), \( [n_1^*(\theta)|\theta] \), as a function \( x_1 \) for \( \theta = 0.2, 0.5 \) and 1. The locally optimal design points for the given \( \theta \) are \( x^*(\theta) = 5 \), 2 and 1, respectively. In the figure it can be seen that for each value \( \theta \) there exists a small interval, with respect to \( x_1 \), where \( n_1^*(\theta) = 100 \). This interval contains the corresponding locally optimal design point \( x^*(\theta) \) and represent the neighborhoods, mentioned previously, where a single stage experiment has greater expected information than any two-stage adaptive experiment. For all other values of \( x_1 \) there exists a two-stage adaptive experiment with greater expected information.

### 3.3.1 Approximation of the locally optimal stage one sample size

One thing to note in Figure 3.1 is how sensitive \( [n_1^*(\theta)|\theta] \) is for different values of \( \theta \). Since \( \theta \) is an unknown quantity and \( x_1 \) is a predetermined constant, a reasonable goal
Figure 3.1: Optimal allocation to stage 1 by $x_1 \in [a, b] = [.25, 10]$. The dashed, solid and dot-dashed lines represent $[n^*_1(\theta) | \theta]$ for $\theta = 0.2, 0.5$ and 1, respectively, using $n = 100$ and $\sigma = 1$. The three dotted vertical dotted lines represent the locally corresponding locally optimal design points for the given $\theta$ values, $x^*(0.2) = 5$, $x^*(0.5) = 2$, $x^*(1) = 1$. 
is to select a sample size allocation that is robust for a range of parameter values. Toward this goal we propose the following procedure:

Step 1: Assign a prior distribution to $\theta$; call it $\nu(\theta)$.

Step 2: Calculate the optimal allocation to stage 1, $n_1^*(\theta)$, conditioned on $\theta$, $[n_1^*(\theta)|\theta]$. To do this divide the domain of $\theta$ into $N$ discrete intervals, where $N$ is selected to achieve a desired precision. Insert the midpoint value of $\theta$ from each of the $N$ intervals into (3.9) and find $n_1^*(\theta)$.

Step 3: Find the weighted average of the $n_1^*(\theta)|\theta$ values found in Step 2, with respect to the prior $\nu(\theta)$ specified in Step 1. For example if $\nu(\theta)$ is the uniform distribution one simply has to find the mean of the $N$ values of found in Step 2.

In summary, we propose allocating $n_1$ of the $n$ subjects to stage 1 using

$$\tilde{n}_1^* := \tilde{n}_1^*(\nu(\theta)) = \int [n_1^*(\theta)|\theta] \, d\nu(\theta).$$

(3.10)

Using values of $n$, $\sigma$ and $\mathcal{X}$ as before, we find $\tilde{n}_1^*$ for $x_1 = 0.5, 1$ and 2. Using an interval length of 0.01 and the uniform prior with domain $[0.1, 5.0]$, we obtain $\tilde{n}_1^* = 54, 35$ and 23, respectively.

To see the potential benefit in using the above procedure, consider Figure 3.2 which shows a plot of the ratio of $M(\xi_A, \theta)$ from adaptive experiments with $n_1 = \{1, \ldots, n\}$ to $M(\xi_A, \theta)$ from an adaptive experiment using $n_1 = \tilde{n}_1^*$ as the solid line for values of $\theta = 0.5, 1, 2$ and 3 with $x_1 = 1$ and $n$, $\sigma$ and $\mathcal{X}$ as before. The dotted line parallel to the $x$-axis at 1.0 is provided for a reference. At each point along the $x$-axis, if the
solid line is less than 1.0 it indicates that an experiment with first stage sample size equal to \( \tilde{n}_1^* \) has greater information than an experiment with the first stage sample size equal to the corresponding \( n_1 \) at that point on the x-axis and visa versa. For example, in the plot where \( \theta = 1 \), at the point on the x-axis where \( n_1/n = w_1 = 1.0 \), the expected information from an experiment with a single stage has a 12% increase in \( M(\xi_A, \theta) \) compared to a design with \( w_1 = \tilde{n}_1^* \). However, if \( \theta = 0.5, 2, \) or 3 using \( w_1 = 1.0 \) results in a significant loss of information when compared to a design with \( n_1 = \tilde{n}_1^* \). For example, if \( \theta = 3 \) a design with \( w_1 = 1.0 \) then the expected information for a design with \( w_1 = 1.0 \) is 75% less than the expected information in a design with \( n_1 = \tilde{n}_1^* \). If one continues to compare the performance of a design with \( n_1 = \tilde{n}_1^* \) against other values of \( n_1 \) it is evident that \( \tilde{n}_1^* \) is robust selection for different values of \( \theta \).

To compare the alternative selection of \( n_1 = \sqrt{n} \) to \( n_1 = \tilde{n}_1^* \), return to Figure 3.2. In each plot dashed line represents the ratio of the average expected information from an experiment with \( n_1 = \sqrt{n} \) to the average expected information from an experiment with \( n_1 = \tilde{n}_1^* \), respectively. When \( \theta = 0.5, 1 \) and 2 using \( n_1 = \tilde{n}_1^* \) represents an increase in information of 32%, 39% and 16% when compared to \( n_1 = \sqrt{n} \). Only in the case when \( \theta = 3 \) does using \( n_1 = \sqrt{n} \) have greater average expectation than \( n_1 = \tilde{n}_1^* \). The increase is only 7% in this case. Overall, an experiment with \( n_1 = \tilde{n}_1^* \) provides a much more robust experiment, in terms of expected information, than does an experiment with \( n_1 = \sqrt{n} \).
Figure 3.2: Ratio of the average expected information from an experiment with $n_1 = \{1, \ldots, n\}$ to average expected information from an experiment with $n_1 = n_1^*$ (solid line). Ratio of average expected information with $n_1 = \sqrt{n} = 10$ to average expected information from an experiment with $n_1 = n_1^*$ (dashed line). Dotted line at 1.0 is provided and is used represent the average expected information from an experiment with $n_1 = n_1^*$ since it is the value used in the denominator.
Chapter 4

Inference in a Two Stage Adaptive Optimal Design

4.1 Information Alternatives

Section 3.3 shows the usefulness of the average expected information, \( M(\xi_A, \theta) \), from a design perspective. In this section \( M(\xi_A, \theta) \) is compared to common alternatives from an analysis perspective. To contrast the two perspectives, consider that an optimal design prescribes treatment allocations so as to minimize a concave function of information; whereas when analysis is the goal, information measures are used to approximate the variance of \( \hat{\theta}_n \). A summary of notation concerning information measures and their estimates is given in Table 4.1.

The most common alternative to \( M \) used in adaptive optimal design is derived by ignoring the stage dependency induced by selecting \( x_2(\hat{\theta}_{a_1}) \) adaptively and instead
<table>
<thead>
<tr>
<th>Shorthand</th>
<th>Full Notation</th>
<th>Brief Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>$M(\xi_A, \theta)$</td>
<td>Fisher’s information</td>
</tr>
<tr>
<td>$M_{ind}$</td>
<td>$M_{ind}(\xi_A, \theta, \hat{\theta}_n)$</td>
<td>Information measure under independence</td>
</tr>
<tr>
<td>$M^*$</td>
<td>$M(\xi^*, \theta)$</td>
<td>Upper bound on information in adaptive experiment</td>
</tr>
<tr>
<td>$\hat{M}$</td>
<td>$M(\xi_A, \hat{\theta}_n)$</td>
<td>Fisher Information with $\theta$ estimated by $\hat{\theta}_n$</td>
</tr>
<tr>
<td>$\hat{M}_{ind}$</td>
<td>$M_{ind}(\xi_A, \hat{\theta}<em>n, \hat{\theta}</em>{n_1})$</td>
<td>Independent measure with $\theta$ estimated by $\hat{\theta}_n$</td>
</tr>
<tr>
<td>$\hat{M}_{obs}$</td>
<td>$M_{obs}(\xi_A, \hat{\theta}<em>n, \hat{\theta}</em>{n_1})$</td>
<td>Observed information estimate</td>
</tr>
</tbody>
</table>

Table 4.1: Summary of notation for information measures and their estimates.
treat all responses as independent, \(i.e.,\)

\[
M_{\text{ind}} := M_{\text{ind}}(\xi_A, \theta, \hat{\theta}_{n_1}) = \frac{1}{\sigma^2} \left[ w_1 \left( \frac{\partial \eta(x_1, \theta)}{\partial \theta} \right)^2 + w_2 \left( \frac{\partial \eta(x_2(\hat{\theta}_{n_1}), \theta)}{\partial \theta} \right)^2 \right]. \tag{4.1}
\]

The primary advantage of \(M_{\text{ind}}\) is convenience. To understand its drawbacks it is helpful to examine the benchmark

\[
M^* := M(\xi^*, \theta) = \frac{1}{\sigma^2} \left[ w_1 \left( \frac{\partial \eta(x_1, \theta)}{\partial \theta} \right)^2 + w_2 \left( \frac{\partial \eta(x^*(\theta), \theta)}{\partial \theta} \right)^2 \right]. \tag{4.2}
\]

\(M^*\) is a benchmark for two-stage experiments with fixed first stage because it is the upper bound on expected information, \(i.e.,\) \(M^* \geq M\). Further \(M \rightarrow M^*\) provided \(\partial \eta(x, \theta)/\partial \theta\) is bounded and continuous with respect to \(x\). These two points make \(M^*\) a useful benchmark in the design of adaptive optimal experiments.

However, considering (3.2), we have for finite samples

\[
n\text{Var} \left( \hat{\theta}_n \right) \geq M^{-1} \geq M^{*-1} \tag{4.3}
\]

which guarantees that \(M\) is closer to \(n\text{Var} \left( \hat{\theta}_n \right)\) than \(M^*\). Therefore, from an analysis perspective \(M^*\) does not represent an informative benchmark. \(M_{\text{ind}}\) is simply a plug in estimate of \(M^* (x_2(\hat{\theta}_{n_1}) \text{ is inserted in place of } x^*(\theta))\) and for the reasons mentioned the appropriateness of using it as an approximation of \(n\text{Var} \left( \hat{\theta}_n \right)\) is called into question.

The preceding comparisons do not consider that \(M\) and \(M_{\text{ind}}\) are functions of the unknown parameter \(\theta\) and thus have to be estimated. These comparisons provide a
theoretical justification for preferring $M$ to $M_{ind}$. However, it is more important to understand the performance of $M$ evaluated at $\hat{\theta}_n$, $\hat{M}$, compared to popular alternatives. The common alternatives are $\hat{M}_{ind}$, $M_{ind}$ evaluated at $\hat{\theta}_n$, and the observed information measure advocated by Efron and Hinkley (1978):

$$M_{obs} := M_{obs}(\xi_A, \hat{\theta}_n, \hat{\theta}_{n_1}) = -\left[ \frac{\partial^2}{\partial \theta^2} \log f(y_1, y_2|\theta, x_1) \right]_{\theta = \hat{\theta}_n} = M_{ind}(\xi_A, \hat{\theta}_n, \hat{\theta}_{n_1})$$

$$- \frac{1}{\sigma^2} \left[ \sum_{i=1}^{2} w_i[y_i - \eta(x_i, \theta)] \frac{\partial^2}{\partial \theta^2} \eta(x_i, \theta) \right]_{\theta = \hat{\theta}_n}. \quad (4.4)$$

Efron and Hinkley (1978) show that $M_{obs}^{-1}$ is an approximation of $n \text{Var}(\hat{\theta}_n|a)$, where $a$ is some ancillary statistic that effects the precision of $\hat{\theta}_n$. Further they argue that $\text{Var}(\hat{\theta}_n|a)$ is a better representation of the data than $\text{Var}(\hat{\theta}_n)$. They provide the theoretical justification for independent identically distributed observations from the translation family of distributions. Extensions to nontranslation families was done via simulation. It was not extended to include dependent observations. From an examination of (4.4), the argument could be made that $M_{obs}$ has no mechanism to account for dependency. As with $M_{ind}$ the appropriateness of this measure for analysis needs to be clarified.

For a general mean function we have given reasons why $M$ should be preferred to common alternatives from an analysis perspective. In the following section a specific example is discussed and then a simulation study is conducted to compare the information measures for a specific mean function.
4.2 Example: An Exponential Mean Function

In model (3.3) let $\eta(x, \theta) = \exp\{-\theta x\}$, $a = .25$, $b = 10$, where the adaptive procedure described in section 3.2.1 is utilized. The model’s simplicity makes it useful for illustrative purposes; but it is also important in its own right. Fisher (1947) used a variant of this model in which $x$ indicates the number of serial dilutions in a laboratory experiment to illustrate the relationship between information on the mean function $\eta(x, \theta)$ and information on $\theta$. Cochran (1973) elaborated on the experiment that motivated Fisher and used the exponential mean function (as we do) to illustrate statistical complications with nonlinear regression more generally. Complications encountered with model (3.3) evaluated at a single point are likely to exist, or be exaggerated, with more complicated designs and/or more complicated mean functions.

Before examining the numeric comparisons we first walk through the adaptive procedure for this example. First note that the first stage likelihood is given by

$$L(\theta | x_1, y_1) = f(y_1 | x_1, \theta) \propto \exp\left\{-\frac{n_1}{2\sigma^2} \left(\bar{y}_1 - e^{-\theta x_1}\right)^2\right\}.$$ 

The mean function, $\exp\{-\theta x\} \in (0, 1)$, thus the likelihood must be maximized separately for $\bar{y}_1 < 0$, $\bar{y}_1 \in (0, 1)$ and $\bar{y}_1 > 1$:

1. If $\bar{y}_1 \in (0, 1)$, then the MLE is the unique solution to

$$S_1 = \frac{\partial}{\partial \theta} \log f(y_1 | x_1, \theta) = -\frac{n_1}{2\sigma^2} \left(\bar{y}_1 - e^{-\theta x_1}\right) x_1 e^{-\theta x_1} = 0. \quad (4.5)$$

2. If $\bar{y}_1 > 1$, then the left side of (4.5) is a decreasing function of $\theta$ and $x_1$. Thus the MLE of $\theta$ is $\arg\max_{\theta} \exp\{-\theta x_1\} = 0$. 

55
3. If $\bar{y}_1 < 0$, then the left side of (4.5) is a increasing function of $\theta$ and $x$. The MLE of $\theta$ is $\arg\min_{\theta} \exp\{-\theta x_1\} = \infty$. The divergence of the MLE to infinity necessitates the restriction of the search to be less than some predetermined constant, $\overline{\theta}$.

In summary, for the one point design

$$
\hat{\theta}_{n_1} = \begin{cases} 
\frac{-\log \bar{y}_1}{x_1}, & \text{if } \bar{y}_1 \in (e^{-\overline{\theta} x_1}, 1), \\
0, & \text{if } \bar{y}_1 \geq 1, \\
\overline{\theta}, & \text{if } \bar{y}_1 \leq e^{-\overline{\theta} x_1}.
\end{cases}
$$

For this example, the locally optimal design is

$$
x^*(\theta) = \arg\max_{x \in \mathcal{X}} \text{Var}\left\{\frac{1}{n_1} S_1\right\} = \arg\max_{x \in \mathcal{X}} \left(\frac{x_1^2 e^{-2\theta x_1}}{2}\right)^2 = \begin{cases} 
\theta^{-1}, & \text{if } a \leq \frac{1}{\overline{\theta}} \leq b, \\
b, & \text{if } \frac{1}{\overline{\theta}} \geq b, \\
a, & \text{if } \frac{1}{\overline{\theta}} \leq a.
\end{cases}
$$

(4.6)

Since $a$ and $\overline{\theta}$ are predetermined constants, one can simplify the procedure and its analysis by selecting $\overline{\theta} = a^{-1}$. In this case, the adaptively selected stage 2 treatment is

$$
x_2(\hat{\theta}_{n_1}) = \begin{cases} 
\hat{\theta}_{n_1}^{-1}, & \text{if } \bar{y}_1 \in (e^{-a^{-1} x_1}, e^{-b^{-1} x_1}), \\
b, & \text{if } \bar{y}_1 \geq e^{-b^{-1} x_1}, \\
a, & \text{if } \bar{y}_1 \leq e^{-a^{-1} x_1}.
\end{cases}
$$

(4.7)
Further, the average expected information, (3.8), is

\[
M(\xi_A, \theta) = \frac{1}{\sigma^2} \left( w_1 x_1^2 e^{-2\theta x_1} + w_2 \pi_a a^2 e^{-2\theta a} + w_2 \pi_b b^2 e^{-2\theta b} + w_2 E_{\bar{y}_1} \left( \frac{-x_1}{\log \bar{y}_1} \right)^2 e^{-2\theta \left( \frac{-x_1}{\log \bar{y}_1} \right)} \cdot I(e^{-a^{-1}x_1} < \bar{y}_1 < e^{-b^{-1}x_1}) \right),
\]

where \( \pi_a = \Phi \left( \sqrt{n_1} \left( e^{-a^{-1}x_1} - e^{-\theta x_1} \right) / \sigma \right) \) and \( \pi_b = 1 - \Phi \left( \sqrt{n_1} \left( e^{-b^{-1}x_1} - e^{-\theta x_1} \right) / \sigma \right) \).

### 4.3 The Final MLE \( \hat{\theta}_n \) of \( \theta \)

Since responses are independent conditional on treatment and \( x_2 \) is an onto function of \( y_1 \), the likelihood is

\[
\mathcal{L}(\theta|y_1, y_2, x_1) \propto \exp \left\{ -\frac{1}{2} \left( n_1 [\bar{y}_1 - e^{-\theta x_1}]^2 + n_2 [\bar{y}_2 - e^{-\theta x_2}]^2 \right) \right\}
\]

which has the same form as when \( x_2 \) is fixed except now the second stage mean is a random function of \( \bar{y}_1 \). Note that, because of this, the joint density of \((y_1, y_2)\) is no longer a member of the exponential family.

The distributions of \( \bar{y}_1 \) and \( \bar{y}_2 \) determine the distribution of \( \hat{\theta}_n \) and hence inference on \( \hat{\theta}_n \). Although, \( \bar{y}_1 \) is normally distributed, the boundaries of the design space and the adaptive selection of \( x_2 \) result in \( \bar{y}_2 \) following a mixture distribution:

\[
f_{\bar{y}_2}(\bar{y}_2) = \pi_a f_{\bar{y}_2|x_2}(\bar{y}_2|x_2 = a) + \pi_b f_{\bar{y}_2|x_2}(\bar{y}_2|x_2 = b) + \int_a^b f_{\bar{y}_2|x_2}(\bar{y}_2|x_2) f_{x_2}(x_2) \, dx_2.
\]

where \( f_{x_2}(x_2) \) is found through the transformation given in (4.7). Further details can
be found in appendix B.3.

For the remainder of this section the values $n = 100$, $w_1 = 0.20$, $x_1 = 2.0$, $\theta = 1.0$, $a = 0.25$, and $b = 10$ are used, unless explicitly stated otherwise, in a simulation to compare the information alternatives discussed in Section 4.1.

Figures 4.1 and 4.2 show histograms of $\bar{y}_1$ and $\bar{y}_2$, respectively, from a simulation of model (3.3) with 10,000 iterations for $n = 100$, $\theta = 1$, $a = .25$, $b = 10$, $x_1 = 2$ and $w_1 = 0.2$. Certainly, $\bar{y}_1$ is normally distributed as assumed. However, the distribution of $\bar{y}_2$ can clearly be seen to be the described mixture distribution.

After the second stage the MLE, $\hat{\theta}_n$, using all the data can be found by solving

$$\frac{1}{n} S = w_1 (\bar{y}_1 - e^{-x_1 \theta}) x_1 e^{-x_1 \theta} + w_2 (\bar{y}_2 - e^{-x_2 \theta}) x_2 e^{-x_2 \theta} = 0 \tag{4.9}$$

subject to boundary conditions, i.e., if $\tilde{\theta}$ is the unique solution to (4.9), then

$$\hat{\theta}_n = \begin{cases} \tilde{\theta} & \text{if } \tilde{\theta} \in (0, a^{-1}) \\ 0 & \text{if } \tilde{\theta} \leq 0 \\ a^{-1} & \text{if } \tilde{\theta} \geq a^{-1}. \end{cases}$$

The upper bound $a^{-1}$ is necessary to guarantee $E[\hat{\theta}_n] < \infty$ for finite sample sizes.

**Theorem 3.** For model 3.3 with $\eta(x, \theta) = e^{-\theta x}$ if $\theta \in (0, a^{-1})$ then $\hat{\theta}_n \longrightarrow \theta$ as $n \longrightarrow \infty$, for $w_i$ fixed.

For proof of this theorem see Appendix B.1. The restriction $\theta \in (0, a^{-1})$ is done to ensure that for finite samples the MLE will exist. This result holds in general for $\theta \in (0, \infty)$ by recognizing that $a$ is can be selected as arbitrarily small.
Figure 4.1: Histogram of $\bar{y}_1$. Mean Responses from 10,000 simulations of Model (3.3) for $n = 100$, $\theta = 1.0$, $\bar{\theta} = 0.1$, $\hat{\theta} = 4.0$, $x_1 = 2$ and $w_1 = 0.2$. 

59
Figure 4.2: Histogram of $\bar{y}_2$ from Stage 2. Mean Responses from 10,000 simulations of Model (3.3) for $n = 100$, $\theta = 1.0$, $\bar{\theta} = 0.1$, $\hat{\theta} = 4.0$, $x_1 = 2$ and $w_1 = 0.2$. 
Theorem 4. For model 3.3 with \( \eta(x, \theta) = e^{-\theta x} \) if \( \theta \in (0, a^{-1}) \) then \( \sqrt{n}(\hat{\theta}_n - \theta) \to N\left(0, \frac{1}{\sigma^2} \left[ w_1x_1e^{-\theta x_1} + w_2x^*e^{-\theta x^*} \right]^{-1} \right) \) as \( n \to \infty \).

For proof of this theorem see Appendix B.2. This result also holds for \( \theta \in (0, \infty) \) using the same argument that \( a \) can be arbitrarily small.

Remark: The results of theorem 3 and 4 extend readily to include any function \( \eta(x, \theta) \) provided standard regularity conditions are satisfied.

4.3.1 Some Comparisons of Estimators of \( \theta \)

In Chapter 2 the MLE from a single stage fixed design was examined. Single stage experiments are significantly less difficult to implement than more complex two-stage adaptive design; thus, it is of interest to compare the maximum likelihood estimates resulting from such experiments. For this section we denote the MLE from a fixed one point design and an adaptive optimal design as \( \hat{\theta}_{ss} \) and \( \hat{\theta}_n \), respectively.

Furthermore, when \( w_1 \) is small and \( x_1 \) is far from the locally optimal design point, \( x^* \), it may be expected that the first stage contributes little information. For this reason we will compare the estimate from a two-stage adaptive design where only the second stage date is used to calculate the MLE, \( \hat{\theta}_{n2} \). This is a common procedure when small pilot studies are used before a more in depth study is conducted.

Asymptotically \( \hat{\theta}_{ss} \), \( \hat{\theta}_{n2} \) and \( \hat{\theta}_n \) are all consistent estimators of \( \theta \); however, the limits of their standard errors differs significantly. Sections 2.1.1, 5.1.2 and 4.3 give the limiting distributions for \( \hat{\theta}_{ss} \), \( \hat{\theta}_{n2} \) and \( \hat{\theta}_n \), respectively. From a comparison of these distributions one can see than \( \hat{\theta}_n \) is more precise, in terms of MSE than the other two asymptotically. However, this is not our main interest in this section; it is our goal to
assess the small sample performance of each estimator. To conduct the comparisons
a set of 10,000 Monte Carlo simulations has been analyzed.

Figures 4.3, 4.4 and 4.5 are histograms from the simulations of \( \hat{\theta}_{ss} \), \( \hat{\theta}_{n_2} \) and \( \hat{\theta}_n \),
respectively. The solid line in each plot represents the true mean \( \theta = 1 \). From a
comparison of the figures it is apparent that the \( \hat{\theta}_n \) is a better estimate of \( \theta \) since it
has less bias, is less skewed and has reduced probability of a boundary value. For this
example the bias of \( \hat{\theta}_{ss} \), \( \hat{\theta}_{n_2} \) and \( \hat{\theta}_n \) is 1.32, 1.19 and 1.09, respectively. The probability
of a boundary, \( \pi_a + \pi_b \) for \( \hat{\theta}_{ss} \), \( \hat{\theta}_{n_2} \) and \( \hat{\theta}_n \) is 0.089, 0.064 and 0.006, respectively. The
standard deviation of \( \hat{\theta}_{ss} \), \( \hat{\theta}_{n_2} \) and \( \hat{\theta}_n \) is 0.927, 0.795 and .457, respectively. For this
example \( \hat{\theta}_n \) outperforms the alternative estimators in every respect.

The individual histograms were provided to help understand the general shape
of the distributions for each estimate separately. For a joint comparison Figure 4.6
plots smooth density histograms of each in a single plot. Again the straight line at
\( \theta = 1.0 \) represents the true parameter. In this figure we can clearly see the advantages
previously discussed of \( \hat{\theta}_n \) when \( n = 100 \). Figure 4.7 shows a plot of the smooth
density histogram for the three estimators when \( n = 1600 \). The performance of \( \hat{\theta}_{ss} \) is
poor in comparison to the other two. \( \hat{\theta}_{n_2} \) and \( \hat{\theta}_n \) perform reasonable similar with the
exception that the standard deviation of \( \hat{\theta}_n \) is less than \( \hat{\theta}_{n_2} \).

In the preceding discussions we evaluated the adaptive procedure at a single fixed
first stage sample size proportion \( w_1 = 0.20 \). We now examine the adaptive procedure
as a function of \( w_1 \). Figure 4.8 plots the biases of \( \hat{\theta}_{ss} \), \( \hat{\theta}_{n_2} \) and \( \hat{\theta}_n \) as functions of \( w_1 \). For
each value \( w_1 = \{0.01, \ldots, .99\} \) a simulation with 10,000 iterations was completed and
\( \hat{\theta}_{n_2} \) and \( \hat{\theta}_n \) calculated for each. Only one simulation of 10,000 was done to calculate
\( \hat{\theta}_{ss} \) since it is fixed with respect to \( w_1 \). This figure shows how dramatic the bias is
Figure 4.3: Histogram of $\hat{\theta}_{ss}$ from a single stage fixed design. Parameter Estimates from 10,000 simulations of Model (3.3) for $n = 100$, $\theta = 1.0$, $\underline{\theta} = 0.1$, $\overline{\theta} = 4.0$, $x_1 = 2$ and $w_1 = 0.2$. 
Figure 4.4: Histogram of $\hat{\theta}_{n_2}$ from Stage 2. Parameter Estimates from 10,000 simulations of Model (3.3) for $n = 100$, $\theta = 1.0$, $\vartheta = 0.1$, $\bar{\theta} = 4.0$, $x_1 = 2$ and $w_1 = 0.2$. 
Figure 4.5: Histogram of $\hat{\theta}_n$ from a two-stage adaptive procedure. Parameter Estimates from 10,000 simulations of Model (3.3) for $n = 100, \theta = 1.0, \bar{\theta} = 0.1, \tilde{\theta} = 4.0$, $x_1 = 2$ and $w_1 = 0.2$. 
Figure 4.6: Smooth density histograms of $\hat{\theta}_{ss}$ (solid line), $\hat{\theta}_{n^2}$ (dotted line) and $\hat{\theta}_{n}$ (dashed line). Parameter Estimates from 10,000 simulations of model (3.3) for $n = 100$, $\theta = 1.0$, $\bar{\theta} = 0.1$, $\bar{\theta} = 4.0$, $x_1 = 2$ and $w_1 = 0.2$. The median, mean and variance of simulated values are shown at the top of each histogram.
Figure 4.7: Smooth density histograms of $\hat{\theta}_{ss}$ (solid line), $\hat{\theta}_{n2}$ (dotted line) and $\hat{\theta}_n$ (dashed line). Parameter Estimates from 10,000 simulations of model (3.3) for $n = 1600$, $\theta = 1.0$, $\theta = 0.1$, $\bar{\theta} = 4.0$, $x_1 = 2$ and $w_1 = 0.2$. The median, mean and variance of simulated values are shown at the top of each histogram.
Figure 4.8: Bias of $\hat{\theta}_{sa}$ (solid line), $\hat{\theta}_{n2}$ (dotted line) and $\hat{\theta}_n$ (dashed line) from model (3.3) plotted by the proportion of subjects in stage 1. Values $n = 100$, $\theta = 1.0$, $\bar{\theta} = 0.1$, $\hat{\theta} = 4.0$, $x_1 = 2$ were used.
reduced when a two-stage adaptive design is used and $\hat{\theta}_n$ is used to estimate $\theta$.

For the same simulations in Figure 4.8, Figure 4.9 compares the probability of a boundary value, $\pi_a + \pi_b$, for the three estimates. Not only does the boundary values effect the variance but they also represent unsatisfactory estimates. The use of two-stage adaptive design is and using $\hat{\theta}_n$ is used to estimate $\theta$ significantly reduces the probability of a boundary value.

4.3.2 Comparison of $n\text{Var}[\hat{\theta}_n]$ to its approximations $M^{-1}$ and $M^{*^{-1}}$.

The usefulness, from an analysis perspective, of any information measure is how accurately it approximates $n\text{Var}[\hat{\theta}_n]$. With this in mind consider Figure 4.10 which shows a plot of the main objects of interest, $n\text{Var}[\hat{\theta}_n]$ along with $M^{-1}$ and $M^{*^{-1}}$ against the proportion allocated to stage 1. Monte Carlo simulations were used to obtain $n\text{Var}[\hat{\theta}_n]$. The terms $M^{-1}$ and $M^{*^{-1}}$ were both obtained numerically. This figure confirms what was predicted in equation (4.3), namely that, $M^{-1}$ is a better approximation to $n\text{Var}[\hat{\theta}_n]$ than $M^{*^{-1}}$. In this plot it appears that neither represent accurate approximations; however, this is simply a result of the small sample size, i.e., as $n$ increases $M^{-1}$ will converge to $n\text{Var}[\hat{\theta}_n]$.

4.3.3 Comparison of $n\text{Var}[\hat{\theta}_n]$ to its approximations $M^{-1}$ and $M_{ind}^{-1}$.

Even though it was shown in section 4.1 and confirmed in the previous subsection that $M^{-1}$ is better than $M_{ind}^{-1}$. $M_{ind}^{-1}$ is significantly more convenient to obtain. Thus it is of interest to assess the impact of this difference. For an exponential mean
Figure 4.9: Probability that parameter estimates are at the boundaries $\hat{\theta}_{ss}$ (solid line), $\hat{\theta}_{n_2}$ (dotted line) and $\hat{\theta}_n$ (dashed line) from model (3.3) plotted by the proportion of subjects in stage 1. Values $n = 100$, $\theta = 1.0$, $\theta = 0.1$, $\bar{\theta} = 4.0$, $x_1 = 2$ were used.
Figure 4.10: Variance approximations by proportion allocated to stage 1. The solid, dashed, and dotted lines represent $M^{*-1}$, $M^{-1}$ and $n\hat{\text{Var}}[\hat{\theta}_n]$ by $w_1$, respectively. Values $x_1 = 2$, $n = 100$ and $\sigma = 1$ were used.
function, $d\eta(x, \theta)/d\theta$ is bounded and continuous with respect to $x$ and thus $M^{-1} - M^{-1}_{\text{ind}}$ converges to 0 in probability. However, the performance for finite samples is of more interest. This comparison is difficult to express due in large part to the dependence of $M^{-1}_{\text{ind}}$ on $\hat{\theta}_n$. For a given experiment either measure could be closer to $n\text{Var}[\hat{\theta}_n]$. Therefore, it is appropriate to consider the probability that $M^{-1}$ is closer to $n\text{Var}[\hat{\theta}_n]$ than $M^{-1}_{\text{ind}}$, i.e.,

$$P\left(\left|M^{-1}_{\text{ind}} - n\text{Var}[\hat{\theta}_n]\right| \leq \left|M^{-1} - n\text{Var}[\hat{\theta}_n]\right|\right).$$

Unfortunately, because no closed form solution for $\text{Var}[\hat{\theta}_n]$ exists, (4.10) cannot be determined exactly through an analytical or numerical procedure. Equation (4.10) can approximated using Monte Carlo simulation to obtain $\text{Var}[\hat{\theta}_n]$. Figure 4.11 shows a plot of the approximated probability in (4.10). In this figure if the solid line is greater than 1/2 it indicates that $M^{-1}$ has greater probability of more accurately approximating $n\text{Var}[\hat{\theta}_n]$ than $M^{-1}_{\text{ind}}$. For $w_1 < 0.15$ this probability is less than 1/2. The under performance of $M^{-1}$ for small $w_1$ is due in large part to boundary probabilities, i.e., $P(x_2(\hat{\theta}_{n_1}) = a \text{ or } b)$ which has a greater effect on $M^{-1}$ than $M^{-1}_{\text{ind}}$. $P(x_2(\hat{\theta}_{n_1}) = a \text{ or } b)$ decrease as $w_1$ increases. For all values of $w_1 > 0.15$ there is greater probability that $M^{-1}$ is a better approximation to $n\text{Var}[\hat{\theta}_n]$ than $M^{-1}_{\text{ind}}$.

### 4.3.4 Comparison of estimates of information $\hat{M}^{-1}$, $\hat{M}^{-1}_{\text{ind}}$ and $\hat{M}^{-1}_{\text{obs}}$

Comparisons in the preceding two subsections were carried out without estimating $\theta$. For an actual experiment the added variability caused by inserting $\hat{\theta}_n$ into the information alternatives will significantly affect the analysis. In this subsection we examine this effect for the proposed estimates of information. Figures 4.12(a), 4.12(b)
Figure 4.11: $P\left(\left|M_{ind}^{-1} - n\text{Var}[\hat{\theta}_n]\right| < \left|M^{-1} - n\text{Var}[\hat{\theta}_n]\right|\right)$ given $\theta = 1$, $x_1 = 2$, $a = .24$, $b = 4$, and $n = 100$ by proportion allocated to stage 1.
Figure 4.12: Variance approximations divided by $n \hat{\text{Var}} \left[ \hat{\theta}_n \right]$ as a function of the proportion allocated to stage 1. The solid lines represent 1.0 is for reference and represents $n \hat{\text{Var}} \left[ \hat{\theta}_n \right]$. The dotted, dot-dashed, and dashed lines are the 25th, 50th and 75th quantiles of the three information measure estimates.
and 4.12(c) plot the 25th, 50th and 75th quantiles of \( \hat{M}^{-1}, \hat{M}_{\text{ind}}^{-1} \) and \( \hat{M}_{\text{obs}}^{-1} \) divided by \( n\hat{\text{Var}}[\hat{\theta}] \) obtained via Monte Carlo approximations as a function of \( w_1 \), respectively. What would be desired from an approximation of \( n\hat{\text{Var}}[\hat{\theta}] \) in this figure would be to see the quantiles centered, symmetric and narrow around 1.0. This would indicate that the approximation and the variance are approximately equal. From this figure there is no indication that the information measure differ as approximations of \( n\hat{\text{Var}}[\hat{\theta}] \).

In Figure 4.12, 10,000 simulations were completed, then for each simulation \( \hat{M}^{-1}, \hat{M}_{\text{ind}}^{-1} \) and \( \hat{M}_{\text{obs}}^{-1} \) were evaluated at the resulting estimates. The resulting quantiles are what is plotted in Figure 4.12. This masks a very concerning difference that can be seen if a different approach is taken. Instead for the following comparisons the 25th, 50th and 75th quantiles of the final MLE, \( \hat{\theta}_n \), from the 10,000 simulations were found. Then \( \hat{M}^{-1}, \hat{M}_{\text{ind}}^{-1} \) and \( \hat{M}_{\text{obs}}^{-1} \) are evaluated at the quantile values of \( \hat{\theta}_n \). Note \( \hat{M}_{\text{ind}}^{-1} \) and \( \hat{M}_{\text{obs}}^{-1} \) depend on \( \hat{\theta}_{n1} \) and \( \hat{\theta}_{n1}, \hat{y}_1 \) and \( \hat{y}_2 \) in addition to \( \hat{\theta}_n \), respectively. In order to be as generous as possible the median for each of these values was used for subsequent comparisons. The performance is much worse than presented when other quantile values are used. It would be expected that using this approach would yield identical plots presented in Figure 4.12. For \( \hat{M}^{-1} \) the two plots are exactly the same except for the scale; however, for \( \hat{M}_{\text{ind}}^{-1} \) and \( \hat{M}_{\text{obs}}^{-1} \) the difference is severe. Figure 4.13 plots the same comparisons presented in Figure 4.12 except \( \hat{M}^{-1}, \hat{M}_{\text{ind}}^{-1} \) and \( \hat{M}_{\text{obs}}^{-1} \) were evaluated at the 25th, 50th and 75th quantiles of \( \hat{\theta}_n \). Once again it is desired to see the quantiles of the approximations centered, symmetric and narrow around 1.0. None of the three approximations are symmetric around 1.0; however, the quantiles \( \hat{M}^{-1} \) forms a much tighter interval than the other two. This is a strong indication of usefulness of \( \hat{M}^{-1} \) as compared to the alternatives as an estimate of \( n\hat{\text{Var}}[\hat{\theta}] \).
Figure 4.13: Variance approximations evaluated at the $25^{th}$, $50^{th}$ and $75^{th}$ quantiles $\hat{\theta}_n$ divided by $n\widehat{\text{Var}}[\hat{\theta}_n]$ as a function of the proportion allocated to stage 1. The solid lines represent 1.0 is for reference and represents $n\widehat{\text{Var}}[\hat{\theta}_n]$. The dotted, dot-dashed, and dashed lines are the $25^{th}$, $50^{th}$ and $75^{th}$ quantiles of the three information measure estimates. Values $\theta = 1$, $x_1 = 2$, $\sigma = 1$, $\mathcal{X} \in (.25, 4)$ and $n = 100$ were used in the simulation.
Figure 4.14: Variance approximations evaluated at the 25th, 50th and 75th quantiles $\hat{\theta}_n$ divided by $n\widehat{\text{Var}} \left[ \hat{\theta}_n \right]$ as a function of the proportion allocated to stage 1. The solid lines represent 1.0 as a reference and represent $n\widehat{\text{Var}} \left[ \hat{\theta}_n \right]$. The dotted, dot-dashed, and dashed lines are the 25th, 50th and 75th quantiles of the three information measure estimates. Values $\theta = 1$, $x_1 = 2$, $\sigma = 0.25$, $\mathcal{X} \in (0.25, 4)$ and $n = 100$ were used in the simulation.
For another illustration of the differences that can be seen when $M^{-1}$, $\hat{M}_{\text{ind}}^{-1}$ and $\hat{M}_{\text{obs}}^{-1}$ are evaluated at the quantile values of $\hat{\theta}_n$ the simulation was repeated with everything the same as previous except $\sigma = .25$. Figure 4.14 plots the same information as Figure 4.13 for this new simulation. From a comparison of the two pictures it can be seen that $\hat{M}^{-1}$ has improved and is approximately centered, symmetric and narrow around 1.0. This is what is expected when the standard deviation, $\sigma$, is decreased. $\hat{M}_{\text{ind}}^{-1}$ and $\hat{M}_{\text{obs}}^{-1}$ perform very poorly, in fact their performance is not better than when $\sigma = 1$.

The question of why the performance of $\hat{M}_{\text{ind}}^{-1}$ and $\hat{M}_{\text{obs}}^{-1}$ differs for the two different simulation approaches needs to be addressed. The explanation is that for certain combinations of $\hat{\theta}_{n_1}$ and $\hat{\theta}_n$, $\hat{M}_{\text{ind}}^{-1}$ and $\hat{M}_{\text{obs}}^{-1}$ may represent a good or a poor approximate of $n\text{Var}[\hat{\theta}]$. However, there is no association between $\hat{\theta}_{n_1}$ and $\hat{\theta}_n$ representing good estimates of $\theta$ and $\hat{M}_{\text{ind}}^{-1}$ and $\hat{M}_{\text{obs}}^{-1}$ representing a good approximate of $n\text{Var}[\hat{\theta}]$. This is a concerning problem not present for $\hat{M}^{-1}$. If $\hat{\theta}_n$ estimates $\theta$ well then $\hat{M}^{-1}$ estimates $n\text{Var}[\hat{\theta}]$ well.

An additional difference can be seen in Figure 4.15 which provides a comparison of the frequency with which

$$
|\hat{M}_i^{-1} - n\text{Var}[\hat{\theta}_n]| > |\hat{M}_j^{-1} - n\text{Var}[\hat{\theta}_n]|,
$$

$\{\hat{M}_i^{-1}, \hat{M}_j^{-1}\} \subset \{\hat{M}_{\text{ind}}^{-1}, \hat{M}^{-1}, \hat{M}_{\text{obs}}^{-1}\}$, $i \neq j$. Each line in Figure 4.15 represents a comparison of two measures, and it can be seen that $\hat{M}^{-1}$ is closer to $n\text{Var}[\hat{\theta}_n]$ than both $\hat{M}_{\text{obs}}^{-1}$ and $\hat{M}_{\text{ind}}$ for nearly all values of $w_1$.

The dotted line in Figure 4.15 is simply an approximation of (4.10) obtained via Monte Carlo simulation and thus the result is expected.
Figure 4.15: Proportion of times one information measure estimate is closer to $n\hat{\text{Var}}[\hat{\theta}_n]$ than another. The dotted line, $\hat{M}^{-1}$ is closer than $\hat{M}_{\text{ind}}^{-1}$, the dashed line, $\hat{M}^{-1}$ is closer than $\hat{M}_{\text{obs}}^{-1}$ and the solid line, $\hat{M}_{\text{obs}}^{-1}$ is closer than $\hat{M}_{\text{ind}}^{-1}$. 
The message that seems to be clear, at least for an exponential mean function, is that $\hat{M}^{-1}$ is more appropriate to use for adaptive designs and has a benefit in accuracy for inference following an adaptive design.
Chapter 5

Two-Stage Adaptive Optimal Design with Fixed First Stage Sample Size

5.1 Asymptotic Properties

We examine three different ways of deriving the asymptotic distribution of the final MLE which may be used for inference at the end of the study. The first is under the assumption that both $n_1$ and $n_2$ are large. The second considers the data from the second stage alone. Finally, assume a fixed first stage sample size and a large second stage sample size.
5.1.1 Large Stage 1 and Stage 2 Sample Sizes

Provided common regularity conditions hold,

\[ \sqrt{n} \left( \hat{\theta}_n - \theta \right) \xrightarrow{D} \mathcal{N} \left( 0, [M(\xi^*, \theta)]^{-1} \right), \quad (5.1) \]

as \( n_1 \to \infty \) and \( n_2 \to \infty \), where as before \( \xi^* = \{(x_1, n_1), (x^*(\theta), n_2)\} \). This result is used to justify the common practice of using \( x^*(\theta)|_{\theta=\hat{\theta}_{n_1}} \) to estimate \( x^*(\theta) \) in order to make inferences about \( \theta \). However, in Chapter 4 it was shown that \( [M(\xi_A, \theta)]^{-1} \) is a better approximation of \( n\text{Var}[\hat{\theta}_n] \) than \( [M(\xi^*, \theta)]^{-1} \) for finite sample sizes. The distribution with \( [M(\xi_A, \theta)]^{-1} \) inserted for \( [M(\xi^*, \theta)]^{-1} \) is also of interest.

5.1.2 Distribution of the MLE if Only Second Stage Data are Considered

Often pilot data are discarded after being used to inform the design of the second stage. The distribution of the MLE found using only the second stage data and treating \( x_2 \) as fixed is

\[ \sqrt{n_2} \left( \hat{\theta}_{n_2} - \theta \right) \xrightarrow{D} \mathcal{N} \left( 0, [M_2(x_2, \theta)]^{-1} \right), \quad (5.2) \]

as \( n_2 \to \infty \), where \( M_2(x_2, \theta) = \sigma^{-2} (\partial \eta(x_2, \theta)/\partial \theta)^2 \). The estimate \( \hat{\theta}_{n_2} \) will likely perform poorly in comparison to \( \hat{\theta}_n \) if \( n_1 \) and \( n_2 \) are relatively the same size but conceivably may perform quite well when \( n_1 \) is much smaller than \( n \). For this reason it represents an informative benchmark distribution.
5.1.3 Fixed First Stage Sample Size; Large Second Stage Sample Size

When the first stage sample size is fixed and the second stage is large we have the following result:

**Theorem 5.** For model (3.3) with $x_2$ as defined in (3.7) if $\frac{\partial}{\partial \theta} \eta(x,\theta) \neq 0$ for all $x \in \mathcal{X}$, $x_2$ an onto function of $\bar{y}_1$, $\theta \in \Theta$, $|\frac{\partial}{\partial \theta} \eta(x,\theta)| < \infty$ and provided common regularity conditions (proof in Appendix 5),

$$\sqrt{n} \left( \hat{\theta}_n - \theta \right) \xrightarrow{D} UQ \quad (5.3)$$

as $n_2 \to \infty$, where $Q \sim \mathcal{N}(0,\sigma^2)$ and $U = \left( \frac{\partial \eta(x_2,\theta)}{\partial \theta} \right)^{-1}$ is a random function of $\bar{y}_1$.

**Remark:** Provided $\frac{\partial \eta(x,\theta)}{\partial \theta}$ is bounded and continuous $UQ$ is the asymptotic distribution of $\sqrt{n} \left( \hat{\theta}_n - \theta \right)$ as $n \to \infty$. The important case for this exposition is presented in Theorem 5. However, the two other potential cases can be shown easily.

Case 1: $n_1 \to \infty, n_2 \to \infty$ and $n \to \infty$. As $n_1 \to \infty, x_2 \to x^*(\theta)$ which implies that $U \to \left[ \partial(x^*(\theta'),\theta)/\partial \theta \right]_{\theta' = \theta}^{-1}$, a constant and thus $UQ$ converges to asymptotic distribution of $\sqrt{n} \left( \hat{\theta}_n - \theta \right)$ given in (5.1.1).

Case 2: $n_1 \to \infty, n_2$ fixed and $n \to \infty$. Just as in case 1, $U \to [M(x^*(\theta),\theta)]^{-1}$, where $M(x^*(\theta),\theta) = [\partial(x^*(\theta'),\theta)/\partial \theta]_{\theta' = \theta}^{-1}$. Note that $M(x^*(\theta),\theta)$ differs from $M(\xi^*,\theta)$ which depends on $x_1$ and $x^*(\theta)$. Therefore $UQ \to \mathcal{N}(0,\sigma^2[M(x^*(\theta),\theta)]^{-1})$. Look back at equation (C.4) in the proof, but now take $n_2$ to be fixed; $\frac{1}{\sqrt{n}} S_2 \to 0$ and $\frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} S_2 \to 0$ and the only term left is

$$\frac{1}{\sqrt{n}} S_1 \quad \frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} S_1.$$
\( \frac{\sqrt{n}}{n} S_1 \to \mathcal{N}(0, [M(x^*(\theta), \theta)]) \) and \( \frac{\sqrt{n}}{\sqrt{\pi} \partial} S_1 \to [M(x^*(\theta), \theta)]^{-1} \) as \( n \to \infty \). Therefore, 
\[ \sqrt{n}(\hat{\theta}_n - \theta) \to \mathcal{N}(0, \sigma^2[M(x^*(\theta), \theta)]^{-1}) \] as \( n \to \infty \) which is equivalent to \( UQ \).

### 5.2 Example: One Parameter Exponential Mean Function

In model (3.3) let \( \eta(x, \theta) = e^{-\theta x} \), where \( x \in \mathcal{X} = [a,b], 0 < a < b < \infty \) and \( \theta \in (0, \infty) \).

The limiting distributions of the MLE in subsections 5.1.1 and 5.1.2 can be derived easily. For the asymptotic distribution of the MLE in 5.1.3 consider the following theorem. For details on the functions \( h, v_1 \) and \( v_2 \) see the proof of theorem 6 in Appendix C.1.

**Theorem 6.** If \( \eta(x, \theta) = e^{-\theta x} \) in model (3.3) then 

\[ \sqrt{n}(\hat{\theta}_n - \theta) \overset{D}{\to} UQ \]

as \( n \to \infty \) where \( UQ \) is defined by

\[
P\{UQ \leq t\} = \begin{cases} 
P\{U \geq t/q\} - \infty < q \leq 0 \} \Phi(q), & \text{if } t \in (-\infty, 0) \\
P\{U \leq t/q\} 0 < q \leq \infty \} [1 - \Phi(q)], & \text{if } t \in (0, \infty), \end{cases}
\]

where \( \Phi(\cdot) \) is the standard normal cumulative distribution function. Let \( \Psi(q) = \)
\[ \Phi(\sqrt{n}(q - \eta(x, \theta))/\sigma) \text{ and } h(s) = s^{-1}e^{\theta s}. \text{ Then if } h(a) < h(b) \]

\[ P\{U \geq t/q| -\infty < q \leq 0\} \Phi(q) = \Phi(t/\sigma h(1/\theta)) \]

\[ + [1 - (\Psi(v_2(h(a))) - \Psi(v_1(h(a))))] \times [\Phi(t/\sigma h(a)) - \Phi(t/\sigma h(1/\theta))] \]

\[ + [\Psi(v_2(h(b))) - \Psi(v_2(h(a)))] \times [\Phi(t/\sigma h(b)) - \Phi(t/\sigma h(a))] \]

and

\[ P\{U \leq t/q|0 < q \leq \infty\} [1 - \Phi(q)] = \Phi(t/\sigma h(b)) \]

\[ + [\Psi(v_2(h(a))) - \Psi(v_1(h(a)))] \times [\Phi(t/\sigma h(1/\theta)) - \Phi(t/\sigma h(a))] \]

\[ + [1 - (\Psi(v_2(h(b))) - \Psi(v_2(h(a))))] \times [\Phi(t/\sigma h(a)) - \Phi(t/\sigma h(b))] . \]

If \( h(b) < h(a) \) then

\[ P\{U \geq t/q| -\infty < q \leq 0\} \Phi(q) = \Phi(t/\sigma h(1/\theta)) \]

\[ + [1 - (\Psi(v_2(h(b))) - \Psi(v_1(h(b))))] \times [\Phi(t/\sigma h(b)) - \Phi(t/\sigma h(1/\theta))] \]

\[ + [\Psi(v_1(h(b))) - \Psi(v_1(h(a)))] \times [\Phi(t/\sigma h(b)) - \Phi(t/\sigma h(a))] \]

and

\[ P\{U \leq t/q|0 < q \leq \infty\} [1 - \Phi(q)] = \Phi(t/\sigma h(a)) \]

\[ + [\Psi(v_2(h(b))) - \Psi(v_1(h(b)))] \times [\Phi(t/\sigma h(1/\theta)) - \Phi(t/\sigma h(b))] \]

\[ + [1 - (\Psi(v_1(h(b))) - \Psi(v_1(h(a))))] \times [\Phi(t/\sigma h(b)) - \Phi(t/\sigma h(a))] . \]
5.2.1 Comparisons of Asymptotic Distributions

First, consider the distribution described in (5.1.1) using $M(\xi_A, \theta)$ in place of $M(\xi^*, \theta)$ and the distribution described in (5.2). When $n_1$ is significantly smaller than $n_2$, $M(\xi_A, \theta)$ and $M(x_2, \theta)$ can differ significantly as a function of $\bar{y}_1$. This is primarily because $M(x_2, \theta)$ is a function of $x_2$, whereas $M(\xi_A, \theta)$ is an average over $\bar{y}_1$. Through simulation it can be seen that a $N\left(0, [M(x_2, \theta)]^{-1}\right)$ is a better approximate distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ than $N\left(0, [M(\xi_A, \theta)]^{-1}\right)$ for only a small interval of $x_2$, and this interval has a very small probability. For these reasons the distribution of the MLE using only the second stage data as described in subsection (5.1.2) is not considered further.

Now for a set of numeric examples consider three distributions: (5.1.1), (5.1.1) using $M(\xi_A, \theta)$ in place of $M(\xi^*, \theta)$ and the distribution of $UQ$ defined in (5.3). An asymptotic distribution can be justified in inference if it is approximately equal to the true distribution. In this case the true distribution is that of $\sqrt{n}(\hat{\theta}_n - \theta)$. However, $\hat{\theta}_n$ does not have a closed form and thus its distribution cannot be obtained analytically or numerically. To approximate this distribution 10,000 Monte Carlo simulations have been completed for each example to create a benchmark distribution.

Figure 5.1 plots the three different candidate approximate distributions, found exactly using numerical methods, together with the distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ approximated using Monte Carlo simulations, for $\theta = 1$, $x_1 = 2$, $\sigma = .5$, $a = .25$, $b = 4$, $n_1 = 5$ and $n = \{30, 1000\}$. Note the y-axis represents $P(T_i \leq t)$, $i = 1, 2, 3$, where $T_1$ is $N\left(0, [M(\xi^*, \theta)]^{-1}\right)$, $T_2$ is $N\left(0, [M(\xi_A, \theta)]^{-1}\right)$ and $T_3$ is $UQ$. When $n = 30$ it is difficult, graphically, to determine if $T_2$ or $T_3$ provides a better approximation for $\sqrt{n}(\hat{\theta}_n - \theta)$. It seems that if $t \in (0, 4)$ the two distributions appear to be ap-
proximately equal; however, when \( t \in (-4, 0) \) \( T_3 \) appears to be closer to the target distribution that. It is fairly clear that for this example \( T_1 \) performs poorly.

When \( n = 1000 \), it is clear that \( T_3 \) is much closer to \( \sqrt{n}(\hat{\theta}_n - \theta) \) than both \( T_1 \) and \( T_2 \). Further, comparing the two plots one can see how the distribution of \( \sqrt{n}(\hat{\theta}_n - \theta) \) has clearly converged to \( UQ \) but still differs from those \( T_1 \) and \( T_2 \) significantly, as predicted by Theorem 5 and 6.

Using only graphics it is difficult to assess which of \( T_1 \), \( T_2 \) and \( T_3 \) is nearest \( \sqrt{n}(\hat{\theta}_n - \theta) \) for a variety of cases. To get a better understanding, the integrated absolute difference of the CDFs of \( T_1 \), \( T_2 \) and \( T_3 \) versus that of \( \sqrt{n}(\hat{\theta}_n - \theta) \) for \( x_1 = 2 \), \( \sigma = 0.5 \), \( a = 0.25 \), \( b = 4 \), \( n = \{5, 10, 15\} \) and \( n = \{30, 50, 100, 400\} \) are presented in Table 5.1. First consider the table where \( \theta = 0.5 \). The locally optimal stage 1 design point is \( x_1 = 2 \) when \( \theta = 0.5 \); as a result this scenario is the most generous to distribution \( T_1 \). However, even for this ideal scenario \( T_3 \) outperforms \( T_1 \) and \( T_2 \) for all values of \( n_1 \). In many cases the difference between \( T_3 \) and \( T_1 \) is quite severe. In this scenario \( T_3 \) outperforms \( T_2 \), however, the differences are not great.

Next examine the results for \( \theta = 1 \) and \( \theta = 1.5 \). Once again \( T_3 \) outperforms \( T_1 \) and \( T_2 \) in all but 2 cases, where in many cases its advantage is quite significant. Also note that \( T_2 \) outperforms \( T_1 \) about half the time when \( \theta = 1 \) and the majority of the time when \( \theta = 1.5 \). This supports our observation that when the distance between \( x_1 \) and \( x^*(\theta) \) increases the performance of \( T_1 \) compared with \( T_2 \) and \( T_3 \) worsens which indicates a lack of robustness for the commonly used distribution \( T_1 \). This lack of robustness is not evident for \( T_1 \) and \( T_2 \).

If \( n_1 \to \infty \), \( T_1 \), \( T_2 \) and \( T_3 \) have the same asymptotic distribution. Although our method is motivated by the scenario where \( n_1 \) is a small pilot study, there is no
Figure 5.1: In each plot the CDF of $\sqrt{n}(\hat{\theta}_n - \theta)$ was obtained via Monte Carlo simulations. The $P(T_1 \leq t)$, where $T_1 \sim N(0, [M(\xi^*, \theta)]^{-1})$. The $P(T_2 \leq t)$, where $T_2 \sim N(0, [M(\xi_A, \theta)]^{-1})$. The $P(T_3 \leq t)$, where $T_3 \sim UQ$. Values $\theta = 1$, $x_1 = 2$, $n_1 = 5$, $\sigma = 0.5$, $a = 0.25$ and $b = 4$ were used.
Table 5.1: Integrated absolute difference of the cumulative distributions ($\times 100$) of $T_1 \sim \mathcal{N}(0, [M(\xi^*, \theta)]^{-1})$, $T_2 \sim \mathcal{N}(0, [M(\xi_A, \theta)]^{-1})$ and $T_3 \sim UQ$ versus the approximate cumulative distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ obtained via Monte Carlo simulations for various $n_1$ and various moderate sizes of $n$. Values $\theta = 1$, $x_1 = 2$, $\sigma = 0.5$, $a = 0.25$ and $b = 4$ were used.
theoretical reason that $T_3$ will not perform competitively when $n_1$ is large. Table 5.2 presents the integrated differences for the distributions $T_2$ and $T_3$ from $\sqrt{n}(\hat{\theta}_n - \theta)$ for $x_1 = 2$, $\theta = 1$, $\sigma = 0.5$, $a = 0.25$, $b = 4$, $n_1 = \{50, 100, 200\}$ and $n = \{400, 1000\}$. $T_1$ is not included in the table since due to the lack of robustness; it can perform better or worse than the other two distributions based on the value of $\theta$. Even with larger values of $n_1$, $T_3$ performs slightly better when $n_1 = 50$ and $100$ and only slightly worse when $n = 200$ indicating that using $T_3$ is robust for moderately large $n_1$.

Through the discussion of Tables 5.1 and 5.2 it was shown that $T_3$ is a better representation of the distribution of the $\sqrt{n}(\hat{\theta}_n - \theta)$ than the alternative. However, when inference about the parameters is the main goal it is important to compare the tail values of the candidate distributions against the true tail values. Table 5.3 compares the values of $\alpha$ and $\beta$ such that $P\{T_i \leq \alpha\} = 0.025$ and $P\{T_i \geq \beta\} = 0.975$, $i = 2, 3$ versus the $P\{\hat{\theta}_n \leq \alpha\} = 0.025$ and $P\{\hat{\theta}_n \leq \alpha\} = 0.975$ for $n = 30$ and $n = 50$, and all other values the same as before. What would represent a good approximate distribution would be to have both quantiles 0.025 and 0.975 be near the true quantiles found via Monte Carlo simulations. A pattern that seems to be present when examining $T_2$ and $T_3$ compared to the true, is that if one is closer to 0.025 then other will be closer to 0.975 for a given scenario. When $T_2$ is closer the the desired quantile it is usually only small improvement over $T_3$; however, when $T_3$ is closer to the desired quantile the improvement over $T_2$ can be significant.
\[ \theta = 1.0 \]

<table>
<thead>
<tr>
<th>( n_1 )</th>
<th>( T_1 )</th>
<th>( T_2 )</th>
<th>( T_3 )</th>
<th>( T_1 )</th>
<th>( T_2 )</th>
<th>( T_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>21</td>
<td>6</td>
<td>18</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>8</td>
<td>21</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>9</td>
<td>20</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>13</td>
<td>9</td>
<td>13</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>11</td>
<td>14</td>
<td>4</td>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Integrated absolute difference of the cumulative distributions \((\times 100)\) of \( T_1 \sim \mathcal{N}(0, [M(\xi^*, \theta)]^{-1}) \), \( T_2 \sim \mathcal{N}(0, [M(\xi, \theta)]^{-1}) \) and \( T_3 \sim UQ \) versus the approximate cumulative distribution of \( \sqrt{n}(\hat{\theta}_n - \theta) \) obtained via Monte Carlo simulations for various \( n_1 \) and various large sizes of \( n \). Values \( \theta = 1, \ x_1 = 2, \ \sigma = 0.5, \ a = 0.25 \) and \( b = 4 \) were used.
Table 5.3: Quantiles from the distribution of $\hat{\theta}_n$ obtained via a Monte Carlo simulation are represented as “True”. Median quantiles for $T_2 \sim \mathcal{N} \left(0, [M(\xi, \theta)]^{-1}\right)$ and $T_3 \sim UQ$ are presented for comparison. $x_1 = 2, \sigma = 0.5, a = 0.25$ and $b = 4$ were used.
Chapter 6

Summary and concluding remarks

6.1 Discussion

There is little research on inference for adaptive designs. Rosenberger, Flournoy, and Durham (1997) give conditions under which the MLE from adaptive designs will be asymptotically normally distributed. Durham, Flournoy, and Rosenberger (1997) recognized that to find Fisher’s information for an adaptive design, an expectation must be taken over the random components in the design.

We motivated the use of Fisher’s, or expected, information in a two-stage adaptive design by showing that minimizing \( M(\xi_A, \theta)^{-1} \) minimizes the lower bound of \( n \text{Var}[\hat{\theta}_n] \). Fisher’s information is then explicitly derived and further explored for a two-stage adaptive optimal design in the context of a nonlinear regression model with normal errors from both a design and an analysis perspective.

From a design perspective a locally optimal stage one sample size is defined and
we recommended a procedure to approximate it. A proposition is given showing that asymptotically the locally optimal stage one sample size is of the order $\sqrt{n}$. For a simulated example where the mean function is an exponential distribution a method for selecting $n_1$ is shown to be a robust stage one sample size for different values of $\theta$ against alternatives including $\sqrt{n}$ which performed poorly by comparison. Finally, a justification for the robustness of a two-stage adaptive optimal design is explained.

From an analysis perspective it is argued that $M^{-1}$ is a better approximation of $n \text{Var}[\hat{\theta}_n]$ than $M^{*-1}$. Thus since $M^{-1}_{\text{ind}}$, a commonly used information measure in adaptive optimal design, can be viewed as simply an estimate of $M^{*-1}$, it is not as appropriate as $M^{-1}$ as an approximation of $n \text{Var}[\hat{\theta}_n]$. These arguments were justified in a numeric example where $\theta$ is assumed to be known.

The case where $\theta$ is known is useful in providing a theoretical justification for preferring $M^{-1}$ to $M^{-1}_{\text{ind}}$, however in practice $\theta$ is unknown and thus it is of interest to understand the performance of the estimate, $\hat{M}^{-1}$ against alternatives. The two alternatives considered were $\hat{M}^{-1}_{\text{ind}}$ and $\hat{M}^{-1}_{\text{obs}}$. Efron and Hinkley (1978) and Lindsay and Li (1997) argue that the observed information measure is to be preferred to Fisher’s information for an analysis. We provide a heuristic argument and a simulated example which indicate that their argument might not extend to adaptive experiments. See Yao and Flournoy (2010) for additional details on $\hat{M}_{\text{obs}}$. Once again we found that $\hat{M}^{-1}$ outperforms both of the previously mentioned alternatives.

In Chapters 2, 3 and 4 when given, the theoretical justification in this paper assumes $n_1, n_2 \to \infty$ as $n \to \infty$. Often in practice a small first stage pilot study may be of interest. In chapter 5 we address this scenario by assuming a finite first stage sample size and a large second stage sample size. We show for a general nonlinear
one parameter regression model with normal errors that the asymptotic distribution of the MLE is a scale mixture distribution. We considered only one parameter for simplicity and clarity of exposition.

For the one parameter exponential mean function, the distribution of the adaptively selected second stage treatment and the asymptotic distribution of the MLE were derived assuming a finite first stage sample size and a large second stage sample size. Then the performance of the normalized asymptotic distribution of the MLE, \( UQ \), is analyzed and compared to popular alternatives for a set of simulations.

The distribution of \( UQ \) is shown to represent a considerable improvement over the other proposed distributions when \( n_1 \) is considerably smaller than \( n \). This is true even when \( n_1 \) is moderately large in size.

Since the optimal choice of \( n_1 \) is shown to be of the order \( \sqrt{n} \) for this model in Section 3.3, the usefulness of these findings could have significant implications for many combinations of \( n_1 \) and \( n \).

Suppose it is desired that \( P\left\{ D_1 \leq \sqrt{n} \left( \hat{\theta} n - \theta \right) \leq D_2 \right\} = 1 - \alpha \), where \( \alpha \) is the desired confidence level and \( \theta \) is the true parameter. If one were to use the large sample approximate distribution given in equation (5.1.1), \( D_1 \) and \( D_2 \) and therefore \( n \), cannot be determined until after stage 1. However, using (5.1.1) with \( M(\xi_A, \theta) \) in place of \( M(\xi^*, \theta) \) or by using \( UQ \) on can compute the overall sample size necessary to solve for \( D_1 \) and \( D_2 \) before stage one is initiated. One could determine \( n \) initially using (5.1.1) with \( M (\xi_A, \theta) \) or \( UQ \) and then update this calculation after stage 1 data is available. Such same size recalculation requires additional theoretical justification and investigation of their practical usefulness.

One additional way to improve inference would be to find biased adjusted esti-
mates $\tilde{\theta}_n$ that are superior to $\hat{\theta}_n$ for finite samples. We have not investigated the impact on inference of estimating the variances in the distributions of $UQ$.

6.2 Future Work

In Chapter 3 the locally optimal stage one sample size is defined and a method to approximate it is proposed for a two-stage adaptive optimal experiment for a model containing a single parameter. A natural extension is to develop a similar concept for a model with a multidimensional parameter vector $\theta$. To do this the adaptive optimal procedure must be extended for such models.

First we present the procedure for a single stage locally optimal design procedure in detail. For some concave function $\phi$ the locally optimal design, found using approximate theory,

$$\xi_\phi^* = \arg\min_{\xi \in \Xi} \phi(M(\xi, \theta)).$$

The design $\xi_\phi^*$ is a function of the unknown parameter vector $\theta$; therefore it must be approximated. Let $\theta'$ be an approximation of $\theta$ based on prior information or expert opinion then

$$\xi_\phi^{\theta'} = \arg\min_{\xi \in \Xi} \phi(M(\xi, \theta')) = \left\{ \lambda_1, \ldots, \lambda_m \atop x_1, \ldots, x_m \right\}$$

is the locally $\phi$-optimal design corresponding to $\theta'$. Then for a finite sample size $n$, $\xi_\phi^{\theta'}$ can be approximated using a rounding procedure, yielding the locally $\phi$-optimal
design based on $\theta'$ for a sample size $n$,

$$\xi_{\theta',n} = \arg\min_{\xi \in \Xi} \phi (M(\xi, \theta')) = \left\{ \frac{n_1}{n}, \ldots, \frac{n_m}{m} \right\}.$$ 

Now the two-stage adaptive optimal procedure with $n_1$, $n_2$ and $n$ fixed would use for the first stage

$$\xi_{\theta',n_1} = \arg\min_{\xi \in \Xi} \phi (M(\xi, \theta')) = \left\{ \frac{n_{11}}{n_1}, \ldots, \frac{n_{1m_1}}{n_1} \right\},$$

where $n_{11}, \ldots, n_{1m_1}$ are the allocations to $x_1, \ldots, x_{m_1}$, respectively. Then the maximum likelihood estimate using the first stage data, $\hat{\theta}_{n_1}$, would be used to find the adaptively selected second stage design,

$$\xi_{\hat{\theta}_{n_1},n_2} = \arg\min_{\xi, \xi_1 \in \Xi} \phi (M(\xi, \hat{\theta}_{n_1})) = \left\{ \frac{\hat{n}_{21}(\hat{\theta}_{n_1})}{n_2}, \ldots, \frac{\hat{n}_{1m_2}(\hat{\theta}_{n_1})}{n_1} \right\},$$

where $\hat{n}_{21}(\hat{\theta}_{n_1}), \ldots, \hat{n}_{1m_2}(\hat{\theta}_{n_1})$ are the allocations to $\hat{x}_1(\hat{\theta}_{n_1}), \ldots, \hat{x}_{m_2}(\hat{\theta}_{n_1})$, respectively. Denote the two-stage adaptive optimal design as $\xi_{\phi,A} = \{\xi_{\theta',n}, \xi_{\hat{\theta}_{n_1},n_2}\}$. Note this procedure is differs from many commonly suggested adaptive optimal procedures, including those in Dragalin and Fedorov (2005), Dragalin, Hsuan, and Padmanabhan (2008) and Dragalin, Fedorov, and Wu (2007), where for each stage only a single treatment level is found since only a single subject is treated at each stage. In these procedures only the treatment is random, whereas in this procedure both the second stage treatment allocations and the corresponding treatment levels are
Now to extend the concept of a locally optimal sample size allocation we must find a definition that incorporates not only the optimal sample size allocation to each stage but also represents the optimal allocation within each stage. Let \( \mathbf{n}_i = (n_{i1}, \ldots, n_{im_i})^T \) represent the sample allocation within the \( i \)th stage and \( \mathbf{n} = (\mathbf{n}_1, \mathbf{n}_2)^T \) represent the total allocation of the entire two stage experiment. Note \( n_i = \sum_{j=1}^{m_i} n_{ij}, \quad i = 1, 2 \) and \( n = n_1 + n_2 \), where \( n, n_1 \) and \( n_2 \) are defined as before. Define the set \( \mathcal{N} \) as all possible combinations of \( \mathbf{n} \) for \( n_{1j} = 0, \ldots, n, \quad j = 1, \ldots, n_{m_1} \) and \( n_{2j} = 0, \ldots, n - n_1, \quad j = 1, \ldots, n_{m_2} \). Define the locally \( \phi \)-optimal sample size allocation for a two-stage adaptive optimal design as

\[
\mathbf{n}^*_\phi = \arg \min_{\mathbf{n} \in \mathcal{N}} \phi \left( \mathcal{M}(\xi_{\phi,A}, \theta) \right).
\]

For a single parameter the only concern is finding the first stage sample size, because of this a straightforward procedure could be proposed. Proposals for approximating \( \mathbf{n}^*_\phi \) are more complicated.

A further extension would then be to extend the preceding concept to include the optimal number of stages. Here we let \( K \) represent the number of stages. This would add a layer of complexity to the previous procedure, since now the goal would be to find the optimal number of stages, \( K, 1 \leq K \leq n \), the optimal number of subjects allocated to each stage \( n_i, \quad i = 1, \ldots, K \) and the optimal allocations to each treatment within each stage \( n_{ij}, \quad j = 1, \ldots, m_i, \quad i = 1, \ldots, K \). This is a very complex problem due to the large number of possible allocations that must be considered.

Next consider an extension motivated by Chapter 5. The motivation for this work is to derive the limiting distribution of the adaptive optimal procedure in Chapter 3.
when the first stage sample size is small for use in inference about the parameter $\theta$. This is somewhat of a reversal of standard optimal design motivation. Traditionally, the limiting, or exact distribution, is used to inform the design and then inference follows directly. A natural extension would be to use the limiting distribution $U_Q$ to inform the design. This would mean finding optimal designs for scale mixture distributions. Finding such procedure would have implications beyond the adaptive optimal procedure examined in this dissertation.
Appendix A

Appendix Relating to Chapter 3

A.1 Proof of Theorem 2

Recall \( \hat{\theta}_{n_1} \) is a function of \( \bar{y}_1 \) and \( x_1 \), thus \( x_2 \left( \hat{\theta}_{n_1} \right) = x_2 \left( \hat{\theta}_{n_1}(x_1, \bar{y}_1) \right) \) and \( \bar{\varepsilon}_1 = \sum_{j=1}^{n_1} \varepsilon_{1j} / n = \bar{y} - \eta(x_1, \theta) \sim \mathcal{N}(0, \sigma^2 / n_1) \). Equation (3.9) can be written

\[
n^*_1(\theta) = \arg\max_{n_1 \in (1, n)} \text{M}(\xi_A, \theta) = \arg\max_{n_1 \in (1, n)} \left[ \frac{n_1}{n} \left( \frac{\partial \eta(x_1, \theta)}{\partial \theta} \right)^2 \right] \nonumber
\]

\[
+ \left( 1 - \frac{n_1}{n} \right) \mathbb{E}_{\bar{\varepsilon}_1} \left[ \left( \frac{\partial \eta(x_2 \left( \hat{\theta}_{n_1}(x_1, \eta(x_1, \theta') + \bar{\varepsilon}_1) \right), \theta) \right)^2 \right]_{\theta' = \theta} .
\]
Then
\[
E_{\tilde{\varepsilon}_1} \left[ \left( \frac{\partial \eta(x_2 \left( \hat{\theta}_{n_1}(x_1, \eta(x_1, \theta') + \varepsilon_1) \right), \theta)}{\partial \theta} \right)^2 \right] \mid_{\theta'=\theta} (A.1)
\]
\[
= \pi_a \left( \frac{\partial \eta(a, \theta)}{\partial \theta} \right)^2 + \pi_b \left( \frac{\partial \eta(b, \theta)}{\partial \theta} \right)^2 + E_{\tilde{\varepsilon}_1} \left[ \left( \frac{\partial \eta(x_2 \left( \hat{\theta}_{n_1}(x_1, \eta(x_1, \theta') + \varepsilon_1) \right), \theta)}{\partial \theta} \right)^2 \cdot I_{a < x_2 \left( \hat{\theta}_{n_1}(x_1, \eta(x_1, \theta) + \varepsilon_1) \right) < b} \right] \mid_{\theta'=\theta} .
\]

Now define \( \frac{1}{\sqrt{n_1}} Z = \bar{\varepsilon}_1 \), where \( Z \sim \mathcal{N}(0, \sigma^2) \) and
\[
f_{x_1, \theta} \left( \frac{1}{\sqrt{n_1}} z \right) = \left( \frac{\partial \eta \left( x_2 \left( \hat{\theta}_{n_1}(x_1, \eta(x_1, \theta') + z/\sqrt{n_1}) \right), \theta \right)}{\partial \theta} \right)^2 \mid_{\theta'=\theta} .
\]

Then by Taylor’s Theorem [cf. Ferguson (1996)]
\[
f_{x_1, \theta} \left( \frac{1}{\sqrt{n_1}} z \right) = f_{x_1, \theta}(0) + \frac{1}{\sqrt{n_1}} z f'_{x_1, \theta}(0) + \frac{1}{n_1} z^2 \int_0^1 \int_0^1 v f''(uvz) du dv ,
\]
holds for all \( z \) and thus expansion can be taken inside the expectation, \( i.e., \)
\[
E_Z \left( f_{x_1, \theta} \left( \frac{1}{\sqrt{n_1}} Z \cdot I_{a < x_2 \left( \hat{\theta}_{n_1}(x_1, \eta(x_1, \theta) + Z/\sqrt{n_1}) \right) < b} \right) \right) = f_{x_1, \theta}(0)
\]
\[
+ \frac{1}{n_1} E_Z \left( Z^2 \int_0^1 \int_0^1 v f''(uvZ) du dv \cdot I_{a < x_2 \left( \hat{\theta}_{n_1}(x_1, \eta(x_1, \theta) + Z/\sqrt{n_1}) \right) < b} \right) = O \left( \frac{1}{n_1} \right) .
\]
Therefore (A.1) is equal to
\[
\pi_a \left( \frac{\partial \eta(a, \theta)}{\partial \theta} \right)^2 + \pi_b \left( \frac{\partial \eta(b, \theta)}{\partial \theta} \right)^2 + \left( \frac{\partial \eta(x^*, \theta)}{\partial \theta} \right)^2 + O \left( \frac{1}{n_1} \right),
\]
i.e., there exists a constant, \( \alpha \neq 0 \), such that (A.1) is asymptotically equivalent to \( \alpha/n_1 \). Thus if we let \( \beta = \left( \frac{\partial \eta(x_1, \theta)}{\partial \theta} \right)^2 \) then
\[
n_1^*(\theta) = \arg \max_{n_1 \in (1, n)} \left[ \frac{n_1}{n} \beta + \frac{\sigma^2}{n_1} \right].
\]
Then the derivative of \( n_1^*(\theta) \), with respect to \( n_1 \), set equal to 0 and solved is
\[
n_1 = \left[ \frac{1}{\alpha} \left( \frac{1}{n} \beta \right) \right]^{-\frac{1}{2}}
\]
which implies that \( n_1^*(\theta) = O(\sqrt{n}) \).

**A.2 Discussion of neighborhoods described in section 3.3**

Let \( \mu(x, \theta) = (\partial \eta(x, \theta)/\partial \theta)^2 \). Now write
\[
n_1^*(\theta) = \arg \max_{n_1 \in (1, n)} \left[ \frac{n_1}{n} \mu(x_1, \theta) + \left( 1 - \frac{n_1}{n} \right) E_{x_2} [\mu(x_2, \theta)] \right].
\]
It is sufficient to show that there is a neighborhood of \( x_1 \) around \( x^*(\theta) \) such that
\[
E_{x_2} [\mu(x_2, \theta)] < \mu(x_1, \theta).
\]  (A.2)
There are no restrictions on $\mu(x, \theta)$ except that $x^*(\theta) = \operatorname{argmax}_x \mu(x, \theta)$ is an interior point of $\mathcal{X}$. Thus there exists a concave function, say $h(x, \theta)$, such that $\operatorname{argmax}_x h(x, \theta) = x^*(\theta)$ and $\mu(x, \theta) \leq h(x, \theta)$ for all $x$. Therefore from Jensen’s Inequality

$$
E_{x_1} [\mu(x_1, \theta)] < E_{x_2} [h(x_2, \theta)] < h(E_{x_2} [x_2], \theta) < \mu(x^*(\theta), \theta).
$$

Thus if $x_1 = x^*(\theta)$, (A.2) holds. Because $\mu(x, \theta)$ is continuous and the equality is strict the existence of the interval around $x^*(\theta)$ is confirmed.
Appendix B

Appendix Relating to Chapter 4

B.1 Proof of Theorem 3

First, we show that if $\theta \in (b^{-1}, a^{-1})$ or $\theta \in (0, b^{-1})$, then

$$
\begin{pmatrix}
\bar{y}_1 \\
\bar{y}_2
\end{pmatrix} \xrightarrow{P_{n \to \infty}}
\begin{pmatrix}
e^{-\theta x_1} \\
e^{-1}
\end{pmatrix}
$$

and

$$
\begin{pmatrix}
\bar{y}_1 \\
\bar{y}_2
\end{pmatrix} \xrightarrow{P_{n \to \infty}}
\begin{pmatrix}
e^{-\theta x_1} \\
e^{-\theta b}
\end{pmatrix}
$$

respectively. That $\bar{y}_1 \to e^{-\theta x_1}$ as $n \to \infty$ in both cases is an application of the strong law of large numbers. To understand the convergence of $\bar{y}_2$ we use its
characteristic function

\[
\psi(y_2(t)) = E(e^{i t y_2}) = \int_{-\infty}^{\infty} e^{i t y_1} \left[ \int_{-\infty}^{\infty} f(y_2|y_1)f(y_1) dy_1 \right] dy_2
\]

\[
= \int_{-\infty}^{\infty} \psi(y_2|y_1)(t) f(y_1) dy_1 = \int_{-\infty}^{\infty} e^{i t e^{-\theta x_2(y_1)} - \frac{t^2}{2\pi^2}} f(y_1) dy_1
\]

\[
= \pi_1 e^{i t e^{-\theta a} - \frac{t^2}{2\pi^2}} + \pi_3 e^{i t e^{-\theta b} - \frac{t^2}{2\pi^2}}
\]

\[
+ E \left[e^{i t e^{\frac{x_1}{\log y_1} - \frac{t^2}{2\pi^2}} I \left( e^{-a^{-1} x_1} < y_1 < e^{-b^{-1} x_1} \right)} \right]
\]

Now, if \( \theta \in (b^{-1}, a^{-1}) \), then \( y_1 \to e^{-\theta x_1} \in (e^{-a^{-1} x_1}, e^{-b^{-1} x_1}) \), \( e^{i t e^{\frac{x_1}{\log y_1} - \frac{t^2}{2\pi^2}}} \) is bounded and continuous, and \( \pi_1, \pi_3 \to 0 \). We conclude that

\[
\lim_{n \to \infty} \psi(y_2(t)) = e^{i t e^{-1}} \implies \tilde{y}_2 \xrightarrow{n \to \infty} e^{-1}.
\] (B.1)

If \( \theta \in (0, b^{-1}) \), then \( y_1 \to (e^{-b^{-1} x_1}, 1) \), \( \pi_1, \pi_3 \to 0 \) and \( \pi_2 \to 1 \), and since \( e^{i t e^{\frac{x_1}{\log y_1} - \frac{t^2}{2\pi^2}}} \leq 1 \), the expectation term in (B.1) is less than or equal to \( \pi_2 \). Hence,

\[
\lim_{n \to \infty} \psi(y_2(t)) = e^{i t e^{-\theta b}} \implies \tilde{y}_2 \xrightarrow{n \to \infty} e^{-\theta b}.
\]

Now consider the following ratio where \( \tilde{\theta} \in (0, a^{-1}) \) and \( \theta \) is the true parameter;

\[
\frac{1}{n} \log \frac{L_n(\tilde{\theta})}{L_n(\theta)} = -\frac{w_1}{2} \left( (\tilde{y}_1 - e^{-\tilde{\theta} x_1})^2 - (\tilde{y}_1 - e^{-\theta x_1})^2 \right)
\]

\[
- \frac{w_2}{2} \left( (\tilde{y}_2 - e^{-\tilde{\theta} x_2(\tilde{y}_1)})^2 - (\tilde{y}_2 - e^{-\theta x_2(\tilde{y}_1)})^2 \right).
\]

105
If $\theta \in (b^{-1}, a^{-1})$, then $x_2 \to \theta^{-1}$ and

$$
\frac{1}{n} \log \frac{L_n(\theta)}{L_n(\theta_1)} = \lim_{n \to \infty} -\frac{w_1}{2} \left( e^{-\theta x_1} - e^{-\tilde{\theta} x_1} \right)^2 - \frac{w_2}{2} \left( e^{-1} - e^{-\tilde{\theta}} \right)^2 < 0
$$

for all $\tilde{\theta} \neq \theta$. If $\theta \in (b^{-1}, a^{-1})$, then $x_2 \to b$ and

$$
\frac{1}{n} \log \frac{L_n(\tilde{\theta})}{L_n(\theta)} = \lim_{n \to \infty} -\frac{w_1}{2} \left( e^{-\theta x_1} - e^{-\tilde{\theta} x_1} \right)^2 - \frac{w_2}{2} \left( e^{-\theta b} - e^{-\tilde{\theta} b} \right)^2 < 0
$$

for all $\tilde{\theta} \neq \theta$. However,

$$
\frac{1}{n} \log \frac{L_n(\hat{\theta}_n)}{L_n(\theta)} = \sup_{\theta \in \Theta} \log \frac{L_n(\tilde{\theta})}{L_n(\theta)} \geq 0
$$

by definition. Thus $\hat{\theta}_n$.

### B.2 Proof of Theorem 4

Since the likelihood has the same form as in 2.1.1 the expansion of Taylor expansion of $\log L_n(\theta)|_{\theta=\hat{\theta}_n}$ around $\log L_n(\theta)$ can still be written as

$$
\log L_n(\theta)|_{\theta=\hat{\theta}_n} = \log L_n(\theta) + \left( \hat{\theta}_n - \theta \right) S + \frac{1}{2} \left( \hat{\theta}_n - \theta \right)^2 \left[ \frac{\partial}{\partial \theta} S \right]_{\theta=\hat{\theta}_n} \quad \text{(B.2)}
$$

where $\tilde{\theta} \in (\theta, \hat{\theta}_n)$. Thus rewriting (B.2) we get

$$
\sqrt{n} \left( \hat{\theta}_n - \theta \right) = \frac{\frac{1}{\sqrt{n}} S}{-\frac{1}{n} \left( \frac{\partial}{\partial \theta} S + \frac{1}{2} \left( \hat{\theta}_n - \theta \right) \left[ \frac{\partial^2}{\partial \theta^2} S \right]_{\theta=\hat{\theta}_n} \right)} \quad \text{(B.3)}
$$
Recall $S = S_1 + S_2$. Then

$$ \frac{1}{\sqrt{n}} S_1 = -\frac{\sqrt{w_1}}{\sigma^2} \sqrt{n_1} (\bar{y}_1 - e^{-\theta x_1}) x_1 e^{-\theta x_1} \sim \mathcal{N} \left( 0, \frac{1}{\sigma^4} x_1^2 e^{-2\theta x_1} \right), \quad (B.4) $$

Now recognize that since the procedure takes place the second stage treatment $x_2 \rightarrow x^*(\theta)$ as $n_1 \rightarrow \infty$. Thus all observations in the second stage are treated at the optimal second stage treatment $x^*(\theta)$, i.e., as $n_1 \rightarrow \infty$

$$ \frac{1}{\sqrt{n}} S_2 \longrightarrow -\frac{\sqrt{w_2}}{\sigma^2} \sqrt{n_2} (\bar{y}_1 - e^{-\theta x^*(\theta)}) x^*(\theta) e^{-\theta x^*(\theta)} $$

$$ \sim \mathcal{N} \left( 0, w_2 \frac{1}{\sigma^4} [x^*(\theta)]^2 e^{-2\theta x^*(\theta)} \right). \quad (B.5) $$

Thus

$$ \frac{1}{\sqrt{n}} S \xrightarrow{n_1 \rightarrow \infty} \mathcal{N} \left( 0, w_1 \frac{1}{\sigma^4} x_1^2 e^{-2\theta x_1} + w_2 \frac{1}{\sigma^4} [x^*(\theta)]^2 e^{-2\theta x^*(\theta)} \right). \quad (B.6) $$

Next consider

$$ \frac{1}{n} \left[ \frac{\partial}{\partial \theta} S_1 \right] = \frac{w_1}{\sigma^2} \left[ (\bar{y}_1 - e^{-\theta x_1}) x_1^2 e^{-\theta x_1} + x_1^2 e^{-2\theta x_1} \right] $$

$$ \sim \mathcal{N} \left( -x_1^2 e^{-2\theta x_1}, \frac{w_1^2}{\sigma^4} x_1^4 e^{-2\theta x_1} \right). \quad (B.7) $$

This implies

$$ \frac{1}{n} \left[ \frac{\partial}{\partial \theta} S_1 \right] \xrightarrow{n_1 \rightarrow \infty} \frac{1}{\sigma^2} x_1^2 e^{-2\theta x_1}. \quad (B.8) $$
by the strong law of large numbers. Using the same logic as before that all observations in the second stage are treated at the optimal second stage treatment \( x^*(\theta) \), i.e., as \( n_1 \to \infty \)

\[
\frac{1}{n} \left[ \frac{\partial}{\partial \theta} S_2 \right] \to \frac{w_2}{\sigma^2} \left[ (\bar{y}_2 - e^{-\theta x^*(\theta)}) [x^*(\theta)]^2 e^{-\theta x^*(\theta)} + [x^*(\theta)]^2 e^{-2\theta x^*(\theta)} \right] \\
\sim \mathcal{N} \left( -[x^*(\theta)]^2 e^{-2\theta x^*(\theta)}, \frac{w_2^2}{\sigma^4} [x^*(\theta)]^4 e^{-2\theta x^*(\theta)} \right). \tag{B.9}
\]

Which implies

\[
\frac{1}{n} \left[ \frac{\partial}{\partial \theta} S_2 \right] \xrightarrow{n \to \infty} \frac{1}{\sigma^2} [x^*(\theta)]^2 e^{-2\theta x^*(\theta)}. \tag{B.10}
\]

Thus

\[
-\frac{1}{n} \left[ \frac{\partial}{\partial \theta} S \right] \xrightarrow{n \to \infty} \frac{1}{\sigma^2} \left( x_1^2 e^{-2\theta x_1} + [x^*(\theta)]^2 e^{-2\theta x^*(\theta)} \right). \tag{B.11}
\]

It is further required that

\[
\frac{\partial^2 S(\tilde{\theta})}{\partial (\tilde{\theta})^2} < \infty. \tag{B.12}
\]

Which is satisfied since

\[
\frac{\partial \eta(x, \tilde{\theta})}{\partial \theta} \frac{\partial^2 \eta(x, \tilde{\theta})}{\partial \theta^2} = x^3 e^{-2\tilde{\theta} x} < 0. \tag{B.13}
\]
Thus by (B.6), (B.11), (B.12) and the result of consistency given in theorem 3 we get

\[ \sqrt{n} \left( \hat{\theta}_n - \theta \right) \xrightarrow{n \to \infty} N \left( 0, \frac{1}{\sigma^2} \left[ w_1 x_1 e^{-\theta x_1} + w_2 x^*(\theta) e^{-\theta x^*(\theta)} \right]^{-1} \right) \]  

(B.14)

\section*{B.3 Distribution of the Second Stage Treatment \( x_2 \)}

The second stage treatment \( x_2 \) is defined by the transformation of \( \bar{y}_1 \)

\[ x_2(\hat{\theta}_{n_1}) = \begin{cases} \hat{\theta}^{-1}_{n_1}, & \text{if } \bar{y}_1 \in \left( e^{-a^{-1}x_1}, e^{-b^{-1}x_1} \right) \\ b, & \text{if } \bar{y}_1 \geq e^{-b^{-1}x_1} \\ a, & \text{if } \bar{y}_1 \leq e^{-a^{-1}x_1}. \end{cases} \]  

(B.15)

This a mixture of discrete and continuous measures with nonzero probability mass at the boundaries of the design space. Therefore

\[ f_{x_2|x_1}(x_2|x_1) = I_{x_2=a} \pi_a + I_{x_2=b} \pi_b + g_{x_2|x_1}(x_2|x_1) I_{a<x_2<b}, \]  

(B.16)

where \( g_{x_2|x_1}(x_2|x_1) \) represents the continuous transform of \( x_2 \) on \( \bar{y}_1 \in \left( e^{-a^{-1}x_1}, e^{-b^{-1}x_1} \right) \) which yields the subdensity function

\[ g_{x_2|x_1}(x_2|x_1) = \frac{x_1}{x_2} \sqrt{\frac{n_1}{2\pi}} e^{-\frac{x_1}{x_2}} \exp \left( -\frac{n_1}{2} \left( e^{-\frac{x_1}{x_2}} - e^{-\theta x_1} \right)^2 \right). \]

For additional details on on the transformation see Yao and Flournoy (2010).
Appendix C

Appendix Relating to Chapter 5

C.1 Proof of Theorem 5

Since the likelihood has the same form as in 2.1.1 the expansion of Taylor expansion of \( \log L_n(\theta) \) around \( \log L_n(\theta) \) can still be written as

\[
\log L_n(\theta) \bigg|_{\theta = \hat{\theta}} = \log L_n(\theta) + \left( \hat{\theta} - \theta \right) S + \frac{1}{2} \left( \hat{\theta} - \theta \right)^2 \left[ \frac{\partial}{\partial \theta} S \right] \bigg|_{\theta = \hat{\theta}}. \tag{C.1}
\]

where \( \tilde{\theta} \in (\theta, \hat{\theta}) \). Thus rewriting (B.2) we get

\[
\sqrt{n} \left( \hat{\theta} - \theta \right) = \frac{1}{\sqrt{n}} S - \frac{1}{n} \left( \frac{\partial}{\partial \theta} S + \frac{1}{2} \left( \hat{\theta} - \theta \right) \left[ \frac{\partial^2}{\partial \theta^2} S \right]_{\theta = \hat{\theta}} \right). \tag{C.2}
\]
It can be shown that \( \hat{\theta}_n \) is consistent for \( \theta \) if \( n_2 \to \infty \) and \( n_1/n \to 0 \) which gives the result

\[
\sqrt{n} \left( \hat{\theta}_n - \theta \right) \approx \frac{1}{\sqrt{n}} S - \frac{1}{n} \frac{\partial}{\partial \theta} S \tag{C.3}
\]

Now decompose the right hand side of (C.3) as

\[
\frac{1}{\sqrt{n}} S - \frac{1}{n} \frac{\partial}{\partial \theta} S = \frac{\sqrt{n} S}{n} \left( S_1 + S_2 \right) - \frac{\sqrt{n} S}{n} \left( \frac{\partial}{\partial \theta} S_1 + \frac{\partial}{\partial \theta} S_2 \right) \]

\[
= \frac{\sqrt{n} S}{n} S_1 - \frac{\sqrt{n} S}{n} \left( \frac{\partial}{\partial \theta} S_1 + \frac{\partial}{\partial \theta} S_2 \right) + \frac{\sqrt{n} S}{n} \left( \frac{\partial}{\partial \theta} S_1 + \frac{\partial}{\partial \theta} S_2 \right) \tag{C.4}
\]

As \( n_2 \to \infty \), \( S_1/\sqrt{n} \to 0 \), \( n_2/n \to 1 \) and \( \frac{1}{n} \frac{\partial}{\partial \theta} S_2 \to 0 \) as \( n \to \infty \). Thus the first term in (C.4) goes to 0 as \( n \to \infty \). Write the second term in (C.4) as

\[
\frac{\sqrt{n} S_2}{n} \left( \frac{\partial}{\partial \theta} S_1 + \frac{\partial}{\partial \theta} S_2 \right) = \left( \frac{1}{\sqrt{n}} S_1 - \frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} S_2 \right)^{-1}
\]

Further as \( n_2 \to \infty \), \( \frac{1}{n} \frac{\partial}{\partial \theta} S_1 \to 0 \) and \( \frac{1}{\sqrt{n}} S_2 \to 0 \),

\[
\frac{1}{\sqrt{n}} S_1 \to 0
\]

and

\[
\frac{1}{\sqrt{n}} S_2 = \frac{1}{\sqrt{n}} \left( \hat{y}_2 - \eta(x_2, \theta) \right) \frac{\partial^2 \eta(x_2, \theta)}{\partial \theta^2} + w_2 \left( \frac{\partial \eta(x_2, \theta)}{\partial \theta} \right)^2
\]

\[
= \frac{1}{\sqrt{n}} \left( \frac{\partial^2 \eta(x_2, \theta)}{\partial \theta^2} \right) + \sqrt{w_2} \frac{\partial \eta(x_2, \theta)}{\partial \theta} \sqrt{\frac{1}{n_2} \left( \hat{y}_2 - \eta(x_2, \theta) \right)} \tag{C.5}
\]
The first term in (C.5) goes to 0. To evaluate the second term, it is important
to recognize that \( \varepsilon_{i2} = \bar{y}_2 - \eta(x_2, \theta) \sim \mathcal{N}(0, \sigma^2/n_2) \) and \( \bar{y}_1 \sim \mathcal{N}(0, \sigma^2/n_1) \) are
independent and thus
\[
\bar{y}_2 - \eta(x_2, \theta) \text{ and } \frac{\partial \eta(x_2, \theta)}{\partial \theta}
\]
are independent. Because of this independence,
\[
\sqrt{n_2} (\bar{y}_2 - \eta(x_2, \theta)) \left( \frac{\partial \eta(x_2, \theta)}{\partial \theta} \right)^{-1} \sim U Q
\]
where \( U \) is a random function of \( \bar{y}_1 \) and \( Q \sim \mathcal{N}(0, \sigma^2) \) as determined by \( \left( \frac{\partial \eta(x_2, \theta)}{\partial \theta} \right)^{-1} \).
Now, with \( \sqrt{w_2} \to 1 \) as \( n_2 \to \infty \) the result follows from an application of Slutsky’s theorem.

C.2 Proof of Theorem 6

Proof: First we find the distribution of \( U \) where \( U = h(z) \) and the random variable
\( z \) is defined by
\[
z = \begin{cases} 
- \frac{x_1}{\log \bar{y}_1}, & \text{if } \bar{y}_1 \in \left( e^{-x_1/a}, e^{-x_1/b} \right) \\
- \frac{x_1}{\log a}, & \text{if } \bar{y}_1 \leq e^{-x_1/a} \\
- \frac{x_1}{\log b}, & \text{if } \bar{y}_1 \geq e^{-x_1/b}
\end{cases}
\]
Figure C.1 illustrates the map from \( U \) to \( z \in [a, b] \) where \( \theta = 1, \sigma^2 = .5, a = .25 \) and \( b = 4 \).
Figure C.1: Map of $z = -x_1 / \log \bar{y}_1$ for $\theta = 1$, $a = .25$ and $b = 4$. 
Lambert’s product log function (cf. Corless, Gonnet, Hare, Jeffrey, and Knuth (1996)) is defined as the solutions to

\[ \ln e^w = c \] (C.6)

for some constant \( c \). Denote the solutions to (C.6) by \( W(w) \). Let

\[ V(c) = \arg_{\bar{y}_1} \left\{ \left( -\frac{x_1}{\log \bar{y}_1} \right)^{-1} \exp \left\{ \frac{\theta - x_1}{\log \bar{y}_1} \right\} = c \right\}. \]

Then

\[ V(c) = \exp \left\{ \frac{\theta x_1}{W(-\theta/c)} \right\}. \]

The \( W \)-function is real valued on \( w \geq -1/e \), single valued at \( w = -1/e \), double valued on \( w \in (-1/e, 0) \). \( U \in \{ \theta e, \max \{ h(a), h(b) \} \} \), \( x_1 \in [a, b] \), \( 0 < a < b < \infty \). Therefore \( V(c) \) is real valued for all \( \theta \in (0, \infty) \). For simplicity define \( v_1 = \min V(c) \) and \( v_2 = \max V(c) \) for a given \( c \).

We present the proof for the cumulative distribution function (CDF) of \( U \) and the CDF of \( UQ \) for the case where \( x^*(\theta) \in [a, b] \) and \( h(a) < h(b) \). The derivation of the distributions under alternative cases are tedious and do not differ greatly from this case.

Note in this case the domain of \( U \) is \( [h(1/\theta) = \theta e, h(b)] \). If \( h(1/\theta) < U < h(a) \),
then

\[ P \{ U \leq t_1 \} = P \{ h(\tilde{y}_1) < t \} \]

\[ = P \{ v_1(t_1) < \tilde{y}_1 < v_2(t_1) \} \]

\[ = \Psi(v_2(t_1)) - \Psi(v_1(t_1)). \]

If \( U = h(a) \), then

\[ P \{ U \leq h(a) \} = \Psi(v_2(h(a))). \]

If \( U \in (h(a), h(b)) \), then

\[ P \{ U \leq t_1 \} = P \{ v_1(t_1) < \tilde{y}_1 < v_2(t_1) \}. \]

However, since \( t_1 < h(a) \) \( P \{ \tilde{y}_1 < v_1(t_1) \} = 0 \),

\[ P \{ U \leq t_1 \} = \Psi(v_2(t_1)). \]

If \( U \geq h(b) \), then

\[ P \{ U \leq h(b) \} = 1. \]
Thus

\[
P\{U \leq t_1\} = \begin{cases} 
0, & \text{if } t_1 \leq h(1/\theta) \\
\Psi(v_2(t_1)) - \Psi(v_2(t_1)), & \text{if } t_1 \in (h(1/\theta), h(a)) \\
\Psi(e^{-x_1/a}), & \text{if } t_1 = h(a) \\
\Psi(v_2(t_1)), & \text{if } t_1 \in (h(a), h(b)) \\
1, & \text{if } h(b) \leq t_1 \leq \infty 
\end{cases}
\]

Figure C.2 plots the CDF of U for \( \theta = 1, x_1 = 2, n_1 = 5, \sigma^2 = .5, a = .25 \) and \( b = 4 \).

The distribution is a piecewise function with discontinuities at the boundary points \( a \) and \( b \).

Now consider the distribution of \( UQ \). Recall \( q \sim \mathcal{N}(0, \sigma^2) \) and \( U \) and \( Q \) are independent. If \( t \in (-\infty, 0) \), then

\[
P\{UQ \leq t\} = P\{U \geq t/q|0 < t/q \leq h(1/\theta)\} P\{0 < t/q \leq h(1/\theta)\} \\
+ P\{U \geq t/q|h(1/\theta) < t/q \leq h(a)\} P\{h(1/\theta) < t/q \leq h(a)\} \\
+ P\{U = h(a)|t/q = h(a)\} P\{t/q = h(a)\} \\
+ P\{U \geq t/q|h(a) < t/q \leq h(b)\} P\{h(a) < t/q \leq h(b)\} \\
+ P\{U = h(b)|t/q = h(b)\} P\{t/q = h(b)\} \\
+ P\{U \geq t/q|h(b) < t/q \leq \infty\} P\{h(b) < t/q < \infty\}.
\]

The distribution is symmetric, thus the derivation of the CDF if \( t \in (0, \infty) \) is analogous.
Figure C.2: CDF of $U$ for $\theta = 1$, $x_1 = 2$, $n_1 = 5$, $\sigma^2 = .5$, $a = .25$ and $b = 4$. 
Bibliography


combinations based on efficacy-toxicity response. *Journal of Statistical Planning
and Inference*, **2**, 352–373.

dose–finding studies based on sigmoid $E_{\text{MAX}}$ model. *Journal of Biopharmaceutical


preoperative doxorubicin-based concurrent chemoradioation and surgical resection for localized extremity and body wall soft tissue sarcomas. *Journal of Clinical Oncology, 22*, 3375–3380.


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.


VITA

I was born and grew up in a small mining town in southeastern Missouri. In 2004 I finished my undergraduate degree in Finance from Missouri State University. After graduation I worked as an accountant for Cerner Corporation, a hospital software company, based in Kansas City. In January of 2008 I joined the Department of Statistics at the University of Missouri to pursue a Ph.D. I began working with my advisor Dr. Nancy Flournoy in the fall of 2010.

In 2008 my wife, LeAnna Lane, and I married and in 2012 we welcomed our first child Mae Pearl Lane.